ManiFeel: Benchmarking and Understanding Visuotactile Manipulation Policy Learning

Abstract

Supervised visuomotor policies have shown strong performance in robotic manipulation but often struggle in tasks with limited visual inputs, such as operations in confined spaces and dimly lit environments, or tasks requiring precise perception of object properties and environmental interactions. In such cases, tactile feedback becomes essential for manipulation. While the rapid progress of supervised visuomotor policies has benefited greatly from high-quality, reproducible simulation benchmarks in visual imitation, the visuotactile domain still lacks a similarly comprehensive and reliable benchmark for large-scale and rigorous evaluation. To address this, we introduce ManiFeel, a reproducible and scalable simulation benchmark designed to systematically study supervised visuotactile policy learning. ManiFeel offers a diverse suite of contact-rich and visually challenging manipulation tasks, a modular evaluation pipeline spanning sensing modalities, tactile representations, and policy architectures, as well as real-world validation. Through extensive experiments, ManiFeel demonstrates how tactile sensing enhances policy performance across diverse manipulation scenarios, ranging from precise contact-driven operations to visually constrained settings. In addition, the results reveal task-dependent strengths of different tactile modalities and identify key design principles and open challenges for robust visuotactile policy learning. Real-world evaluations further confirm that ManiFeel provides a reliable and meaningful foundation for benchmarking and future visuotactile policy development. To foster reproducibility and future research, we will release our codebase, datasets, training logs, and pretrained checkpoints, aiming to accelerate progress toward generalizable visuotactile policy learning and manipulation.

1. Introduction

In recent years, supervised policy learning has made significant strides in robot manipulation, where visual input is used to generate actions for long-horizon and dexterous manipulation tasks [1–4]. However, vision-based policies face notable limitations not only in environments where visual cues are absent or severely degraded, but also in manipulation scenarios that inherently demand precise contact interactions. In cluttered or confined spaces, under low-light conditions, or during complex contact-rich tasks, relying solely on vision becomes insufficient for robust policy execution. In such contexts, tactile feedback provides essential complementary information, capturing contact geometry, force distribution, and subtle interaction dynamics that vision alone cannot perceive.

Combining visual and tactile modalities has therefore become a promising and intuitive direction for learning robotic manipulation, particularly within supervised policy learning frameworks. Recent studies have shown that integrating high-resolution vision-based tactile sensors, such as GelSight [5], can significantly enhance manipulation performance across diverse contact-rich and visually challenging scenarios [6–10]. Other tactile sensing modalities, including tactile arrays [11–13], force-torque sensors [14, 15], tactile 3D point clouds [16], and contact-induced audio signals [17], have likewise demonstrated strong potential for enhancing perception and control in contact-intensive or unstructured environments.

However, unlike visual imitation learning, which benefits from a wide range of established simulation benchmarks that support large-scale experimentation and statistically meaningful comparisons [18–20], visuotactile policy learning lacks comparable infrastructure. This gap is largely attributed to the challenges of accurately simulating high-fidelity tactile sensing. The lack of standardized benchmarking tasks and reproducible simulation environments has hindered the systematic development and evaluation of robust visuotactile manipulation policies. Consequently, evaluating the effectiveness of different approaches in a controlled and consistent manner remains a significant challenge. As a result, current research in supervised visuotactile policy learning lacks sufficient comparative analysis [7, 8, 10], limiting our understanding of its core principles and hindering the exploration of promising future directions.

To address these challenges, we introduce **ManiFeel**, a scalable and reproducible simulation benchmark for su-

pervised visuotactile policy learning. ManiFeel provides a comprehensive testbed for analyzing the strengths and limitations of different sensing modalities, and offering insights into visuotactile policy design. Our contribution can be summarized as follows.

1) **Visuotactile Task Suite**: ManiFeel offers a diverse set of contact-rich manipulation tasks and corresponding human demonstration datasets, covering insertion, assembly, object exploration inside containers, and sorting tasks under visually challenging conditions. These tasks are designed to capture a wide range of contact-rich and visually limited scenarios, enabling systematic evaluation of how different tactile information contributes to multimodal policy learning.

2) **Modular Policy Architecture Design**: ManiFeel modularizes the visuotactile policy pipeline into three key components: sensing modalities, tactile representations, and policies. This modular design facilitates the integration of diverse policy architectures. By enabling flexible combinations and systematic comparisons across these modules, ManiFeel enables in-depth investigation into which model structures offer greater advantages for visuotactile policy learning.

3) **Empirical Analysis of Visuotactile Policy Learning**: We conduct extensive empirical comparisons on diverse tasks across different policy architectures by systematically varying sensing modalities, tactile representations, and policies. These experiments are conducted in both simulation and real-world settings, enabling us to identify how different tactile representations support different types of tasks, uncover effective design principles for visuotactile policy learning, and highlight key challenges and promising directions for future research.

4) **Real-world Deployment and Reproduction**: In addition to constructing simulation benchmarks, we also reproduce the results in real-world settings, comparing different sensing modalities. The correlation between real-world and simulation results further validates the reliability and practical relevance of the ManiFeel simulation benchmark.

5) **Open-Source and Reproducible Benchmark Platform**: To facilitate reproducible research and accelerate progress in visuotactile manipulation, ManiFeel is fully open-sourced, including its codebase, dataset, and pretrained checkpoints. The benchmark provides a standardized and easily extensible platform that enables fair comparisons and systematic evaluations across different models and sensing configurations. By establishing shared baselines and transparent experimental protocols, ManiFeel lowers the entry barrier for new research and promotes cumulative progress toward more generalizable and robust visuotactile manipulation in diverse real-world scenarios.

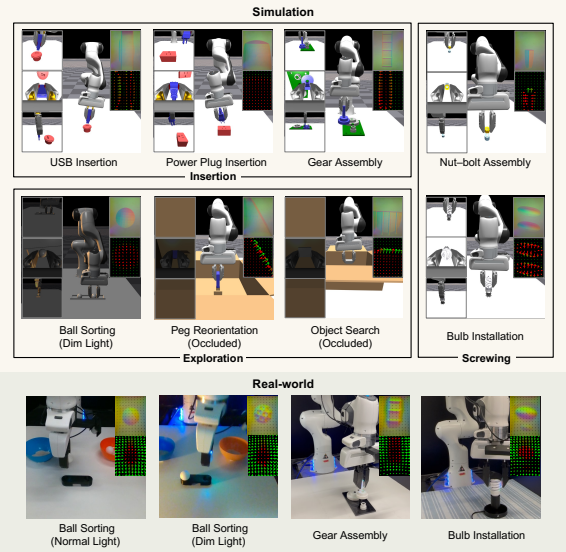


Figure 1. Overview of the manipulation tasks in the ManiFeel benchmark, showing each task's setup along with the available visual modalities (e.g., wrist camera, agentview camera) and tactile observations (e.g., high-resolution tactile images and tactile force-field signals).

2. ManiFeel: Visuotactile Manipulation Learning Benchmark

This section introduces the task suite and modular pipeline design of the ManiFeel benchmark. ManiFeel features a diverse set of manipulation tasks spanning clear-vision conditions to severely degraded or fully occluded visual settings (Section 2.1). The benchmark includes three main task categories: *Insertion*, *Screwing*, and *Exploration*. This results in a total of 13 different task setups, including 9 simulation setups and 4 real-world setups, as shown in Fig. 1. These tasks are designed to vary in their reliance on visual and tactile modalities, providing a comprehensive framework for analyzing policy performance across different sensory challenges. In addition, we describe the modular pipeline design that enables systematic evaluation of different sensing modalities, tactile representations, and policies across the task suite (Section 2.2).

2.1. Task Suite and Dataset

2.1.1. Task and Environment Design

We design a suite of diverse manipulation tasks to systematically evaluate supervised visuotactile policy learning under varying sensory modalities and environmental conditions. The tasks are organized into three main categories: *Insertion*, *Screwing*, and *Exploration* manipulation tasks (see Fig. 1). *Insertion* consists of classical assembly tasks, ranging from peg-in-hole insertion to gear assembly. *Screwing* includes nut-and-bolt threading and light-bulb installation.

Exploration involves tasks where visual input is occluded or limited, such as object reorientation and object search within a box, as well as object sorting under dim lighting, requiring the robot to rely primarily on tactile feedback. All task environments are implemented in the IsaacGym simulator [21], and one representative task from each category is replicated in the real-world setup.

2.1.2. Observation modalities

ManiFeel is developed on the IsaacGym platform [21] with the integration of the TacSL framework [22] to simulate robotic manipulators equipped with tactile fingers. TacSL supports high-fidelity simulation of vision-based tactile sensors based on the GelSight R1.5 design [23]. The tactile modality includes two complementary observation channels. The first, tactile RGB images (TacRGB), are high-resolution tactile images that capture surface deformation patterns on the elastomer contact area, providing rich information about the contacted object's surface texture and geometry. The second, tactile force field (TacFF), represents spatial force-field data that describe distributed normal and shear forces across the contact surface, encoding both contact pressure and directional force information. Together, these tactile representations enable fine-grained perception of contact geometry and interaction dynamics during manipulation.

For vision modality, the framework supports multiple camera configurations, such as front, side, or wrist-mounted views, and allows for occluded or degraded conditions to emulate realistic manipulation challenges.

2.1.3. Teleoperated Demonstration Data Collection

We collect human demonstrations for supervised policy learning using a 3Dconnexion SpaceMouse[1]. Tactile feedback plays a critical role during data collection, especially for tasks requiring precise contact reasoning or operating under visual occlusion and degraded lighting conditions. In these settings, operators rely heavily on real-time tactile observations displayed on a monitor to perceive subtle contact events for effective object manipulation, classification, and exploration.

2.2. Modular Pipeline Design

In this section, we present the modular pipeline design that forms the basis of our benchmark. The benchmark is designed to systematically evaluate the performance of visuotactile policies which are composed of different sensing modalities, tactile representations, and policies, across a broad set of tasks. This modular structure enables flexible combinations and fair comparisons, ultimately facilitating a deeper understanding of the design principles behind effective visuotactile policy learning. Moreover, the modularity of our design provides strong flexibility for integrating future advancements, such as new sensing modalities, novel tactile representation learning methods, and emerging policy architectures.

Choice of Modalities. To investigate the contribution of different sensing modalities, ManiFeel enables direct comparison across three sensing configurations: **vision-only**, **vision with TacRGB**, and **vision with TacFF**. ManiFeel evaluates pairwise combinations of vision with a single tactile modality (either TacRGB or TacFF) to ensure clear and interpretable comparisons. By isolating each tactile modality, we can analyze how TacRGB and TacFF individually complement visual perception across different tasks and scenarios. Combining vision with both tactile modalities would introduce confounding interactions, making it difficult to attribute performance gains to a specific tactile representation and to understand its task-specific contribution. Additionally, we do not consider a tactile-only baseline because most benchmark tasks require visual information for global localization and pre-contact alignment, and our focus is to understand how tactile sensing augments a vision-based policy rather than replaces it.

Regarding the visuotactile sensing fusion scheme, ManiFeel employs a multimodal observation encoder architecture that integrates visual, tactile, and proprioceptive inputs, as illustrated in Fig. 2. Formally, at each timestep t, the multimodal observation includes both the current and previous observations from each modality, expressed as

$$\mathbf{o}_t = \{\mathbf{I}^{\text{vis}}_{t-1:t},\ \mathbf{I}^{\text{tac}}_{t-1:t},\ \mathbf{S}^{\text{ff}}_{t-1:t},\ \mathbf{o}^{\text{prop}}_{t-1:t}\}, \tag{1}$$

where $\mathbf{I}^{\text{vis}} \in \mathbb{R}^{H\times W\times 3}$ denotes the visual observation from an RGB camera, $\mathbf{I}^{\text{tac}} \in \mathbb{R}^{h\times w\times 3}$ represents the high-resolution tactile image (TacRGB), $\mathbf{S}^{\text{ff}} \in \mathbb{R}^{r\times c\times 3}$ corresponds to the tactile force-field map (TacFF), encoding normal force and two-dimensional shear forces at each discretized contact point, and $\mathbf{o}^{\text{prop}} \in \mathbb{R}^7$ denotes robot proprioception, consisting of the end-effector position and orientation represented in quaternion form. All modalities are normalized to the range $[-1, 1]$ before being passed to their respective encoders. Each sensing modality is processed by a dedicated encoder: f_{vis} for visual observations, f_{tac} for tactile RGB, and f_{ff} for force-field tactile feedback. At each time step t, the encoder receives both the current and previous observations, denoted as $(\cdot)_{t-1:t}$, and computes the corresponding latent embeddings: $\mathbf{z}^{\text{vis}}_t = f_{\text{vis}}(\mathbf{I}^{\text{vis}}_{t-1:t}) \in \mathbb{R}^{d_{\text{vis}}}$, $\mathbf{z}^{\text{tac}}_t = f_{\text{tac}}(\mathbf{I}^{\text{tac}}_{t-1:t}) \in \mathbb{R}^{d_{\text{tac}}}$, $\mathbf{z}^{\text{ff}}_t = f_{\text{ff}}(\mathbf{S}^{\text{ff}}_{t-1:t}) \in \mathbb{R}^{d_{\text{ff}}}$.

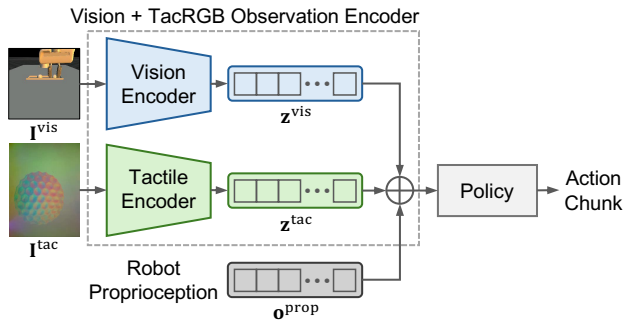
For a given modality configuration (e.g., vision with TacRGB), the encoded latent embeddings are concatenated to form the fused multimodal representation:

$$\mathbf{z}_t = \mathbf{z}^{\text{vis}}_t \oplus \mathbf{z}^{\text{tac}}_t \oplus \mathbf{o}^{\text{prop}}_t \in \mathbb{R}^{d_{\text{vis}}+d_{\text{tac}}+7}, \tag{2}$$

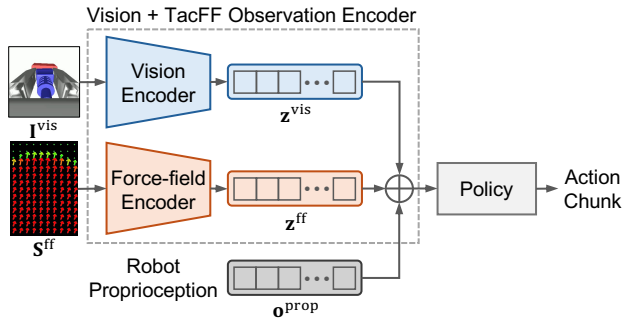
or, for the TacFF configuration,

$$\mathbf{z}_t = \mathbf{z}^{\text{vis}}_t \oplus \mathbf{z}^{\text{ff}}_t \oplus \mathbf{o}^{\text{prop}}_t \in \mathbb{R}^{d_{\text{vis}}+d_{\text{ff}}+7}, \tag{3}$$

[1] https://3dconnexion.com/us/product/spacemouse-compact/



(a) Vision + TacRGB observation encoder.



(b) Vision + TacFF observation encoder.

Figure 2. Multimodal observation encoder architectures used in ManiFeel. Each configuration fuses visual observations ($\mathbf{I}^{\text{vis}}$), tactile feedback ($\mathbf{I}^{\text{tac}}$ or $\mathbf{S}^{\text{ff}}$), and robot proprioception ($\mathbf{o}^{\text{prop}}$) through separate modality encoders. The encoded features ($\mathbf{z}^{\text{vis}}$, $\mathbf{z}^{\text{tac}}$, and $\mathbf{z}^{\text{ff}}$) are concatenated and passed to the policy network to generate action chunks.

where $\oplus$ denotes feature concatenation. The fused representation $\mathbf{z}_t$ is then passed to the policy network to generate the action chunk.

Tactile Representations. While the choice of modality defines the type of tactile input (e.g., TacRGB or TacFF), ManiFeel also supports the integration and comparison of different tactile encoders (f_{tac}) that extract latent embeddings ($\mathbf{z}^{\text{tac}}$) from high-resolution tactile images ($\mathbf{I}^{\text{tac}}$). In this study, we benchmark four representative tactile representations: (1) ResNet18 trained from scratch [1, 24], (2) T3 [10], a pretrained encoder using large-scale tactile datasets, (3) UniT [8], a lightweight representation trained with generalizability, and (4) AnyTouch [25], a generalizable tactile representation learned across diverse sensors and modalities. These comparisons aim to reveal whether pretrained or carefully designed tactile encoders offer meaningful advantages in visuotactile policy learning.

Policies. ManiFeel also supports benchmarking different policy architectures for visuotactile manipulation. In this study, we compare three representative policy types commonly used in robot learning and analyze their integration with different choices of visuotactile inputs: Diffusion Policy [1], Equivariant Diffusion Policy [26], and Flow Matching [27]. These policies differ in their generative modeling approaches and model architectures. Our goal is to understand how each type of policy performs under different input modalities, and whether certain policies are more compatible with the inclusion of tactile sensing. Specifically, we study how tactile signals influence performance across policy architectures.

3. Experiments and Results

We conduct a comprehensive set of experiments to evaluate the effectiveness of diverse visuotactile policies across diverse manipulation tasks using the ManiFeel simulation benchmark and real-world settings. Our study systematically benchmarks different sensory modalities, tactile representations, and policies under varying task conditions within the supervised policy learning framework. Note that all evaluated sensing configurations also include proprioceptive feedback (i.e., the robot end-effector's pose). For simulation results, success rates are reported as the primary evaluation metric across all experiments. Specifically, success rates are averaged over the last ten epochs of training, across 50 different environment initializations, and over three random seeds. This results in each reported success rate being the average of 1500 evaluations, providing strong statistical significance. For real-world experiments, we perform 15 consecutive rollouts and record the success rate. The initial environment conditions are varied for each rollout. Through extensive experiments, we aim to address four central questions in supervised visuotactile policy learning:

1. How does tactile sensing enhance performance in contact-rich and visually constrained manipulation tasks?
2. How do different tactile modalities contribute to policy performance across diverse tasks?
3. How consistent are the observed behaviors and outcomes between simulation and real-world environments?
4. Lessons learned and design principles for developing efficient and robust visuotactile policies

These questions guide our analysis toward understanding the role of tactile sensing in multimodal policy learning and identifying key challenges and future directions for advancing multimodal robotic manipulation.

3.1. Modality Benchmarking

Setup. This section evaluates policy performance under three sensing configurations, vision only, vision with TacRGB, and vision with TacFF, across nine manipulation tasks, including four insertion tasks, two screwing tasks, and three exploration tasks, as defined in the ManiFeel simulation benchmark (see Section 2.1). In this setup, wrist-mounted camera views are used for the insertion and screwing task categories, while external occluded camera views

Table 1. Success rates (mean ± std) averaged over three seeds across tasks and sensing configurations. Best in each column is bold.

Sensing Configuration	**Insertion**				**Screwing**		**Exploration**		
	Peg Insertion	USB Insertion	Power Plug Insertion	Gear Assembly	Nut–bolt Assembly	Bulb Installation	Peg Reorientation (Occluded)	Object Search (Occluded)	Ball Sorting (Dim)
Vision only	0.14 ± 0.01	0.49 ± 0.03	0.51 ± 0.02	0.63 ± 0.02	0.80 ± 0.03	0.76 ± 0.02	0.29 ± 0.02	0.52 ± 0.04	0.57 ± 0.03
Vision + TacRGB	0.21 ± 0.02	0.49 ± 0.02	0.51 ± 0.01	0.57 ± 0.02	0.78 ± 0.03	0.72 ± 0.04	$\mathbf{0.39 \pm 0.02}$	$\mathbf{0.69 \pm 0.01}$	$\mathbf{0.72 \pm 0.02}$
Vision + TacFF	$\mathbf{0.40 \pm 0.02}$	$\mathbf{0.56 \pm 0.01}$	$\mathbf{0.58 \pm 0.02}$	$\mathbf{0.66 \pm 0.01}$	$\mathbf{0.81 \pm 0.03}$	$\mathbf{0.81 \pm 0.02}$	0.28 ± 0.02	0.52 ± 0.02	0.60 ± 0.02

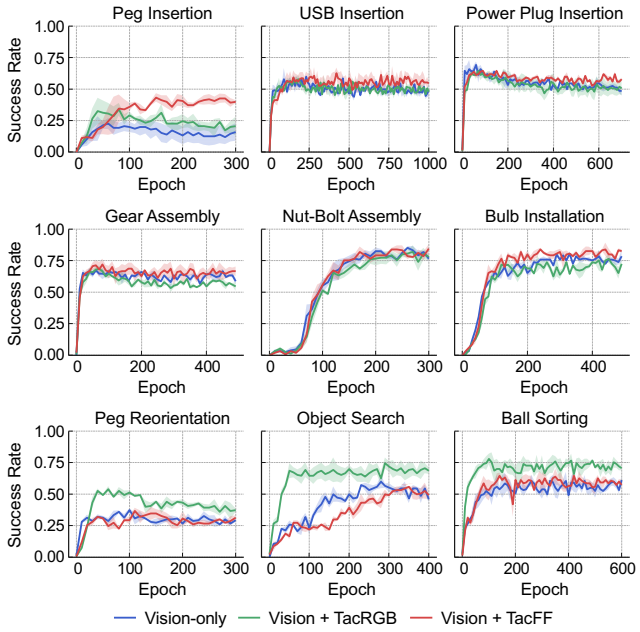


Figure 3. Learning curves comparing the three different sensing configurations (i.e., Vision-only, Vision + TacRGB, and Vision + TacFF) across 9 simulation manipulation tasks. Each plot shows success rates over training epochs. Note that all sensing configurations also include robot proprioception inputs.

are employed for the exploration tasks, particularly for peg reorientation and object search. The ball sorting task is conducted under dim lighting conditions to simulate visually degraded perception. This setup emulates the limited visual information available to humans when operating in similarly vision-occluded or low-light environments and highlights the challenges faced by visual perception in such scenarios. All policies use a ResNet18 tactile encoder trained from scratch, and the policy is based on the Diffusion Policy [1], a widely adopted and well-established framework for supervised policy learning.

Results. Table 1 reports the success rates of the policy evaluations for the three aforementioned sensing configurations across the nine simulation benchmark tasks. Corresponding learning curves for all tasks under these configurations are shown in Fig. 3. Each reported success rate represents the average of approximately 1500 evaluations (3 random seeds, 50 environment initializations, and the last 10 training epochs), ensuring strong statistical significance. Even small percentage-point differences, therefore, reflect meaningful performance gains. In the insertion category, peg insertion success increases by 26 percentage points when TacFF is added, showing that force-based tactile feedback is essential for precise alignment and control. In the screwing category, bulb installation improves by about 5 percentage points with TacFF, indicating that tactile force cues enhance control during tightening. For exploration tasks, TacRGB provides the greatest benefit, improving object search by approximately 17 percentage points and ball sorting under dim lighting by around 15 percentage points compared to vision only. These results address the first research question by demonstrating that tactile sensing provides complementary feedback that enables the robot to perform more reliably both contact-rich manipulation tasks, such as insertion and screwing, and perception-challenging exploration tasks, such as peg reorientation, object search, and ball sorting, in visually-degraded environments.

However, not all tactile information contributes equally across task types. The performance gain of each tactile modality depends strongly on the nature of the task and the characteristics of the required manipulation skill. Tasks involving precise contact alignment, geometric fitting, and force regulation, such as insertion and screwing, benefit the most from TacFF feedback. This representation encodes distributed normal and shear forces across the contact surface, providing physically grounded cues that help the policy maintain stable contact and perform fine alignment. In contrast, tasks that rely more on perceiving surface texture and geometry, such as object search or ball sorting, depend more heavily on TacRGB signals. In these cases, high-resolution tactile images capture detailed surface deformation and shape information. For instance, in ball sorting under dim lighting, the success rate increases from 57% (vision only) to 72% with TacRGB, outperforming TacFF (60%). These modality-specific trends address the second research question, showing that the impact of tactile sensing depends not only on its inclusion but also on the type of tactile information and the manipulation context in which it is applied.

That said, adding more sensory data does not always

Table 2. Success rates (mean ± std) averaged over 3 seeds across tasks and modalities. Results are shown for three policy architectures (DP, EquiDP, FM). Bold indicates the best performance within each task and architecture.

Task	DP			EquiDP			FM		
	Vision only	Vision + TacRGB	Vision + TacFF	Vision only	Vision + TacRGB	Vision + TacFF	Vision only	Vision + TacRGB	Vision + TacFF
Gear Assembly	0.63 ± 0.02	0.57 ± 0.02	**0.66 ± 0.01**	0.70 ± 0.02	0.66 ± 0.02	**0.74 ± 0.03**	0.38 ± 0.01	0.37 ± 0.01	**0.39 ± 0.02**
Bulb Installation	0.76 ± 0.02	0.72 ± 0.04	**0.81 ± 0.02**	0.94 ± 0.01	0.94 ± 0.01	**0.95 ± 0.02**	0.60 ± 0.03	0.63 ± 0.03	**0.66 ± 0.02**
Object Search	0.52 ± 0.04	**0.69 ± 0.01**	0.52 ± 0.02	0.44 ± 0.08	**0.67 ± 0.04**	0.46 ± 0.07	0.16 ± 0.03	**0.44 ± 0.03**	0.15 ± 0.03
Ball Sorting	0.43 ± 0.03	**0.73 ± 0.03**	0.50 ± 0.03	0.21 ± 0.03	**0.46 ± 0.02**	0.31 ± 0.04	0.41 ± 0.03	**0.69 ± 0.01**	0.43 ± 0.03

guarantee improved performance. In certain cases, such as the gear assembly task, the vision-only policy (63%) achieves higher performance than the vision + TacRGB configuration (57%). When the additional sensory input does not provide task-relevant or well-structured information, it can introduce noise and distract the policy during training. Complex tactile image patterns may cause the model to overfit or attend to visually detailed but semantically weak features, ultimately degrading performance. In contrast, TacFF provides a more compact and physically meaningful description of contact interaction that the policy can more easily associate with action outcomes. Consequently, integrating TacFF generally leads to consistent improvements, while TacRGB may sometimes reduce performance when the task does not require detailed surface information.

Together, these results reveal how tactile sensing enhances robotic policy learning in contact-rich or visually ambiguous conditions while demonstrating the importance of selecting and integrating modalities appropriately. They highlight a key challenge for future visuotactile policy learning: effective multimodal integration must emphasize structured, task-relevant tactile representations to fully realize the benefits of touch in robotic manipulation.

3.2. Policy Benchmarking

Setup. In this section, we benchmark three policy architectures: Diffusion Policy (DP) [1], Equivariant Diffusion Policy (EquiDiff) [26], and Flow Matching (FM) [27], across four representative manipulation tasks selected from the three task categories: gear assembly, bulb installation, object search, and ball sorting. Each policy is evaluated under the three aforementioned sensing configurations. For tactile input, we use a ResNet18 encoder trained from scratch, while for visual input, DP and FM employ a standard ResNet18 encoder and EquiDiff uses an equivariant ResNet18 [26].

Results. As shown in Table 2, all three policy architectures exhibit consistent trends across tasks when tactile sensing is incorporated. The corresponding learning curves for these policy configurations are provided in Fig. 4. Adding TacFF generally improves performance in contact-rich insertion and screwing tasks, while TacRGB provides greater benefit in exploration tasks where visual input is limited. For instance, in gear assembly, TacFF increases success by about 3 percentage points for DP and 4 percentage points for EquiDP compared to their vision-only baselines, reflecting the importance of force-based tactile feedback for precise alignment and contact regulation. In contrast, TacRGB yields the highest gains in object search and ball sorting, improving EquiDP vision-only performance by 23 and 25 percentage points, and FM vision-only performance by 28 percentage points in both tasks, respectively. These improvements confirm the advantage of high-resolution tactile images in providing detailed surface information that compensates for degraded visual cues.

Across architectures, FM performs worse than DP and EquiDP in most benchmarking tasks. Between the diffusion-based methods, EquiDP consistently achieves higher success rates in insertion and screwing tasks, such as bulb installation and gear assembly. This advantage likely arises because these tasks rely heavily on visual feedback for coarse localization and alignment before tactile sensing becomes critical for fine contact adjustment. The equivariant encoder in EquiDP effectively captures spatial symmetries from wrist-mounted camera observations, allowing it to leverage visual geometry more efficiently. In such tasks, improvements in vision-only performance often translate proportionally to gains in visuotactile policies. In contrast, exploration tasks, including object search and ball sorting, depend more on tactile sensing for object recognition and classification under occluded or low-light conditions. In these cases, EquiDP offers limited benefits, as its strength in visual geometric reasoning is less applicable to the tactile-driven fusion required for these scenarios. As a result, EquiDP sometimes performs slightly worse than the standard DP, which demonstrates stronger robustness when visual cues are unreliable and achieves more effective fusion of high-fidelity tactile sensing for visuotactile policies.

Additionally, as discussed in the modality benchmarking results in Section 3.1, adding TacRGB may negatively affect performance when its high-dimensional deformation images provide limited task-relevant information, particularly in contact-rich manipulation tasks. A similar effect appears

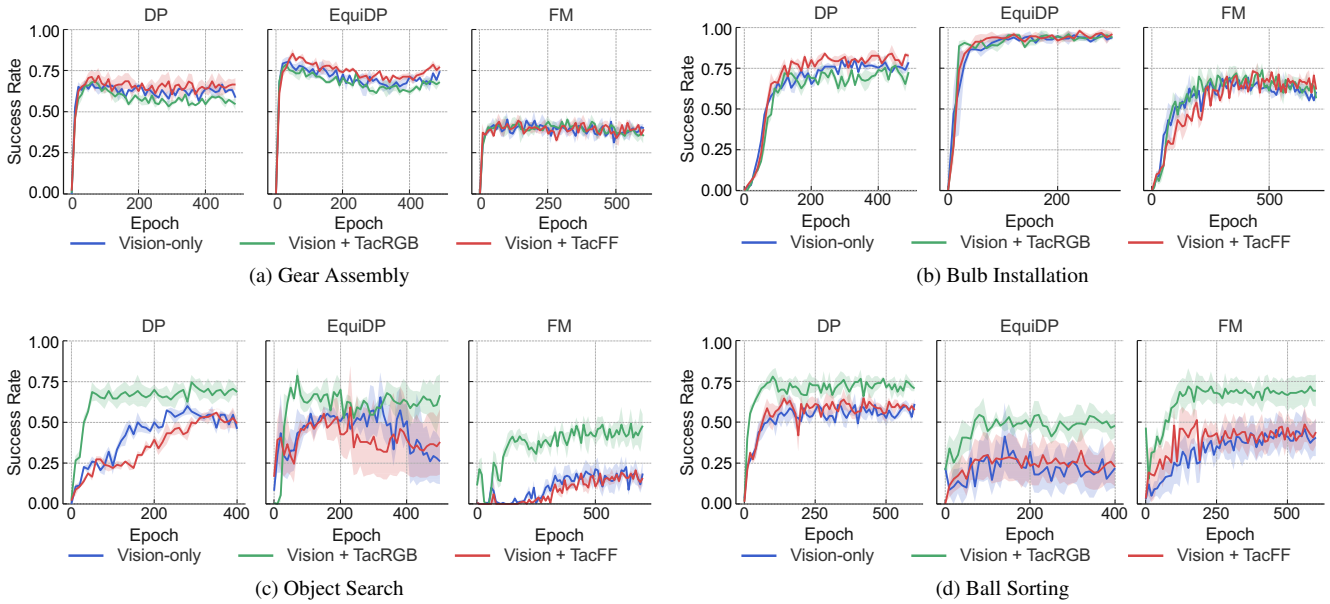


Figure 4. Learning curves comparing the three different sensing configurations (i.e., Vision-only, Vision + TacRGB, and Vision + TacFF) across different policy architectures, for four representative tasks: (a) Gear Assembly, (b) Bulb Installation, (c) Object Search, and (d) Ball Sorting

in bulb installation, where the drop is most pronounced for DP. This suggests that DP is more sensitive to the variability introduced by TacRGB, whereas EquiDP and FM exhibit greater robustness due to their equivariant spatial encoding and smoother optimization dynamics. However, this robustness does not override a fundamental modality-task mismatch. When the tactile modality does not supply the type of information the task relies on, architectural differences may reduce sensitivity but cannot guarantee improved performance or fully eliminate degradation, for example, as shown in the gear assembly task where all architectures still suffer degradation. This highlights the importance of prioritizing the selection of tactile modalities whose sensing characteristics align closely with the demands of the manipulation skill before considering policy architecture design.

Overall, these results confirm that tactile sensing enhances performance across architectures and tasks, but the extent of improvement depends on both the tactile modality and the policy design.

3.3. Tactile Representation Benchmarking

Setup. We present simulation benchmarking results for vision–TacRGB policies equipped with different tactile representations. These tactile representations encode high-resolution tactile images (TacRGB) into feature-rich embeddings. The evaluation is conducted across three representative manipulation tasks: gear assembly, object search, and ball sorting. The ResNet18 encoder is trained from scratch, while UniT [8], T3 [10], and AnyTouch [25] are pre-trained tactile encoders that are fine-tuned during policy training.

Table 3. Success rates (mean $\pm$ std) for different tactile representation encoders across representative manipulation tasks in simulation. Results are averaged over three seeds. Best performance in each task is highlighted in bold.

Task	ResNet18	UniT	T3	AnyTouch
Gear Assembly	0.57 ± 0.02	$\mathbf{0.61 \pm 0.01}$	0.57 ± 0.02	0.28 ± 0.02
Object Search	0.69 ± 0.01	0.58 ± 0.03	$\mathbf{0.71 \pm 0.03}$	0.59 ± 0.02
Ball Sorting	0.73 ± 0.03	$\mathbf{0.80 \pm 0.05}$	0.68 ± 0.01	0.74 ± 0.02

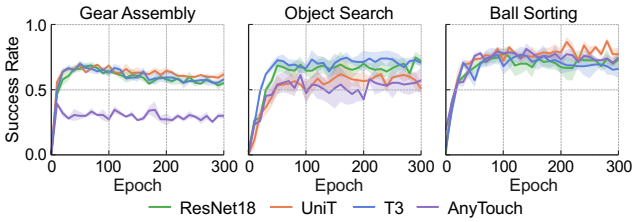


Figure 5. Learning curves across vision–TacRGB policies using different tactile representations, for three representative tasks, namely gear assembly, object search, and ball sorting.

Results. Table 3 reports the success rates of vision-TacRGB policies with different tactile representation encoders across the three representative tasks. The corresponding learning curves are shown in Fig. 5. The re-

sults indicate that pre-trained tactile encoders such as UniT and T3, which are specifically designed for high-resolution tactile image processing, consistently outperform the baseline ResNet18 trained from scratch. This improvement is especially pronounced in tasks requiring dynamic contact reasoning or texture-dependent recognition, such as object search and ball sorting, where UniT and T3 achieve the highest success rates. Among the evaluated representations, UniT shows the most balanced performance across all tasks, including the force-sensitive gear assembly. AnyTouch demonstrates strong performance in texture-based reasoning tasks but struggles in tasks that demand fine force-sensitive control, suggesting that current pre-trained tactile encoders may still be specialized toward specific sensing contexts.

Overall, the results emphasize the importance of leveraging pre-trained tactile representations to enhance visuotactile policy learning. However, the performance variation across tasks also reveals that a unified tactile encoder capable of generalizing across diverse manipulation scenarios remains an open challenge. While pretrained visual encoders have demonstrated strong transferability in policy learning [2, 28, 29], this trend does not consistently extend to tactile representations. Although several tactile representation models have been specifically designed for policy learning, their performance remains inconsistent across tasks and modalities. This highlights the ongoing need for more robust and generalizable tactile representation models that can effectively support policy learning in diverse contact conditions. Developing such representations that can seamlessly adapt to both spatial contact interactions and texture-dependent tactile cues will be crucial for advancing robust visuotactile robot learning. We envision ManiFeel as a reproducible testbed to facilitate this progress, similar to how standardized benchmarks in vision have accelerated the development of transferable visual representations.

3.4. Real-world Evaluation

Setup. To assess the validity of the proposed ManiFeel simulation benchmark, we examine the consistency between simulated and real-world performance across three representative tasks: gear assembly, bulb installation, and ball sorting under both normal and dim lighting conditions. All policies are trained from scratch using real-world datasets collected specifically for the real-world settings. We benchmark the three sensing configurations introduced earlier (vision only, vision + TacRGB, and vision + TacFF), using ResNet18 encoders trained from scratch and Diffusion Policy as the policy backbone. Experiments are conducted using a Franka Emika Panda manipulator[2] equipped with a parallel gripper integrated with a GelSight R1.5 tactile sensor [23]. In the real setup, the tactile RGB signal ($\mathbf{I}^{\text{tac}}$) is directly captured as raw high-resolution images from the GelSight R1.5 sensor, providing detailed surface deformation patterns within the contact region. To obtain the tactile force-field map ($\mathbf{S}^{\text{ff}}$), we estimate local contact forces by combining marker-based optical flow and depth reconstruction following [23]. Specifically, two-dimensional marker displacements encode tangential shear components, while the reconstructed surface depth provides the normal indentation magnitude, together forming a spatial map that encodes distributed normal and shear forces across the contact surface. Both tactile modalities are normalized to the range $[-1, 1]$ before being passed to their respective encoders, ensuring consistency in dynamic range between real and simulated TacRGB and TacFF signals.

[2] https://franka.de/products/franka-research-3

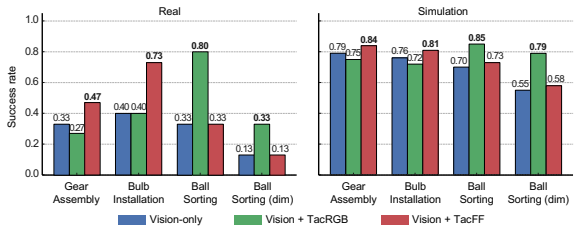


Figure 6. Success rates for real and simulated environments across different sensing configurations and scenarios.

Results. Figure 6 shows the comparison between simulated and real environments across three sensing configurations. The overall trends observed in simulation are consistently reproduced in the real world, demonstrating that ManiFeel's task suite captures representative and transferable manipulation behaviors. In both domains, incorporating tactile sensing substantially improves policy performance, particularly in contact-rich and visually constrained scenarios. For instance, in gear assembly, TacFF improves success by 14 percentage points over the vision-only baseline (from 33% to 47%) in real experiments, reflecting similar improvements in simulation. Likewise, TacFF achieves the highest success in bulb installation (73% in real, 81% in sim), confirming the importance of force-based tactile feedback for accurate tightening and contact regulation.

Overall, these results confirm that ManiFeel produces consistent and interpretable performance trends between simulation and reality, validating that our simulation benchmark effectively reflects the nature of real-world visuotactile manipulation. While reproducing identical quantitative results across domains is inherently impractical and beyond the scope of this work, the consistency of relative trends demonstrates that ManiFeel provides a reliable and meaningful foundation for benchmarking and future visuotactile policy development.

References

[1] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Proceedings of Robotics: Science and Systems*, 2023.

[2] 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," *arXiv preprint arXiv:2402.10329*, 2024.

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

[4] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," *arXiv preprint arXiv:2401.02117*, 2024.

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

[6] L. Wang, J. Zhao, Y. Du, E. H. Adelson, and R. Tedrake, "Poco: Policy composition from and for heterogeneous robot learning," in *Proceedings of Robotics: Science and Systems*, 2024.

[7] 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 *Proceedings of Conference on Robot Learning*, 2024.

[8] Z. Xu, R. Uppuluri, X. Zhang, C. Fitch, P. G. Crandall, W. Shou, D. Wang, and Y. She, "UniT: Data efficient tactile representation with generalization to unseen objects," 2025. [Online]. Available: https://arxiv.org/abs/2408.06481

[9] 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.

[10] J. Zhao, Y. Ma, L. Wang, and E. H. Adelson, "Transferable tactile transformers for representation learning across diverse sensors and tasks," in *Proceedings of Conference on Robot Learning*, 2024.

[11] I. Guzey, Y. Dai, B. Evans, S. Chintala, and L. Pinto, "See to touch: Learning tactile dexterity through visual incentives," in *Proceedings of International Conference on Robotics and Automation*, 2024.

[12] T. Lin, Y. Zhang, Q. Li, H. Qi, B. Yi, S. Levine, and J. Malik, "Learning visuotactile skills with two multi-fingered hands," in *Proceedings of International Conference on Robotics and Automation*, 2025.

[13] R. Bhirangi, V. Pattabiraman, E. Erciyes, Y. Cao, T. Hellebrekers, and L. Pinto, "Anyskin: Plug-and-play skin sensing for robotic touch," *arXiv preprint arXiv:2409.08276*, 2024.

[14] W. Yang, A. Angleraud, R. S. Pieters, J. Pajarinen, and J.-K. Kämäräinen, "Seq2seq imitation learning for tactile feedback-based manipulation," in *Proceedings of IEEE International Conference on Robotics and Automation*, 2023, pp. 5829–5836.

[15] Y. Hou, Z. Liu, C. Chi, E. Cousineau, N. Kuppuswamy, S. Feng, B. Burchfiel, and S. Song, "Adaptive compliance policy: Learning approximate compliance for diffusion guided control," *arXiv preprint arXiv:2410.09309*, 2024.

[16] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li, "3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing," in *Proceedings of Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=bk28WlkqZn

[17] Z. Liu, C. Chi, E. Cousineau, N. Kuppuswamy, B. Burchfiel, and S. Song, "Maniwav: Learning robot manipulation from in-the-wild audio-visual data," *arXiv preprint arXiv:2406.19464*, 2024.

[18] 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 *Proc. Conf. Robot Learn.*, 2022, pp. 1678–1690.

[19] A. Mandlekar, S. Nasiriany, B. Wen, I. Akinola, Y. Narang, L. Fan, Y. Zhu, and D. Fox, "Mimicgen: A data generation system for scalable robot learning using human demonstrations," *arXiv preprint arXiv:2310.17596*, 2023.

[20] S. Tao, F. Xiang, A. Shukla, Y. Qin, X. Hinrichsen, X. Yuan, C. Bao, X. Lin, Y. Liu, T.-k. Chan *et al.*, "Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai," *arXiv preprint arXiv:2410.00425*, 2024.

[21] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, and G. State, "Isaac gym: High performance gpu-based physics simulation for robot learning," 2021. [Online]. Available: https://arxiv.org/abs/2108.10470

[22] I. Akinola, J. Xu, J. Carius, D. Fox, and Y. Narang, "Tacsl: A library for visuotactile sensor simulation and learning," *IEEE Transactions on Robotics*, vol. 41, pp. 2645–2661, 2025.

[23] S. Wang, Y. She, B. Romero, and E. Adelson, "GelSight Wedge: Measuring high-resolution 3D contact geometry with a compact robot finger," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2021, pp. 6468–6475.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016.

[25] R. Feng, J. Hu, W. Xia, T. Gao, A. Shen, Y. Sun, B. Fang, and D. Hu, "Anytouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors," *arXiv preprint arXiv:2502.12191*, 2025.

[26] D. Wang, S. Hart, D. Surovik, T. Kelestemur, H. Huang, H. Zhao, M. Yeatman, J. Wang, R. Walters, and R. Platt, "Equivariant diffusion policy," in *8th Annual Conference on Robot Learning*, 2024.

[27] Q. Zhang, Z. Liu, H. Fan, G. Liu, B. Zeng, and S. Liu, "Flowpolicy: Enabling fast and robust 3d flow-based policy via consistency flow matching for robot manipulation," 2024. [Online]. Available: https://arxiv.org/abs/2412.04987

[28] J. Pari, N. M. Shafiullah, S. P. Arunachalam, and L. Pinto, "The surprising effectiveness of representation learning for visual imitation," in *Proceedings of Robotics: Science and Systems*, 2022.

[29] J. Barreiros, A. Beaulieu, A. Bhat, R. Cory, E. Cousineau, H. Dai, C.-H. Fang, K. Hashimoto, M. Z. Irshad, M. Itkina *et al.*, "A careful examination of large behavior models for multitask dexterous manipulation," *arXiv preprint arXiv:2507.05331*, 2025.