Multi-Resolution Tactile Imitation Learning for Contact-Rich Robotic Manipulation

Rickmer Krohn[1,3,4], **Erik Helmut**[2], **Niklas Funk**[2],
Jan Peters[2,3,4,5], **Vignesh Prasad**[1,3,4] **and Georgia Chalvatzaki**[1,3,4]

[1] Interactive Robot Perception & Learning, TU Darmstadt
[2] Intelligent Autonomous Systems, TU Darmstadt
[3] Hessian AI, [4] Robotics Institute Germany
[5] German Research Center for AI (DFKI)

https://mitas-touch.github.io

Abstract: Touch sensing is beneficial for solving a wide variety of manipulation tasks. While there exists a wide range of tactile sensors with different properties, exploiting the fusion of multiple heterogeneous tactile sensors to improve manipulation learning remains underexplored. We present Multi-Resolution Tactile Sensing (MiTaS), a representation framework that leverages multiple tactile sensors operating at different temporal resolutions in order to solve complex contact-rich manipulation tasks. We propose an effective architecture using modality-specific convolutional stems and transformer-based fusion that effectively integrates information from an RGB camera stream, a vision-based GelSight Mini sensor and a high-frequency event-based Evetac sensor. This multisensory representation then conditions a flow-matching policy for solving downstream tasks. Our systematic study across five contact-rich manipulation tasks demonstrates the effectiveness of multi-resolution tactile features in imitation learning wherein MiTaS achieves an average success rate of 80 %, while vision-only (35.5 %) and visuotactile (55.5 %, 63 %) baselines cannot solve the task reliably.



Figure 1: MiTaS combines an RGB camera (blue), a GelSight Mini (red) and an event-based Evetac sensor (yellow) mounted on the gripper. Our method fuses these multi-resolution tactile streams to condition a flow-matching policy, enabling improved control in contact-rich manipulation tasks such as key in lock.

10th Conference on Robot Learning (CoRL 2026), Austin TX, USA.

1 Introduction

Humans rely on multi-modal touch sensing to accomplish delicate manipulation tasks, seamlessly integrating information from diverse mechanoreceptors operating at different temporal and spatial resolutions [1]. Similarly, robotic systems can benefit from heterogeneous tactile sensing to understand objects, tasks, and contact dynamics, enabling precise manipulation capabilities.

RGB vision-based tactile sensors such as GelSight [2, 3, 4] have shown remarkable results in manipulation by providing high-resolution spatial information about local contact geometry, deformation, and surface structure [5, 6, 7, 8, 9, 10, 11]. However, such frame-based tactile sensing is less suited to capturing very fast contact changes, such as impact, incipient slip, high-frequency vibration. These changes may occur between frames or be smoothed out by conventional image-based processing. Recent advances in tactile sensing technology have also introduced event-based tactile cameras such as Evetac [12] or GelEvent [13], which offer high temporal resolution to capture rapid contact dynamics that may otherwise go unnoticed by traditional vision-based sensors. These high-frequency signals can detect critical events such as early contact, incipient slip, vibrations during insertion, and subtle variations during sliding contact.

Given the complementary strengths of classic vision-based and event-based sensors, a crucial question arises: can robotic systems benefit from leveraging heterogeneous tactile sensing capabilities to better understand intricate contact dynamics and improve manipulation learning?

In this work, we present **M**ulti-Resolution **Ta**ctile **S**ensing (MiTaS), a comprehensive representation framework to condition a flow matching imitation learning policy to leverage multiple tactile sensors, operating at different temporal resolutions, to solve complex manipulation tasks. Our key insight is that different tactile modalities capture complementary aspects of contact-rich interactions: standard vision-based sensors provide high spatial resolution of contact geometry, while event-based sensors capture high-frequency temporal dynamics. We introduce a effective architecture that extracts and fuses heterogeneous tactile features through modality-specific CNN-stems and transformer-based fusion. Additionally, we propose a multi-tactile dropout scheme that boosts the performance of a classic visuotactile policy by leveraging Evetac during training, without requiring it at inference.

Across five contact-rich manipulation tasks, policies using multiple heterogeneous tactile sensors consistently outperform vision-only baselines. MiTaS achieves an average success rate of 80%, whereas vision-only policies fail to solve the tasks reliably, highlighting the benefit of combining complementary tactile sensors for both dynamic and fine-grained manipulation. An analysis of attention scores throughout task execution reveals how sensor importance shifts during different phases of manipulation, validating our multi-resolution sensing hypothesis. To the best of our knowledge, this is the first work to combine a classic vision-based tactile sensor with an event-based tactile sensor, thereby showing how complementary tactile features enable complex contact-rich manipulation.

Our contributions are the following: (**1**) A effective architecture for extracting and fusing heterogeneous tactile sensor readings operating at different temporal resolutions; (**2**) A multi-tactile dropout scheme that improves visuotactile policy performance on contact-rich tasks where reactivity to small perturbations is key; (**3**) A multi-resolution tactile dataset combining egocentric RGB, vision-based, and event-based tactile sensing across five challenging contact-rich manipulation tasks.

2 Related Work

Tactile Sensing for Robotic Manipulation. Touch feedback has become increasingly important for contact-rich robotic manipulation [14], particularly in settings where local contact state cannot be reliably inferred from vision alone [5, 15, 16, 17, 18, 19, 20, 7, 21]. Different families of tactile sensors have been successfully used to improve manipulation policies. Vision-based tactile sensors, such as GelSight [2], provide dense spatial information about contact geometry [22, 23], while event-based tactile sensors such as Evetac [12] and GelEvent [13] capture high-frequency contact changes.

As tactile sensing has gained traction, reactive diffusion policies have incorporated visuotactile feedback for contact-rich manipulation [24], demonstrating the value of multi-modal conditioning. Large-scale datasets such as Touch100k [25] have enabled touch-centric multimodal representation learning by providing diverse vision-language-touch data. Recent work has explored representations that go beyond pixel-level tactile information [26] and developed tactile-conditioned diffusion policies for force-aware manipulation [27, 7]. Pretraining approaches such as VITaL [28] have demonstrated that visuotactile pretraining benefits both tactile and non-tactile manipulation policies. The principle that "touch begins where vision ends" [29] has motivated generalizable policies for contact-rich manipulation. Self-supervised contrastive pretraining [30] has further advanced multimodal visuotactile representation learning. Deep reinforcement learning with visuotactile feedback [8] has enabled learning of complex manipulation policies, while safe self-supervised learning [31] has made real-world industrial insertion feasible. Tactile sensing also guides object exploration [32, 33, 34]. Recent vision-tactile-language-action models [35] incorporate preference learning for insertion tasks, and world models such as DreamTacVLA [36] enable learning to predict future tactile feedback in simulation [37] and real world. Unified representations across static and dynamic tactile sensors [38] have further demonstrated the potential of multisensory tactile learning. Along similar lines, transferable tactile transformers [39] have shown promise for learning representations across diverse sensors and tasks. Despite these significant advancements in tactile-driven manipulation, most current frameworks rely on a single touch sensor modality. Drawing inspiration from the multimodal nature of human touch, this work bridges this gap by exploring how fusing two distinct tactile sensors with complementary temporal resolutions can improve the robustness of learning robotic manipulation.

Multisensory Learning for Manipulation. The integration of vision and touch has been explored extensively for improving manipulation capabilities [17, 20, 40, 41, 42]. Fusing multiple sensors with stage guidance has been applied to complex tasks such as pouring and peg insertion [43]. The combination of vision and tactile via CNN-stems and masked autoencoding improves visualtactile and vision policies [44]. Sparsh-X [26] is training a representation on the multimodal Digit360 [45] sensor via Self- Supervised Learning. The Fusion of Vision, Audio and Touch via a self-attention layer improves single-modality performance [19, 46]. Visual-force-tactile fusion [47] has enabled gentle intricate insertion tasks by leveraging complementary information from multiple sensing modalities. Multi-modal human tactile demonstrations [9] have been leveraged to improve contact-rich manipulation through imitation learning. Nevertheless, these approaches focus on cross-modal integration (e.g., vision plus audio) rather than on interplay among tactile-sensing modalities. Despite the promise of multisensory learning, fusing co-located tactile modalities with complementary temporal resolutions to condition manipulation policies remains largely unexplored, a gap that MiTaS directly addresses.

3 Multi-Resolution Tactile Sensing

In this section, we present Multi-Resolution Tactile Sensing (MiTaS), a novel framework that leverages multiple heterogeneous tactile sensors operating at different temporal resolutions to learn manipulation policies through imitation. Our proposed approach consists of three main components: (1) a effective architecture utilizing modality-specific CNN stems that encode diverse sensor inputs into unified token representations, (2) a flow-matching policy for action generation and visuomotor control conditioned on these tokens, and (3) a multi-tactile dropout scheme that leverages all available sensors during training to improve downstream policy performance when only a subset of sensors is available at inference time. To the best of our knowledge this is the first work to combine a classic vision-based tactile sensor with an event-based tactile sensor, showing how complementary tactile features enable complex contact-rich manipulation. Figure 2 provides an overview of the MiTaS encoder architecture and policy training.

For training the multisensory visuomotor policies we consider three complementary sensing modalities: egocentric RGB from a wrist-mounted camera, GelSight Mini tactile images [2], and high-rate Evetac tactile frames [12]. At each timestep, the observation is $o_t = \left[O_t^{\text{vision}}, O_t^{\text{gelsight}}, O_t^{\text{evetac}}\right]$,

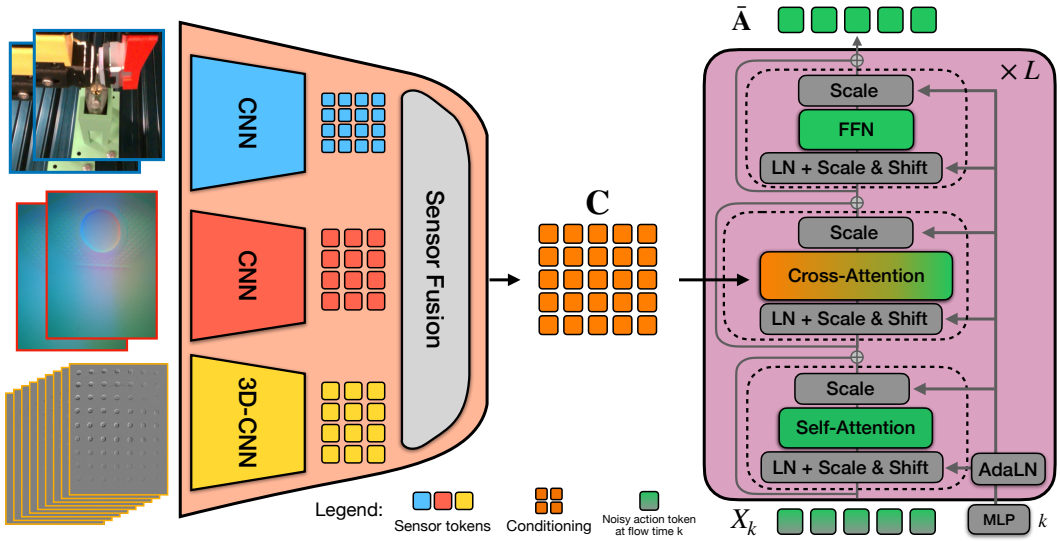


Figure 2: Overview of the MiTaS architecture: Modality-specific CNN-stems encode Vision, GelSight, and Evetac sensors into token embeddings. These tokens are fused via transformer-based self-attention to condition a flow matching policy. The policy does not receive the robot state and must infer the delta position prediction solely from sensor readings.

where $O_t^{\text{vision}} \in \mathbb{R}^{2\times3\times128\times128}$ is recorded at 25 Hz and captures the scene and hand-object configuration, $O_t^{\text{gelsight}} \in \mathbb{R}^{2\times3\times120\times160}$ is recorded at 25 Hz and captures local contact geometry, and $O_t^{\text{evetac}} \in \mathbb{R}^{16\times120\times160}$ stacks 16 event-based frames per timestep from the 200 Hz sensor to capture fast contact dynamics. Vision and GelSight observations are normalized to $[0, 1]$, while Evetac is normalized to $[-\frac{1}{2}, \frac{1}{2}]$ so the default grey frame corresponds to 0 (no event activity) rather than 0.5. These input observations are then passed into our proposed MiTaS Architecture, and we subsequently leverage a flow-matching generative model [48, 49] for predicting a chunk of end-effector delta actions. We want to highlight that since the policy is neither conditioned on robot state nor expressed in absolute position space, it must infer the appropriate relative motion directly from high-dimensional multimodal sensory input. In the following, we first detail the MiTaS encoding architecture, then the policy parameterization for action generation, and finish with our proposed multi-tactile dropout procedure.

3.1 Visual Multi-Tactile Encoding Architecture

Extracting and fusing features from heterogeneous sensor modalities is essential to effectively learn multimodal imitation learning policies. We propose a transformer-based architecture to handle multimodal data during encoding and using cross-attention to condition the policy. Cross-attention allows for efficient integration of high-dimensional data, without introducing a bottleneck on the perception side. The Encoder consists of sensor stems, positional / modality encoding, and sensor fusion. An overview of the Architecture is shown in Figure 2.

Sensor Stems: Each modality is tokenised by a dedicated CNN-stem to extract a meaningful sensor-specific representation. The wrist-camera stacks two RGB frames along the channel axis and applies four strided 2D convolutions, producing a $N_{\text{vis}} = 16 \times 16$ grid of D=256-dimensional tokens. The GelSight stem mirrors this design at the rectangular 120×160 resolution and yields a $N_{\text{gs}} = 12\times16$ token grid. Evetac's 16-frame stack is processed by a four-layer 3D-convolutional stem that jointly compresses time and space into a $N_{\text{evt}} = 12 \times 16$ grid, reflecting the higher temporal density of that modality. The 3D-CNN uses temporal kernels of 3–3–3–2 and strides of 2 in time and space through the first three stages, then a final $(2\times4\times5)$ convolution layer that collapses time to form the token features. The embeddings of the tactile modalities follow the spatial dimension of the sensor observation and allow associations between the different tactile readings later on. Furthermore, early convolutions result in overlapping receptive fields, providing spatially coherent features and have been shown to stabilize training [50].

Positional and modality encoding: Each sensor stem outputs a grid of D-dimensional tokens per sensor; after flattening, every token $\boldsymbol{z}_i$ is augmented with a learned position vector $\boldsymbol{p}_{g(i)}$ that encodes its location within that sensor's token grid, and a modality-specific embedding $\boldsymbol{e}_{s(i)}$ where $s(i) \in \{\text{vision}, \text{gelsight}, \text{evetac}\}$ identifies the source sensor, i.e., $\boldsymbol{z}'_i = \boldsymbol{z}_i + \boldsymbol{p}_{g(i)} + \boldsymbol{e}_{s(i)}$. This separates *where* a token lies in the spatial layout from *which* modality produced it. Per-token positional encodings and modality-specific segment embeddings are learned jointly with the policy.

Sensor Fusion: Rather than pooling each modality into a single global embedding, we retain the full grid of spatially resolved tokens per sensor. The augmented tokens $\boldsymbol{z}'_i$ from Vision (N_{vis}), GelSight (N_{gs}), and Evetac (N_{evt}) are concatenated into a visual multi-tactile sequence of $N_{cond} = 644$ tokens per observation step, preserving fine-grained visual and tactile detail such as local contact geometry. This sequence is processed by a transformer encoder [51] consisting of multi-head self-attention, MLP layers, pre-normalization, and residual connections. The fusion transformer applies full self- and cross-attention across sensors, leading to rich multisensory features that form the conditioning for the policy-head.

3.2 Action Generation through Flow Matching

Next, we utilize the visual multi-tactile embeddings for robot action generation. Following recent advances in learning robotic manipulation policies from expert demonstrations [52, 53], we leverage a conditional flow matching generative model for policy learning [48]. We define our policy output as a fixed-length sequence of future robot commands $\boldsymbol{A_t} = [\boldsymbol{a}_t, \boldsymbol{a}_{t+1}, \ldots, \boldsymbol{a}_{t+H-1}] \in \mathbb{R}^{H \times d_a}$, with horizon H and action dimension d_a. The individual actions are *delta* commands in the robot local frame, i.e., $\boldsymbol{a} = \Delta\boldsymbol{p} = [\Delta x, \Delta y, \Delta z, \Delta\psi] \in \mathbb{R}^4$. This formulation controls translation and the yaw rotation around the z-axis, while keeping pitch and roll fixed. The control loop executes only the first $n_{\text{act}} \leq H$ predicted steps before replanning, improving reactivity while retaining a temporally coherent training target. The policy head is a Diffusion-Transformer (DiT) that regresses a time-conditioned velocity field along a straight noise-to-action path.

Flow-matching training and inference: During training, we start with sampling an initial random noise vector $\boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and we also sample a desired normalized action chunk from our available demonstrations, i.e., $\boldsymbol{x}_1 = \bar{\boldsymbol{A}}$. Following the conditional flow-matching paradigm [49], we define a straight-line probability path $\boldsymbol{x}_k$ and its corresponding constant target velocity field $\boldsymbol{u}_k$ for $k \in [0, 1]$ as: $\boldsymbol{x}_k = k \cdot \boldsymbol{x}_1 + (1-k) \cdot \boldsymbol{x}_0$, and $\boldsymbol{u}_k = \mathrm{d}\boldsymbol{x}_k / \mathrm{d}t$. The training objective is to optimize the network parameters θ such that its output $\hat{\boldsymbol{v}}_\theta(\boldsymbol{x}_k, k, \boldsymbol{C})$ that is conditioned on the current intermediate action state $\boldsymbol{x}_k$, the flow time k, and the vision and multi-tactile tokens $\boldsymbol{C}$ accurately regresses this target velocity field, i.e., we minimize mean-squared error $\|\hat{\boldsymbol{v}}_\theta(\boldsymbol{x}_k, k, \boldsymbol{C}) - \boldsymbol{u}_k\|^2$ over all the flow time steps across all demonstrations. During inference, we generate actions by integrating the learned velocity field via explicit Euler integration starting from an initial noise vector, $\boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and iterate $\boldsymbol{x}_{k+\Delta k} \leftarrow \boldsymbol{x}_k + \Delta k\, \hat{\boldsymbol{v}}_\theta(\boldsymbol{x}_k, k, \boldsymbol{C})$ for $k \in [0, 1)$ with a fixed $\Delta k = 1/K$.

DiT action head: The denoiser network ultimately outputs $\hat{\boldsymbol{v}}_\theta(\boldsymbol{x}_k, k, \boldsymbol{C})$, thereby operating on an intermediate action state $\boldsymbol{x}_k \in \mathbb{R}^{H \times d_a}$. Each horizon slot is linearly projected to a D-dimensional token and summed with a learned positional embedding over the chunk, yielding $\boldsymbol{X}_k \in \mathbb{R}^{H \times D}$. A stack of L DiT blocks updates $\boldsymbol{X}_k$ while conditioning on multisensory tokens $\boldsymbol{C} \in \mathbb{R}^{N_{\text{cond}} \times D}$ produced by the perception encoder; queries come from the action tokens and keys/values from $\boldsymbol{C}$, and we use cross-attention without an intermediate projection since both streams share dimension D. Each block applies (i) multi-head self-attention over the H action tokens, (ii) multi-head cross-attention to $\boldsymbol{C}$, and (iii) a token-wise MLP, using residual connection throughout. We condition each block on t via AdaLN-Zero [54, 55] to provide a strong, stable time-dependent modulation while initializing the network near an identity mapping (details in Appendix A.3). The last layer maps the updated tokens back to the final output $\hat{\boldsymbol{v}}_\theta(\boldsymbol{x}_k, k, \boldsymbol{C}) \in \mathbb{R}^{H \times d_a}$.

3.3 Multi-Tactile Dropout

Recent works have shown that training a policy on multisensory data can improve vision-based policies, either with pretraining [44, 28, 56] or masking observations [21]. Building on these insights, we introduce a Multi-Tactile Dropout (MTD) scheme designed to train a visuotactile policy (utilizing vision and GelSight) while exploiting the high-frequency Evetac sensor exclusively during the training phase. Each training batch is constructed by sampling half of the sequences with Evetac observations and the remaining half without. This dropout procedure allows the high-frequency temporal dynamics captured by Evetac to shape the representation space of the sensor-fusion network. Crucially, because the model must simultaneously learn to generate accurate actions in the absence of Evetac, this procedure effectively regularizes the standard visuotactile features to compensate for the missing features extracted from Evetac observations implicitly. The Evetac encoder is omitted entirely during deployment, leaving a streamlined vision-and-gelsight policy that benefits from cross-modal feature alignment established during the training phase.

4 Experimental Evaluation

We evaluate MiTaS on five challenging contact-rich manipulation tasks that require tactile feedback: Gear assembly, Board Wiping, Lamp Installation, Key in Lock, and Lightbulb Connection. Our experiments investigate: (1) overall performance compared to a strong multisensory baseline and vision-only approaches, (2) the advantage of our fusion architecture over two alternative fusion strategies, (3) the contribution of individual sensors through ablation studies, (4) the impact of Multi-Tactile Dropout, and (4) the dynamic importance of different sensors throughout task execution.

Hardware Platform and Data Collection: This Project uses a 7-DoF Franka Panda robot with an attached Robotis RH-P12-RN gripper. All sensors are mounted on the gripper with 3D-printed holders. Vision is obtained by a wrist-mounted Intel RealSense D435 camera. Gelsight Mini and Evetac are attached to one finger each and form the parallel gripper. Our pipeline is built on Nautilus [57] and controls the robot with a Cartesian impedance controller. We collect 30 expert demonstrations by teleoperating the robot via a Spacemouse. The control frequency is 15Hz, and sensor frames are uniformly sampled in a 0.06 s window to form the sensor history per observation step.

Baselines and Sensor ablation: We compare the MiTaS architecture against Sparsh-X [26], a multimodal tactile transformer baseline. Sparsh-X uses 8 unimodal and 4 fusion transformer layers, in which attention bottlenecks [58] are used as a fusion mechanism instead of self-attention. Instead of pretrained models, we deploy our CNN-stems for PolyTouch [46], which combines sensor features via cross-attention and uses the class tokens to condition the policy head. Q-former [59] fuses CNN-stem sensor features via 32 learnable query-tokens, which then form the multisensory latent space. All models are trained end-to-end with the policy objective using the same DiT action-head. Therefore, the models only differ in architecture and sensor settings to shape the multisensory conditioning of the flow matching policy. For the sensor ablation, we simply omit the encoding of the respective sensor(s). Vision-only models resemble the Vision Transformer architecture [60], following the sensor stems from MiTaS and Sparsh-X, we call them ViT-CNN and ViT.

4.1 Task Setups

Gear Assembly [61] requires inserting a cylindrical gear between two placed gears. The gear must be aligned in position and orientation to seat on the holder. The gear-teeth can cause it to jam, which is resolved by adjusting orientation and applying fine-grained position changes. **Board Wiping** entails wiping a marked line from a fixed board. Contact must be maintained along the line throughout the wipe. Insufficient pressure fails to remove the mark, while excessive pressure tilts or wedges the sponge and breaks the planar contact. **Lamp Installation** [62] requires to align a 3D-printed bulb in a socket in order to screw its threaded base into it. Once the threads are engaged, the bulb is rotated about its vertical axis to drive the threads. Loss of axial alignment can cause it to jam before seating. **Key in Lock** involves inserting a key fully into a real lock. The key must align with the keyway in

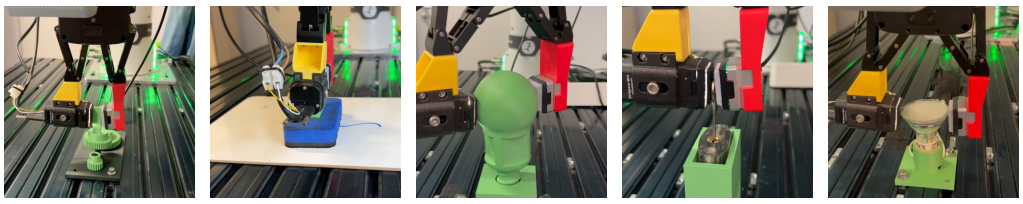


Figure 3: Representative images of the five contact-rich manipulation tasks in order from left to right: Gear Assembly, Board Wiping, Lamp Installation, Key in Lock, and Lightbulb Connection.

both position and orientation. The long and tight insertion path causes even small angular deviations to bind the key against the internal walls. Precise alignment must hold across the full movement until the key is seated. **Lightbulb Connection** consists of grasping a GU10 LED bulb and connecting it to its socket. The two posts at the base of the bulb must align with the socket slots to enter, after which the bulb is rotated about its vertical axis until the post heads slide into their locking slots. See Figure 3 for an overview of the tasks and Figure 6 (Appendix) for full task rollouts.

4.2 Results

We first evaluate overall task performance against the multisensory baseline Sparsh-X and vision-only models. Next, we conduct a sensor fusion study that compares our fusion transformer against alternative strategies for combining visual and tactile features. Followed by a sensor ablation study to showcase the benefits of multi-resolution tactile sensing compared to visuotactile models. To isolate the effect of sensor encoding, all models share the same flow-matching policy head and are trained on the same 30 demonstrations, differing only in which and how the sensor inputs are encoded into conditioning tokens. Lastly, we evaluate our multi-tactile dropout scheme against the sensor ablation from above. We evaluate every policy over 20 trials and report the success rate. The main results in Figure 4 are averaged over two independent datasets, totaling 40 trials per model, which yields a more robust comparison. A table with all results can be found in the Appendix A.8.

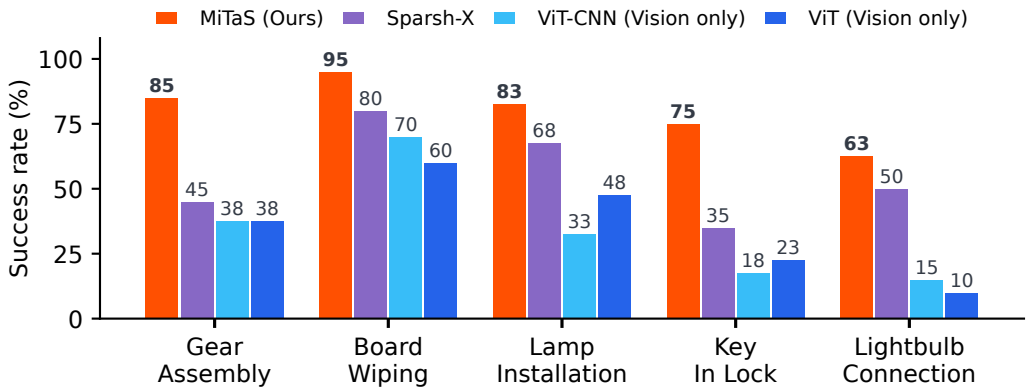


Figure 4: MiTaS outperformes the multimodal baseline Sparsh-X and two vision-only ViT-baselines in every task. Vision-only models failed, due to occlusion and complex contact dynamics, highlighting the need for tactile sensing in contact-rich manipulation.

Figure 4 shows the success rates of MiTaS, Sparsh-X, and both vision-only variants ViT and ViT-CNN over five manipulation tasks. MiTaS, with an average success rate of 80%, is able to solve all tasks consistently, while Sparsh-X (avg. 55.5%), with the same sensor setting, is not reliable in every task. Remarkably, Sparsh-X inserted the key multiple times but never completed the task, getting stuck in the keyway. Both vision-only models (ViT 35.5%; ViT-CNN 34.5%) have a low performance across all tasks and are not able to consistently solve Key in Lock and Lightbulb Connection, due to tight insertions dynamics and visual occlusions. Multi-resolution tactile feedback paired with suitable encoding enables a high success rate with only 30 demonstrations in these complex contact-rich manipulation tasks.

Sensor Fusion Study: We compare the fusion architecture from our main results against PolyTouch, which uses cross-attention, and Q-Former, which uses learnable queries for sensor fusion. Table 1 compares the results of the four sensor fusion architectures across 20 trials, trained on the same dataset using Vision, Gelsight and Evetac as observations. PolyTouch and Q-former have a slightly better average performance than Sparsh-X, while our approach with a high number of conditioning tokens and full self-attention fusion obtains the best performance. The results indicate that, next to the fusion mechanism, the shape of the conditioning is important for imitation learning.

Table 1: Sensor Fusion Architecture Study.

Task	**MiTaS**	**Sparsh-X**	**Polytouch**	**Q-former**
Gear	**80%**	40%	55%	50%
Board	**100%**	75%	70%	75%
Lamp	**85%**	65%	55%	75%
Key	**75%**	60%	55%	60%
Lightbulb	**60%**	45%	55%	55%
Avg.	**80%**	57%	58%	63%

Sensor Ablation Study: We compare the full sensor setting against the Vision and Gelsight setting. Additionally, we have trained the MiTaS model on Vision and Evetac investigating its performance without Gelsight grounding. Table 2 shows the sensor ablation results. Visual multi-tactile models achieve the highest success rate regardless of model design, confirming that heterogenous tactile sensors provide complementary information required to solve contact-rich manipulation task. Still, simple vision-tactile policies are able to solve most of the task over half of the time. Evetac as the sole tactile sensor provides useful information in certain task, but fails completely when object in hand location is needed. Gelsight, on the other hand, can align the gear and key from various initial states (see Appendix A.7).

Table 2: Sensor Ablation Study over 20 trials.

	MiTaS (Ours)			**Sparsh-X**	
Task	V+G+E	V+G	V+E	V+G+E	V+G
Gear	**90%**	45%	0%	50%	25%
Board	**90%**	70%	80%	85%	50%
Lamp	**80%**	65%	65%	70%	55%
Key	**75%**	55%	0%	10%	20%
Lightbulb	**65%**	20%	40%	55%	15%
Avg.	**80%**	51%	37%	54%	33%

Multi-Tactile Dropout: Table 3 shows the impact of our proposed multi-tactile dropout, leveraging the Evetac sensor during training and not requiring it during policy evaluation. It boosted MiTaS performance in 3/5 tasks, while being helpful in 2/5 tasks for Sparsh-X. In tasks, like Key in Lock, the dropout scheme lowers performance as it disturbs Gelsight's features for in-hand localization. In Board Wiping, both visuotactile models benefit from Evetac information obtained through MTD, aligning with the insights from the sensor ablation (MiTaS V+E) that Evetac features are beneficial in such dynamic tasks. Using Evetac during inference on a multi-tactile dropout trained model increases performance slightly (see Appendix A.8).

Table 3: Multi-Tactile Dropout Ablation at V+G.

Task	**MiTaS (Ours)**			**Sparsh-X**		
	V+G	MTD	Δ	V+G	MTD	Δ
Gear	45%	**55%**	+10	25%	15%	−10
Board	70%	**85%**	+15	50%	**85%**	+35
Lamp	65%	60%	−5	55%	**70%**	+15
Key	**55%**	20%	−35	20%	5%	−15
Lightbulb	20%	**45%**	+25	15%	10%	−5

4.3 Attention Analysis with Sensor Visualization

To understand how MiTaS uses the different sensors over the course of a task, we visualize the cross-attention from the executed action token to each sensor modality during a successful episode of the Lamp Installation task.

Figure 5 shows how the policy attends on vision during reaching, on Gelsight for the prolonged screwing motion and on all sensors during insertion. The improved performance and attention show that evetac is not driving the representation, but offers relevant features during crucial stages of the task. We observed recovery behaviour, when it was jammed during Gear assembly and resolving ambiguity in the Lighbulb Installation task, where Gelsight could not differentite between hovering closly over the socket or being fully inserted. Extended Attention analysis can be found in the Appendix A.9.

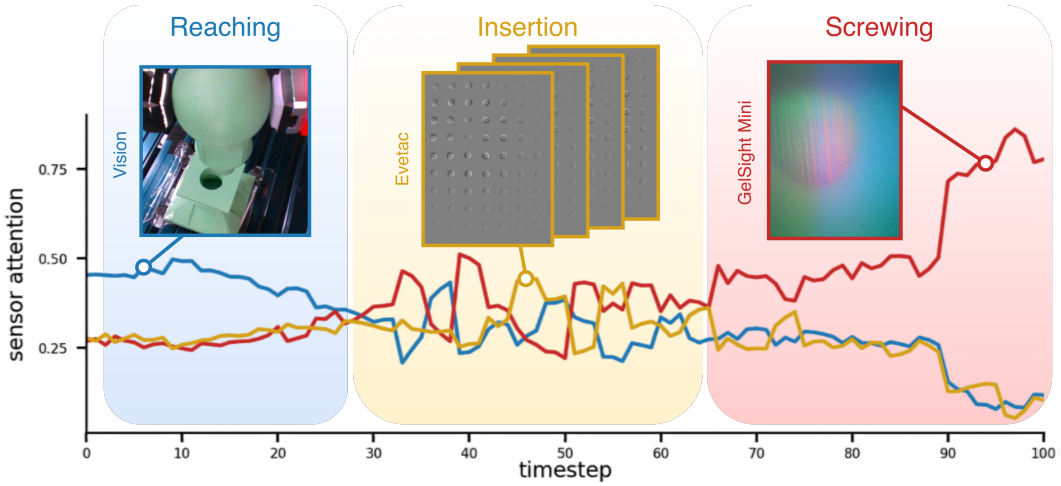


Figure 5: Attention analysis with sensor readings in Lamp Installation.

5 Limitations

While MiTaS demonstrates strong performance across our contact-rich task suite, several limitations remain. First, our investigation of multi-resolution fusion focuses exclusively on a specific pair of optical tactile sensors: the frame-based GelSight and the event-based Evetac. While this setup allowed for an exact and detailed analysis of complementary spatial and temporal resolutions, further research is required to evaluate how our architectural framework scales when integrating other heterogeneous sensor pairs or alternative tactile technologies. Second, our data collection pipeline relies on teleoperation via a Spacemouse, which fundamentally constrains scalability. Transitioning to alternative demonstration sources, such as human video or markerless hand tracking without a physical robot in the loop, could drastically scale up data collection. Furthermore, our policy is currently limited to 4-DoF actions by choice, though the framework itself imposes no fundamental limit on action dimensionality. While a 4-DoF space sufficed for our target contact-rich tasks, scaling to full 6-DoF control would likely require incorporating proprioceptive robot state information, which is currently omitted from our policy's observation space. Finally, the 15 Hz replanning frequency of our control loop places an upper bound on the framework's reactivity. Integrating our multi-resolution sensing framework with recent advances in highly reactive policy architectures offers a promising direction to further improve policy performance.

6 Conclusion

In this work, we propose Multi-Resolution Tactile Sensing (MiTaS), a novel imitation learning framework that leverages multiple tactile sensors operating at different temporal resolutions to solve complex manipulation tasks. Specifically, MiTaS encodes Vision, GelSight, and Evetac frames via CNN-stems and transformer fusion to condition a flow matching policy. Our results demonstrate how vision combined with heterogeneous tactile features obtained at low and high frequencies and spatial resolutions enables fine-grained manipulation, even under heavy occlusion, in complex contact-rich manipulation tasks. MiTaS achieves 80%, while the next best sensor-fusion architecture has a success rate of 63% and vision-only policies fail to solve the tasks consistently. This highlights the impact of our multisensory architecture for performing contact-rich tasks. Furthermore, we have shown how Multi-Tactile Dropout can boost a classic visuotactile policy in certain tasks. In future work, we want to investigate event-based tactile features for representation learning on a larger scale to improve widespread vision-based tactile sensors for manipulation. We further plan to investigate more reactive policies that asynchronously fuse vision and high-frequency tactile signals.

Acknowledgments

This research is funded by the German Research Foundation (DFG) Emmy Noether Programme (CH 2676/1-1), the EU's Horizon Europe project ARISE (Grant no.: 101135959), the German Federal Ministry of Education and Research (BMBF) project "RiG" (Grant no.: 16ME1001) and the European Research Council (ERC) project "SIREN" (Grant No.: 101163933). The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). The hardware is funded by the German Research Foundation (DFG).

References

[1] R. Johansson and J. Flanagan. Tactile Sensory Control of Object Manipulation in Humans. *The Senses: A Comprehensive Reference*, 2020.

[2] W. Yuan, S. Dong, and E. H. Adelson. GelSight: High-Resolution Robot Tactile Sensors for Estimating Geometry and Force. *Sensors*, 2017.

[3] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, et al. Digit: A Novel Design for a Low-Cost Compact High-Resolution Tactile Sensor with Application to in-hand Manipulation. *IEEE Robotics and Automation Letters (RA-L)*, 2020.

[4] B. Ward-Cherrier, N. Pestell, L. Cramphorn, B. Winstone, M. E. Giannaccini, J. Rossiter, and N. F. Lepora. The Tactip Family: Soft Optical Tactile Sensors with 3D-Printed Biomimetic Morphologies. *Soft Robotics*, 2018.

[5] S. Dong, D. K. Jha, D. Romeres, et al. Tactile-RL for Insertion: Generalization to Objects of Unknown Geometry. In *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2021.

[6] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, and S. Levine. The Feeling of Success: Does Touch Sensing Help Predict Grasp Outcomes?, 2017.

[7] E. Helmut, N. Funk, T. Schneider, C. de Farias, and J. Peters. Tactile-Conditioned Diffusion Policy for Force-Aware Robotic Manipulation. In *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2026.

[8] J. Hansen, F. Hogan, D. Rivkin, D. Meger, M. Jenkin, and G. Dudek. Visuotactile-RL: Learning Multimodal Manipulation Policies with Deep Reinforcement Learning. In *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2022.

[9] K. Yu, Y. Han, Q. Wang, V. Saxena, D. Xu, and Y. Zhao. MimicTouch: Leveraging Multimodal Human Tactile Demonstrations for Contact-rich Manipulation. In *Proc. of The 8th Conference on Robot Learning*, 2024.

[10] M. Yang, A. Church, Y. Lin, C. J. Ford, H. Li, E. Psomopoulou, D. A. Barton, N. F. Lepora, et al. AnyRotate: Gravity-Invariant In-Hand Object Rotation with Sim-to-Real Touch. In *Proc. of the 9th Conference on Robot Learning*, 2025.

[11] B. Romero, H.-S. Fang, P. Agrawal, and E. Adelson. EyeSight Hand: Design of a Fully-Actuated Dexterous Robot Hand with Integrated Vision-Based Tactile Sensors and Compliant Actuation. In *Proc. of the IEEE Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2024.

[12] 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*, Oct. 2024.

[13] D. Yin, S. Lu, J. Yang, Y. Zhang, Z. Dai, D. Nan, B. Cai, S. He, and X. Chen. GelEvent—A Novel High-Speed Tactile Sensor With Event Camera. *IEEE Transactions on Instrumentation and Measurement*, 2025.

[14] 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 Transactions on Robotics*, 36(6): 1619–1634, 2020.

[15] 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*, 2018.

[16] 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 *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023.

[17] H. Qi, B. Yi, S. Suresh, M. Lambeta, Y. Ma, R. Calandra, and J. Malik. General In-Hand Object Rotation with Vision and Touch. In *Proc. of The 7th Conference on Robot Learning*. PMLR, 2023.

[18] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li. 3D-ViTac: Learning Fine-Grained Manipulation with Visuo-Tactile Sensing. In *Proc. of The 9th Conference on Robot Learning*. PMLR, 2025.

[19] 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 *Proc. of the 6th Conference on Robot Learning*, 2022.

[20] 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 Transactions on Robotics*, 41:946–959, 2024.

[21] N. Funk, C. Chen, T. Schneider, G. Chalvatzaki, R. Calandra, and J. Peters. On The Importance of Tactile Sensing for Imitation Learning: A Case Study on Robotic Match Lighting. *IEEE Robotics and Automation Letters (RA-L)*, 2025.

[22] J. Chen, A. Kshirsagar, F. Heller, M. G. Andreu, B. Belousov, T. Schneider, L. P. Y. Lin, K. Doerschner, K. Drewing, and J. Peters. Investigating Active Sampling for Hardness Classification with Vision-Based Tactile Sensors, 2025.

[23] A. Böhm, T. Schneider, B. Belousov, A. Kshirsagar, L. Lin, K. Doerschner, K. Drewing, C. A. Rothkopf, and J. Peters. What Matters for Active Texture Recognition With Vision-Based Tactile Sensors. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[24] H. Xue, J. Ren, W. Chen, G. Zhang, Y. Fang, G. Gu, H. Xu, and C. Lu. Reactive Diffuison Policy: Slow-Fast Visual-Tactile Policy Learning for Contact-Rich Manipulation . In *Proc. of Robotics: Science and Systems*, 2025.

[25] N. Cheng, J. Xu, C. Guan, J. Gao, W. Wang, Y. Li, F. Meng, J. Zhou, B. Fang, and W. Han. Touch100k: A Large-Scale Touch-Language-Vision Dataset For Touch-Centric Multimodal Representation. In *Proc. of Information Fusion*, 2025.

[26] C. Higuera, A. Sharma, T. Fan, C. K. Bodduluri, B. Boots, M. Kaess, M. Lambeta, T. Wu, Z. Liu, F. R. Hogan, and M. Mukadam. Tactile Beyond Pixels: Multisensory Touch Representations for Robot Manipulation. In *Proc. of The 9th Conference on Robot Learning*, 2025.

[27] E. Helmut, L. Dziarski, N. Funk, B. Belousov, and J. Peters. Learning Force Distribution Estimation for the GelSight Mini Optical Tactile Sensor Based on Finite Element Analysis. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

[28] A. George, S. Gano, P. Katragadda, and A. B. Farimani. VITaL Pretraining: Visuo-Tactile Pretraining for Tactile and Non-Tactile Manipulation Policies, Sept. 2024. arXiv:2403.11898 [cs].

[29] Z. Zhao, S. Haldar, J. Cui, L. Pinto, and R. Bhirangi. Touch Begins Where Vision Ends: Generalizable Policies for Contact-Rich Manipulation, 2025. arXiv:2506.13762 [cs].

[30] V. Dave, F. Lygerakis, and E. Rueckert. Multimodal Visual-Tactile Representation Learning Through Self-Supervised Contrastive Pre-Training. In *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2024.

[31] L. Fu, H. Huang, L. Berscheid, H. Li, K. Goldberg, and S. Chitta. Safe Self-Supervised Learning in Real of Visuo-Tactile Feedback Policies for Industrial Insertion, 2023. arXiv:2210.01340 [cs].

[32] Z. Yi, R. Calandra, F. Veiga, H. van Hoof, T. Hermans, Y. Zhang, and J. Peters. Active tactile object exploration with Gaussian processes. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[33] T. Schneider, G. Duret, C. de Farias, R. Calandra, L. Chen, and J. Peters. Tactile MNIST: Benchmarking Active Tactile Perception, 2025.

[34] T. Schneider, C. de Farias, R. Calandra, L. Chen, and J. Peters. APPLE: Toward General Active Perception via Reinforcement Learning. In *International Conference on Learning Representations*, 2026.

[35] C. Zhang, P. Hao, X. Cao, X. Hao, S. Cui, and S. Wang. VTLA: Vision-Tactile-Language-Action Model with Preference Learning for Insertion Manipulation, May 2025. arXiv:2505.09577 [cs].

[36] G. Ye, Z. Zhang, X. Zhao, S. Wu, H. Lu, S. Lu, and H. Liu. Learning to Feel the Future: DreamTacVLA for Contact-Rich Manipulation, 2025. arXiv:2512.23864 [cs].

[37] TacEx: GelSight Tactile Simulation in Isaac Sim – Combining Soft-Body and Visuotactile Simulators, author=Duc Huy Nguyen and Tim Schneider and Guillaume Duret and Alap Kshirsagar and Boris Belousov and Jan Peters, 2024.

[38] 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. In *Proc. of the International Conference on Learning Representations*, 2025.

[39] J. Zhao, Y. Ma, L. Wang, and E. Adelson. Transferable Tactile Transformers for Representation Learning Across Diverse Sensors and Tasks. In *Proc. of The 8th Conference on Robot Learning*, 2024.

[40] J. Lenz, T. Gruner, D. Palenicek, T. Schneider, and J. Peters. Analysing the Interplay of Vision and Touch for Dexterous Insertion Tasks. In *CoRL 2024 Workshop on Learning Robot Fine and Dexterous Manipulation: Perception and Control*, 2024.

[41] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg. Making Sense of Vision and Touch: Self-Supervised Learning of Multimodal Representations for Contact-Rich Tasks. In *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2019.

[42] 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*, 2020.

[43] R. Feng, D. Hu, W. Ma, and X. Li. Play to the Score: Stage-Guided Dynamic Multi-Sensory Fusion for Robotic Manipulation. In *Proc. of The 8th Conference on Robot Learning*, 2024.

[44] C. Sferrazza, Y. Seo, H. Liu, Y. Lee, and P. Abbeel. The Power of the Senses: Generalizable Manipulation from Vision and Touch through Masked Multimodal Learning. In *Proc. of the IEEE Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2024.

[45] M. Lambeta, T. Wu, A. Sengul, V. R. Most, N. Black, K. Sawyer, R. Mercado, H. Qi, A. Sohn, B. Taylor, N. Tydingco, G. Kammerer, D. Stroud, J. Khatha, K. Jenkins, K. Most, N. Stein, R. Chavira, T. Craven-Bartle, E. Sanchez, Y. Ding, J. Malik, and R. Calandra. Digitizing Touch with an Artificial Multimodal Fingertip, 2024.

[46] J. Zhao, N. Kuppuswamy, S. Feng, B. Burchfiel, and E. Adelson. Polytouch: A Robust Multi-Modal Tactile Sensor for Contact-Rich Manipulation Using Tactile-Diffusion Policies. In *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2025.

[47] P. Jin, B. Huang, W. W. Lee, T. Li, and W. Yang. Visual-Force-Tactile Fusion for Gentle Intricate Insertion Tasks. *IEEE Robotics and Automation Letters (RA-L)*, 2024.

[48] K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. R. Equi, C. Finn, N. Fusai, M. Y. Galliker, D. Ghosh, L. Groom, K. Hausman, b. ichter, S. Jakubczak, T. Jones, L. Ke, D. LeBlanc, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, J. Tanner, Q. Vuong, H. Walke, A. Walling, H. Wang, L. Yu, and U. Zhilinsky. $\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization. In *Proc. of The 9th Conference on Robot Learning*, 2025.

[49] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow Matching for Generative Modeling. In *Proc. of the International Conference on Learning Representations*, 2023.

[50] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick. Early Convolutions Help Transformers See Better. *Advances in Neural Information Processing Systems*, 2021.

[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, 2017.

[52] J. Urain, A. Mandlekar, Y. Du, N. Muhammad "Mahi" Shafiullah, D. Xu, K. Fragkiadaki, G. Chalvatzaki, and J. Peters. A Survey on Deep Generative Models for Robot Learning From Multimodal Demonstrations. *IEEE Transactions on Robotics*, 2026.

[53] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. *The Intl. Journal of Robotics Research*, 2025.

[54] W. Peebles and S. Xie. Scalable Diffusion Models with Transformers. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2023.

[55] S. Dasari, O. Mees, S. Zhao, M. K. Srirama, and S. Levine. The Ingredients for Robotic Diffusion Transformers. In *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2025.

[56] R. Krohn, V. Prasad, G. Tiboni, and G. Chalvatzaki. Self-Supervised Multisensory Pretraining for Contact-Rich Robot Reinforcement Learning. *IEEE Robotics and Automation Letters*, 2026.

[57] Y. Jin, J. Guo, X. Jia, Y. Deng, Z. Li, H. Liu, W. Liao, V. Prasad, M. Franzius, G. Neumann, et al. Nautilus: From One Prompt to Plug-and-Play Robot Learning. 2026.

[58] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun. Attention Bottlenecks for Multimodal Fusion. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2021.

[59] J. Li, D. Li, S. Savarese, and S. Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proc. of the 40th International Conference on Machine Learning*, 2023.

[60] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021.

[61] B. Tang, M. A. Lin, I. Akinola, A. Handa, G. S. Sukhatme, F. Ramos, D. Fox, and Y. Narang. IndustReal: Transferring Contact-Rich Assembly Tasks from Simulation to Reality. In *Robotics: Science and Systems*, 2023.

[62] M. Heo, Y. Lee, D. Lee, and J. J. Lim. FurnitureBench: Reproducible Real-World Benchmark for Long-Horizon Complex Manipulation. In *Robotics: Science and Systems*, 2023.

[63] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng. Fourier Features Let Networks Learn High Frequency Functions In Low Dimensional Domains. *Advances in Neural Information Processing Systems*, 2020.

A Appendix

This appendix provides supplementary material that supports the experimental evaluation presented in the main text.

A.1 Controller

We use a Cartesian pose controller implemented as a Cartesian-to-joint target generator coupled with a low-level joint impedance servo. At each control step (15Hz), the input, provided either by a SpaceMouse teleoperation interface or a learned policy, specifies a target end-effector pose in a reduced Cartesian space (x, y, z, ψ), while roll and pitch are held fixed to values latched at reset. The Cartesian pose error is converted to a bounded Cartesian increment and mapped to joint motion using a Jacobian-based inverse kinematics module with joint-limit handling and nullspace regularization. This yields a desired 7-DoF joint configuration $\mathbf{q}_d \in \mathbb{R}^7$ that is tracked by a joint impedance controller running at 1kHz, applying $\tau = K_q(\mathbf{q}_d - \mathbf{q}) + K_{qd}(\dot{\mathbf{q}}_d - \dot{\mathbf{q}})$, resulting in compliant tracking of the commanded Cartesian targets.

Table 4: Controller parameters.

Parameter	Value
Outer-loop command rate	15 Hz
Command space	Cartesian position (xyzyaw)
Max linear increment per step	0.075 m
Max rotational increment per step	0.15 rad
Max joint increment per step	0.2 rad (per joint)
Nullspace gain	0.025
Regularization weight	1×10^{-2}
Max Cartesian velocity-IK iterations	300
Max nullspace iterations	300
Min. distance to joint limits	0.3 rad
Impedance servo rate	1000 Hz
Joint stiffness K_q	$[40, 30, 50, 25, 35, 25, 10]$
Joint damping K_{qd}	$[4, 6, 5, 5, 3, 2, 1]$

A.2 Parameter Count and Inference

We evaluate all nine policy variants on an NVIDIA GeForce RTX 4080, measuring inference speed over 100 runs per model and reporting the average runtime with 10 flow-matching integration steps (see Table 5). For each model we also report the number of trainable parameters, split into the perception encoder (Perc.) and the total model size. Every model is using the same policy head with 13,701,508 parameters.

Table 5: Parameter count and inference benchmark.

Model	Sensors	MTD	#params Perc.	#params Total	Tok/obs	ms	Hz
ViT-CNN	V	–	3,079,232	16,780,740	260	35.5	227.4
ViT	V	–	4,051,328	17,752,836	260	34.4	232.2
Sparsh	V+G	–	9,398,744	23,100,252	448	39.2	203.5
Sparsh	V+G	✓	15,500,056	29,201,564	448	38.9	208.4
Sparsh	V+G+E	–	15,500,056	29,201,564	640	54.6	153.4
MiTaS	V+E	–	8,769,280	22,470,788	452	35.6	223.7
MiTaS	V+G	–	5,074,776	18,776,284	452	35.3	224.1
MiTaS	V+G	✓	10,764,824	24,466,332	452	36.1	230.1
MiTaS	V+G+E	–	10,764,824	24,466,332	644	43.5	180.5

A.3 AdaLN-Zero time conditioning.

Each DiT block is modulated by the time embedding $\boldsymbol{\tau}(t)$ using AdaLN-Zero [54, 55, 63]. For a sublayer input $\mathbf{Y}$, we apply adaptive layer normalization $\mathrm{AdaLN}(\mathbf{Y};\boldsymbol{\beta},\boldsymbol{\gamma}) = \mathrm{LN}(\mathbf{Y})\odot(1+\boldsymbol{\gamma})+\boldsymbol{\beta}$, where $(\boldsymbol{\beta},\boldsymbol{\gamma})$ are produced from $\boldsymbol{\tau}(t)$. A per-block MLP outputs $(\boldsymbol{\beta},\boldsymbol{\gamma},g)$ triplets for self-attention, cross-attention, and the MLP, and each residual update is gated as $\mathbf{Y} \leftarrow \mathbf{Y} + g\, f(\mathrm{AdaLN}(\mathbf{Y};\boldsymbol{\beta},\boldsymbol{\gamma}))$. We initialize the AdaLN projection and the final output head to zero so the network starts near an identity mapping and learns time-conditioning smoothly.

A.4 Symbols and Hyperparameters

Table 6: Symbols of the MiTaS pipeline.

Symbol	Meaning
H	Action horizon
d_a	Action dimension
$\mathbf{A} \in \mathbb{R}^{H \times d_a}$	Action chunk
$\bar{\mathbf{A}} \in \mathbb{R}^{H \times d_a}$	Normalized action chunk
n_{act}	Number of executed actions per replanning step ($n_{\text{act}} \leq H$)
$O^{\text{vision}} \in \mathbb{R}^{2\times3\times128\times128}$	Wrist-camera observation
$O^{\text{gelsight}} \in \mathbb{R}^{2\times3\times120\times160}$	GelSight tactile observation
$O^{\text{evetac}} \in \mathbb{R}^{16\times120\times160}$	EveTac observation
$\mathbf{x}_0 \in \mathbb{R}^{H \times d_a}$	Initial noise sample, $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$\mathbf{x}_1 \in \mathbb{R}^{H \times d_a}$	Data endpoint for flow matching ($\mathbf{x}_1 = \bar{\mathbf{a}}$)
k	Continuous flow time, $k \in [0, 1]$
$\mathbf{x}_k \in \mathbb{R}^{H \times d_a}$	On-path state returned by the conditional flow matcher
$\mathbf{u}_k \in \mathbb{R}^{H \times d_a}$	Target conditional velocity returned by the flow matcher
$\mathbf{C} \in \mathbb{R}^{N_{\text{cond}} \times D}$	Conditioning tokens from the multisensory encoder
N_{cond}	Number of conditioning tokens
N_{vis}	Number of Vision tokens
N_{gs}	Number of GelSight tokens
N_{evt}	Number of Evetac tokens
D	Token embedding width (shared by $\mathbf{C}$ and action tokens)
$\mathbf{z}_i \in \mathbb{R}^D$	Sensor token
$\mathbf{p}_{g(i)} \in \mathbb{R}^D$	Learned position embedding for token-grid location $g(i)$
$\mathbf{e}_{s(i)} \in \mathbb{R}^D$	Learned modality embedding for source modality $s(i)$
$s(i) \in \{\text{vision}, \text{gelsight}, \text{evetac}\}$	Modality identifier for token i
$\mathbf{z}'_i \in \mathbb{R}^D$	Augmented token, $\mathbf{z}'_i = \mathbf{z}_i + \mathbf{p}_{g(i)} + \mathbf{e}_{s(i)}$
$\boldsymbol{\tau}(k) \in \mathbb{R}^{d_k}$	Time embedding of k
$v_\theta(\mathbf{x}_k, k, \mathbf{C})$	DiT network predicting velocity / vector field
$\hat{\mathbf{v}}_\theta \in \mathbb{R}^{H \times d_a}$	Predicted velocity chunk
m_h	Validity mask for timestep h (episode padding mask)
$\mathcal{L}_{\text{FM}}$	Flow-matching training loss (masked MSE)
K	Number of Euler integration steps at inference
Δk	Euler step size, $\Delta k = 1/K$
$\Delta\mathbf{p} \in \mathbb{R}^4$	Delta position and delta yaw: $[\Delta x, \Delta y, \Delta z, \Delta\psi]$

Table 7: Hyperparameters.

Hyperparameter	Value	Hyperparameter	Value
Batch size	64	Vision normalization	$[0, 1]$
Action horizon H	16	Gelsight normalization	$[0, 1]$
Executed steps per replan n_{act}	1–3	Evetac normalization	$[-\frac{1}{2}, \frac{1}{2}]$
Observation window n_{obs}	4	Embedding width D	256
Action normalization	minmax	DiT depth L	10
Action target	$\Delta\mathbf{p} \in \mathbb{R}^4$	Attention heads	8
Time embedding dim d_t	128	MLP ratio	4
Time embedding	sinusoidal	Dropout	0.0
Flow-matching noise σ	0.0	FM inference steps K	10
Policy learning rate	1×10^{-4}	Encoder learning rate	1×10^{-5}
Weight decay	1×10^{-6}	Adam betas	$(0.95, 0.999)$
Learning-rate schedule	cosine	Gradient clipping	1.0

A.5 Task Progress

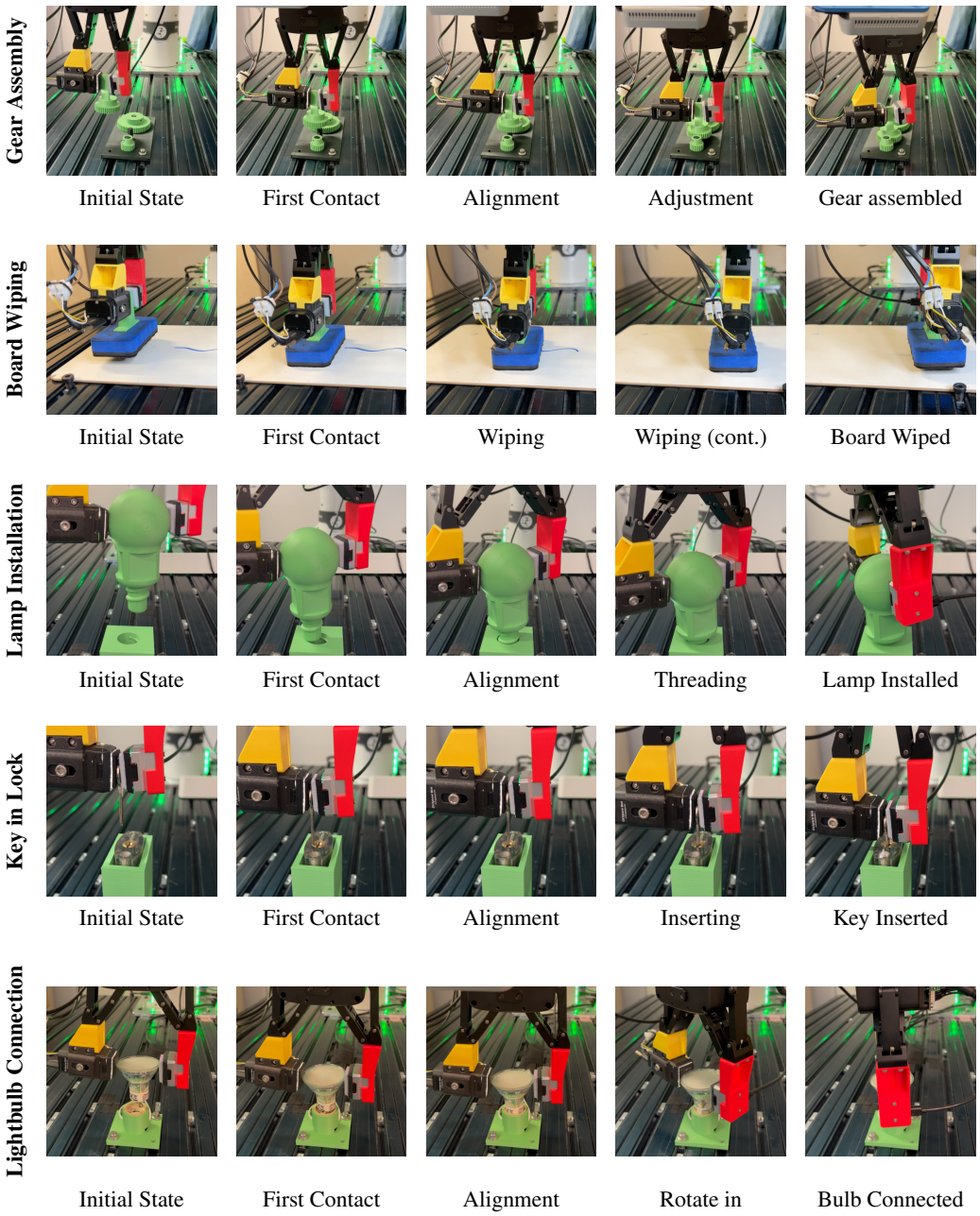


Figure 6: Task progress across five manipulation tasks. Each row shows representative frames from a successful rollout, progressing from the initial scene configuration (left) to task completion (right). The tasks span a range of contact-rich skills, including Gear Assembly, Board Wiping, Lamp Installation (FurnitureBench), Key in Lock, and Lightbulb Connection.

A.6 Failure Cases

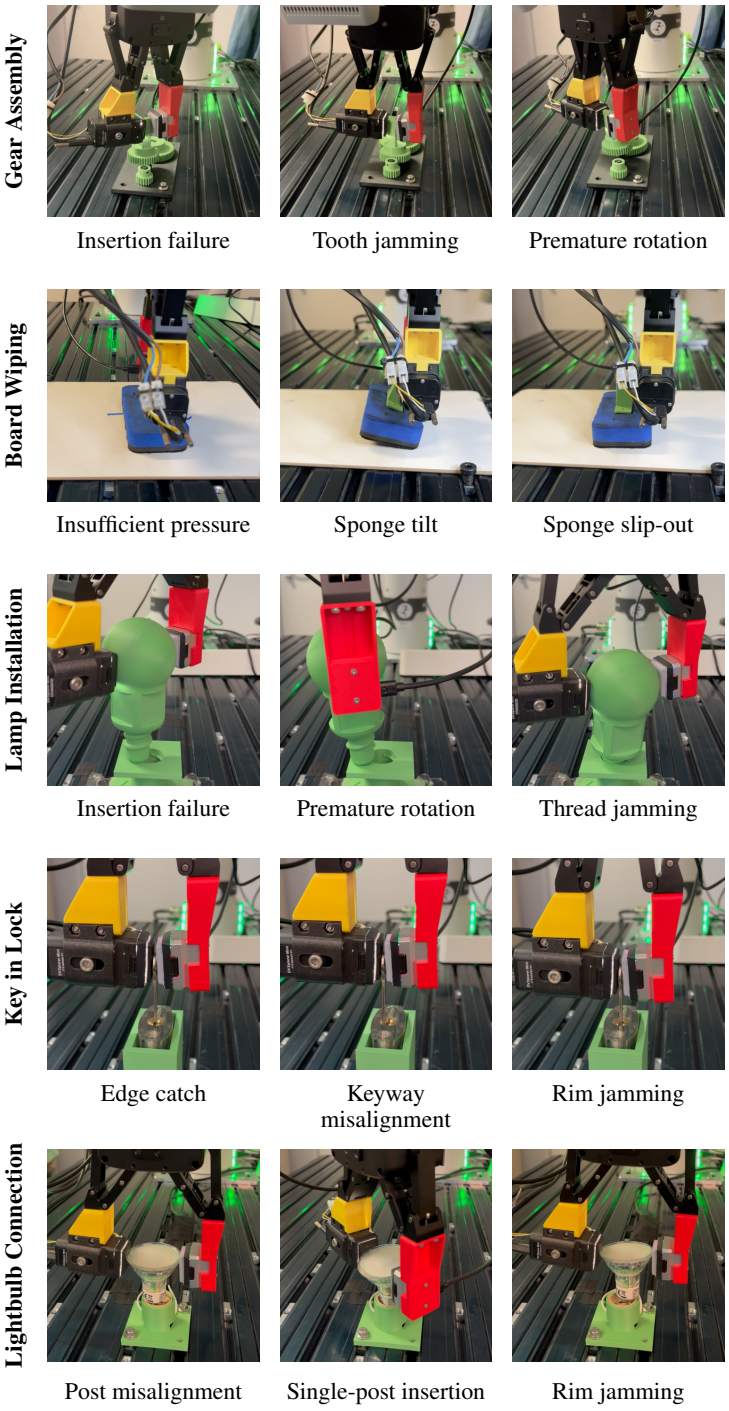


Figure 7: Examples of failure cases across the five manipulation tasks. Each row shows representative frames from a failed rollout. These examples are illustrative and not exhaustive.

A.7 In-Hand Distribution of Initial States

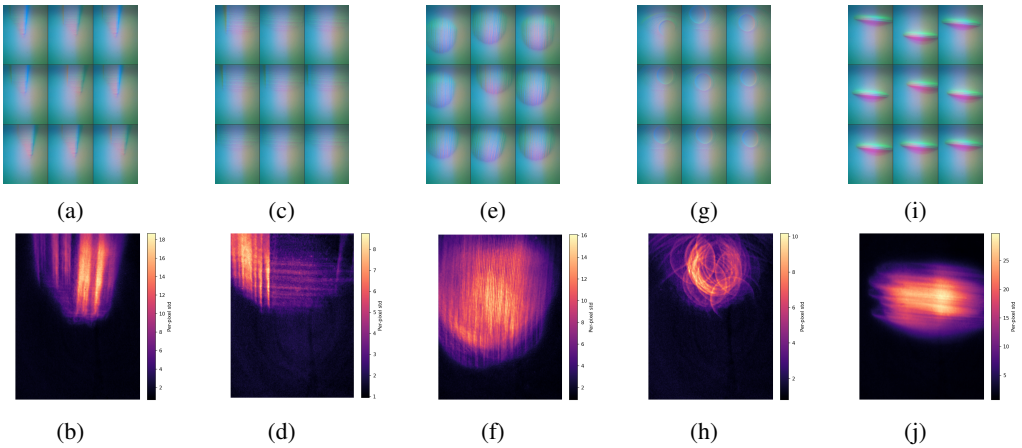


Figure 8: In-hand GelSight distributions after reset across Gear Assembly (a-b), Board Wiping (c-f), Lamp Install (e-f), Key In Lock (g-h) and Lighbulb Connect (i-j). Each column shows a 3×3 montage on top and the per-pixel standard deviation at the botton of the respective training dataset.

A.8 Results overview

Table 8 summarizes all 1700 policy evaluation runs across both datasets (20 trials per model and task), pooling the results of models evaluated on both. Table 9 reports policies trained on the first dataset: the main models, a sensor ablation, and an ablation of multi-tactile dropout (MTD) training. Table 10 reports policies trained on a second, independently collected dataset: the main models, MTD inference with and without evetac observation, and additional sensor-fusion architectures (PolyTouch with cross-attention and Q-Former with learnable queries).

Table 8: Policy performance across all five manipulation tasks, reported as successes/trials. Unless marked, each entry is out of 20 trials. †Evaluated on two independent datasets with successes and trials pooled, i.e., out of 40.

Model	Sensors	MTD	Gear Assembly	Board Wiping	Lamp Installation	Key in Lock	Lightbulb Connection	Total	Avg.
ViT-CNN†	V	–	15/40	28/40	13/40	7/40	6/40	69/200	34.5%
ViT†	V	–	15/40	24/40	19/40	9/40	4/40	71/200	35.5%
MiTaS	V+E	–	0/20	16/20	13/20	0/20	8/20	37/100	37%
Sparsh-X	V+G	–	5/20	10/20	11/20	4/20	3/20	33/100	33%
MiTaS	V+G	–	9/20	14/20	13/20	11/20	4/20	51/100	51%
Sparsh-X	V+G	✓	3/20	17/20	14/20	1/20	2/20	37/100	37%
MiTaS†	V+G	✓	22/40	33/40	23/40	17/40	18/40	113/200	56.5%
MiTaS	V+G+E	✓	12/20	17/20	13/20	11/20	10/20	63/100	63%
Q-Former	V+G+E	–	10/20	15/20	15/20	12/20	11/20	63/100	63%
Poly	V+G+E	–	11/20	14/20	11/20	11/20	11/20	58/100	58%
Sparsh-X†	V+G+E	–	18/40	32/40	27/40	14/40	20/40	111/200	55.5%
MiTaS (Ours)†	V+G+E	–	**34/40**	**38/40**	**33/40**	**30/40**	**25/40**	**160/200**	**80%**

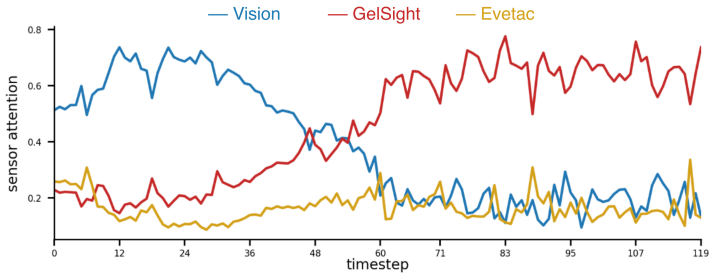
Table 9: Policy performance of the first dataset across all manipulation tasks (success rate in %). Bold: best in column. Avg.: mean over the five evaluated tasks.

Model	Sensors	Co-Train	Gear Assembly	Board Wiping	Lamp installation	Key in Lock	Lightbulb Connection	Avg.
ViT-CNN	V	–	40%	70%	20%	0%	0%	26%
ViT	V	–	40%	60%	55%	0%	0%	31%
MiTaS	V+E	–	0%	80%	65%	0%	40%	37%
Sparsh	V+G	–	25%	50%	55%	20%	15%	33%
MiTaS	V+G	–	45%	70%	65%	55%	20%	51%
Sparsh	V+G	✓	15%	85%	70%	5%	10%	37%
MiTaS	V+G	✓	55%	85%	60%	20%	45%	53%
Sparsh	V+G+E	–	50%	85%	70%	10%	55%	54%
MiTaS (Ours)	V+G+E	–	**90%**	**90%**	**80%**	**75%**	**65%**	**80%**

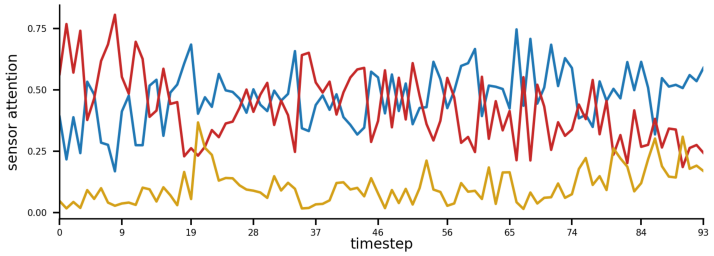
Table 10: Policy performance of the second dataset across all manipulation tasks (success rate in %, 20 trials per task). Bold: best in column. Avg.: mean over the five tasks

Model	Sensors	MTD	Gear Assembly	Board Wiping	Lamp Installation	Key in Lock	Lightbulb Connection	Avg.
ViT-CNN	V	–	35%	70%	45%	35%	30%	43%
ViT	V	–	35%	60%	40%	45%	20%	40%
MiTaS	V+G	✓	55%	80%	55%	65%	45%	60%
MiTaS	V+G+E	✓	60%	85%	65%	55%	50%	63%
Q-Former	V+G+E	–	50%	75%	75%	60%	55%	63%
Poly	V+G+E	–	55%	70%	55%	55%	55%	58%
Sparsh-X	V+G+E	–	40%	75%	65%	60%	45%	57%
MiTaS (Ours)	V+G+E	–	**80%**	**100%**	**85%**	**75%**	**60%**	**80%**

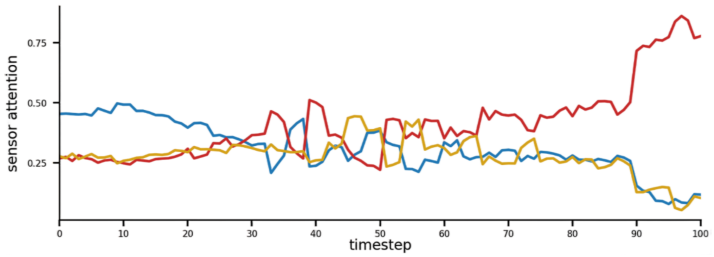
A.9 Attention analysis



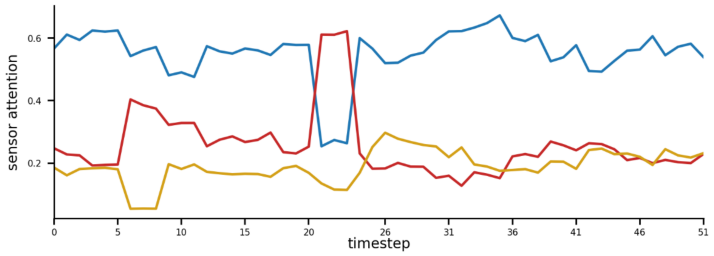
(a) Gear Assembly



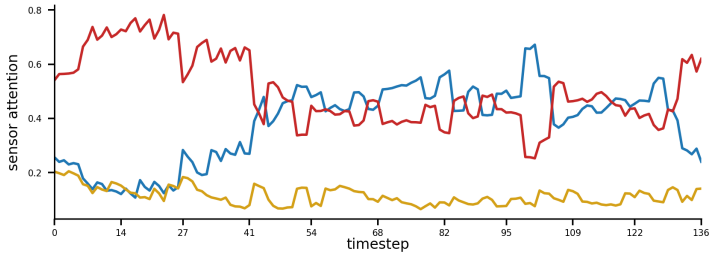
(b) Board Wiping



(c) Lamp Installation



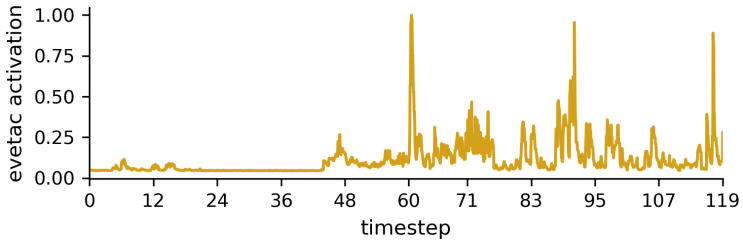
(d) Key in Lock



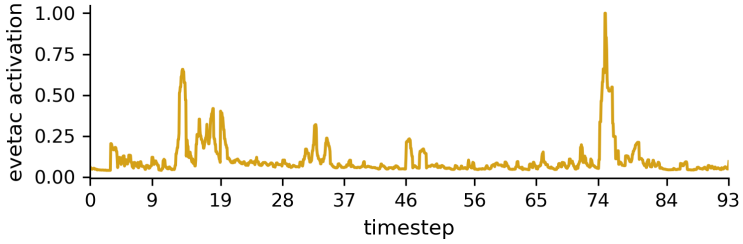
(e) Lightbulb Connection

Figure 9: For each task we plot the cross-attention from the executed action token to each sensor modality (Vision, GelSight, Evetac) over the course of a successful episode.

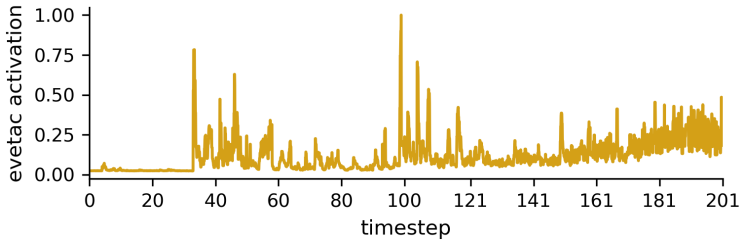
A.10 Evetac Activation



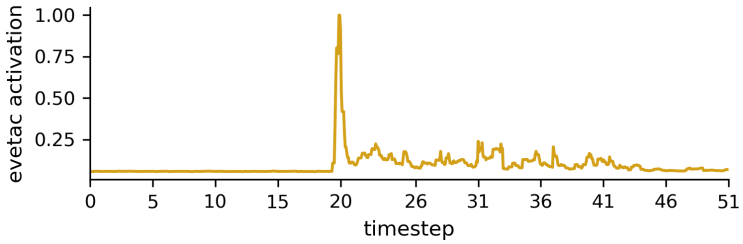
(a) Gear Assembly



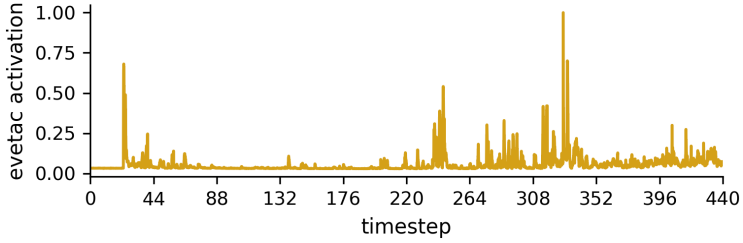
(b) Board Wiping



(c) Lamp Installation



(d) Key in Lock



(e) Lightbulb Connection

Figure 10: We define the Evetac activation as the L1-distance to the base image in pixel-space. Evetac activations differ for each task and task-stage indicating its usefullness to the policy. During reaching, beginning of a trajectory, the activation is low, it spikes by contact and provides high frequency features during object interaction.