TAC2PIX: IMAGE-SPACE VISUO-TACTILE FUSION FOR DEXTEROUS MANIPULATION

Anonymous authors
Paper under double-blind review

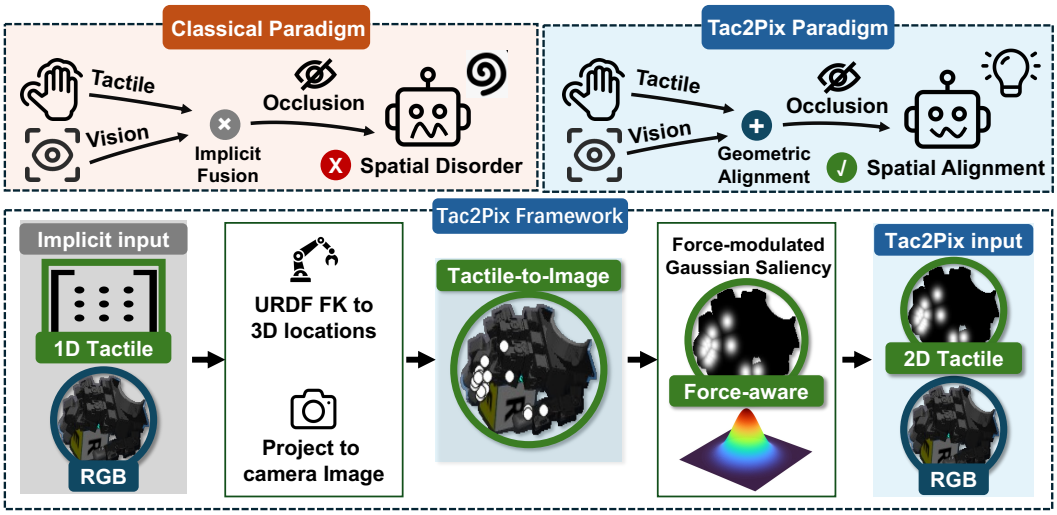


Figure 1: **Overview of Tac2Pix**. Classical tactile-vision fusion relies on implicit multimodal embeddings that often lose spatial correspondence under occlusion. Our **Tac2Pix** paradigm explicitly projects tactile contacts onto image-space saliency maps, producing a force-aware and spatially aligned representation for robust dexterous manipulation.

ABSTRACT

Vision and touch are naturally complementary for dexterous manipulation. Yet they are represented in fundamentally different forms: visual observations are dense and spatially structured, whereas tactile measurements are sparse, heterogeneous, and robot-centric. This mismatch makes it challenging to establish explicit spatial correspondence between the two modalities, particularly when they are fused in latent space. However, each tactile measurement is associated with a known location on the robot's kinematic chain, offering a localization prior that could bridge this representational gap. This raises a natural question: *can this localization prior be used to construct a shared representation that explicitly aligns touch with vision?* We introduce **Tac2Pix**, an image-space visuo-tactile fusion framework that projects sparse tactile measurements onto the RGB image plane using forward kinematics and camera calibration, then renders them as force-aware saliency maps. By expressing touch in the same pixel space as vision, Tac2Pix directly leverages pretrained 2D visual representations. We further employ zero-initialized input adaptation to preserve the original visual pathway at the start of policy learning. To assess whether this geometry-based interface yields consistent policy gains, we evaluate Tac2Pix on three simulated and three real-world dexterous manipulation tasks. In simulation, Tac2Pix consistently improves visuo-tactile policy learning across tasks, policy architectures, and visual backbones. In real-world experiments, it improves average success under visual occlusion by 26.7 percentage points over the strongest latent visuo-tactile fusion baseline, with gains persisting under both random and physical occlusions. Together, these results support image space as an effective interface for fusing sparse touch with pretrained visual representations. Project page: Tac2Pix.github.io.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Dexterous robotic manipulation benefits from complementary sensory modalities. Vision provides dense spatial observations and can leverage powerful pretrained representations for robotic control (Radosavovic et al., 2023; Majumdar et al., 2023). Touch, in contrast, directly measures physical interaction, including contact events and force magnitudes, and remains informative when visual evidence is incomplete or occluded (Calandra et al., 2018; Luo et al., 2026a). However, tactile sensing is sparse, heterogeneous, and distributed over the robot embodiment, making it difficult to integrate with image-centric visual representations. Crucially, the spatial structure of robot touch is already known: each tactile measurement is anchored to a specific location on the robot kinematic chain. Yet most visuo-tactile policies do not explicitly exploit this geometry when relating touch to vision, leaving the correspondence between the two modalities to be learned from limited task demonstrations.

Existing approaches address this mismatch mainly in two ways. Implicit fusion methods encode vision and touch separately and combine their representations through concatenation, modulation, or attention (Feng et al., 2025a; Su et al., 2024). Although flexible, these methods must infer touch–vision correspondence from task data while operating on sensor-specific tactile representations. A second line of work introduces explicit spatial structure by lifting tactile signals into 3D representations (Wu et al., 2025; Huang et al., 2025a). While effective for spatial reasoning, such approaches typically require depth sensing or dedicated 3D processing, making it less direct to reuse the rapidly improving ecosystem of pretrained 2D visual representations.

This motivates a third alternative: grounding touch directly in the native image space of RGB observations. Rather than requiring the policy to infer cross-modal correspondence from demonstrations through latent-space fusion, we use robot kinematics and camera calibration to establish this correspondence explicitly. We introduce **Tac2Pix**, an image-space visuo-tactile fusion framework that converts heterogeneous tactile measurements into a shared, pixel-aligned representation. Tac2Pix first normalizes tactile sensor responses, then uses forward kinematics and camera calibration to project each measurement onto the RGB image plane, where it is rendered as a force-aware saliency cue. Together, these steps yield a unified image-space representation of touch that makes its spatial correspondence with vision explicit before feature learning.

By expressing tactile information in the same pixel coordinate system as RGB observations, Tac2Pix can directly leverage the spatial structure and pretrained representations of standard 2D visual backbones. Inspired by ControlNet (Zhang et al., 2023), we introduce a zero-initialized input adaptation mechanism that adds a tactile saliency channel to the visual encoder while preserving its pretrained RGB pathway at initialization. Because fusion occurs in image space, the same tactile representation can be used across different visual encoders and downstream policy architectures through lightweight input adaptation. Tac2Pix thus provides a lightweight, broadly compatible interface between sparse tactile sensing and pretrained visual representations.

We evaluate Tac2Pix on three simulated and three real-world dexterous manipulation tasks, spanning four policy architectures and multiple visual backbones. Tac2Pix achieves strong and consistent performance across these settings, with particularly pronounced gains under degraded visual observations. In real-world experiments, Tac2Pix increases the average success rate under visual occlusion by 26.7 percentage points over the strongest implicit fusion baseline, with gains persisting under both random and physical occlusions.

Our main contributions are summarized as follows:

- We propose **Tac2Pix**, a tactile-enhanced policy pipeline that converts heterogeneous tactile measurements into force-aware, pixel-aligned saliency maps for policy learning.
- We introduce mechanisms for (1) establishing explicit touch–vision correspondence before feature learning, reducing the need to learn this alignment from task demonstrations, and for (2) integrating tactile saliency maps with pretrained 2D visual representations through zero-initialized input adaptation.
- We validate Tac2Pix across six dexterous manipulation tasks, four policy architectures, and multiple visual backbones, demonstrating consistent gains in simulation and real-world manipulation, particularly under visual occlusion.

2 RELATED WORKS

Implicit Visuo-Tactile Fusion. Tactile sensing provides direct measurements of physical interaction and has been shown to improve robotic manipulation in contact-rich settings, including dexterous manipulation (Guzey et al., 2023; Lin et al., 2025), regrasping (Calandra et al., 2018), slip and contact-state estimation (Dong et al., 2019; She et al., 2021), and grasp outcome prediction (Calandra et al., 2017). As a complementary modality to vision, touch is particularly valuable when visual observations are occluded, degraded, or insufficient for inferring contact states (Li et al., 2023; Sferrazza et al., 2024; Luo et al., 2026b). Most existing visuo-tactile policies combine the two modalities only after separate feature extraction. For instance, tactile features may be concatenated with visual embeddings (Chen et al., 2026), used to modulate visual representations through tactile-conditioned parameters (Feng et al., 2025a; Zhu et al., 2025), or mixed with visual tokens through attention-based modules (Chen et al., 2023b; Pattabiraman et al., 2024). While flexible, these approaches leave touch–vision spatial correspondence implicit, requiring it to be learned from limited task demonstrations. In contrast, Tac2Pix establishes this correspondence through geometric projection, expressing tactile measurements explicitly in the same pixel coordinate system as RGB and enabling the integration of both modalities within a shared visual encoder.

Spatially Grounded Tactile Representations. Recent work has increasingly incorporated explicit spatial structure into tactile representations for dexterous manipulation. Robot Synesthesia (Yuan et al., 2024) represents tactile contacts as point clouds, preserving their 3D spatial organization for visuo-tactile in-hand manipulation. 3DTacDex (Wu et al., 2025) introduces a canonical 3D tactile representation that preserves both contact geometry and force information while reducing variation induced by hand configuration. 3D-ViTac (Huang et al., 2025a) further places visual and dense tactile observations in a unified 3D representation space, explicitly retaining their geometric relationships. Complementarily, SaTA (Huang et al., 2026) anchors tactile features to the hand kinematic frame through forward kinematics, providing a stable reference for geometric contact reasoning. Together, these works demonstrate the importance of preserving spatial structure in tactile perception. However, they represent touch in dedicated 3D, hand-centric, or sensor-layout-aware spaces. Tac2Pix instead grounds tactile measurements directly in the native coordinate system of RGB observations. This provides a straightforward interface with existing 2D visual representations, without requiring an additional 3D or hand-centric processing pipeline.

Scalable Tactile Representations and Pretrained Visual Backbones. Learning general tactile representations remains challenging because tactile sensors vary widely in signal type, spatial layout, sampling characteristics, calibration, and noise profiles (Yuan et al., 2017; Bhirangi et al., 2024; Huang et al., 2025b). Prior work has learned representations from specific tactile modalities, such as high-resolution vision-based tactile images (Lambeta et al., 2020; Dave et al., 2024) and sparse force measurements (Xue et al., 2024), as well as from paired visual-tactile observations for cross-modal alignment (Lee et al., 2019). Recent methods further seek reusable tactile representations across sensors and interaction types (Feng et al., 2025b; Rodriguez et al., 2025). However, unlike RGB images, tactile signals lack a standardized spatial format amenable to large-scale pretraining pipelines, making it difficult to directly reuse the spatial inductive biases and pretrained representations of standard visual backbones (He et al., 2016; Dosovitskiy et al., 2021). Tac2Pix therefore uses image space as a common interface, transforming heterogeneous tactile measurements into pixel-aligned saliency maps that directly leverage the spatial inductive biases and pretrained representations of existing 2D visual backbones.

3 METHODOLOGY

As shown in Figure 2, we propose Tac2Pix, a lightweight visuo-tactile policy learning framework that converts sparse, heterogeneous tactile measurements into visually aligned saliency maps. We first present the problem formulation in Sec. 3.1. We then introduce a visuo-tactile alignment representation in Sec. 3.2, which maps contact information to image-space saliency cues through the robot kinematic chain and camera projection. Finally, leveraging the proposed representation, we introduce a zero-initialized input adaptation in Sec. 3.3 that integrates tactile saliency into pretrained 2D visual encoders while preserving their original visual pathway at initialization.

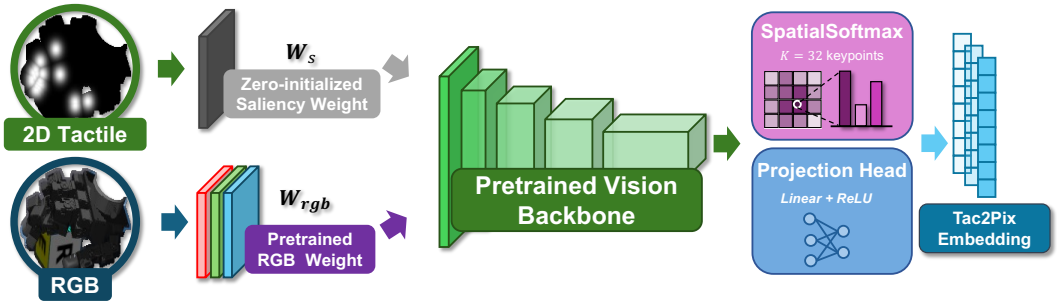


Figure 2: **The Tac2Pix architecture.** Tac2Pix extends a pretrained RGB visual encoder with a zero-initialized saliency channel, allowing projected tactile cues to be fused in the image domain while preserving the original visual representation at initialization.

3.1 Problem Formulation

Our goal is to acquire tactile manipulation skills via imitation learning. We follow the standard visuotactile imitation learning paradigm, in which a policy $\pi_\theta(\mathbf{a}_t \mid \mathbf{o}_{t-k:t})$ is learned from expert demonstrations $\mathcal{D} = \{(\mathbf{o}_t, \mathbf{a}_t)\}_{t=1}^{N}$. Here, $\mathbf{a}_t$ denotes the expert action at time step t, and $\mathbf{o}_t = \{\mathbf{I}_t, \mathbf{s}_t, \mathbf{f}_t\}$ denotes the multimodal robot observation, comprising an RGB image $\mathbf{I}_t \in \mathbb{R}^{H\times W\times 3}$, robot proprioception $\mathbf{s}_t$, and low-dimensional tactile measurements $\mathbf{f}_t$.

Unlike approaches that learn cross-modal alignment from latent features, Tac2Pix constructs an aligned visuo-tactile observation explicitly in image space. Specifically, the raw tactile measurements $\mathbf{f}_t$ are transformed into an image-space saliency map $\mathbf{S}_t \in \mathbb{R}^{H\times W\times 1}$, which is concatenated with the RGB image to form an augmented RGB-Saliency observation:

$$\mathbf{X}_t = \mathrm{Concat}(\mathbf{I}_t, \mathbf{S}_t) \in \mathbb{R}^{H\times W\times 4}. \tag{1}$$

The learning objective is therefore to train a policy $\pi_\theta(\mathbf{a}_t \mid \mathbf{X}_{t-k:t}, \mathbf{s}_{t-k:t}, \mathbf{f}_{t-k:t})$ using standard visual imitation learning losses. By embedding tactile information in the same spatial domain as visual observations, this formulation establishes explicit cross-modal spatial correspondences while avoiding ad hoc architectures for heterogeneous tactile modalities.

3.2 Force-Aware Tactile-to-Pixel Projection

Tac2Pix converts heterogeneous tactile observations into a common image-space representation by addressing two sources of inconsistency: differences in sensor response scale and differences in sensor layout. We first normalize tactile magnitudes across sensor types, and then use robot kinematics and camera geometry to place each tactile response at its corresponding image location.

Specifically, let $\mathbf{f}_t = \{f_{i,t}\}_{i=1}^{M}$ denote the tactile readings from M sensing nodes. Since different tactile sensors may operate at different response scales, we normalize each reading as $\tilde{f}_{i,t} = \tanh\left(\gamma f_{i,t}/F^i_{\mathrm{limit}}\right)$, where F^i_{limit} is the sensor-specific saturation limit and γ controls the scaling. The normalized responses are then placed into a common spatial frame through kinematic projection.

We next localize each normalized tactile response according to the physical position of its sensing node. Each tactile node is associated with a fixed local position $\mathbf{L}_i$ on its corresponding robot link. Given the robot proprioceptive state $\mathbf{s}_t$, its 3D position in the world frame is obtained through forward kinematics:

$$\mathbf{P}_{i,t} = \mathrm{FK}(\mathbf{s}_t, \mathbf{L}_i), \tag{2}$$

where $\mathbf{P}_{i,t} \in \mathbb{R}^3$. For camera view c, let $\mathbf{R}^c \in \mathbb{R}^{3\times 3}$ and $\mathbf{t}^c \in \mathbb{R}^3$ denote the camera extrinsics, and $\mathbf{K}^c \in \mathbb{R}^{3\times 3}$ the intrinsic matrix. The corresponding image coordinate $\mathbf{p}^c_{i,t} = [u^c_{i,t}, v^c_{i,t}]^\top$ is obtained by

$$[u^c_{i,t}, v^c_{i,t}, 1]^\top \sim \mathbf{K}^c\left(\mathbf{R}^c\mathbf{P}_{i,t} + \mathbf{t}^c\right). \tag{3}$$

Projected nodes outside the image bounds are discarded, and the remaining valid nodes form $\mathcal{V}^c_t$. Because these projected responses remain spatially sparse, we render them into a dense saliency

map $\mathbf{S}_t^c \in \mathbb{R}^{H\times W\times 1}$ using force-aware Gaussian kernels:

$$\mathbf{S}_t^c(u,v) = \max_{i\in\mathcal{V}_t^c}\left[\tilde{f}_{i,t}\exp\left(-\frac{(u-u_{i,t}^c)^2+(v-v_{i,t}^c)^2}{2\sigma_{i,t}^2}\right)\right]. \tag{4}$$

The spatial support of each tactile response is modulated by its normalized magnitude:

$$\sigma_{i,t} = \sigma_{\min} + \tilde{f}_{i,t}\left(\sigma_{\max} - \sigma_{\min}\right). \tag{5}$$

The tactile magnitude therefore modulates both response strength and spatial support, allowing stronger contacts to exert a broader influence on nearby visual features. We use max aggregation so that overlapping projected sensors do not artificially amplify the response through summation.

Together, these operations express heterogeneous tactile readings as a common image-space representation for subsequent visuo-tactile learning. The geometric alignment itself is deterministic and contains no learned parameters. The learning process is used purely for developing the downstream policy that infers action.

3.3 Zero-Initialized Backbone Adaptation

After constructing the observation, we integrate the projected tactile saliency into a pretrained 2D visual encoder. Rather than introducing a separate tactile encoder, we augment the visual backbone's input with an additional saliency channel. This design preserves the spatial correspondence established in Sec. 3.2 while allowing pretrained RGB visual representations to be reused.

Let $\mathcal{E}_{\text{rgb}}$ denote the pretrained input projection of a visual encoder and $\mathcal{E}_s$ the newly introduced projection for the saliency channel. For camera view c, the initial visual feature is written as

$$\mathbf{z}_t^c = \mathcal{E}_{\text{rgb}}(\mathbf{I}_t^c) + \mathcal{E}_s(\mathbf{S}_t^c), \qquad \mathcal{E}_s = \mathbf{0} \text{ at initialization.} \tag{6}$$

The pretrained RGB projection is kept unchanged, whereas the parameters associated with the additional saliency input are initialized to zero. Consequently, the Tac2Pix encoder is functionally equivalent to the original pretrained visual encoder at initialization. During policy learning, the saliency parameters are optimized jointly with the downstream policy, allowing spatially aligned tactile cues to be incorporated without disrupting the pretrained visual pathway at the start of training.

This formulation is independent of a particular visual backbone. For convolutional encoders, the additional saliency input corresponds to extending the input-channel weights of the first convolutional layer; for patch-based encoders, the same principle is applied to the input patch projection. The resulting per-view visual features are then combined with robot proprioception over the observation horizon to form the conditioning sequence $\mathbf{g}_{t-T_o+1:t}$ used by the downstream policy.

4 Experiments

We evaluate the proposed Tac2Pix framework to determine whether image-space tactile grounding improves visuo-tactile imitation learning. We focus on three key research questions:

RQ1: Does Tac2Pix consistently improve visuo-tactile policy learning across tasks, policy architectures, and visual backbones? (Sec. 4.1)

RQ2: Is Tac2Pix effective on real-world dexterous manipulation tasks, and how does it compare with alternative spatial tactile representations? (Sec. 4.2)

RQ3: How do saliency rendering, fusion and initialization, and touch–image alignment affect Tac2Pix performance? (Sec. 4.3)

We conduct experiments in both simulation and real-world environments. Policies are trained on demonstrations collected under normal, unobstructed visual observations. They are then evaluated under two conditions: *normal*, with full RGB input, and *occluded*, where a preprogrammed black mask is applied to task-relevant image regions. The main evaluation considers normal and fixed-occlusion conditions; additional robustness studies further evaluate random and physical occlusions. The occlusion mask is used only during evaluation, with the same mask size used across tasks.

Table 1: **Simulation success rates (%).** We evaluate multiple visuo-tactile fusion mechanisms across downstream policy architectures and visual conditions. Tac2Pix denotes our image-aligned tactile saliency representation. The best and second-best performances in each column within each policy architecture are highlighted with lavender and blue backgrounds, respectively; ties share the same color.

Policy Architecture	Fusion Mechanism	Pick-and-Place			Cube-Push			Rotate-Cross		
		Normal	Occlud.	Avg.	Normal	Occlud.	Avg.	Normal	Occlud.	Avg.
Behavior Clone MLP (Zhang et al., 2018)	Vision-Only	7.4	0.0	3.7	33.3	0.0	16.7	10.0	2.0	6.0
	Concat	13.2	0.0	6.6	41.7	0.0	20.8	28.0	6.0	17.0
	FiLM	6.6	0.0	3.3	50.0	3.3	26.7	8.0	0.0	4.0
	CLiP	3.3	0.0	1.7	38.3	3.3	20.8	28.0	2.0	15.0
	Cross-Attn	3.3	8.3	5.8	45.0	21.7	33.3	18.0	6.0	12.0
	Ours (Tac2Pix)	15.7	0.8	8.3	53.3	6.7	30.0	38.0	8.0	23.0
Action Chunk Transformer (Zhao et al., 2023)	Vision-Only	58.7	0.0	29.3	75.0	1.7	38.3	66.0	16.0	41.0
	Concat	67.8	5.8	36.8	80.0	0.0	40.0	92.0	20.0	56.0
	FiLM	36.4	0.8	18.6	71.7	13.3	42.5	60.0	18.0	39.0
	CLiP	38.0	19.0	28.5	60.0	48.3	54.2	68.0	40.0	54.0
	Cross-Attn	67.8	6.6	37.2	68.3	23.3	45.8	84.0	26.0	55.0
	Ours (Tac2Pix)	71.9	23.1	47.5	85.0	46.7	65.8	90.0	44.0	67.0
Diffusion Policy (Chi et al., 2025)	Vision-Only	71.9	7.4	39.7	96.7	25.0	60.8	78.0	26.0	52.0
	Concat	72.7	14.9	43.8	90.0	25.0	57.5	80.0	38.0	59.0
	FiLM	71.9	13.2	42.6	95.0	38.3	66.7	64.0	42.0	53.0
	CLiP	62.0	24.8	43.4	85.0	43.3	64.2	64.0	48.0	56.0
	Cross-Attn	42.1	34.7	38.4	71.7	46.7	59.2	70.0	52.0	61.0
	Ours (Tac2Pix)	77.7	52.1	64.9	90.0	43.3	66.7	92.0	40.0	66.0
SmolVLA (Shukor et al., 2025)	Vision-Only	62.8	34.7	48.8	76.7	10.0	43.3	82.0	64.0	73.0
	Concat	63.6	28.9	46.3	61.7	30.0	45.8	84.0	64.0	74.0
	FiLM	57.0	28.1	42.6	68.3	31.7	50.0	78.0	60.0	69.0
	CLiP	55.4	32.2	43.8	68.3	11.7	40.0	84.0	50.0	67.0
	Cross-Attn	55.4	38.0	46.7	70.0	26.7	48.3	72.0	64.0	68.0
	Ours (Tac2Pix)	69.4	36.4	52.9	71.7	31.7	51.7	84.0	72.0	78.0

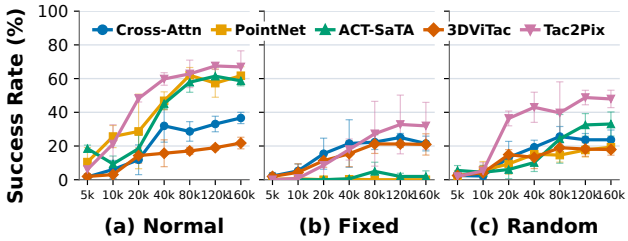


Figure 3: **Learning curves on simulated Pick-and-Place.** Curves show validation success with standard deviation.

Table 2: **Compatibility across visual backbones.** Using different visual encoders, reported as Normal/Occluded.

Method	R3M	VIP	ResNet
Vision-Only	46.3/9.9	48.8/9.9	58.7/0.0
Cross-Attn	49.6/4.1	27.3/25.6	67.8/6.6
Tac2Pix	41.3/28.9	63.6/27.3	71.9/23.1

In simulation, we evaluate three dexterous manipulation tasks: pick-and-place, cube-push, and rotate-cross. Together, these tasks test object localization, contact-rich interaction, and robustness under partial observability. For real-world evaluation, we deploy the learned policies on an xArm6 equipped with a LEAP Hand (Shaw et al., 2023). The hand includes 12 joint-mounted FSR sensors and 4 fingertip TwinTac sensors (Huang et al., 2025b). Visual observations are captured by two calibrated RealSense D435 cameras. We evaluate three real-world tasks: pick-and-place, open-drawer, and flip-box, under both normal and occluded observation conditions. Details on demonstrations, tactile simulation, occlusion masks, and evaluation initialization ranges are provided in Appendix A.

4.1 Simulation Evaluations

To answer RQ1, we evaluate whether Tac2Pix provides consistent benefits across tasks, policy architectures, visual backbones, and visual conditions in simulation. We consider three manipulation tasks and four policy backbones: Behavior Cloning (BC) MLP (Zhang et al., 2018), Action Chunking Transformer (ACT) (Zhao et al., 2023), Diffusion Policy (DP) (Chi et al., 2025), and SmolVLA (Shukor et al., 2025). For each policy class, we compare Tac2Pix with vision-only input, concatenation, FiLM (Perez et al., 2018), CLIP-style alignment (Radford et al., 2021), and cross-attention (Jaegle et al., 2021).



Figure 4: **Real-world tasks and occlusion protocol.** We evaluate Tac2Pix on three real-world dexterous manipulation tasks by attaching a fixed task-relevant visual mask after the highlighted interaction stage and comparing policy observations with and without tactile saliency.

Performance across tasks and policies. Table 1 reports success rates for the three simulation tasks under normal and occluded visual conditions. Overall, across tasks and policy architectures, Tac2Pix achieves the best or second-best performance in most settings. We make two main observations. First, simply adding tactile inputs does not consistently improve over vision-only policies. Concat, FiLM, CLIP-style fusion, and Cross-Attn improve some settings but degrade others. This is consistent with recent dexterous-manipulation studies showing that naive tactile conditioning can introduce irrelevant tactile interference or degrade existing policy capabilities (Zhang et al., 2026; Niu et al., 2026). This suggests that explicitly projecting tactile cues into image space provides a useful geometric prior, avoiding the need to recover touch–vision correspondence implicitly during policy learning. Second, Tac2Pix shows the strongest overall performance under visual occlusion. It ranks first or second in most occluded settings and achieves the highest average occluded success across all evaluated fusion mechanisms.

Beyond final success rates, Figure 3 evaluates simulated Pick-and-Place over three training seeds under normal, fixed, and random occlusion. Tac2Pix maintains its advantage throughout training and across seeds, while remaining effective when the occlusion pattern changes. Together, these results indicate that the observed gains are consistent across policy architectures, optimization runs, and visual conditions rather than being tied to a particular initialization or fixed mask.

Compatibility across visual backbones. To isolate the effect of the visual representation, we fix the task to simulated *Pick-and-Place* and use ACT as the downstream policy, while replacing only the visual encoder with R3M (Nair et al., 2022), VIP (Ma et al., 2023), and ResNet. As shown in Table 2, Tac2Pix achieves the highest average success across normal and occluded conditions for all three backbones, as well as the highest success under occlusion in each case. These results show that the benefit of image-space tactile grounding persists across the evaluated pretrained visual representations. Such compatibility also offers a path for Tac2Pix to benefit from stronger future visual backbones without redesigning the tactile representation.

4.2 Real-World Deployment

To answer RQ2, we deploy Diffusion Policy on three real-world tasks under normal and visually degraded observations. Figure 4 illustrates the task setups, tactile saliency, and the stages at which visual occlusion is introduced. To assess whether the benefits of image-space tactile grounding transfer across different real-world manipulation behaviors, we compare Tac2Pix with the baselines on all three tasks under both visual conditions. As shown in Table 3, Tac2Pix achieves the highest average success across *Pick-and-Place*, *Open Drawer*, and *Flip Box*, with substantially larger gains

Table 3: **Real-world experiments.** We evaluate policies across three tasks under both normal and severely occluded visual conditions. Ours maintains high robustness when vision is compromised.

Method	Pick & Place		Open Drawer		Flip Box		Average (%)	
	Normal	Occluded	Normal	Occluded	Normal	Occluded	Normal	Occluded
Vision-Only	9/20	0/20	12/20	4/20	13/20	2/20	56.7	10.0
Concat	9/20	1/20	10/20	6/20	14/20	1/20	55.0	13.3
Cross-Attn	4/20	0/20	6/20	4/20	11/20	11/20	35.0	25.0
Ours (Tac2Pix)	9/20	7/20	14/20	10/20	17/20	14/20	66.7	51.7

Table 4: **Real-world robustness beyond fixed visual occlusion.** Success rates (mean $\pm$ std, %) over three independently trained seeds under normal, fixed, random, and physical occlusion.

Method	Normal	Fixed Occ.	Random Occ.	Physical Occ.
Vision-Only	41.7 ± 4.7	6.7 ± 9.4	23.3 ± 15.5	25.0 ± 4.1
Cross-Attn	25.0 ± 4.1	6.7 ± 4.7	10.0 ± 4.1	21.7 ± 6.2
3D-ViTac (Huang et al., 2025a)	5.0 ± 4.1	0.0 ± 0.0	0.0 ± 0.0	1.7 ± 2.4
SaTA (Huang et al., 2026)	13.3 ± 2.4	0.0 ± 0.0	5.0 ± 0.0	3.3 ± 2.4
Ours (Tac2Pix)	$\mathbf{46.7 \pm 2.4}$	$\mathbf{25.0 \pm 7.1}$	$\mathbf{28.3 \pm 8.5}$	$\mathbf{31.7 \pm 4.7}$

when task-relevant visual information is occluded. This suggests that explicit image-space tactile grounding is particularly useful when visual evidence becomes incomplete.

To further test whether the advantage of Tac2Pix extends beyond the fixed-mask setting, we conduct a three-seed real-world evaluation on *Pick-and-Place* and compare against adapted 3D-ViTac and SaTA (Table 4). We additionally introduce *Random Occ.*, which varies the occluded image region during execution, and *Physical Occ.*, which introduces human-induced occlusion directly in front of the camera. Tac2Pix achieves the highest success across all four conditions, including normal, fixed, random, and physical occlusion. Importantly, its advantage persists over both implicit fusion and alternative spatial tactile representations, indicating that directly grounding touch in image space remains effective beyond a particular occlusion pattern or fusion architecture.

4.3 Ablation on Tac2Pix Design Choices

To answer RQ3, we ablate three key design choices in Tac2Pix: (1) the tactile saliency rendering strategy, to examine whether alternative rendering variants yield better performance; (2) the visuo-tactile fusion architecture, to assess the effectiveness of the proposed fusion mechanism; and (3) spatial alignment between visual and tactile inputs, to evaluate robustness to projection errors. We conduct all ablations using Diffusion Policy on the pick-and-place task and compare success rates across design variants.

Ablation on Tac2Pix rendering variants. Our proposed force-aware Tac2Pix represents projected contacts as force-modulated Gaussian kernels. To validate this rendering design, we compare it with two alternatives in simulation: (a) *RGB Overlay*, which renders tactile cues directly on the RGB image while preserving a standard three-channel input; and (b) *Binary Tac2Pix*, which appends a fourth saliency channel but assigns a constant intensity to all active contacts, thereby encoding contact location without force magnitude.

Table 5 summarizes the results. Under normal, unobstructed vision, the performance differences among the representations are relatively minor. Under occlusion, however, these differences become more pronounced: both tactile-augmented variants substantially outperform the vision-only baseline. In particular, binary Tac2Pix demonstrates that encoding contact location alone already provides a strong geometric prior for manipulation. Nevertheless, force-aware Tac2Pix achieves the highest success rates in both visual settings, suggesting that continuous force magnitude conveys additional interaction information beyond binary contact.

Ablation on fusion architecture. We examine the effects of fusion stage and initialization through three variants: (1) *Late Fusion*, which encodes RGB and saliency separately before pooling and concatenating their features; (2) *Intermediate Fusion*, which encodes saliency with a lightweight convolutional network and adds its features to an intermediate ResNet feature map; and (3) *Early*

Table 5: **Ablation on tactile saliency rendering.** We compare RGB overlay, binary contact saliency, and force-aware under both conditions.

Variant	Normal	Occ.	Avg.
Vision-Only	71.9	7.4	39.7
RGB Overlay	66.9	28.1	47.5
Binary Tac2Pix	70.2	44.6	57.4
Ours (Force-aware)	**77.7**	**52.1**	**64.9**

Table 6: **Ablation on fusion stage and initialization.** We compare late, intermediate, and assess the effect of zero initialization.

Variant	Normal	Occ.	Avg.
Late Fusion	64.5	14.0	39.3
Intermediate Fusion	69.4	26.4	47.9
Early w/o Zero Init.	49.6	21.5	35.6
Ours (Early)	**77.7**	**52.1**	**64.9**

Table 7: **Sensitivity to spatial misalignment.** Success under controlled translational offsets of the tactile saliency map while keeping RGB unchanged, evaluated in simulation and the real world.

Setting	Condition	0 px	25 px	50 px	100 px
Sim	Normal	77.7%	71.9%	71.1%	67.8%
	Occluded	52.1%	27.3%	31.4%	18.2%
Real	Normal	9/20	8/20	5/20	5/20
	Occluded	7/20	4/20	3/20	3/20

Fusion w/o Zero Initialization, which retains our architecture but uses standard initialization for the added saliency-channel weights.

As shown in Table 6, intermediate fusion outperforms late fusion under both visual conditions, with a larger advantage under occlusion. Removing zero initialization reduces early-fusion success under normal vision and occlusion. The full model achieves the highest success in both settings, supporting the combination of early tactile injection and zero-initialized input adaptation.

Ablation on spatial alignment. Finally, to assess the importance of touch–image alignment, Table 7 quantifies Tac2Pix's sensitivity to spatial misalignment. To this end, we introduce controlled pixel shifts (Δ_x, Δ_y) to tactile saliency map. The offset is fixed within each trial and randomly sampled across trials: for saliency map $S \in \mathbb{R}^{H\times W}$, we construct

$$S_\Delta(x, y) = S\big((x - \Delta_x) \bmod W,\ (y - \Delta_y) \bmod H\big), \tag{7}$$

while leaving the RGB image unchanged. The results show that performance generally decreases as spatial misalignment increases, with a larger drop under visual occlusion than under normal vision. The same qualitative trend appears in simulation and the real world, confirming that correct touch–image correspondence becomes particularly important when visual evidence is degraded.

Overall, the experiments support a reusable design principle: known contact geometry can be expressed in a visual form that standard policy models can consume directly. Tac2Pix shows that this representation yields practical benefits in both simulation and real-world manipulation, particularly when vision is unreliable.

5 Conclusion

In this paper, we introduced Tac2Pix, a visuo-tactile framework that projects tactile contacts onto the 2D image plane as force-aware saliency maps. Through zero-initialized input adaptation, Tac2Pix injects explicit spatial priors into standard visual encoders while preserving pretrained visual representations. Across simulated and real-world experiments, Tac2Pix consistently improves visuo-tactile imitation learning, particularly when vision is degraded. These results support image space as a simple, broadly compatible interface for integrating sparse touch with visual representations. Tac2Pix currently relies on known sensor geometry and calibrated camera parameters, and its 2D projection does not preserve full depth information. Future work will explore calibration-robust projection. More broadly, Tac2Pix offers a path to connect touch with the rapidly evolving visual learning community, enabling robot policies to leverage advances in visual backbones without large-scale tactile pretraining.

AI USE STATEMENT

In this work, we used generative AI tools to search for and summarize relevant literature, provide feedback on research methods and experimental design, suggest experimental parameters, write and edit software code, clean and reformat data, support data analysis, and improve the readability of the manuscript. We did not use generative AI tools to develop theoretical models or conceptual frameworks, propose or refine hypotheses, generate synthetic datasets, formulate mathematical claims, contribute essential steps to mathematical proofs, write proofs, or interpret the results. Translation was not applicable to this work. The authors reviewed all AI-assisted work and verified and tested AI-generated code. We take responsibility for the final content of this work, including any text, claims, or artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of our work. We provide detailed experimental protocols, baseline adaptations, ablation settings, and policy/training configurations in the appendix (Secs. A–E). Our data generation, training, and inference pipelines are built on the open-source LeRobot framework. We plan to release the modular Tac2Pix implementation together with the configurations to reproduce our experiments upon acceptance.

REFERENCES

Raunaq Bhirangi, Venkatesh Pattabiraman, Enes Erciyes, Yifeng Cao, Tess Hellebrekers, and Lerrel Pinto. Anyskin: Plug-and-play skin sensing for robotic touch, 2024. URL https://arxiv.org/abs/2409.08276.

Remi Cadene, Simon Aliberts, Francesco Capuano, Michel Aractingi, Adil Zouitine, Pepijn Kooijmans, Jade Choghari, Martino Russi, Caroline Pascal, Steven Palma, et al. Lerobot: An open-source library for end-to-end robot learning. *arXiv preprint arXiv:2602.22818*, 2026.

Roberto Calandra, Andrew Owens, Manu Upadhyaya, Wenzhen Yuan, Justin Lin, Edward H. Adelson, and Sergey Levine. The feeling of success: Does touch sensing help predict grasp outcomes? In Sergey Levine, Vincent Vanhoucke, and Ken Goldberg (eds.), *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pp. 314–323. PMLR, 13–15 Nov 2017.

Roberto Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, Edward H. Adelson, and Sergey Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3(4):3300–3307, October 2018. ISSN 2377-3774. doi: 10.1109/lra.2018.2852779. URL http://dx.doi.org/10.1109/LRA.2018.2852779.

Linghao Chen, Yuzhe Qin, Xiaowei Zhou, and Hao Su. Easyhec: Accurate and automatic hand-eye calibration via differentiable rendering and space exploration. *IEEE Robotics and Automation Letters*, 8(11):7234–7241, November 2023a. ISSN 2377-3774. doi: 10.1109/lra.2023.3315551. URL http://dx.doi.org/10.1109/LRA.2023.3315551.

Sirui Chen, Chen Wang, Kaden Nguyen, Li Fei-Fei, and C Karen Liu. Arcap: Collecting high-quality human demonstrations for robot learning with augmented reality feedback. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8291–8298. IEEE, 2025.

Xitong Chen, Yifeng Pan, Min Li, and Xiaotian Ding. Dexvitac: Collecting human visuo-tactile-kinematic demonstrations for contact-rich dexterous manipulation, 2026. URL https://arxiv.org/abs/2603.17851.

Yizhou Chen, Mark Van der Merwe, Andrea Sipos, and Nima Fazeli. Visuo-tactile transformers for manipulation. In Karen Liu, Dana Kulic, and Jeff Ichnowski (eds.), *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 2026–2040. PMLR, 14–18 Dec 2023b.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Vedant Dave, Fotios Lygerakis, and Elmar Rueckert. Multimodal visual-tactile representation learning through self-supervised contrastive pre-training. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8013–8020. IEEE, 2024.

Siyuan Dong, Daolin Ma, Elliott Donlon, and Alberto Rodriguez. Maintaining grasps within slipping bounds by monitoring incipient slip. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 3818–3824. IEEE, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Ruoxuan Feng, Di Hu, Wenke Ma, and Xuelong Li. Play to the score: Stage-guided dynamic multisensory fusion for robotic manipulation. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard (eds.), *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 340–363. PMLR, 06–09 Nov 2025a.

Ruoxuan Feng, Jiangyu Hu, Wenke Xia, Tianci Gao, Ao Shen, Yuhao Sun, Bin Fang, and Di Hu. Anytouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors, 2025b. URL https://arxiv.org/abs/2502.12191.

Irmak Guzey, Ben Evans, Soumith Chintala, and Lerrel Pinto. Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play. In Jie Tan, Marc Toussaint, and Kourosh Darvish (eds.), *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 3142–3166. PMLR, 06–09 Nov 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Binghao Huang, Yixuan Wang, Xinyi Yang, Yiyue Luo, and Yunzhu Li. 3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing, 2025a. URL https://arxiv.org/abs/2410.24091.

Jialei Huang, Yang Ye, Yuanqing Gong, Xuezhou Zhu, Yang Gao, and Kaifeng Zhang. Spatially anchored tactile awareness for robust dexterous manipulation, 2026. URL https://arxiv.org/abs/2510.14647.

Xiyan Huang, Zhe Xu, and Chenxi Xiao. Twintac: A wide-range, highly sensitive tactile sensor with real-to-sim digital twin sensor model, 2025b. URL https://arxiv.org/abs/2509.10063.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pp. 4651–4664. PMLR, 2021.

Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, Dinesh Jayaraman, and Roberto Calandra. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3): 3838–3845, July 2020. ISSN 2377-3774. doi: 10.1109/lra.2020.2977257. URL http://dx.doi.org/10.1109/LRA.2020.2977257.

Michelle A Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 International conference on robotics and automation (ICRA)*, pp. 8943–8950. IEEE, 2019.

Hongyu Li, Snehal Dikhale, Soshi Iba, and Nawid Jamali. Vihope: Visuotactile in-hand object 6d pose estimation with shape completion. *IEEE Robotics and Automation Letters*, 8(11):6963–6970, 2023. doi: 10.1109/LRA.2023.3313941.

Pei Lin, Yuzhe Huang, Wanlin Li, Jianpeng Ma, Chenxi Xiao, and Ziyuan Jiao. Pp-tac: Paper picking using tactile feedback in dexterous robotic hands. *arXiv preprint arXiv:2504.16649*, 2025.

Shengcheng Luo, Xiaoyang Chen, Hong Ying, Xiaoying Zhou, Jiaming Jiang, Haoran Guo, Wanlin Li, Ziyuan Jiao, and Chenxi Xiao. Touch2robot: Robot touch in the human demonstration loop, 2026a. URL https://arxiv.org/abs/2609.24660.

Shengcheng Luo, Xiyan Huang, Zhe Xu, Wanlin Li, Ziyuan Jiao, and Chenxi Xiao. Blind dexterous grasping via real2sim2real tactile policy learning, 2026b. URL https://arxiv.org/abs/2606.11767.

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training, 2023. URL https://arxiv.org/abs/2210.00030.

Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Tingfan Wu, Jay Vakil, Pieter Abbeel, Jitendra Malik, Dhruv Batra, Yixin Lin, Oleksandr Maksymets, Aravind Rajeswaran, and Franziska Meier. Where are we in the search for an artificial visual cortex for embodied intelligence? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 655–677. Curran Associates, Inc., 2023.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation, 2022. URL https://arxiv.org/abs/2203.12601.

Dantong Niu, Zhuoyang Liu, Zekai Wang, Boning Shao, Zhao-Heng Yin, Anirudh Pai, Yuvan Sharma, Stefano Saravalle, Ruijie Zheng, Jing Wang, Ryan Punamiya, Mengda Xu, Yuqi Xie, Yunfan Jiang, Letian Fu, Konstantinos Kallidromitis, Matteo Gioia, Junyi Zhang, Jiaxin Ge, Haiwen Feng, Fabio Galasso, Wei Zhan, David M. Chan, Yutong Bai, Roei Herzig, Jiahui Lei, Li Fei-Fei, Ken Goldberg, Jitendra Malik, Pieter Abbeel, Yuke Zhu, Danfei Xu, Linxi Fan, and Trevor Darrell. T-rex: Tactile-reactive dexterous manipulation, 2026. URL https://arxiv.org/abs/2606.17055.

Venkatesh Pattabiraman, Yifeng Cao, Siddhant Haldar, Lerrel Pinto, and Raunaq Bhirangi. Learning precise, contact-rich manipulation through uncalibrated tactile skins, 2024. URL https://arxiv.org/abs/2410.17246.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In Karen Liu, Dana Kulic, and Jeff Ichnowski (eds.), *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 416–426. PMLR, 14–18 Dec 2023.

Samanta Rodriguez, Yiming Dou, William van den Bogert, Miquel Oller, Kevin So, Andrew Owens, and Nima Fazeli. Contrastive touch-to-touch pretraining. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5857–5863. IEEE, 2025.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Carmelo Sferrazza, Younggyo Seo, Hao Liu, Youngwoon Lee, and Pieter Abbeel. The power of the senses: Generalizable manipulation from vision and touch through masked multimodal learning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9698–9705, 2024. doi: 10.1109/IROS58592.2024.10802719.

Kenneth Shaw, Ananye Agarwal, and Deepak Pathak. Leap hand: Low-cost, efficient, and anthropomorphic hand for robot learning. *arXiv preprint arXiv:2309.06440*, 2023.

Yu She, Shaoxiong Wang, Siyuan Dong, Neha Sunil, Alberto Rodriguez, and Edward Adelson. Cable manipulation with a tactile-reactive gripper. *The International Journal of Robotics Research*, 40(12-14):1385–1401, 2021.

Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, Simon Alibert, Matthieu Cord, Thomas Wolf, and Remi Cadene. Smolvla: A vision-language-action model for affordable and efficient robotics, 2025. URL https://arxiv.org/abs/2506.01844.

Entong Su, Chengzhe Jia, Yuzhe Qin, Wenxuan Zhou, Annabella Macaluso, Binghao Huang, and Xiaolong Wang. Sim2real manipulation on unknown objects with tactile-based reinforcement learning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9234–9241. IEEE, 2024.

Tianhao Wu, Jinzhou Li, Jiyao Zhang, Mingdong Wu, and Hao Dong. Canonical representation and force-based pretraining of 3d tactile for dexterous visuo-tactile policy learning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6786–6792. IEEE, 2025.

Yueh-Hua Wu, Jiashun Wang, and Xiaolong Wang. Learning generalizable dexterous manipulation from human grasp affordance. In *Conference on robot learning*, pp. 618–629. PMLR, 2023.

Zhe Xu, Feiyu Zhao, Xiyan Huang, and Chenxi Xiao. Etac: A lightweight and efficient tactile simulation framework for learning dexterous manipulation. *arXiv preprint arXiv:2604.20295*, 2026.

Zhengrong Xue, Han Zhang, Jingwen Cheng, Zhengmao He, Yuanchen Ju, Changyi Lin, Gu Zhang, and Huazhe Xu. Arraybot: Reinforcement learning for generalizable distributed manipulation through touch. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16744–16751. IEEE, 2024.

Wenzhen Yuan, Siyuan Dong, and Edward H. Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors (Basel, Switzerland)*, 17, 2017. URL https://api.semanticscholar.org/CorpusID:3474913.

Ying Yuan, Haichuan Che, Yuzhe Qin, Binghao Huang, Zhao-Heng Yin, Kang-Won Lee, Yi Wu, Soo-Chul Lim, and Xiaolong Wang. Robot synesthesia: In-hand manipulation with visuotactile sensing, 2024. URL https://arxiv.org/abs/2312.01853.

Kaidi Zhang, Heng Zhang, Zhengtong Xu, Zhiyuan Zhang, Md Rakibul Islam Prince, Xiang Li, Xiaojing Han, Yuhao Zhou, Arash Ajoudani, and Yu She. Tacvla: Contact-aware tactile fusion for robust vision-language-action manipulation, 2026. URL https://arxiv.org/abs/2603.12665.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Xi Chen, Ken Goldberg, and Pieter Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 5628–5635. Ieee, 2018.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware, 2023. URL https://arxiv.org/abs/2304.13705.

Xinyue Zhu, Binghao Huang, and Yunzhu Li. Touch in the wild: Learning fine-grained manipulation with a portable visuo-tactile gripper. *ArXiv*, abs/2507.15062, 2025. URL https://api.semanticscholar.org/CorpusID:280270301.

APPENDIX OVERVIEW

This appendix provides additional details on the experimental setup, simulation and real-world task configurations, qualitative analyses, fusion mechanisms, and downstream policy training.

A. Experimental Setup and Task Details describes the real-world hardware platform, teleoperation interface, observation and action spaces, tactile sensing layout, and demonstration collection protocol. It also details the simulation and real-world task configurations, including saliency generation, task initialization ranges, and demonstration statistics.

B. Extended Discussions and Visualizations provides additional analyses of the proposed tactile saliency representation, including depth ambiguity in 2D projection, inference efficiency compared with 3D baselines, and Grad-CAM visualizations under visual occlusion.

C. Saliency Fusion Ablation Details describes details of different saliency fusion architectures mentioned in Sec 4.3.

D. Training Details of Fusion Mechanisms describes the implementation details of the evaluated visuo–tactile fusion baselines, including cross-attention fusion, FiLM modulation, and CLIP-based vision–tactile contrastive pretraining.

E. Training Details of Downstream Policies summarizes the downstream policy backbones used in our experiments, including BC-MLP, ACT, and Diffusion Policy, together with their architectures, training objectives, and hyperparameters.

A EXPERIMENTAL SETUP AND TASK DETAILS

The complete physical setup, including the tactile sensor layout and camera viewpoints, is shown in Fig. 5. We use a unified teleoperation interface for both simulated and real-world data collection. The operator controls the xArm6 and LEAP Hand using a Meta Quest 3 headset for wrist tracking and visual feedback, together with a Manus Quantum Metaglove for finger-motion capture. Motion retargeting and data recording are built on the open-source ARCap framework (Chen et al., 2025). For each task, we collect task-successful expert demonstrations using visual teleoperation, while robot tactile signals are synchronously recorded during execution. Although the operator does not receive tactile or haptic feedback, the recorded tactile observations capture the robot's contact state along the expert trajectory and are used as additional policy inputs during learning. Unless otherwise stated, evaluation uses a separate set of initial states that expands the demonstration initialization range to test robustness beyond the training distribution.

All policies use the same observation and action interface in simulation and the real world. At each timestep, the observation contains proprioception, two RGB camera views, and tactile readings. The proprioceptive state is $s_t = [q_t^{\text{arm}}, q_t^{\text{hand}}] \in \mathbb{R}^{22}$, including 6 arm joints and 16 hand joints. Each RGB image is center-cropped and resized to 224×224. For Tac2Pix policies, we additionally provide one saliency map per camera view at the same resolution. Unless otherwise stated, all models are trained on NVIDIA A40 GPUs and deployed on an NVIDIA RTX 4090 GPU.

The tactile observation consists of 32 taxel readings from four TwinTac fingertip sensors (Huang et al., 2025b) and 12 FSR readings distributed across the hand: $\tau_t = [\tau_t^{\text{taxel}}, \tau_t^{\text{fsr}}] \in \mathbb{R}^{44}$, where $\tau_t^{\text{taxel}} \in \mathbb{R}^{32}$ and $\tau_t^{\text{fsr}} \in \mathbb{R}^{12}$. The policy outputs a 22-dimensional target command, $a_t = [a_t^{\text{arm}}, a_t^{\text{hand}}] \in \mathbb{R}^{22}$. All demonstrations and policy rollouts are represented at 20 Hz.

A.1 SIMULATION TASK DETAILS

Saliency generation in simulation. In simulation, tactile signals are generated using a tactile simulator built on ETac (Xu et al., 2026). Simulation uses the same 44-node tactile layout as the real robot, consisting of 32 TwinTac taxels and 12 FSR nodes. Each node is attached to its corresponding hand link with a fixed local offset, and its 3D position is obtained through forward kinematics from the simulated joint state. The resulting 44 scalar tactile readings are provided identically to all

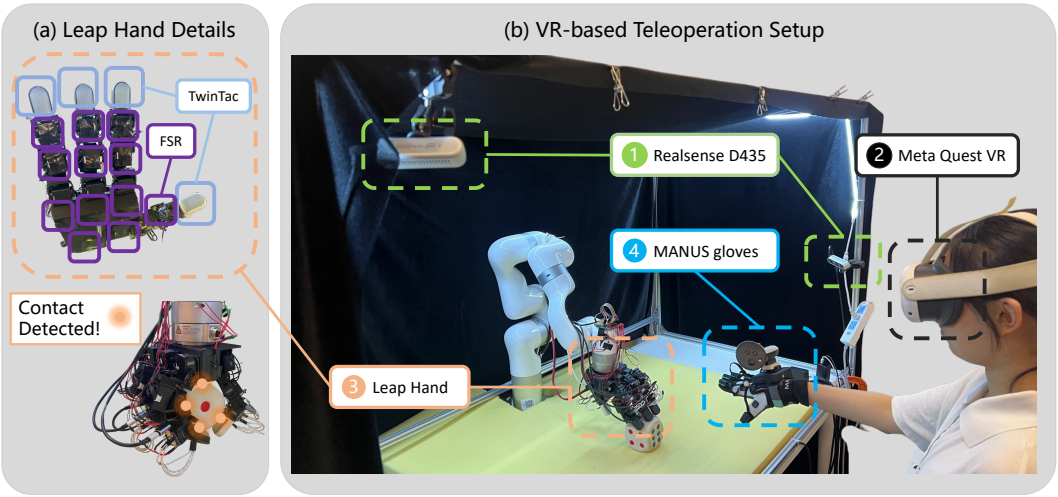


Figure 5: **Real-world teleoperation and deployment platform.**

methods. For Tac2Pix, these same nodes are projected into each camera view and rendered using the force-dependent Gaussian formulation in Sec. 3.2. We use $\sigma_{\min} = 4$ and $\sigma_{\max} = 12$ in all experiments.

Pick and Place. The Pick-and-Place task evaluates whether a policy can grasp a cube-like object, lift it from the tabletop, and drop it into a target region. We collect 53 expert demonstrations, totaling 20,212 frames. During demonstration collection, the cube's initial position is uniformly sampled from a 12×12 cm region, as shown in Fig. 6d. For evaluation, we test policies on 121 initial states, including both in-distribution and out-of-distribution (OOD) cases; the OOD split expands the sampling region to 16×16 cm.

Cube Push. The Cube-Push task evaluates whether a policy can perform planar, contact-rich manipulation by pushing a cube into a target slot. We collect 32 expert demonstrations, totaling 9,500 frames. During demonstration collection, the cube is initialized uniformly within a 2×8 cm area on the tabletop, as shown in Fig. 6e. For evaluation, we test policies on 60 initial configurations, with cube positions uniformly sampled from a 2×16 cm area.

Rotate Cross. The Rotate-Cross task evaluates whether a policy can rotate a valve-like object through sustained contact. We collect 25 expert demonstrations, totaling 11,084 frames. During demonstration collection, the cross is initialized uniformly within a 10×10 cm area on the tabletop, as shown in Fig. 6f. For evaluation, we test policies on 50 initial configurations, with cross positions uniformly sampled from a 12×12 cm area.

A.2 Real Tasks Details

Real-world saliency generation. Real-world policies use the observation and action definition described in Sec. A. For Tac2Pix policies, real-world saliency is computed from the physical tactile sensing nodes on the hand: 4×8 TwinTac fingertip taxels and 12 FSR sensors, yielding 44 projected tactile nodes in total. At each timestep, the 3D position of each tactile node is computed from the robot forward kinematics and its fixed local offset on the corresponding hand link. We calibrate the camera extrinsics using EasyHEC (Chen et al., 2023a) and use the calibrated RealSense intrinsics for image projection. The projected tactile nodes are then rendered into image-space saliency maps using the force normalization, Gaussian rendering, and max-aggregation procedure described in Sec. 3.2. This produces one saliency map for each camera view.

Pick and Place. The task configuration is similar to that in simulation. The policy must grasp a randomly initialized cube on one side of the workspace and drop it into a specified bowl on the other side. We collect 48 expert demonstrations for this task, corresponding to 23,820 frames, with

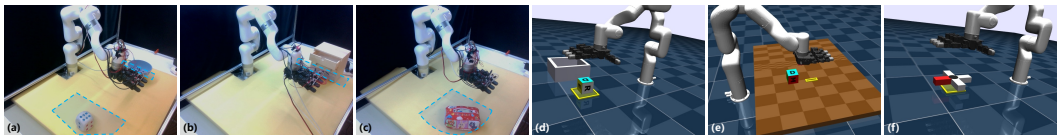


Figure 6: **Initialization workspace of operational objects.** The blue areas in real tasks and yellow areas in simulation tasks are initial workspace of operation objects. (a) Pick and place (Real). It is noteworthy that the bowl where the cube should be lifted in, is also randomly placed initially; (b) Open drawer. (Real); (c) Flip box(Real); (d) Pick and place (Sim); (e) Cube Push (Sim); (f) Rotate cross (Sim).

initializations shown in Fig. 6a. During demonstration collection, the cube is initialized randomly within a 20×30 cm area on the tabletop.

Open Drawer. The Open-Drawer task evaluates whether the dexterous hand can hook its fingers onto the edge of a partially opened drawer and pull it outward to complete the drawer-opening motion. We collect 48 expert demonstrations for this task, corresponding to 11,848 frames, with initializations shown in Fig. 6b. During demonstration collection, the drawer pose is initialized randomly within a 5×30 cm area on the tabletop.

Flip Box. The Flip-Box task evaluates whether the dexterous hand can use its fingers to manipulate a tissue box twice along its long edge, rotating it by a total of 180 degrees until its bottom face points upward. We collect 26 expert demonstrations for this task, corresponding to 12,125 frames, with initializations shown in Fig. 6c. During demonstration collection, the tissue box is initialized within a sector-shaped region with a radius of 35 cm on one side of the tabletop.

Relation between the two real-world evaluations. Table 3 reports the original single-seed evaluation across all three real-world tasks, with 20 trials per task and condition. Table 4 expands the Pick-and-Place evaluation to three independently trained seeds and additionally includes random and physical occlusion settings. The original Pick-and-Place run is included as one of the three seeds in the expanded evaluation.

Occlusion protocols. We evaluate three visual-degradation settings. *Fixed Occlusion* uses the task-specific software mask from the main evaluation, with the mask activated after the predefined interaction stage and kept fixed for the remainder of the rollout. *Random Occlusion* uses the same mask size and activation stage, but relocates the mask every 20 frames among predefined task-relevant image regions. The complete mask schedule is shared across methods for matched evaluation. *Physical Occlusion* introduces dynamic human-induced occlusion directly in front of the camera, without digitally modifying the RGB stream.

Adaptation of spatial tactile baselines. We adapt 3D-ViTac by replacing its dense tactile arrays with our 44 sparse sensing nodes. Each node is represented by its tactile reading and its 3D position computed through forward kinematics. We combine these tactile points with 512 points from camera depth images, giving priority to points near the tactile sensors and within the interaction workspace. Following the original framework, we use an adapted PointNet++ encoder to process the combined points and a diffusion policy to predict actions.

We replace SaTA's tactile-image encoder with a shared MLP over the eight taxel readings from each TwinTac fingertip. We retain spatial anchoring through Fourier encoding and FiLM, using fingertip poses computed by forward kinematics in the palm frame. The 12 FSR readings form an additional global token, which joins the four fingertip tokens, RGB features, and proprioception in the ACT policy. The lower performance of the adapted 3D-ViTac and SaTA baselines likely reflects a sensing–representation mismatch: their original formulations exploit richer spatial tactile observations, whereas our setting provides sparse scalar readings.

Table 8: **Efficiency comparison per step.** Denoising, pre-denoising, and overall time costs are reported in milliseconds.

Model	Denoising Time (ms)	Pre-denoising Latency (ms)	Overall Time (ms)
Vision-Only	64.26 ± 0.01	10.10 ± 5.46	74.36
Concat	64.23 ± 0.00	10.23 ± 4.24	74.46
FiLM	64.24 ± 0.00	11.60 ± 3.31	75.84
CLiP	66.39 ± 0.00	7.81 ± 2.01	74.20
Point Cloud	76.72 ± 0.00	95.12 ± 7.83	171.84
Cross-Attn	64.56 ± 0.00	15.13 ± 3.96	79.69
Ours (Tac2Pix)	64.24 ± 0.00	21.06 ± 4.54	85.30

B Extended Discussions and Visualizations

In this section, we provide additional analyses and qualitative visualizations to discuss depth ambiguity in our projection model (Sec. B.1), the efficiency analysis (Sec. B.2), and the policy's attention mechanism under visual occlusion (Sec. B.3).

B.1 Addressing Depth Ambiguity in 2D Projection

A limitation of our kinematic projection model is its limited expressivity in representing the full 3D contact location. In particular, contacts occurring on the side of the object facing away from the camera may still project onto the 2D foreground. This creates depth ambiguity: from a single 2D saliency map, the model cannot determine whether a contact lies on the front or rear surface of the object, nor can it precisely distinguish which front or rear tactile sensor is activated.

However, this ambiguity has little effect on overall performance across the tasks we evaluate, as reflected in Tables 1 and 3. We attribute this robustness to three factors. First, the policy has access to the 3D kinematic configuration of the robot fingers through proprioception, which provides information beyond the projected saliency map. Second, visual observations are aggregated from multiple camera viewpoints, reducing reliance on any single ambiguous projection. Third, our network is robust to contact offsets, as analyzed in Table 7. As a result, the network is not required to infer the complete 3D contact state from a single 2D view alone, and the limited expressivity of the projection model does not significantly degrade performance.

B.2 Computational Efficiency

For real-time deployment, low inference latency is essential for high-frequency closed-loop control. We compare Tac2Pix with the baselines across different inference stages. Beyond comparing visuo-tactile fusion mechanisms, this analysis contextualizes the efficiency of image-space tactile grounding against an alternative explicit 3D grounding strategy.

Tac2Pix projects tactile contacts onto the 2D image plane and reuses a standard visual encoder. In contrast, a natural alternative is to introduce an explicit 3D geometric branch, where depth-derived point-cloud features are encoded separately and fused with RGB and tactile representations at the latent level. While this provides additional 3D spatial context, it typically requires depth input, point-cloud preprocessing, and a separate 3D feature extractor, which can introduce non-negligible latency in closed-loop deployment. As a reference, in addition to the baselines used in the main experiments, we include a 3D visuo-tactile policy following (Wu et al., 2023). This policy uses RGB, depth, and tactile inputs by concatenating embeddings from pretrained modality-specific encoders. Specifically, it uses ResNet-18 for image features and PointNet (Qi et al., 2017) for point-cloud features.

The efficiency comparison results are shown in Tab. 8, we first measure the denoising-stage inference cost after the global condition has been constructed. In this setting, the main difference is the conditional vector passed to the diffusion model. Most policies require about 64–66 ms per control step. However, the 3D policy has a substantially longer conditioning vector due to the concatenated PointNet features, increasing its inference time to 76.72 ms. In contrast, Tac2Pix and the vision-only baseline require only 64.24 ms and 64.26 ms, respectively.

We also report pre-processing latency in Tab. 8, including all preprocessing and feature extraction before diffusion denoising. The 3D policy incurs the largest preprocessing cost, 95.12 ± 7.83 ms, dominated by point-cloud processing: farthest point sampling and PointNet feature extraction together take 68.23 ms. Tac2Pix is much faster, with a preprocessing latency of 21.06 ± 4.54 ms; saliency generation itself takes only 6.14 ± 1.89 ms. These results show that forward-kinematics-based saliency computation is lightweight enough for real-time deployment and does not introduce a prohibitive latency burden.

Overall, Tac2Pix offers a favorable efficiency–robustness trade-off. It preserves the denoising-stage speed of standard 2D visual diffusion policies while adding only modest preprocessing overhead for tactile saliency construction.

B.3 Feature Attention Visualizations Under Occlusion

To illustrate how the policy uses tactile cues under visual occlusion, we visualize feature attention with Grad-CAM (Selvaraju et al., 2017) in Fig. 7. Red indicates stronger model attention, while blue indicates weaker attention. For each task, we compare Grad-CAM maps across models under normal and occluded settings.

Under normal visibility, highlighted regions typically appear around the object and the robot hand. Under occlusion, baseline models tend to shift attention away from task-relevant regions. In contrast, Tac2Pix continues to focus near the hand, suggesting that it learns to rely on projected tactile cues when visual cues are unreliable.

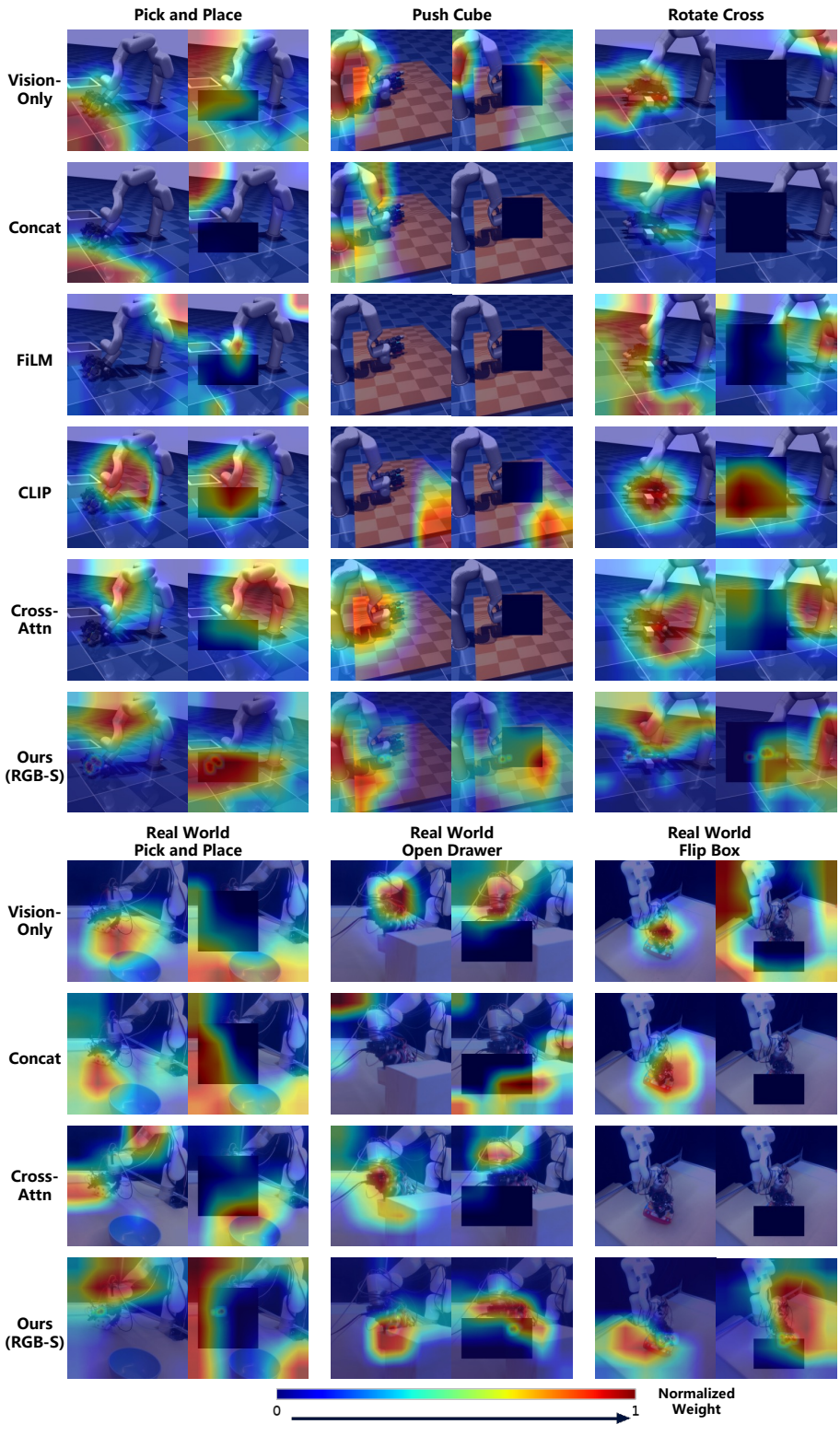


Figure 7: **Grad-CAM result of tasks in simulation and real-world.**

C Saliency Fusion Ablation Details

We provide implementation details for the saliency-fusion ablation in Sec. 4.3. All variants use the same Diffusion Policy backbone, training hyperparameters, RGB observations, proprioceptive states, and rendered saliency maps. The only difference is where the saliency stream is injected into the visual conditioning pipeline, which illustrated in Fig. 8.

Early fusion without zero initialization. This variant uses exactly the same Tac2Pix architecture and training settings as the full model, but initializes the added saliency-channel weights using the standard initialization of the visual backbone rather than zeros. Thus, the only changed factor is the initialization strategy.

Late fusion. The late-fusion variant treats the saliency map as an additional image-like modality. For each camera view, the RGB image and its corresponding saliency map are encoded by separate ResNet-18 encoders with the same spatial-softmax pooling interface. The saliency input is expanded to a 3-channel image to match the standard RGB encoder interface. The pooled RGB features, pooled saliency features, and other state inputs are concatenated after visual encoding to form the global condition for the diffusion U-Net. Thus, this variant increases the conditioning dimension but does not allow RGB and saliency features to interact before spatial pooling.

Intermediate fusion. The intermediate-fusion variant injects saliency features inside the ResNet visual encoder. The saliency map is first passed through a lightweight mask-projection branch, implemented as convolutional layers that map the single-channel saliency input to the channel dimension of an intermediate ResNet feature map. The projected saliency feature is then added to the corresponding RGB feature map through a zero-initialized projection layer before the remaining ResNet blocks are applied. This design follows a ControlNet-style residual conditioning formulation: the initial network behaves similarly to the RGB-only encoder, while the saliency residual is learned during fine-tuning (Zhang et al., 2023). Compared with late fusion, this variant allows feature-level interaction between the two streams, but the saliency signal is introduced only after the earliest visual convolutions.

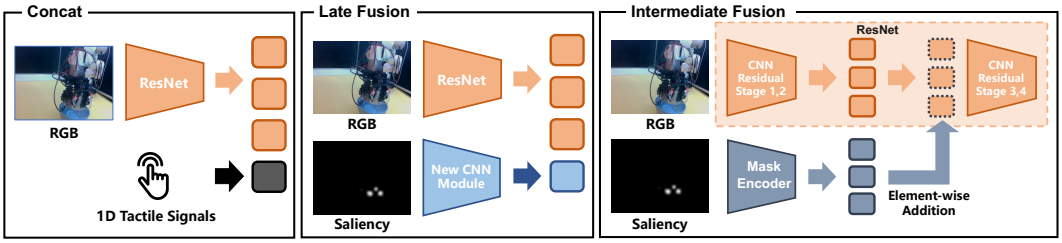


Figure 8: **Fusion architecture details.** (a) `concat`, where all features are concatenated to form the global conditioning vector. (b) Late-fusion and (c) intermediate-fusion variants.

D Training Details of Fusing Mechanisms

Cross-attention fusion. `cross-attn` is implemented following (Jaegle et al., 2021). It represents multi-view images, tactile measurements, and the proprioceptive state as modality-specific tokens. All modality tokens are projected into a shared 128-dimensional hidden space and augmented with learnable view or modality embeddings. A set of four learnable latent queries attends to these tokens through 4-head cross-attention, followed by latent self-attention and a feed-forward block with a dropout rate of 0.1. The resulting latent features are mean-pooled, concatenated with skip projections of the raw proprioceptive state and tactile vectors, and mapped to a 256-dimensional global conditioning vector for each observation step.

FiLM fusion. `FiLM` injects tactile information through two tactile-conditioned feature-wise modulation blocks, applied separately to the proprioceptive state feature and the concatenated multi-view visual feature. For a feature stream of dimension d, the corresponding FiLM block takes

Table 9: Downstream policy and training hyperparameters used in our configurations.

Policy	Domain	Obs. steps	Pred. horizon	Exec. steps	Batch size	Train steps	Optimizer	Learning rate	Weight decay
BC-MLP	Sim	1	1	1	64	120K	AdamW	1×10^{-4}	1×10^{-4}
ACT	Sim	1	16	8	64	120K	AdamW	1×10^{-5}	1×10^{-4}
DP	Sim/Real	4	24	8	64	120K	Adam	1×10^{-4}	1×10^{-6}
SmolVLA	Sim	1	50	10	8	120K	AdamW	1×10^{-4}	1×10^{-10}
3D-ViTac	Sim/Real	4	24	16	64	120K	AdamW	1×10^{-4}	1×10^{-4}
SaTA	Sim/Real	1	16	8	64	120K	AdamW	1×10^{-5}	1×10^{-4}

the tactile vector f_t as input and predicts $2d$ modulation channels using a lightweight projection, $\mathrm{Linear}(\mathrm{Mish}(f_t))$. The output is split into a scale term $\gamma(f_t) \in \mathbb{R}^d$ and a bias term $\beta(f_t) \in \mathbb{R}^d$, and the feature is modulated as $\tilde{x} = \gamma(f_t) \odot x + \beta(f_t)$. The final diffusion condition is formed by concatenating the tactile-modulated state feature, the tactile-modulated visual feature, and the tactile vector over the observation horizon.

CLIP-based fusion. `CLIP` is implemented following (Radford et al., 2021) using a two-stage design. First, in the VT-CLIP pretraining stage, the vision encoder uses one ResNet-18 backbone per camera view, concatenates the 512-dimensional per-view features, and projects them to a 256-dimensional visual feature. The tactile encoder takes a flattened tactile-history window of length 5 and maps it through a three-layer MLP with Mish activations to a 256-dimensional tactile feature. Both modalities are further projected to a 128-dimensional normalized contrastive space and trained with a symmetric CLIP loss. Second, during policy training, the pretrained VT-CLIP encoders are used as frozen feature extractors. The extracted features are concatenated with the proprioceptive state and flattened across the observation horizon to form the global conditioning vector.

E Training Details of Downstream Policies

We evaluate four downstream policy backbones: BC-MLP (Zhang et al., 2018), ACT (Zhao et al., 2023), DP (Chi et al., 2025), and SmolVLA (Shukor et al., 2025) implemented based on Lerobot (Cadene et al., 2026). All policies use the same observation interface and action space if not mentioned in extra. Important training hyperparameters used in our configuration can be found in Table 9. For controlled comparison, we use the same 120K optimization-update budget across policy architectures, which was sufficient for performance to largely stabilize in our experiments.

BC-MLP. `BC-MLP` is a single-step behavior cloning policy (Zhang et al., 2018). Each vector modality is projected to a 256-dimensional feature. The per-modality features are concatenated across the observation history and passed through an MLP with hidden dimensions $[512, 256]$, ReLU activation, dropout 0.1, and an MSE action regression loss.

ACT. `ACT` predicts an action chunk with a transformer encoder-decoder policy (Zhao et al., 2023). The transformer uses a 512-dimensional hidden state, 8 attention heads, 4 encoder layers, and 1 decoder layer. Following the original ACT formulation, we enable the conditional VAE branch during training. The VAE encoder takes the ground-truth action chunk together with the proprioceptive state and encodes them into a 32-dimensional latent variable. The policy decoder is then conditioned on this latent variable, visual features, and the proprioceptive state to reconstruct the action chunk. The policy is trained with a L1 masked action reconstruction loss plus a KL regularization term. During inference, the VAE encoder is not used and the latent variable is set to the prior mean.

DP. `DP` predicts an action trajectory using a conditional 1D denoising U-Net (Chi et al., 2025). The diffusion U-Net uses channel dimensions $[512, 1024, 2048]$, a kernel size of 5, group normalization with 8 groups, a 128-dimensional diffusion-step embedding, and 100 DDPM denoising steps. During deployment, we execute the first 8 actions of each predicted action chunk and then replan with a new prediction.

SmolVLA. `SmolVLA` (Shukor et al., 2025) combines a pretrained SmolVLM2 backbone with a flow-matching action expert. In our Tac2Pix implementation, the policy is conditioned on multi-view RGB images, their corresponding tactile saliency maps, the proprioceptive state, and a fixed task identifier. Each saliency map is processed by a zero-initialized patch projection, whose output is added to the corresponding RGB patch embeddings. The pretrained vision-language backbone parameters remain frozen, while the saliency projection, action expert, and input/output projection layers are jointly optimized using the flow-matching objective.