Tactile Modality Fusion for Vision-Language-Action Models

Charlotte Morissette[1,2], Amin Abyaneh[1,2], Wei-Di Chang[1,2], Anas Houssaini[1,2], David Meger[1,2], Hsiu-Chin Lin[1,2], Jonathan Tremblay[3], and Gregory Dudek[1,2]

[1] McGill University, CAN [2] Mila - Québec AI Institute, CAN [3] NVIDIA, USA

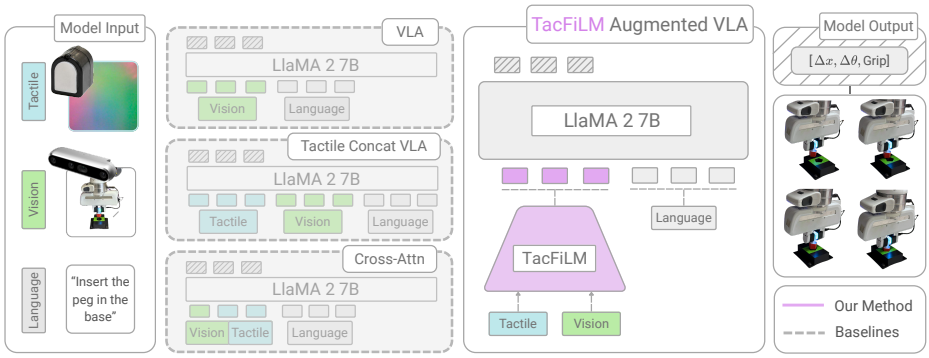

Fig. 1: TacFiLM Overview We present TacFiLM, a lightweight modality-fusion approach for integrating visual-tactile signals into VLA models. The left panel shows the model inputs, including tactile, visual, and language modalities. In grey, baseline approaches. To the right, we show our proposed TacFiLM-augmented VLA. The rightmost boxes show model outputs and rollouts.

Abstract. We propose TacFiLM, a lightweight modality-fusion approach that integrates visual-tactile signals into vision-language-action (VLA) models. While advances in VLAs have introduced robot policies that are both generalizable and semantically grounded, these models mainly rely on vision-based perception. Vision alone, however, cannot capture the complex interaction dynamics that occur during contact-rich manipulation, including contact forces, surface friction, compliance, and shear. While recent attempts to integrate tactile signals into VLA models often increase complexity through token concatenation or large-scale pretraining, the heavy computational demands of behaviour models necessitate lightweight fusion strategies. To address these challenges, TacFiLM outlines a post-training finetuning approach that conditions intermediate visual features on pretrained tactile representations using feature-wise linear modulation (FiLM). Experimental results on insertion and drawer opening tasks demonstrate consistent improvements in success rate, direct task performance, completion time, and force stability across both in-distribution and out-of-distribution tasks. Together, these results support our method as an effective approach to integrating tactile signals into VLA models, improving contact-rich manipulation behaviours.
Project page: `https://charliem7.github.io/projects/TacFilm/`

1 Introduction

In robotics, complex manipulation remains an open problem, with contact-rich tasks often exhibiting poor robustness to failure modes. Although existing approaches aimed at solving contact-rich tasks continue to rely predominantly on vision-based perception (31; 37; 46), humans achieve remarkable dexterity by integrating visual, tactile, and proprioceptive sensory feedback (33). In particular, tactile signals provide critical information about object geometry, surface friction, and contact forces, complementing vision during physical interactions where occlusion and millimetre-scale adjustments limit visual feedback. Vision-based tactile sensors such as DIGIT (38) and GelSight (65) capture this information as images, enabling integration with vision-based architectures.

Recent work has begun integrating tactile signals into vision-language-action (VLA) models through finetuning or multimodal pretraining (11; 12; 30; 68). These approaches face two challenges. First, VLA models require substantial data and compute for training and adaptation, creating demand for lightweight fusion strategies that remain accessible within a post-training finetuning paradigm. Second, the common baseline of feature concatenation often requires training separate tactile encoders and appends additional tokens to the VLA input, increasing sequence length and computational cost (58) while risking performance degradation as context grows (51; 59).

To tackle these problems, we propose TacFiLM, a novel lightweight modality fusion approach that integrates visual-tactile information into pretrained VLA models via feature-wise linear modulation (FiLM) (47), as seen in Figure 1. We adopt a lightweight FiLM-based adaptation strategy rather than introducing additional cross-modal fusion modules, such as cross-attention (25; 71), because it enables parameter-efficient adaptation without extensive multimodal pretraining. Our method leverages pretrained tactile representations and image conditioning, enabling tactile integration without increasing token sequence length or retraining large model components. In contrast to concatenation-based fusion strategies that append tactile embeddings to visual or language tokens, TacFiLM conditions intermediate visual features on tactile embeddings. This preserves pretrained visual-language priors while incorporating tactile signals.

We evaluate our approach through real-robot experiments across over 1,000 rollouts on a diverse set of contact-rich insertion and pulling tasks, spanning both in-distribution and out-of-distribution settings. Our method consistently improves success rate, direct insertion/opening percentage, force stability and execution efficiency. Notably, for in-distribution tasks, our method achieves a 100% success rate on the 3mm clearance Circle-Peg task and improves over the next-best baseline by up to 50% on selected tasks. In the out-of-distribution setting, TacFiLM maintains strong performance with 100% success rate on the 3mm clearance peg insertion tasks and improves HDMI cable plugging success by 30%. It additionally reduces excessive interaction forces, requiring approximately one-third of the force applied by the baseline methods on select tasks.

Overall, we show that FiLM-based fusion improves task performance while enhancing sensitivity to contact dynamics and reducing applied forces. Our main contributions are summarized as follows:

- TacFiLM, a novel modality fusion approach that integrates tactile signals through image conditioning.
- Comprehensive experiments showing that TacFiLM improves success rates by up to 50% with shorter episodes and reduced contact forces compared to concatenation and cross-attention-based fusion.
- An investigation of the use of pretrained tactile encoders such as Sparsh (26) and T3 (72) in fusing tactile signals into VLA models.

2 Related Work

2.1 Vision-Language-Action (VLA) Models

Motivated by the success of large language models (7; 13; 56; 57; 60) and vision-language models (2; 9; 10; 32; 41; 42; 49), early approaches exploring the incorporation of semantic reasoning into robotics focused on high-level planning, often leaving action generation and execution to separate low-level controllers (1; 17; 50). Others leveraged VLMs for robotics tasks such as affordance prediction, success detection, and representation learning, demonstrating improved semantic grounding and generalization across tasks (18; 35; 44; 53; 69). VLA models were subsequently introduced to unify semantic reasoning and action generation by leveraging semantic representations while directly grounding them in low-level robot control.

At a high level, VLA models aim to ground the semantic representations learned by vision-language models into physical action policies. These models typically combine a visual encoder, a projector, and a language model backbone to jointly process visual and textual tokens to generate output actions. Models such as OpenVLA (37), $\pi_{0.5}$ (31), RT-1-X/RT-2-X (46), and many others (4; 6; 36; 54; 55; 73), have been at the forefront of these advances.

Despite this progress, most existing VLA models primarily focus on vision and language modalities, with limited exploration of additional sensory inputs such as touch. This leaves open questions regarding how complementary modalities, particularly tactile signals, can be effectively incorporated into VLA frameworks for contact-rich manipulation.

2.2 Tactile Sensing in Robot Learning

The development of vision-based tactile sensors such as GelSight (65), DIGIT (38), and STS (28) has significantly advanced the integration of tactile perception in robotic manipulation. These sensors use an embedded camera to capture tactile information by recording the deformation of a gel membrane. Prior research has established tactile sensing as a complementary modality to vision in

manipulation tasks, enhancing perception and control in contact-rich interactions (5; 39). Their high spatial resolution allows these sensors to capture slip, contact geometry, and deformation with high fidelity (28; 38; 65). This has made them a natural choice for integration into robotic end-effectors. They are often used to solve contact-rich robotic tasks such as insertions, grasping, and in-hand manipulation, where tactile feedback provides critical information about physical interactions (8; 14; 15; 23; 27; 38; 48; 52; 61; 66).

Prior work has begun exploring the integration of tactile signals into foundation models, such as VLA models. At a high level, these approaches can be divided into two categories: those that incorporate tactile information during finetuning, and those that rely on additional multimodal pretraining to align tactile representations. The post-training finetuning approaches introduce tactile inputs during model finetuning, often by encoding tactile signals as additional tokens that are concatenated with the original VLM inputs. These methods rely on attention-based fusion during action generation (3; 24; 30; 40; 63; 64; 68). Conversely, the other category of approaches focuses on learning shared vision-tactile representations through large-scale pretraining or contrastive learning to align modalities prior to policy learning (11; 12; 21; 25; 34; 62; 70; 71). However, as large behaviour models demand increasing amounts of data and compute for training and finetuning, we are motivated to explore lightweight modality fusion strategies within the post-training finetuning paradigm.

Within tactile fusion approaches, the most prevalent methods are concatenation and cross-attention. Concatenation-based methods integrate tactile embeddings by appending them to visual and language tokens, enabling the transformer to reason across all modalities (24; 63; 64; 68). Cross-attention-based methods instead allow visual and tactile representations to attend to one another through dedicated attention layers (25; 71). Although these approaches have demonstrated promising results, their implementation either increases the token sequence length or requires training additional attention parameters, both increasing computational overhead. Our approach, in contrast, proposes a lightweight fusion strategy that conditions visual features on tactile information, maintains token lengths, and requires minimal additional training.

To support post-training finetuning strategies, pretrained tactile representations are important. Similar to pretrained visual backbones, these models learn tactile embeddings that encode tactile features such as contact, geometry, and deformation. Methods such as Sparsh (26), T3 (72), and others (19; 20; 22; 43; 62) have begun to explore self-supervised and large-scale pretraining approaches for learning transferable tactile representations.

Our approach builds on these advances by investigating how pretrained tactile representations can be effectively integrated into VLA policies within a post-training finetuning method. Our method further addresses limitations of naive fusion approaches, which often require training additional tactile encoders and increasing token sequence length.

3 Methodology

TacFiLM outlines a tactile modality fusion approach that conditions VLAs on visuotactile representations. Unlike prior work, we aim for tactile integration without increasing the number of input tokens or requiring task-specific encoders. The former is achieved by employing a lightweight conditioning strategy as explained in Section 3.1, while the latter is the direct result of using pretrained tactile representations as seen in Section 3.2. Consistent with the design philosophy of generalist VLAs, our fusion strategy of pretrained tactile representations in TacFiLM reduces the need for extensive retraining and repeated sensor-specific data collection.

3.1 Policy Architecture

Figure 2 illustrates the overall architecture of our tactile-conditioned VLA policy. The model builds upon the OpenVLA-OFT framework (36), which combines a fused SigLIP (67) and DINOv2 (45) visual backbone, a lightweight MLP projector, and a decoder-only Llama2 7B language model (57). To leverage tactile signals effectively within the VLA policy, we propose a modality fusion approach that integrates pretrained tactile embeddings into the VLA backbone. In contrast to approaches that jointly train modality encoders with the policy (24; 30; 63; 68), TacFiLM leverages pretrained tactile representations and aims to preserve the visual-language priors learned by the original VLA model. Specifically, we propose conditioning visual features with tactile information through a FiLM-based fusion approach.

At each time step, the image input and language prompts are processed following the standard VLA pipeline. Input images are encoded into patch-level visual embeddings by the fused vision backbone. In parallel, tactile observations are encoded using a pretrained tactile representation model. This tactile embedding is then used to condition intermediate vision representations.

The resulting tactile-conditioned visual features are projected into the language model input space and concatenated with text tokens before being processed by the decoder-only LLM. The decoder's final hidden states are then passed to an MLP action head, which directly regresses continuous robot actions using an L1 regression objective. This design enables the model to jointly reason over visual, tactile, and language information while predicting continuous action chunks for robotic manipulation.

FiLM-Based Fusion Feature-wise Linear Modulation conditions a backbone on an auxiliary signal by learning per-channel scale and shift (γ,β) parameters from the new input and applying a feature-wise affine transformation to intermediate activations (47). This auxiliary signal is a visuotactile image encoded with a pretrained tactile representation discussed above. At each time step t, tactile images are encoded, and the resulting patch features are averaged to obtain a pooled tactile embedding z_t. For each selected ViT block n in the VLAs visual

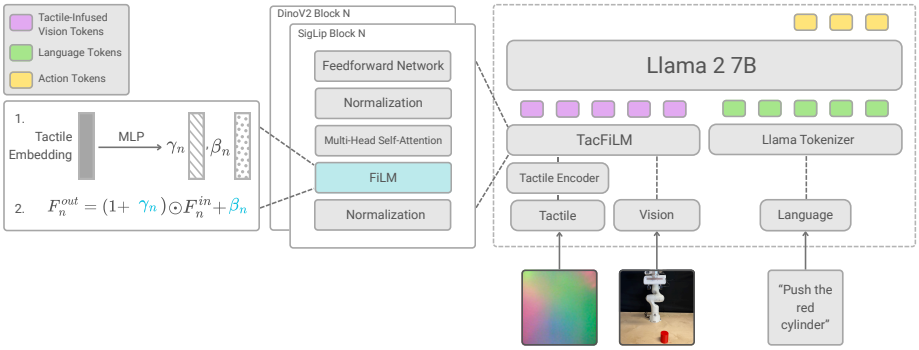

Fig. 2: TacFiLM's modality fusion pipeline. Tactile embeddings are integrated into the vision backbone immediately preceding the multi-head attention layers. The resulting multimodal tokens, combined with language inputs, serve as the basis for action generation within the Llama backbone.

encoders, DINOv2 and SigLIP, an MLP projects z_t to γ_n and β_n FiLM parameters. We apply FiLM after normalization and before multi-head self-attention as a feature-wise affine modulation of the intermediate visual features F_n as shown in the following equation:

$$\text{FiLM}(F_n|\gamma_n, \beta_n) = F_n \odot (1 + \gamma_n) + \beta_n. \tag{1}$$

The model architecture is shown in Figure 2. Following the design principles from OpenVLA-OFT, γ and β are applied to the entire feature map. We initialize γ and β to zero, so conditioning starts near identity. We opt for a FiLM approach because it is computationally lightweight, provides a low-dimensional, inspectable global tactile bias, eliminates the need to append additional tokens to the language backbone, and integrates cleanly into ViT blocks. By default, TacFiLM applies FiLM conditioning to all ViT blocks in the visual encoder; we ablate this choice in Section 4.3.

3.2 Pretrained Tactile Representations

Our fusion approach is agnostic to the choice of tactile encoder. We evaluate two architectures with different pretraining objectives, T3 and Sparsh, to validate this flexibility. Both produce fixed-dimensional embeddings compatible with our FiLM conditioning mechanism.

T3 The T3 model (72) is a tactile representation framework that extends to various visuotactile sensors and multiple downstream tasks. The architecture includes sensor-specific encoders, task-specific decoders, and a shared transformer

trunk. Each sensor is associated with its own independent encoder, while downstream tasks are managed by specific decoders, enabling transfer across sensors and tasks. For our application, we retain only the sensor encoder and shared transformer trunk. Each sensor encoder is implemented using a Vision Transformer (ViT) (16), which processes tactile images as patch tokens. The encoder maps raw tactile observations into latent feature representations. These features are subsequently processed by a shared ViT-based transformer trunk, which refines the embeddings into unified tactile representations. All tactile frames are resized to 224×224 before being processed by the encoder.

Sparsh The Sparsh model (26) is a pretrained tactile representation framework that learns generalizable features from large-scale tactile data. This method adopts a ViT architecture (16) trained using various self-supervised learning (SSL) objectives. Sparsh processes tactile images through a ViT encoder, where image inputs are tokenized into patch embeddings and passed through transformer blocks to produce latent tactile representations. Different Sparsh variants correspond to the three SSL paradigms they propose: Sparsh-MAE, Sparsh-IJEPA and Sparsh-DINO. We investigate all three. **Sparsh-MAE** uses a masked autoencoding objective where an encoder learns contextual representations of masked images, enabling a decoder to reconstruct the masked regions. **Sparsh-IJEPA** uses a joint-embedding predictive objective, learning representations by predicting latent features of masked regions rather than reconstructing raw inputs. **Sparsh-DINO** uses a self-distillation objective, where a student network learns to predict the representations of a teacher network. For all Sparsh variants, image preprocessing is identical. Two tactile frames, separated by five time steps, are concatenated channel-wise. The background is removed, and the resulting images are resized to 224×224.

3.3 Training

Despite large-scale pretraining, off-the-shelf VLAs often lack the precision required for specialized tasks. This motivates the use of post-training, specifically efficient finetuning recipes such as Low-Rank Adaptation (LoRA) (29), to adapt the generalist model to specific domains without losing its pre-trained representations. We build on the observation that parameter-efficient finetuning provides an effective mechanism for adapting generalist VLAs, and extend this approach to integrate tactile signals into these models. We LoRA-finetune the linear layers of the TacFiLM-augmented VLA, namely, the OpenVLA-OFT and tactile backbone linear layers, while training the model's FiLM layers from scratch. By freezing most of the model and updating only a targeted subset of weights, we inject novel visuotactile modalities into the policy. In doing so, we aim to retain the base VLA's semantic understanding while leveraging rich tactile feedback for contact-rich manipulation.

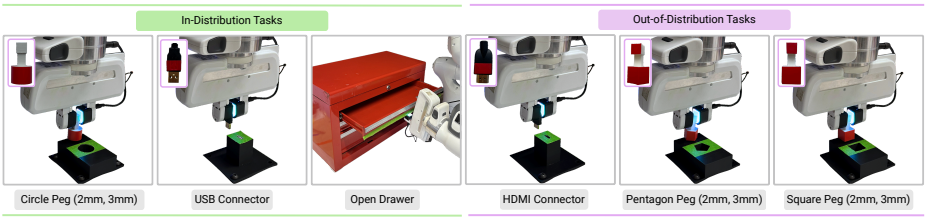

Fig. 3: Task definitions. Insertion tasks differ in peg or connector shape and clearance but share the goal of successful insertion. Open-drawer consists of hooking the gripper under the drawer and pulling it open.

4 Experiments

We conduct a series of experiments to evaluate the proposed modality fusion approach and the use of pretrained tactile representations in large pretrained models. Our experiments are guided by the following research questions: **Q1:** How do different pretrained tactile encoders (T3, Sparsh variants) influence downstream policy performance? **Q2:** How do different modality fusion mechanisms compare in integrating tactile signals into VLA models? **Q3:** Does the proposed TacFiLM improve task success, execution efficiency, and contact sensitivity?

4.1 Experiment Setup

Tasks and Data Collection Our experimental evaluation focuses on a diverse suite of insertion and pulling tasks, depicted in Figure 3, and designed to test contact-rich manipulation capabilities. The benchmark of peg insertions and cable plugging spans multiple object geometries, varying peg and connector shapes, and introduces different levels of difficulty by varying insertion clearances. These variations are important because they create an environment in which direct insertions are difficult and the policy must often rely on tactile feedback for fine-grained adjustments. The benchmark also includes a drawer-opening task, extending the evaluation beyond insertion to further contact-rich scenes.

In each task, the objective is strictly defined: the robot must either fully insert the object into the target base or fully open the drawer, whether directly or through one or more corrective adjustments. We do not evaluate the model on its grasping abilities, and as a result, the robot begins all trajectories with the object in hand for insertion tasks.

To evaluate our approach on a real-world setup, we collected expert demonstrations by teleoperating a Franka Emika Panda arm to complete the previously defined tasks. High-level robot control and teleoperation were implemented via Polymetis[1], while low-level commands were transmitted through libfranka over

[1] `facebookresearch.github.io/fairo/polymetis`

the Franka Control Interface (FCI) at a control frequency of 1 kHz. The FCI provided real-time feedback of the robot's state to the controller. An operator controlled the end-effector pose using a 3Dconnexion SpaceMouse, whose 6-DoF inputs were mapped to Cartesian pose commands with tunable gains. For each task, we collected 80 demonstrations, each of approximately 70 steps. During data collection, we recorded time-aligned robot observations at 10 Hz, including joint positions and velocities, end-effector pose, gripper width and status, RGB and tactile images from an Intel RealSense camera and a DIGIT sensor, respectively, and executed actions. A fixed natural language description was appended to each demonstration. For insertion tasks, prompts followed the template: "Insert the [colour] [shape] peg into the [colour] base". For the drawer-opening task, the prompt was the following: "Hook the gripper under the green handle of the top drawer and pull it open".

Evaluation Methods To address the research questions outlined above, we compare TacFiLM against the following baseline methods. 1) **OpenVLA-OFT:** OpenVLA-OFT is the base VLA model upon which all other approaches are built. The model combines a fused SigLIP (67) and DINOv2 (45) visual encoder, an MLP projector and a decoder-only Llama2 7B (57) backbone. It is trained on visual observations and language prompts, serving as a vision-only baseline with which to compare our approach. 2) **TactileConcat:** We also consider feature concatenation, as implemented in prior work (3; 24; 30; 40; 64; 68). In our implementation, tactile images are embedded using the pretrained T3 or Sparsh models described above. The resulting tactile features are then projected to the VLM's input space with a learned two-layer MLP. The resulting tactile tokens are concatenated with the VLM language and image tokens, which are then passed as input to the model. 3) **Cross-Attn:** Lastly, we implement a cross-attention-based fusion strategy as proposed in prior work (25; 71). Our implementation follows PolyTouch's architecture (71). Tactile images are first embedded using the pretrained T3 or Sparsh models and projected into the visual feature space. Cross-attention is then applied after the vision backbone through six stacked cross-attention blocks with residual connections, in which visual patch embeddings serve as queries that attend to the projected tactile embeddings as keys/values.

To compare our approach against the tactile fusion mechanisms proposed in prior work, we implement each baseline on a shared backbone, OpenVLA-OFT (36), so that differences in performance can be attributed to the fusion strategy rather than the underlying architecture.

We further evaluate the effects of FiLM integration location through an ablation study, proposing variants of our model, TacFiLM, in which FiLM is integrated only into a subset of ViT blocks.

Evaluation Metrics To evaluate the performance of the methods listed above on our tasks, we compare task success rate (%), percentage of direct insertions/openings, average maximum force exerted (N) and average time to task

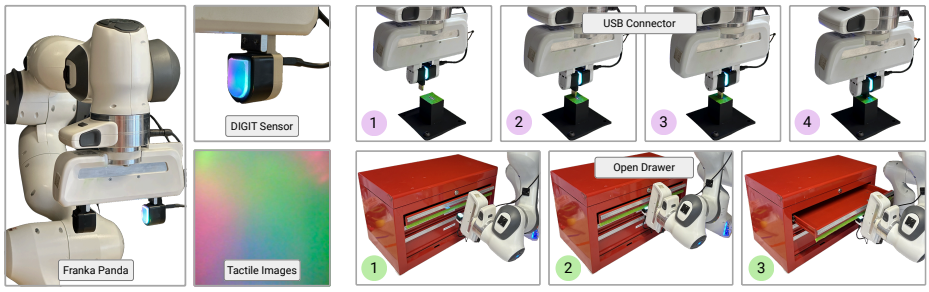

Fig. 4: Experiment setup (left) and sample rollouts (right). Franka parallel grippers, which the robot uses to hold the pegs, have a DIGIT tactile sensor installed on them. The rollouts demonstrate USB connector insertions and open-drawer tasks.

completion across all rollouts (s). These metrics were selected to capture different aspects of manipulation performance. Success rate reflects overall task reliability, including both direct and recovered behaviour, while the percentage of direct insertions/openings measures the model's ability to achieve accurate alignment and grip without resorting to recovery behaviours. The average maximum force allows us to assess whether the policy accounts for contact dynamics during task execution. Finally, average task completion time reflects execution efficiency. We report mean values across trials to characterize overall expected performance. All deployed methods were trained to 80k steps. A task is deemed successful if the object in hand is fully inserted into the base or the drawer is fully opened. We distinguish direct insertions/opening (first attempt) from recovered insertions/openings (after adjustment).

Hardware We run all deployments on the Franka Emika Panda[2], a robot arm with 7 degrees-of-freedom. The arm is equipped with a parallel two-finger gripper, the Franka hand, on which is mounted a DIGIT visuotactile sensor, as seen in Figure 4.

4.2 Experimental Results

We evaluate method performance on both in-distribution and out-of-distribution tasks to assess task-specific performance and generalization capabilities. In total, we conduct over 1,000 rollouts: 480 for in-distribution evaluation (30 per method), 300 for out-of-distribution (15 per method), and an additional 240 for ablation studies.

In-Distribution Tasks We compare the performance of TacFiLM against three baselines for four tasks: circle-peg insertion with a 3mm clearance, circle-peg in-

[2] `https://franka.de/documents`

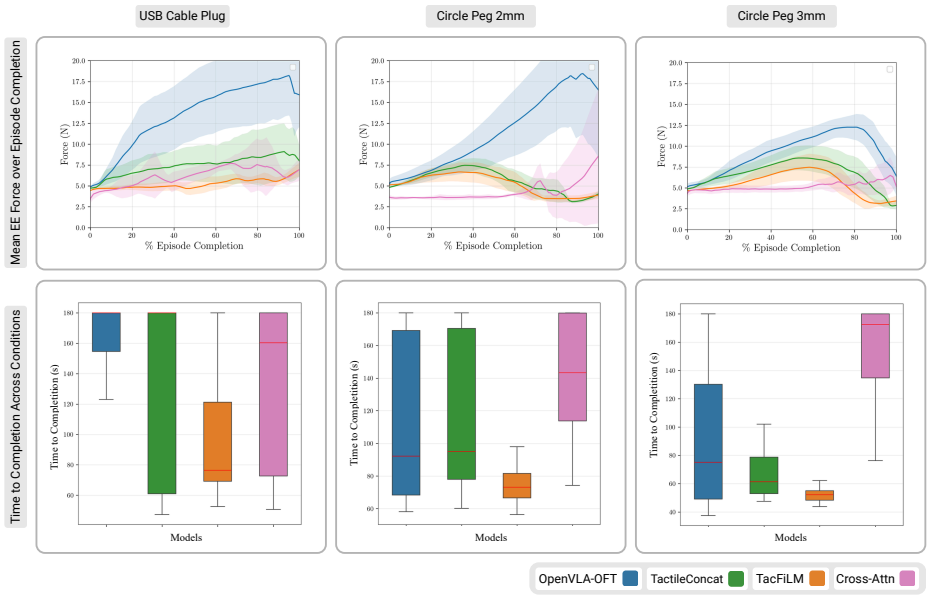

Fig. 5: Force and task completion time analysis. Top row: Average force measurements across successfully recovered insertions for the in-distribution (ID) tasks. Bottom row: Task completion times across different methods for ID insertion tasks. The results demonstrate that tactile-aware methods prevent excessive force application while TacFiLM also significantly reduces task completion time.

sertion with a 2mm clearance, USB cable plugging and drawer opening. Table 1 shows the overall results for the in-distribution tasks. We observe that for the easiest task, circle-peg insertion with the highest clearance (3mm), TacFiLM and TactileConcat significantly outperform the vision-only and cross-attention models, achieving similar success rates. However, we notice that the percentage of direct insertions is considerably higher for our method, highlighting the model's ability to leverage tactile signals without degrading the vision ones. Further, both average maximum force and average completion time are lower for TacFiLM. This suggests that our modality fusion approach allows for time-efficient task execution as well as improved sensitivity to contact dynamics. This trend continues for the other three tasks, but we notice that the overall performance of the TactileConcat approach, when compared against TacFiLM, decreases as the tasks get harder. The insertion results can be visualized in Figure 5.

Out-of-Distribution Tasks In Table 1, we further evaluate the generalization capabilities of our model and the baseline approaches on out-of-distribution tasks, including square-peg insertion (2mm and 3mm), pentagon-peg insertion (2mm and 3mm) and HDMI cable plugging. For the higher-clearance insertion

tasks, we observe results similar to the in-distribution setting, where TacFiLM and TactileConcat outperform the vision-only and cross-attention baselines. Our method again achieves the highest success rate and greatly improves the percentage of direct insertions, while also producing the lowest average maximum force and completion time. For the more challenging 2mm clearance insertion tasks, TacFiLM and TactileConcat methods exhibit comparable success rates. However, we observe a notable increase in force magnitude for TactileConcat, which exerts significantly higher forces relative to its in-distribution results. In contrast, TacFiLM maintains a low force and high direct insertion rate across its episodes. The HDMI cable plugging task further highlights these differences. While the vision-only and TactileConcat approaches exhibit near-zero percent success rates, TacFiLM significantly improves both success and direct insertion rates while maintaining a lower completion time. Average force values are comparable across methods. It is noteworthy that the cross-attention baseline consistently underperforms across all tasks, with only a slight relative improvement on the HDMI cable plugging task. We hypothesize that this behaviour is attributable to the limited amount of task-specific training data. Unlike TacFiLM, cross-attention introduces additional trainable interactions between visual and tactile representations, which may require substantially more data to learn effective multimodal correspondences.

Overall, these results indicate that FiLM conditioning yields more robust generalization and maintains sensitivity to contact dynamics under distribution shifts. In contrast, feature concatenation may be more susceptible to variations in contact geometry and exhibit lower sensitivity to contact dynamics, particularly in the 2mm clearance tasks.

4.3 Ablation Study

FiLM Integration Stage By default, TacFiLM applies FiLM conditioning to all ViT blocks (denoted AllFiLM in Table 2). We ablate this choice to determine whether FiLM can be applied to only a subset of blocks. Table 2 summarizes the results obtained when FiLM layers are introduced at different depths of the visual backbone (all, early, middle and late blocks). EarlyFiLM, MiddleFiLM and LateFiLM correspond to model variants where FiLM is integrated into a third of the ViT blocks at varying depths. We evaluate the fusion location results on the 3mm circle-peg insertion task and the OOD 3mm pentagon-peg task. Across all models, we observe comparable results, with high performance, suggesting that FiLM-based tactile conditioning is effective regardless of integration depth. The EarlyFiLM model, however, achieves a significantly higher percentage of direct insertions but yields comparable results on the other metrics. Despite this difference, all ablation variants perform similarly in the OOD generalization task. These findings suggest that applying FiLM conditioning to a limited number of ViT blocks suffices for tactile integration, keeping computational overhead low.

Camera Condition We further conduct experiments to examine the effects of camera conditions on model performance and evaluate whether tactile readings

Table 1: In-distribution and out-of-distribution evaluation results. TacFiLM is compared with the baselines across diverse insertion and pulling tasks (30 rollouts per method for ID, 15 for OOD). We report success rates alongside safety-critical metrics such as maximum insertion force and the rate of direct insertions. TacFiLM consistently outperforms baselines in both reliability and execution efficiency while maintaining lower peak contact forces.

Task	Method	Success (%)	Direct (%)	Avg. Max Force (N)	Avg. Time (s)
In-Distribution Tasks					
Circle-Peg (3mm)	OpenVLA-OFT	86.67	3.33	14.94 ± 4.66	92.24 ± 48.10
	TactileConcat	96.67	16.67	9.19 ± 3.45	75.11 ± 37.28
	Cross-Attn	63.33	10.00	8.16 ± 7.18	155.13 ± 31.59
	TacFiLM	**100.00**	**36.67**	**7.64** ± 2.63	**52.03** ± 5.02
Circle-Peg (2mm)	OpenVLA-OFT	66.67	**23.33**	15.09 ± 12.69	110.44 ± 46.76
	TactileConcat	73.33	0.00	8.72 ± 2.25	114.80 ± 44.44
	Cross-Attn	60.00	20.00	10.38 ± 10.08	139.16 ± 36.57
	TacFiLM	**86.67**	**23.33**	**7.22** ± 2.00	**87.11** ± 37.59
USB-Cable-Plug	OpenVLA-OFT	33.33	0.00	15.01 ± 9.09	164.52 ± 27.06
	TactileConcat	43.33	6.67	12.96 ± 5.04	135.11 ± 56.82
	Cross-Attn	33.33	0.00	26.23 ± 14.98	134.58 ± 52.96
	TacFiLM	**73.33**	**33.33**	**10.15** ± 5.47	**99.71** ± 46.43
Open-Drawer	OpenVLA-OFT	33.33	33.33	13.64 ± 0.54	152.65 ± 39.37
	TactileConcat	26.67	26.67	**9.74** ± 0.89	141.66 ± 35.93
	Cross-Attn	20.00	20.00	17.40 ± 6.58	165.64 ± 28.61
	TacFiLM	**86.67**	**73.33**	10.84 ± 1.87	**94.33** ± 46.43
Average	OpenVLA-OFT	58.10	12.38	14.94 ± 9.16	126.72 ± 51.34
	TactileConcat	64.76	10.48	10.27 ± 4.12	113.04 ± 52.31
	Cross-Attn	48.00	12.00	13.43 ± 12.62	149.92 ± 39.01
	TacFiLM	**86.67**	**37.14**	**8.65** ± 3.80	**81.72** ± 38.00
Out-of-Distribution Tasks					
Circle-Peg (3mm)					
Square-Peg (3mm)	OpenVLA-OFT	93.33	0.00	18.31 ± 8.84	**51.60** ± 5.62
	TactileConcat	93.33	13.33	9.34 ± 5.56	69.77 ± 23.40
	Cross-Attn	60.00	6.67	6.49 ± 1.14	165.27 ± 18.74
	TacFiLM	**100.00**	**46.67**	**5.37** ± 0.41	52.95 ± 4.68
Pentagon-Peg (3mm)	OpenVLA-OFT	46.67	0.00	27.39 ± 17.81	83.48 ± 22.55
	TactileConcat	**100.00**	20.00	10.43 ± 5.36	76.21 ± 17.65
	Cross-Attn	53.33	6.67	9.16 ± 5.22	147.99 ± 35.01
	TacFiLM	**100.00**	**33.33**	**7.51** ± 2.88	**53.15** ± 5.89
Circle-Peg (2mm)					
Square-Peg (2mm)	OpenVLA-OFT	66.67	0.00	34.30 ± 11.18	**64.54** ± 10.64
	TactileConcat	**86.67**	6.67	27.72 ± 8.36	91.83 ± 10.64
	Cross-Attn	53.33	6.67	26.24 ± 12.36	156.33 ± 28.07
	TacFiLM	80.00	**40.00**	**7.06** ± 1.36	111.61 ± 38.31
Pentagon-Peg (2mm)	OpenVLA-OFT	60.00	0.00	21.62 ± 17.87	**80.91** ± 18.05
	TactileConcat	73.33	0.00	23.70 ± 10.29	116.56 ± 30.55
	Cross-Attn	46.67	6.67	20.65 ± 14.01	145.28 ± 34.90
	TacFiLM	**86.67**	**20.00**	**10.50** ± 7.72	104.61 ± 37.60
USB-Cable-Plug					
HDMI-Cable-Plug	OpenVLA-OFT	6.67	0.00	**10.71** ± 8.93	166.88 ± 38.29
	TactileConcat	13.33	0.00	11.18 ± 5.08	174.59 ± 13.84
	Cross-Attn	33.33	0.00	33.82 ± 12.79	133.97 ± 39.33
	TacFiLM	**66.67**	**6.67**	11.54 ± 3.88	**116.87** ± 45.46
Average	OpenVLA-OFT	54.67	0.00	22.46 ± 15.75	89.48 ± 46.05
	TactileConcat	73.33	8.00	16.47 ± 10.54	105.79 ± 43.16
	Cross-Attn	49.33	5.33	19.27 ± 14.62	149.77 ± 33.72
	TacFiLM	**86.67**	**29.33**	**8.40** ± 4.71	**87.84** ± 42.69

Table 2: Ablation study and occlusion tests. Top) Comparison of four FiLM integration locations within the vision backbone. Bottom) Robustness evaluation under varied camera conditions demonstrates that TacFiLM achieves the highest success rate despite visual interference. Methods are evaluated over 15 rollouts (240 total).

Task	Method	Success (%)	Direct (%)	Avg. Max Force (N)	Avg. Time (s)
FiLM Integration Stage					
Circle-Peg (3mm) (In-Distribution)	AllFiLM	**100.00**	36.67	7.64 ± 2.63	**52.03** ± 5.02
	EarlyFiLM	93.33	**60.00**	**6.69** ± 0.92	60.85 ± 8.26
	MiddleFiLM	**100.00**	26.67	7.14 ± 1.71	53.88 ± 5.08
	LateFiLM	**100.00**	23.33	8.47 ± 2.71	54.74 ± 7.34
Pentagon-Peg (3mm) (Out-of-Distribution)	AllFiLM	**100.00**	33.33	**7.51** ± 2.88	**53.15** ± 5.89
	EarlyFiLM	**100.00**	33.33	9.96 ± 4.49	77.40 ± 21.68
	MiddleFiLM	**100.00**	**53.33**	8.99 ± 4.49	53.57 ± 6.89
	LateFiLM	**100.00**	40.00	9.42 ± 3.31	68.57 ± 27.47
Camera Condition					
Circle-Peg (3mm) (80% Dimmed)	OpenVLA-OFT	93.33	0.00	16.29 ± 9.85	73.50 ± 8.41
	TactileConcat	86.67	26.67	11.15 ± 9.21	78.03 ± 30.16
	Cross-Attn	53.33	0.00	9.29 ± 5.96	159.96 ± 29.15
	TacFiLM	**100.00**	**26.67**	**8.62** ± 2.13	**67.79** ± 6.25
Circle-Peg (3mm) (50% Frames)	OpenVLA-OFT	73.33	0.00	15.70 ± 12.39	113.19 ± 40.39
	TactileConcat	80.00	**40.00**	14.64 ± 13.91	71.52 ± 34.51
	Cross-Attn	46.67	0.00	**7.53** ± 1.25	161.10 ± 30.70
	TacFiLM	**100.00**	26.67	8.12 ± 1.90	**51.68** ± 5.33

can compensate for degraded visual inputs. These results can be visualized in the bottom portion of Table 2. Under dimmed lighting conditions (80% reduction), TacFiLM maintains a 100% success rate, outperforming both OpenVLA-OFT, TactileConcat and Cross-Attn. Additionally, our method achieves the lowest average maximum force and execution time. Under the second camera condition, a partially frozen stream (50% frame updates), we notice a performance decrease for the vision-only baseline, whose success rate drops to 73.33%. In contrast, TacFiLM again achieves 100% success and maintains the lowest force and completion time of all baseline methods. Although TactileConcat achieves a higher percentage of direct insertions in this setting, it exerts a higher force magnitude and slower execution time relative to TacFiLM. For Cross-Attn, both camera conditions show decreased performance relative to other baselines as well as its performance on the regular circle-peg (3mm) task, demonstrating poor robustness to changes in camera condition. Overall, these results indicate that our approach improves robustness under visual degradation, allowing the policy to maintain performance when visual information is unreliable.

Tactile Encoders Evaluation To select the tactile encoder for our main experiments, we evaluate several pretrained models on binary classification tasks drawn from T3 (72) and Sparsh (26) benchmarks that isolate tactile signatures relevant to insertion. We assess the learned representations on binary classification tasks by training an MLP classifier on the tactile embeddings. These tasks were selected because they isolate tactile signatures characteristic of the contact-rich phases in insertion tasks. Tasks one and two, *Rotation-High* and

Rotation-Low, evaluate the model's ability to detect whether the peg held by the gripper is rotated relative to its initial position, thereby capturing sensitivity to changes in contact geometry. Task three, *Contact*, assesses whether the model can distinguish between contact and non-contact states on the peg, testing the encoder's ability to capture force and deformation-related features. To further test the pretrained models' force-encoding capacity, we evaluate their performance on the continuous force regression task from the Sparsh TacBench benchmark (26). Table 3 compares the state-of-the-art encoders across these four tasks. Based on the higher performance observed in these evaluations, we adopt Sparsh-DINO as the visuotactile backbone for all subsequent experiments.

Table 3: Comparison of pretrained visuotactile representations. Comparison of four pretrained tactile representations across three binary classification tasks and one TacBench task.

Dataset	T3	Sparsh-IJEPA	Sparsh-MAE	Sparsh-DINO
Classification Tasks				
Rotation-High (%)	92.73	99.15	**99.36**	**99.36**
Rotation-Low (%)	83.09	96.44	96.64	**98.42**
Contact (%)	73.31	85.08	93.92	**95.39**
Average	83.04	93.56	96.64	**97.72**
TacBench Task				
Force Estimation (RMSE)	58.64	40.27	36.61	**36.09**

5 Conclusion

In this work, we investigated the integration of tactile signals into VLA models via post-training finetuning approaches. We evaluated multiple pretrained tactile representations and demonstrated that they allow for effective visuotactile policy adaptation without requiring task-specific encoder training. We proposed TacFiLM, an image conditioning fusion method that uses feature-wise linear modulation to integrate tactile signals, and showed that it improves task performance, increases direct insertion rates, and reduces both completion time and interaction forces. Importantly, our empirical results indicate that feature-level tactile conditioning yields more stable generalization behaviour than concatenation- and cross-attention-based fusion as well as vision-only models.

5.1 Limitations

A broader evaluation of diverse manipulation tasks would further strengthen our results. However, without precise visuotactile simulators, extending the task set proved difficult, as the experiments had to be conducted in a real-world setup, which required extensive data collection and time-consuming rollouts. Additionally, while our FiLM-based fusion approach is designed around the OpenVLA-OFT architecture, extending it to other VLA backbones such as $\pi_{0.5}$ remains future work.

Bibliography

[1] Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., et al.: Do as i can, not as i say: Grounding language in robotic affordances. arXiv preprint arXiv:2204.01691 (2022)

[2] Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)

[3] Bi, J., Ma, K.Y., Hao, C., Shou, M.Z., Soh, H.: Vla-touch: Enhancing vision-language-action models with dual-level tactile feedback. arXiv preprint arXiv:2507.17294 (2025)

[4] Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al.: $\pi 0$: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164 (2024)

[5] Blake, R., Sobel, K.V., James, T.W.: Neural synergy between kinetic vision and touch. Psychological science **15**(6), 397–402 (2004)

[6] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al.: Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817 (2022)

[7] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)

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

[9] Chen, X., Djolonga, J., Padlewski, P., Mustafa, B., Changpinyo, S., Wu, J., Ruiz, C.R., Goodman, S., Wang, X., Tay, Y., et al.: Pali-x: On scaling up a multilingual vision and language model. arXiv preprint arXiv:2305.18565 (2023)

[10] Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A.J., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al.: Pali: A jointly-scaled multilingual language-image model. arXiv preprint arXiv:2209.06794 (2022)

[11] Cheng, N., Xu, J., Guan, C., Gao, J., Wang, W., Li, Y., Meng, F., Zhou, J., Fang, B., Han, W.: Touch100k: A large-scale touch-language-vision dataset for touch-centric multimodal representation. Information Fusion p. 103305 (2025)

[12] Cheng, Z., Zhang, Y., Zhang, W., Li, H., Wang, K., Song, L., Zhang, H.: Omnivtla: Vision-tactile-language-action model with semantic-aligned tactile sensing. arXiv preprint arXiv:2508.08706 (2025)
[13] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. Journal of machine learning research **24**(240), 1–113 (2023)
[14] Cui, L., Zhao, Z., Xie, S., Zhang, W., Han, Z., Zhu, Y.: Vi-tacman: Articulated object manipulation via vision and touch. arXiv preprint arXiv:2510.06339 (2025)
[15] Dong, S., Jha, D.K., Romeres, D., Kim, S., Nikovski, D., Rodriguez, A.: Tactile-rl for insertion: Generalization to objects of unknown geometry. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 6437–6443. IEEE (2021)
[16] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
[17] Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al.: Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378 (2023)
[18] Du, Y., Konyushkova, K., Denil, M., Raju, A., Landon, J., Hill, F., De Freitas, N., Cabi, S.: Vision-language models as success detectors. arXiv preprint arXiv:2303.07280 (2023)
[19] Feng, R., Hu, J., Xia, W., Gao, T., Shen, A., Sun, Y., Fang, B., Hu, D.: Anytouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors. arXiv preprint arXiv:2502.12191 (2025)
[20] Feng, R., Zhou, Y., Mei, S., Zhou, D., Wang, P., Cui, S., Fang, B., Yao, G., Hu, D.: Anytouch 2: General optical tactile representation learning for dynamic tactile perception. arXiv preprint arXiv:2602.09617 (2026)
[21] George, A., Gano, S., Katragadda, P., Farimani, A.B.: Vital pretraining: Visuo-tactile pretraining for tactile and non-tactile manipulation policies. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 258–264. IEEE (2025)
[22] Gupta, H., Mo, Y., Jin, S., Yuan, W.: Sensor-invariant tactile representation. arXiv preprint arXiv:2502.19638 (2025)
[23] Hansen, J., Hogan, F., Rivkin, D., Meger, D., Jenkin, M., Dudek, G.: Visuotactile-rl: Learning multimodal manipulation policies with deep reinforcement learning. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 8298–8304. IEEE (2022)
[24] Hao, P., Zhang, C., Li, D., Cao, X., Hao, X., Cui, S., Wang, S.: Tla: Tactile-language-action model for contact-rich manipulation. arXiv preprint arXiv:2503.08548 (2025)
[25] Heng, L., Geng, H., Zhang, K., Abbeel, P., Malik, J.: Vitacformer: Learning cross-modal representation for visuo-tactile dexterous manipulation. arXiv preprint arXiv:2506.15953 (2025)

[26] Higuera, C., Sharma, A., Bodduluri, C.K., Fan, T., Lancaster, P., Kalakrishnan, M., Kaess, M., Boots, B., Lambeta, M., Wu, T., et al.: Sparsh: Self-supervised touch representations for vision-based tactile sensing. arXiv preprint arXiv:2410.24090 (2024)
[27] Hogan, F.R., Bauza, M., Canal, O., Donlon, E., Rodriguez, A.: Tactile regrasp: Grasp adjustments via simulated tactile transformations. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2963–2970. IEEE (2018)
[28] Hogan, F.R., Jenkin, M., Rezaei-Shoshtari, S., Girdhar, Y., Meger, D., Dudek, G.: Seeing through your skin: Recognizing objects with a novel visuotactile sensor. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1218–1227 (2021)
[29] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
[30] Huang, J., Wang, S., Lin, F., Hu, Y., Wen, C., Gao, Y.: Tactile-vla: unlocking vision-language-action model's physical knowledge for tactile generalization. arXiv preprint arXiv:2507.09160 (2025)
[31] Intelligence, P., Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., et al.: π0.5: a vision-language-action model with open-world generalization. arXiv preprint arXiv:2504.16054 (2025)
[32] Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
[33] Johansson, R.S., Flanagan, J.R.: Coding and use of tactile signals from the fingertips in object manipulation tasks. Nature Reviews Neuroscience **10**(5), 345–359 (2009)
[34] Jones, J., Mees, O., Sferrazza, C., Stachowicz, K., Abbeel, P., Levine, S.: Beyond sight: Finetuning generalist robot policies with heterogeneous sensors via language grounding. arXiv preprint arXiv:2501.04693 (2025)
[35] Karamcheti, S., Nair, S., Chen, A.S., Kollar, T., Finn, C., Sadigh, D., Liang, P.: Language-driven representation learning for robotics. arXiv preprint arXiv:2302.12766 (2023)
[36] Kim, M.J., Finn, C., Liang, P.: Fine-tuning vision-language-action models: Optimizing speed and success. arXiv preprint arXiv:2502.19645 (2025)
[37] Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)
[38] Lambeta, M., Chou, P.W., Tian, S., Yang, B., Maloon, B., Most, V.R., Stroud, D., Santos, R., Byagowi, A., Kammerer, G., 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 **5**(3), 3838–3845 (2020)

[39] Lee, M.A., Zhu, Y., Srinivasan, K., Shah, P., Savarese, S., Fei-Fei, L., Garg, A., Bohg, J.: 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)

[40] Li, J., Wu, T., Zhang, J., Chen, Z., Jin, H., Wu, M., Shen, Y., Yang, Y., Dong, H.: Adaptive visuo-tactile fusion with predictive force attention for dexterous manipulation. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3232–3239. IEEE (2025)

[41] Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)

[42] Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)

[43] Ma, W., Cao, X., Zhang, Y., Zhang, C., Yang, S., Hao, P., Fang, B., Cai, Y., Cui, S., Wang, S.: Cltp: Contrastive language-tactile pre-training for 3d contact geometry understanding. arXiv preprint arXiv:2505.08194 (2025)

[44] Nair, S., Rajeswaran, A., Kumar, V., Finn, C., Gupta, A.: R3m: A universal visual representation for robot manipulation. arXiv preprint arXiv:2203.12601 (2022)

[45] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

[46] O'Neill, A., Rehman, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., Jain, A., et al.: Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 6892–6903. IEEE (2024)

[47] Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)

[48] Qi, H., Yi, B., Suresh, S., Lambeta, M., Ma, Y., Calandra, R., Malik, J.: General in-hand object rotation with vision and touch. In: Conference on Robot Learning. pp. 2549–2564. PMLR (2023)

[49] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)

[50] Shah, D., Osiński, B., Levine, S., et al.: Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In: Conference on robot learning. pp. 492–504. pmlr (2023)

[51] Sharma, A., Saxon, M., Wang, W.Y.: Losing visual needles in image haystacks: Vision language models are easily distracted in short and long contexts. arXiv preprint arXiv:2406.16851 (2024)

[52] She, Y., Wang, S., Dong, S., Sunil, N., Rodriguez, A., Adelson, E.: Cable manipulation with a tactile-reactive gripper. The International Journal of Robotics Research **40**(12-14), 1385–1401 (2021)

[53] Shridhar, M., Manuelli, L., Fox, D.: Cliport: What and where pathways for robotic manipulation. In: Conference on robot learning. pp. 894–906. PMLR (2022)

[54] Stone, A., Xiao, T., Lu, Y., Gopalakrishnan, K., Lee, K.H., Vuong, Q., Wohlhart, P., Kirmani, S., Zitkovich, B., Xia, F., et al.: Open-world object manipulation using pre-trained vision-language models. arXiv preprint arXiv:2303.00905 (2023)

[55] Team, O.M., Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Kreiman, T., Xu, C., et al.: Octo: An open-source generalist robot policy. arXiv preprint arXiv:2405.12213 (2024)

[56] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arxiv 2023. arXiv preprint arXiv:2302.13971 **10** (2023)

[57] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)

[58] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[59] Wang, Z., Yu, W., Ren, X., Zhang, J., Zhao, Y., Saxena, R., Cheng, L., Wong, G., See, S., Minervini, P., et al.: Mmlongbench: Benchmarking long-context vision-language models effectively and thoroughly. arXiv preprint arXiv:2505.10610 (2025)

[60] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)

[61] Wilson, A., Jiang, H., Lian, W., Yuan, W.: Cable routing and assembly using tactile-driven motion primitives. arXiv preprint arXiv:2303.11765 (2023)

[62] Yang, F., Feng, C., Chen, Z., Park, H., Wang, D., Dou, Y., Zeng, Z., Chen, X., Gangopadhyay, R., Owens, A., et al.: Binding touch to everything: Learning unified multimodal tactile representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26340–26353 (2024)

[63] Yu, J., Liu, H., Yu, Q., Ren, J., Hao, C., Ding, H., Huang, G., Huang, G., Song, Y., Cai, P., et al.: Forcevla: Enhancing vla models with a force-aware moe for contact-rich manipulation. arXiv preprint arXiv:2505.22159 (2025)

[64] Yu, S., Lin, K., Xiao, A., Duan, J., Soh, H.: Octopi: Object property reasoning with large tactile-language models. arXiv preprint arXiv:2405.02794 (2024)

[65] Yuan, W., Dong, S., Adelson, E.H.: Gelsight: High-resolution robot tactile sensors for estimating geometry and force. Sensors **17**(12), 2762 (2017)
[66] Yuan, Y., Che, H., Qin, Y., Huang, B., Yin, Z.H., Lee, K.W., Wu, Y., Lim, S.C., Wang, X.: Robot synesthesia: In-hand manipulation with visuotactile sensing. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 6558–6565. IEEE (2024)
[67] Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
[68] Zhang, C., Hao, P., Cao, X., Hao, X., Cui, S., Wang, S.: Vtla: Vision-tactile-language-action model with preference learning for insertion manipulation. arXiv preprint arXiv:2505.09577 (2025)
[69] Zhang, X., Ding, Y., Amiri, S., Yang, H., Kaminski, A., Esselink, C., Zhang, S.: Grounding classical task planners via vision-language models. arXiv preprint arXiv:2304.08587 (2023)
[70] Zhang, Z., Ma, J., Yang, X., Wen, X., Zhang, Y., Li, B., Qin, Y., Liu, J., Zhao, C., Kang, L., et al.: Touchguide: Inference-time steering of visuomotor policies via touch guidance. arXiv preprint arXiv:2601.20239 (2026)
[71] Zhao, J., Kuppuswamy, N., Feng, S., Burchfiel, B., Adelson, E.: Polytouch: A robust multi-modal tactile sensor for contact-rich manipulation using tactile-diffusion policies. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 104–110. IEEE (2025)
[72] Zhao, J., Ma, Y., Wang, L., Adelson, E.H.: Transferable tactile transformers for representation learning across diverse sensors and tasks. arXiv preprint arXiv:2406.13640 (2024)
[73] Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: Conference on Robot Learning. pp. 2165–2183. PMLR (2023)