Robotics: Science and Systems 2026
Sydney, Australia, July 13–17, 2026

ViTacFormer: Learning Cross-Modal Representation for Visuo-Tactile Dexterous Manipulation

Liang Heng*[1,2,3] Haoran Geng*†[1] Kaifeng Zhang[3]
Pieter Abbeel[1] Jitendra Malik[1]
[1]University of California, Berkeley [2]Peking University [3]Sharpa
*Equal Contribution † Project Lead

roboverseorg.github.io/ViTacFormerPage/

***Abstract*—Dexterous manipulation is a cornerstone capability for robotic systems aiming to interact with the physical world in a human-like manner. Although vision-based methods have advanced rapidly, tactile sensing remains crucial for fine-grained control—particularly in unstructured or visually occluded settings. We present ViTacFormer, a representation-learning approach that couples a cross-attention encoder to fuse high-resolution vision and touch with an autoregressive tactile-prediction head that anticipates future contact signals. Building on this architecture, we devise an easy-to-challenging curriculum that steadily refines the visual-tactile latent space, boosting both accuracy and robustness. The learned cross-modal representation drives imitation learning for multi-fingered hands, enabling precise and adaptive manipulation. Across a suite of challenging real-world benchmarks, our method achieves approximately 50% higher success rates than prior state-of-the-art systems. To our knowledge, it is also the first to autonomously complete long-horizon dexterous manipulation tasks that demand highly precise control with an anthropomorphic hand—successfully executing up to 11 sequential stages and sustaining continuous operation for 2.5 minutes.**

I. Introduction

Recent years have seen rapid advances in robotic manipulation[13, 10, 32, 11, 35, 55, 12, 21, 22, 7], with behavior cloning [38, 37, 40, 15, 16] emerging as a promising method for high-precision tasks in real-world settings. However, most existing work remains limited to simple hand configurations[58] and exhibits poor generalization—largely due to the underutilization of tactile sensing[46, 41, 42, 54], which is essential for fine-grained control.

Some works employ cross-attention [25, 8] and curriculum learning [31] for visuo-tactile fusion. Others focus on replicating the success of self-supervised learning to learn representations for tactile signals [48, 9]. While some studies have begun integrating tactile feedback into dexterous manipulation [23, 30], the learned tactile representations are often shallow. Consequently, there is still a lack of an effective model that learns cross-modal representations for visuo-tactile dexterous manipulation [34, 51]. We address this limitation with ViTacFormer, a unified visuo-tactile framework that enables fine-grained, generalizable manipulation through deep cross-modal representation learning.

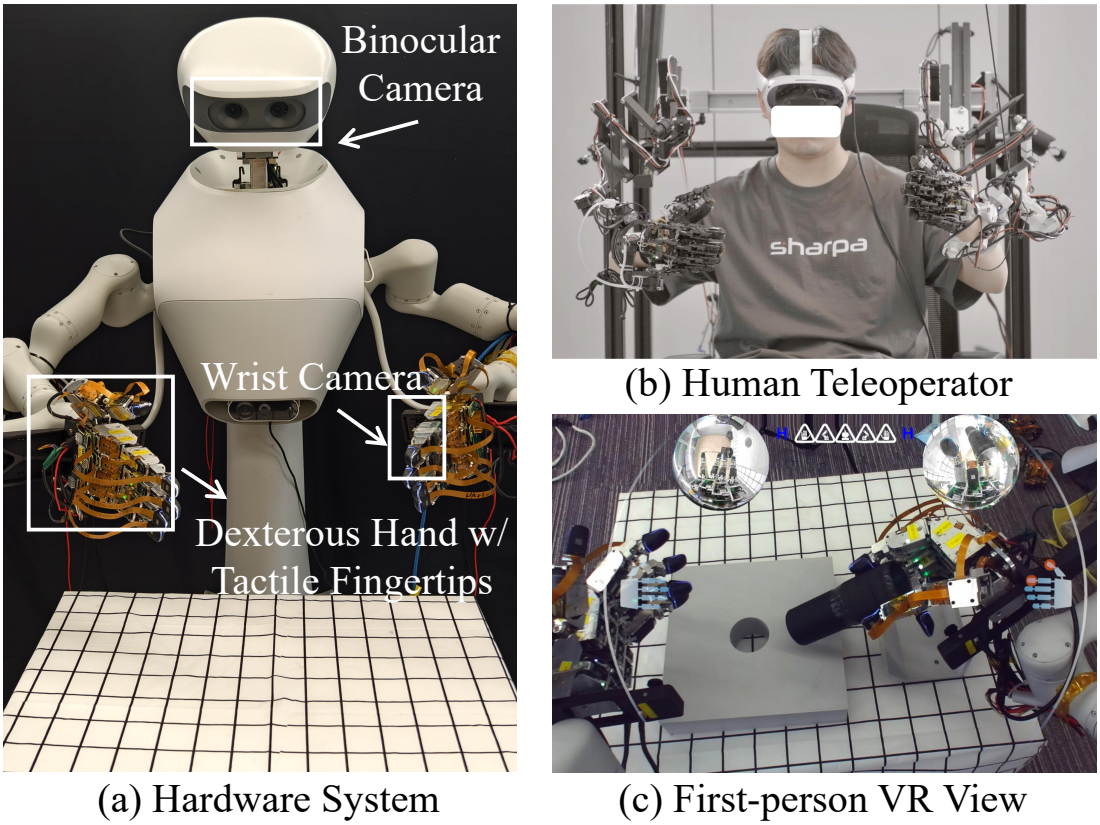
We address this gap with **ViTacFormer**, a unified visuo-tactile framework for dexterous manipulation. Our key idea is a cross-modal representation built with cross-attention layers that fuse visual and tactile cues at every stage of the policy. Crucially, we argue, and empirically confirm, that **predicting future tactile states** is more informative than merely perceiving current ones. ViTacFormer therefore adds a dedicated tactile-prediction head that forces the shared latent space to encode actionable touch dynamics, and then auto-regressively leverage the predicted future tactile signals for generating actions.

Experiments show that learning representations from predicted tactile signals in an autoregressive manner is challenging. To address this, we propose a two-phase curriculum: during the first 75% of training, we use ground-truth tactile inputs to stabilize the representation learning; in the final 25%, we transition to predicted tactile signals, promoting robust cross-modal reasoning.

To evaluate ViTacFormer, we construct the first comprehensive real-world benchmark for visuo-tactile dexterous manipulation, spanning both short- and long-horizon tasks. Across all benchmarks, ViTacFormer improves the success rate by roughly **50%** over strong baselines and is, to our knowledge, the first system to successfully complete very long-horizon dexterous manipulation tasks on a real robot, achieving **11 sequential stages** and sustaining continuous manipulation for over **2.5 minutes**.

In summary, our contributions include:

- A real-world experimental setup featuring bi-manual dexterous robotic hands, a teleoperation system, a high-quality dataset for real-world dexterous manipulation, and a comprehensive benchmark suite for evaluating visuo-tactile manipulation performance.
- A novel multimodal representation learning framework that integrates cross-attention for effective modality fusion, employs autoregressive modeling to forecast tactile signals, and introduces a tailored curriculum to enhance policy learning and generalization.
- We demonstrate strong results showcasing versatile and dexterous manipulation capabilities, including success in complex, long-horizon tasks. Our method outperforms strong baselines by approximately 50% in success rate and, to our knowledge, is the first to achieve very long-horizon dexterous manipulation on a real robot, complet-



(a) Hardware System (b) Human Teleoperator (c) First-person VR View

Fig. 1. An overview of our system hardware and teleoperation setup. (a) Our hardware system setup. (b) Teleoperator with exoskeleton gloves and VR headset. (c) VR interface with binocular and wrist views, and tactile feedback overlay.

ing 11 sequential stages with 2.5 minutes of continuous operation.

II. Related Work

A. Dexterous Manipulation

Dexterous manipulation has emerged as a critical research frontier, with applications in tasks like grasping [42, 54, 41, 17], in-hand manipulation [49, 14, 33, 44, 4], in-hand orientation [3], articulated object manipulation [1, 20, 5, 53, 12, 10, 11], and deformable object manipulation [26, 45, 59, 43]. Meanwhile, behavior cloning (BC) [38, 37, 40, 27, 24, 29, 28, 56] empowers dexterous manipulation with an end-to-end general solution.

Among BC models, diffusion policy (DP) [6] leverages diffusion models [39, 19] to learn the expert actions conditioned on robot observations. Since diffusion models [39, 19] are good at capturing the multi-modalities from diverse data inputs, diffusion policy shows promising results on robotic applications [58]. 3D diffusion policy (DP3) [52] ulitizes 3D point clouds as robot observations. It is more generalizable compared to DP since the learned representation captures geometric information from 3D data. Action chunking transformer (ACT) [57] views BC model as a conditional variational auto-encoder. It learns the multi-modal information from diverse expert data inputs. Empirical studies [58] show that ACT [57] outperforms DP [6] when the collected data is limited.

Our ViTacFormer is built on top of ACT [57], leveraging the advantage of capturing multi-modalities in diverse expert data inputs. It offers a cross-modal representation learning for visuo-tactile dexterous manipulation. The learned representation enables precise and adaptive manipulation on multifingered dexterous hands.

B. Manipulation with Tactile Signals

Prior works with tactile signals focus mainly on learning tactile representations for robotics. Some of these works focus on leveraging force values as tactile signals. For example, [23] builds some proxy tasks such as predicting robot optical flow to extract tactile representations. HATO [30] sets up a bimanual dexterous visuo-tactile manipulation system with a diffusion policy [6] to learn the expert behaviors. Recent approaches have also introduced cross-attention mechanisms [25, 8] and curriculum strategies [31] to enhance multi-sensory fusion. Other works focus on leveraging self-supervised learning to extract rich representations specifically from high-resolution tactile images [48, 47, 50, 18, 9]. Among these works, contrastive learning [48] and masked auto-encoding [36] are two popular streams to extract the representation from raw tactile images.

However, there is still a lack of an effective model that learns cross-modal representations for visuo-tactile dexterous manipulation [34, 51]. Our ViTacFormer proposes a cross-

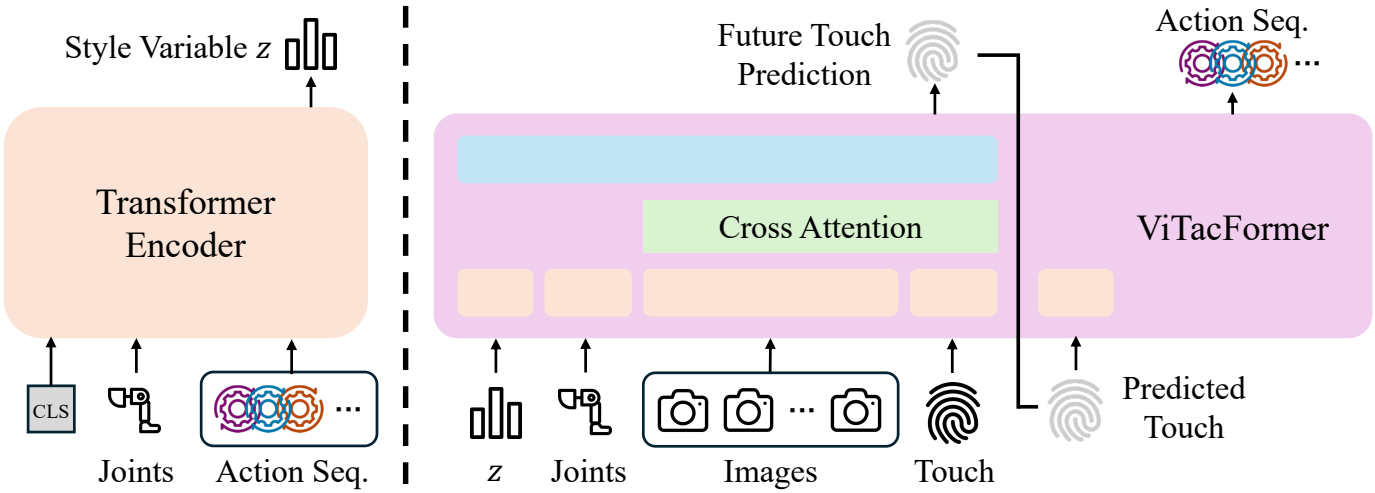


Fig. 2. The neural network architecture for ViTacFormer is a conditional variational auto-encoder. Left: a transformer-based encoder maps action sequence and robot proprioception to action style variable z. Right: a transformer-based encoder-decoder uses style variable z, robot proprioception (joints), and visuo-tactile observations to auto-regressively predict future tactile signals and generate actions.

attention-based auto-regressive model for future tactile forecasting and action generation. Empirical studies show that ViTacFormer unlocks the power of visuo-tactile representation for dexterous manipulation. In particular, ViTacFormer is capable of mastering long-horizon dexterous robotic tasks.

III. Problem Formulation and Hardware Setup

A. Problem Formulation

We address imitation learning for dexterous bi-manual manipulation. Given N expert trajectories $\mathcal{D} = \{\tau_i\}_{i=1}^{N}$, where each $\tau_i = \{(o_t^i, a_t^i)\}_{t=1}^{T_i}$ consists of multimodal observations o_t^i, and corresponding actions a_t^i. Here, the multimodal observations o_t^i include robot proprioception j_t^i, visual observations, v_t^i and tactile observations h_t^i from tactile fingertips.

The goal is to learn a policy π_θ that maps observations to actions: $a_t = \pi_\theta(o_t)$. The policy π_θ is trained to imitate expert behavior and is evaluated in task space, measuring success on manipulation tasks under diverse and long-horizon conditions.

B. Hardware Setup

Our hardware system consists of two Realman robot arms, each equipped with a SharpaWave dexterous hand (Fig.1(a)). Each hand is anthropomorphic, featuring 5 digits with 17 degrees of freedom (DoFs). Note that it is the developing version of SharpaWave. Visual observations are captured using two wrist-mounted fisheye cameras for close-up task views and a top-mounted ZED Mini stereo camera for global scene awareness. Tactile sensing is enabled by high-resolution (320×240) tactile sensors embedded in the fingertips, developed by Sharpa. For policy learning, we extract 3-axis force and torque readings from each of the 10 fingertips to capture contact dynamics efficiently.

We adopt a custom exoskeleton-based teleoperation system to collect high-quality visuo-tactile demonstrations (Fig.1(b)). The operator wears a pair of mechanical exoskeleton gloves that are mechanically coupled to the SharpaWave hands, faithfully capturing finger joint motions. A VR headset provides immersive visual feedback through a first-person interface that integrates multimodal sensory input (Fig.1(c)). The interface combines (i) a stereo top-down view from the ZED Mini, (ii) wrist-mounted local views from both arms, and (iii) real-time tactile overlays on the fingertips that highlight contact activations. This unified perception setup enables the operator to intuitively control both hands in complex, contact-rich tasks. All data streams — including RGB frames, joint states, and compressed tactile maps — are time-synchronized and logged to construct multimodal expert trajectories.

IV. Method

In section IV-A, we introduce a cross-attention-based multimodal integration framework that fuses the visual and tactile observation inputs. In section IV-B, we present autoregressive modeling with tactile forecasting, which better generates actions with predicted future tactile signals. In section IV-C, we summarize the network architecture and learning procedure for our ViTacFormer.

A. Cross-Attention-Based Multimodal Integration

Visual observations and tactile signals share similar semantic information. Traditional neural network architecture fuses visual and tactile observation inputs as naive token fusion. These models don't take the relevant information between visual and tactile observations into consideration.

Cross-attention is a mechanism commonly used in transformers, particularly in tasks involving multi-modal data or interacting with external knowledge. It allows the model to attend to different parts of two input sequences simultaneously, enabling it to capture interactions between them. Con-

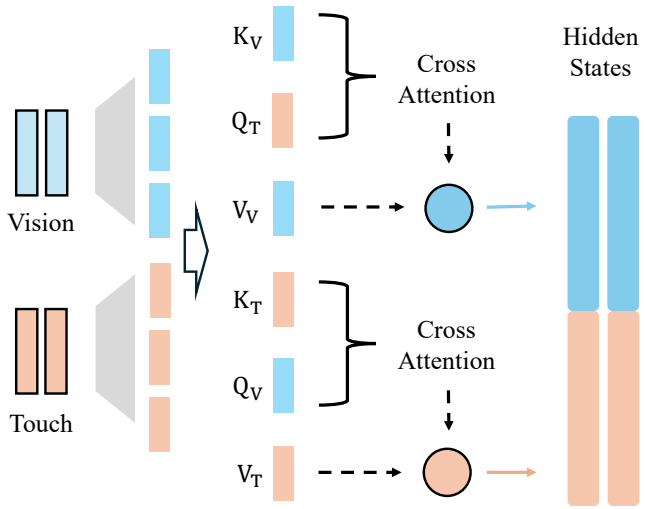


Fig. 3. Cross-attention-based multimodal integration between visual and tactile observations.

sequently, cross-attention-based multimodal integration motivates the agent to capture dependencies between diverse data inputs.

In ViTacFormer, we apply cross attention to visual and tactile observations for learning better representations. This motivates the extraction of the relevant semantic information between visual and tactile signals. Fig. 3 shows the neural network architecture of multimodal integration based on cross-attention. The keys and values from visual observations are calculated with the queries from tactile signals and vice versa. Finally, the cross-attention-based features are concatenated into hidden states for further learning.

B. Auto-Regressive Modeling with Tactile Signal Forecasting

Forecasting future tactile signals motivates the agent to be aware of the change in contact signals. In detail, it motivates the latent representations involving potential future outcomes. Auto-regressively leveraging these predicted future tactile signals motivates the agent to use this prior contact knowledge for better generating actions.

To take advantages above, we formulate the action-generating procedure in two steps. First, we predict the future tactile tokens with style variables z, current robot proprioception (joints), and visuo-tactile observations. We quantitatively validate this forecasting with an average normalized L1 error of $\approx 0.08 \pm 0.02$. This low error confirms the model captures essential contact dynamics, and its effectiveness is further justified by our ablation study (Section V-D), where removing this prediction significantly degrades performance. Next, we concatenate the predicted future tactile signals with the current input tokens for generating actions. Note that we conduct cross-attention-based multimodal integration twice between visuo-tactile signals, both in predicting future tactile signals and generating actions.

In practice, the tactile forecasting module requires sufficient training steps to minimize prediction error. Directly using inaccurate predicted signals as inputs in the early stages prevents the policy from learning valid visuo-tactile associations. To address this, we employ a two-phase curriculum strategy, inspired by Scheduled Sampling [2]. During the first 75% of epochs, we utilize ground-truth tactile tokens to ensure the policy captures correct action dependencies based on accurate states. In the final 25% of epochs, we transition to using predicted tactile signals. This second phase is crucial for adapting the policy to the minor prediction errors inevitably encountered during real-world inference, thereby ensuring robust deployment.

C. Neural Network Architecture and Learning Procedure

Fig. 2 shows the neural network architecture of our ViTacFormer. This architecture is basically a conditional variational auto-encoder. On the left of Fig. 2, there is a transformer-based encoder. It maps the robot's proprioception (joints) and expert action sequence into a style variable z. On the right of Fig. 2, there is a transformer-based encoder-decoder. First, it extracts the representation from visual and tactile observations with a cross-attention-based multimodal integration framework. Next, it auto-regressively predicts the future tactile signals and thus generates the actions with predicted future tactile signals. The style variable z is sampled from expert demonstrations during training, while it is set to zero during inference, following ACT[57].

We find that a combination of supervision between arm end-effectors' position and arm/hand joint angles (JA) is more effective for dexterous manipulation than arm/hand JA supervision only. The training loss of ViTacFormer is shown as:

$$\mathcal{L} = w_1 \cdot \mathcal{L}_{KL} + w_2 \cdot \mathcal{L}_{JA} + w_3 \cdot \mathcal{L}_{tactile} + w_4 \cdot \mathcal{L}_{arm}, \quad (1)$$

where $w_{1,2,3,4}$ are hyper-parmeters, $\mathcal{L}_{KL}$ is KL divergence between the action style variables and Gaussian distribution, $\mathcal{L}_{JA}$ is the L_1 loss based on predicted action and ground truth action, $\mathcal{L}_{tactile}$ is the L_1 loss based on future tactile signals and ground truth. In particular, $\mathcal{L}_{arm}$ is:

$$\mathcal{L}_{arm} = \lambda_1 \cdot \mathcal{L}_{position} + \lambda_2 \cdot \mathcal{L}_{rotation}, \quad (2)$$

where $\lambda_{1,2}$ are hyper-parameters, $\mathcal{L}_{arm}$ indicates the supervision based on arm end-effectors, $\mathcal{L}_{position}$ is the L_2 loss between arm end-effector's position, and $\mathcal{L}_{rotation}$ is the L_1 loss between arm end-effector's rotation. Empirically, we find $\mathcal{L}_{arm}$ is very useful in training dexterous manipulation skills.

D. Implementation Details

Input Modalities. Our model processes three synchronized modalities. *Visual Input:* We utilize four camera views: a stereo pair (180×320) from a top-mounted ZED Mini and two wrist-mounted fisheye views (256×280). All frames are encoded via a vision backbone. *Proprioception:* The robot's state is a 58-dimensional vector $[7, 17, 7, 17, 2]$ representing the dual arms, dexterous hands, and neck. We use a temporal horizon of 6

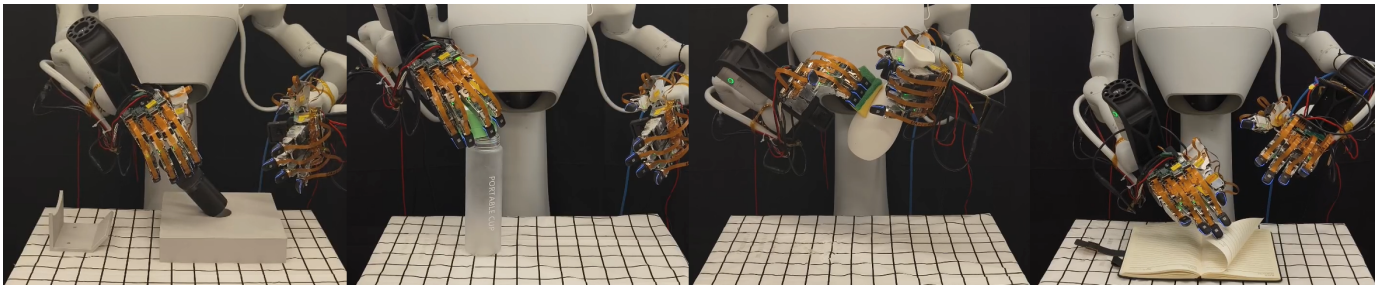


Fig. 4. Four short-horizon visuo-tactile manipulation tasks: Peg Insertion, Cap Twist, Vase Wipe, and Book Flip.

frames, resulting in a $(6, 50)$ input shape. *Tactile Input:* Each fingertip provides 3-axis force and torque data (20 channels total). We process 18 frames of raw signals concatenated with frame-wise deltas, yielding a final tensor of shape $[18, 120]$.

Action Output. The policy generates high-frequency action sequences with shape $(100, 50)$ per rollout. The 100-frame horizon supports fine-grained dexterous motion. During deployment, the policy runs at 10Hz, and we apply temporal smoothing to the predicted trajectory for stable execution.

Training Setup. We train each task using 50 expert demonstrations on 2 NVIDIA H20 GPUs. Short-horizon tasks converge within 12 hours, while long-horizon tasks require up to 2 days. The model is optimized using Adam (lr=$1e^{-4}$, batch size=128). The loss function combines KL divergence on latent style, L1 losses on predicted actions and tactile signals, and auxiliary supervision on end-effector poses.

V. Experiment

In this section, we evaluate the effectiveness of our proposed ViTacFormer. The experiments are designed to answer two questions: (1) How does our algorithm perform compared to other state-of-the-art imitation learning algorithms? The results are presented in section V-C. (2) Is each component of our algorithm effective? The results are introduced in section V-D.

A. Benchmark and Environment Setup

In this section, we introduce the tasks and environment setup. The tasks we conduct include 4 simple dexterous manipulation tasks and a very long-horizon visuo-tactile task. Fig. 4 shows the 4 simple dexterous manipulation tasks, each presenting unique sensory challenges: **Peg Insertion** involves inserting a peg under visual occlusion with a tight 0.5mm clearance; **Cap Twist** requires precise rotational control to unscrew a cap; **Vase Wipe** necessitates following the contour of the vase body to thoroughly wipe off the ink; and **Book Flip** demands applying the appropriate friction force to separate and flip individual pages. We also conduct our ViTacFormer on a very long-horizon task, i.e., making hamburgers.

We conduct algorithm comparison on all tasks and ablation study on four simple tasks. These tasks range from easy to complex dexterous manipulation. The results show that our ViTacFormer outperforms other state-of-the-art imitation learning algorithms by over 50% success rates. The ablation study illustrates that each component in our ViTacFormer improves the manipulation performance. Note that we use only 50 trajectories per task for training in our experiments. This highlights the high sample efficiency of our framework, as it achieves robust generalization to spatial perturbations with limited expert demonstrations.

B. Metrics and Baselines

Metrics We evaluate the algorithms with two established metrics: human normalized score and success rates. Note that the success rates may not reflect the dexterous manipulation process in detail, especially for long-horizon manipulation tasks. We define a new metric to measure the dexterous manipulation performance.

We propose Human Normalized Score (HNS) to evaluate the manipulation process in detail. First, we split the manipulation process into several stages. In each stage, we evaluate the process with $0 - 3$ raw scores. Finally, we normalize the score for a fair comparison. The HNS score is presented as:

$$\text{HNS} = \frac{\sum_{i=1}^{N} w_i \cdot s_i}{3 * \sum_{i=1}^{N} w_i}, \tag{3}$$

which N represents the number of stages in a certain manipulation task, w_i represents tactile reliance in stage i, indicating how strongly this stage depends on tactile feedback, and s_i counts from 0 to 3, which is the raw score for stage i measuring its success level.

Baselines Diffusion Policy (DP) [6] is good at mimicking expert behaviors by capturing multi-modalities from training data with a diffusion model [19]. Additionally, HATO [30] adds the tactile signals as conditions for diffusion policy [6]. ACT [57] builds a conditional variational auto-encoder for learning expert behaviors from demonstrations. ACTw/T [57] adds the tactile signal in its input tokens. Empirical studies [58] show that ACT [57] outperforms DP [6] with limited training data.

In this section, we use DP [6], HATO [30], ACT [57], ACTw/T [57] as our baselines. Among these baselines, DP and ACT are without tactile inputs. HATO and ACTw/T take tactile signals with a naive token fusion.

Fig. 5. Successful model rollout on long-horizon task, i.e., making hamburger. We show the successful model rollout with keyframes in 11 stages. The first row represents the robot hand turning the brand to "open". The second row represents the robot hand shoveling meat to bread. The third row represents the robot hand assembling the hamburger. The fourth row represents the robot hand handing over the hamburger to the plate. The fifth row represents the robot hand turning the brand to "close".

C. Algorithm Comparison

Question 1: How does ViTacFormer perform compared to other SoTA imitation learning algorithms?

To test the effectiveness of our ViTacFormer, we conduct experiments on four short-horizon dexterous manipulation tasks. We test each algorithm on a certain task with inference 10 times. If the results show that ViTacFormer achieves higher success rates compared to SoTA imitation learning baselines, we could prove the efficacy of our ViTacFormer.

Tab. I shows the success rates on four short-horizon dexterous manipulation tasks. Our ViTacFormer achieves the best performance compared to other SoTA imitation learning algorithms. In particular, ViTacFormer outperforms other baselines over 50% success rates, therefore almost solving these tasks. On the other hand, tactile observation input greatly improves the manipulation performance since ACTw/T [57] and HATO [30] outperform ACT [57] and DP [6], respectively.

Question 2: How does ViTacFormer perform on complex long-horizon manipulation tasks?

TABLE I
SUCCESS RATE COMPARISON ON FOUR SHORT-HORIZON DEXTEROUS MANIPULATION TASKS. OUR VITACFORMER ACHIEVES OVER 50% SUCCESS RATES COMPARED TO THE BASELINES.

Task	Peg Insertion	Cap Twist	Vase Wipe	Book Flip
DP	2/10	0/10	3/10	1/10
ACT	4/10	4/10	3/10	2/10
HATO	4/10	1/10	4/10	3/10
ACTw/T	6/10	6/10	4/10	4/10
Ours	**10**/10	**10**/10	**9**/10	**9**/10

To show that our ViTacFormer is effective on long-horizon tasks, we conduct experiments on an 11-stage task, i.e., making hamburgers. To the best of our knowledge, ViTacFormer is the first systems to complete very long-horizon dexterous manipulation tasks on a real robot with a single imitation

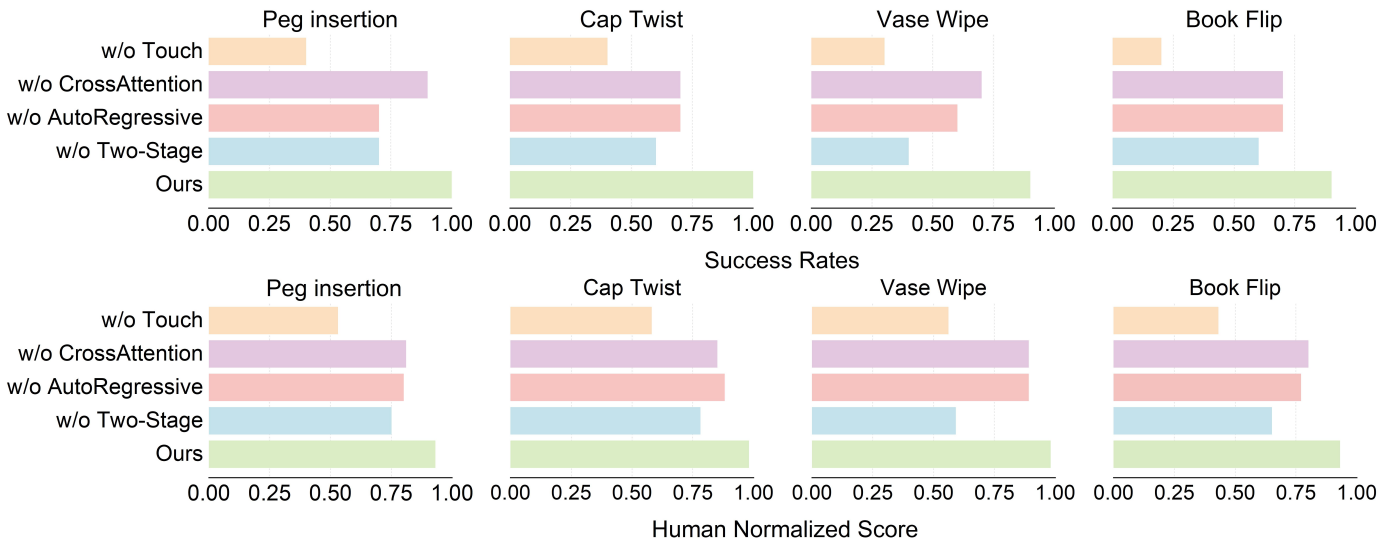


Fig. 6. Ablation study. Performance comparison by removing different components from the full ViTacFormer (Ours).

TABLE II
HUMAN EVALUATION SCORE COMPARISON ON A VERY LONG-HORIZON DEXTEROUS MANIPULATION TASK. VITACFORMER SHOWS PROMISING RESULTS ON THIS LONG-HORIZON TASK.

Stage	1	2	3	4	5	6	7	8	9	10	11	Overall
ACT	2.4	2.5	**1.9**	**2.0**	0.7	2.2	1.6	2.8	2.2	2.2	0.7	0.61
ACT w./T	2.6	**3.0**	1.8	1.6	2.0	2.3	**2.2**	**2.9**	2.0	1.8	1.4	0.72
Ours	**2.9**	**3.0**	**1.9**	1.8	**2.7**	**2.9**	2.0	2.8	**2.4**	**2.5**	**3.0**	**0.88**

learning model. Fig. 5 shows a successful model rollout of our ViTacFormer. Our ViTacFormer masters 11 stages of making hamburgers. We compare with ACT and ACT w./T as the main long-horizon baselines, where ACT w./T additionally uses tactile inputs with naive token fusion. We also evaluate DP and HATO, but they rarely complete the full sequence in this long-horizon setting, so we focus the main comparison on ACT-based baselines.

Tab. II shows the human normalized score (HNS) for each stage in the task, i.e., making hamburgers. HNS is used as a diagnostic metric to evaluate the quality of each stage. Note that once the score is less than 1 for one stage, the model fails on this task. In experiments, we correct the mistake only when a stage fails completely (score below 1) for further stage testing. Specifically, stage 1 corresponds to the first row of Fig. 5; Stage 2-4 corresponds to the second row; Stage 5-7 corresponds to the third row; Stage 8-10 corresponds to the fourth row; and stage 11 corresponds to the fifth row. Adding tactile input with naive fusion improves ACT w./T from 0.61 to 0.72 overall HNS, showing that tactile sensing itself is useful. Our ViTacFormer further improves the overall HNS to **0.88** and achieves higher scores in most stages, indicating that the performance gain comes not only from tactile input, but also from our predictive visuo-tactile representation learning.

We additionally report two success-rate metrics for this long-horizon task. When success is defined as completing every stage with a non-zero score, ACT, ACT w./T, and ViTacFormer achieve 10%, 40%, and **80%**, respectively. Under the stricter criterion of completing the full hamburger-making task without any human intervention, the success rates are 0%, 10%, and **70%**, respectively. This indicates that ViTacFormer improves both stage-wise execution quality and strict end-to-end autonomy.

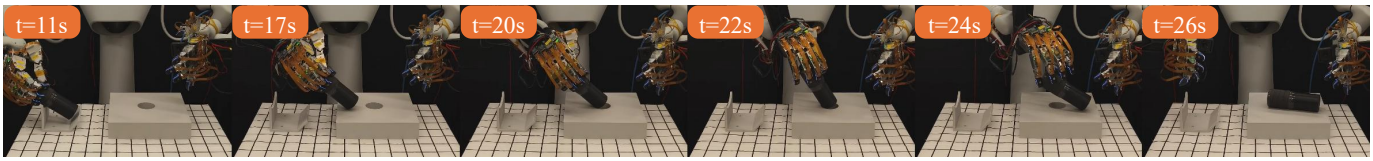
Specifically, in stage 5 (grasping lettuce), the object is soft and deformable, so tactile sensing is required to detect contact and stabilize the grasp, while in stage 11 (flipping the sign at the end), the task demands precise timing and force control under occlusion, where tactile input enables accurate triggering of the flipping motion. These cases show that such tasks cannot be reliably solved by vision alone, and further benefit from predictive visuo-tactile modeling.

D. Ablation Study

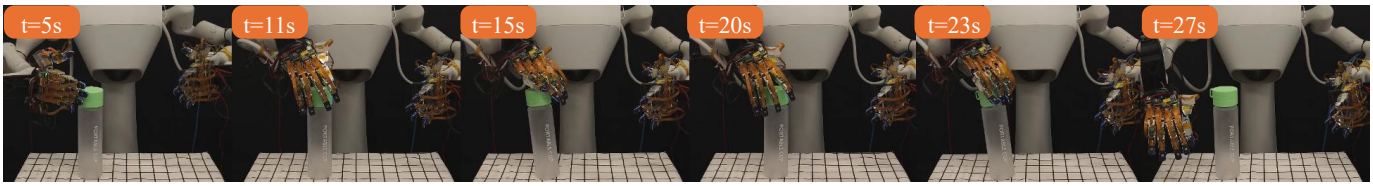
Question 3: How does each component in our ViTacFormer contribute to the baseline?

To rigorously evaluate the contribution of each architectural component, we conduct an ablation study by systematically removing modules from the full ViTacFormer model. Fig. 6 illustrates the quantitative comparison. We analyze the specific impact of each component below:

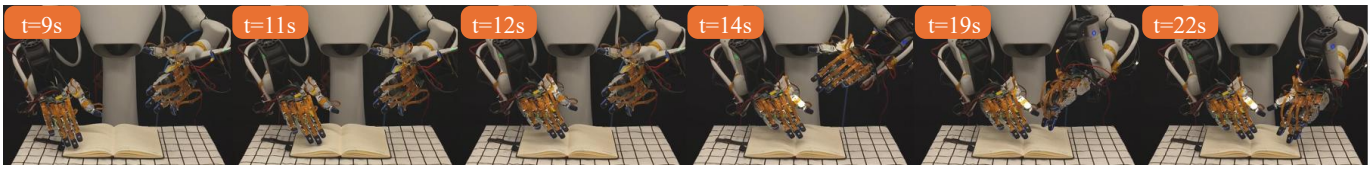
- **Effect of Cross-Attention (w/o CrossAtten):** Removing the cross-attention module and replacing it with naive concatenation leads to a significant performance drop, particularly in the *Peg Insertion* and *Cap Twist* tasks. Without the attention mechanism, the model fails to



(a) Peg Insertion Failure



(b) Cap Twist Failure



(c) Book Flip Failure

Fig. 7. Failure Study. The first row is peg insertion failure, the second row is cap twist failure, and the third row is book flip failure.

dynamically weigh the importance of tactile features when visual features are ambiguous (e.g., during occlusion). This suggests that the dense interaction between modalities is crucial for fine-grained control.

- **Effect of Tactile Forecasting (w/o AutoRegressive):** Excluding the auto-regressive tactile prediction head impairs the temporal modeling of contact dynamics. In dynamic tasks like *Vase Wipe*, we observed that the ablated model often loses contact with the surface or applies inconsistent force. The forecasting objective effectively forces the latent representation to encode not just the *current* state, but the *trend* of contact, serving as a strong regularization for smooth manipulation.
- **Effect of Curriculum Learning (w/o Two-Stage):** Training without the two-phase curriculum (directly using predicted tactile signals or ground truth throughout) results in unstable convergence. The "w/o Two-Stage" variant shows higher variance in success rates. The curriculum acts as a necessary bridge, allowing the policy to first learn valid kinematics before adapting to the noisy nature of predicted tactile signals.

Overall, the full ViTacFormer achieves the best results, validating that each proposed component—modality fusion, predictive modeling, and curriculum training—is essential for solving complex visuo-tactile manipulation tasks.

Question 4: How does the failure occur in baselines?

There are several failure modes from the baselines. Evaluating these failure cases helps us understand the effective factors in our ViTacFormer. Fig. 7 shows the classical failure cases from the baseline ACT w/ Touch. The established tasks are highly tactile-dependent. Consequently, how to leverage the tactile signals in imitation learning is of great significance.

Fig. 7 shows the peg insertion failure. When the peg is moved to the hole, the robot hand isn't aware of the position of the hole and thus fails to insert the peg into the hole. This failure shows the importance of predicting future tactile signals in ViTacFormer. Predicting the future tactile tokens motivates the dexterous hand to be aware of the temporal force difference. When the peg is near the hole, the temporal force difference would be changed, therefore showing the hole is nearby. Consequently, it could improve the robustness of inserting the peg into the hole in ViTacFormer.

Fig. 7 shows the cap twist failure. The robot hand fails to twist the cap. The reason can be traced to the fact that the hand isn't aware of whether the cap is open or closed. It motivates the importance of the autoregressive architecture of our ViTacFormer. Auto-regressively predicting the future tactile signals and leveraging these signals simplifies reasoning about the actions under such complex situations.

Fig. 7 shows the book flip failure. The dexterous hand isn't aware of the book. It flips the book in the air. This originates from the lack of visual and tactile observation fusion. In our ViTacFormer, we use cross-attention-based multimodal integration to fuse the visual and tactile observations. This improves the performance of dexterous manipulation.

VI. Limitation and Future Work

Limitations. Due to the inherent constraints of imitation learning, our policy lacks the capability to autonomously generalize to novel tasks unseen during training. It still depends on human teleoperation for data collection, which is time-consuming and labor-intensive. While our method demonstrates strong performance in long-horizon and challenging

visuo-tactile manipulation tasks, its stability can be affected in scenarios where tactile feedback is extremely noisy or ambiguous. This is primarily due to the limitations of the current sensor resolution and the representation learning process.

Future Work. In the future, we plan to address these limitations by exploring two directions. First, we aim to integrate Sim-to-Real transfer learning to reduce the reliance on real-world human demonstrations. By training in a physics-rich simulation with rendered tactile images, we can scale up data collection significantly. Second, we plan to investigate more generalizable tactile representations, such as foundation models for touch, to enhance robustness against sensor noise and improve adaptation to novel objects.

VII. Conclusion

We present ViTacFormer, a unified visuo-tactile framework for dexterous robotic manipulation that leverages deep cross-modal representation learning. By fusing vision and touch at every stage of the policy and incorporating predictive tactile modeling, ViTacFormer enables robust, fine-grained control across a diverse set of manipulation tasks. Our curriculum-based training strategy further enhances representation stability, allowing the system to effectively reason over predicted tactile signals. Empirical results demonstrate that ViTacFormer significantly outperforms strong baselines—achieving approximately 50% higher success rates—and is the first to complete long-horizon dexterous tasks on a real robot. We believe this work opens new possibilities for generalizable, high-precision robotic manipulation through the principled integration of vision and touch.

References

[1] Chen Bao, Helin Xu, Yuzhe Qin, and Xiaolong Wang. Dexart: Benchmarking generalizable dexterous manipulation with articulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21190–21200, June 2023.

[2] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks, 2015. URL https://arxiv.org/abs/1506.03099.

[3] Tao Chen, Jie Xu, and Pulkit Agrawal. A system for general in-hand object re-orientation. In *Conference on Robot Learning*, pages 297–307. PMLR, 2022.

[4] Yuanpei Chen, Yiran Geng, Fangwei Zhong, Jiaming Ji, Jiechuang Jiang, Zongqing Lu, Hao Dong, and Yaodong Yang. Bi-dexhands: Towards human-level bimanual dexterous manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):2804–2818, 2024. doi: 10.1109/TPAMI.2023.3339515.

[5] Yuanpei Chen, Chen Wang, Yaodong Yang, and C. Karen Liu. Object-centric dexterous manipulation from human motion data, 2024. URL https://arxiv.org/abs/2411.04005.

[6] 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*, page 02783649241273668, 2023.

[7] Yufei Ding, Haoran Geng, Chaoyi Xu, Xiaomeng Fang, Jiazhao Zhang, Songlin Wei, Qiyu Dai, Zhizheng Zhang, and He Wang. Open6dor: Benchmarking open-instruction 6-dof object rearrangement and a vlm-based approach. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7359–7366, 2024. doi: 10.1109/IROS58592.2024.10802733.

[8] Ruoxuan Feng, Di Hu, Wenke Ma, and Xuelong Li. Play to the score: Stage-guided dynamic multi-sensory fusion for robotic manipulation, 2024. URL https://arxiv.org/abs/2408.01366.

[9] Letian Fu, Gaurav Datta, Huang Huang, William Chung-Ho Panitch, Jaimyn Drake, Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, and Ken Goldberg. A touch, vision, and language dataset for multimodal alignment. *arXiv preprint arXiv:2402.13232*, 2024.

[10] Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. *arXiv preprint arXiv:2211.05272*, 2022.

[11] Haoran Geng, Ziming Li, Yiran Geng, Jiayi Chen, Hao Dong, and He Wang. Partmanip: Learning cross-category generalizable part manipulation policy from point cloud observations. *arXiv preprint arXiv:2303.16958*, 2023.

[12] Haoran Geng, Songlin Wei, Congyue Deng, Bokui Shen, He Wang, and Leonidas Guibas. Sage: Bridging semantic and actionable parts for generalizable articulated-object manipulation under language instructions, 2023.

[13] Haoran Geng, Feishi Wang, Songlin Wei, Yuyang Li, Bangjun Wang, Boshi An, Charlie Tianyue Cheng, Haozhe Lou, Peihao Li, Yen-Jen Wang, Yutong Liang, Dylan Goetting, Chaoyi Xu, Haozhe Chen, Yuxi Qian, Yiran Geng, Jiageng Mao, Weikang Wan, Mingtong Zhang, Jiangran Lyu, Siheng Zhao, Jiazhao Zhang, Jialiang Zhang, Chengyang Zhao, Haoran Lu, Yufei Ding, Ran Gong, Yuran Wang, Yuxuan Kuang, Ruihai Wu, Baoxiong Jia, Carlo Sferrazza, Hao Dong, Siyuan Huang, Yue Wang, Jitendra Malik, and Pieter Abbeel. Roboverse: Towards a unified platform, dataset and benchmark for scalable and generalizable robot learning, 2025. URL https://arxiv.org/abs/2504.18904.

[14] Ankur Handa, Arthur Allshire, Viktor Makoviychuk, Aleksei Petrenko, Ritvik Singh, Jingzhou Liu, Denys Makoviichuk, Karl Van Wyk, Alexander Zhurkevich, Balakumar Sundaralingam, Yashraj Narang, Jean-Francois Lafleche, Dieter Fox, and Gavriel State. Dextreme: Transfer of agile in-hand manipulation from simulation to reality. *arXiv*, 2022.

[15] Liang Heng, Xiaoqi Li, Shangqing Mao, Jiaming Liu, Ruolin Liu, Jingli Wei, Yu-Kai Wang, Yueru Jia, Chenyang Gu, Rui Zhao, Shanghang Zhang, and Hao

Dong. Rwor: Generating robot demonstrations from human hand collection for policy learning without robot, 2025. URL https://arxiv.org/abs/2507.03930.

[16] Liang Heng, Jiadong Xu, Yiwen Wang, Xiaoqi Li, Muhe Cai, Yan Shen, Juan Zhu, Guanghui Ren, and Hao Dong. Imagine2act: Leveraging object-action motion consistency from imagined goals for robotic manipulation, 2025. URL https://arxiv.org/abs/2509.17125.

[17] Liang Heng, Yihe Tang, Jiajun Xu, Henghui Bao, Di Huang, and Yue Wang. Humdex: Humanoid dexterous manipulation made easy, 2026. URL https://arxiv.org/abs/2603.12260.

[18] Carolina Higuera, Akash Sharma, Chaithanya Krishna Bodduluri, Taosha Fan, Patrick Lancaster, Mrinal Kalakrishnan, Michael Kaess, Byron Boots, Mike Lambeta, Tingfan Wu, et al. Sparsh: Self-supervised touch representations for vision-based tactile sensing. *arXiv preprint arXiv:2410.24090*, 2024.

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[20] Taoran Jiang, Liqian Ma, Yixuan Guan, Jiaojiao Meng, Weihang Chen, Zecui Zeng, Lusong Li, Dan Wu, Jing Xu, and Rui Chen. Dexsim2real^2: Building explicit world model for precise articulated object dexterous manipulation, 2024. URL https://arxiv.org/abs/2409.08750.

[21] Yuxuan Kuang, Junjie Ye, Haoran Geng, Jiageng Mao, Congyue Deng, Leonidas Guibas, He Wang, and Yue Wang. Ram: Retrieval-based affordance transfer for generalizable zero-shot robotic manipulation. *arXiv preprint arXiv:2407.04689*, 2024.

[22] Yuxuan Kuang, Haoran Geng, Amine Elhafsi, Tan-Dzung Do, Pieter Abbeel, Jitendra Malik, Marco Pavone, and Yue Wang. Skillblender: Towards versatile humanoid whole-body loco-manipulation via skill blending. *arXiv preprint arXiv:2506.09366*, 2025.

[23] Michelle A Lee, Yuke Zhu, Peter Zachares, Matthew Tan, Krishnan Srinivasan, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Learning multimodal representations for contact-rich tasks. *IEEE Transactions on Robotics*, 36(3):582–596, 2020.

[24] Chenxuan Li, Jiaming Liu, Guanqun Wang, Xiaoqi Li, Sixiang Chen, Liang Heng, Chuyan Xiong, Jiaxin Ge, Renrui Zhang, Kaichen Zhou, and Shanghang Zhang. A self-correcting vision-language-action model for fast and slow system manipulation, 2025. URL https://arxiv.org/abs/2405.17418.

[25] Hao Li, Yizhi Zhang, Junzhe Zhu, Shaoxiong Wang, Michelle A Lee, Huazhe Xu, Edward Adelson, Li Fei-Fei, Ruohan Gao, and Jiajun Wu. See, hear, and feel: Smart sensory fusion for robotic manipulation, 2022. URL https://arxiv.org/abs/2212.03858.

[26] Sizhe Li, Zhiao Huang, Tao Chen, Tao Du, Hao Su, Joshua B. Tenenbaum, and Chuang Gan. Dexdeform: Dexterous deformable object manipulation with human demonstrations and differentiable physics, 2023. URL https://arxiv.org/abs/2304.03223.

[27] Xiaoqi Li, Liang Heng, Jiaming Liu, Yan Shen, Chenyang Gu, Zhuoyang Liu, Hao Chen, Nuowei Han, Renrui Zhang, Hao Tang, Shanghang Zhang, and Hao Dong. 3DS-VLA: A 3d spatial-aware vision language action model for robust multi-task manipulation. In *9th Annual Conference on Robot Learning*, 2025. URL https://openreview.net/forum?id=dT45OMevL5.

[28] Xiaoqi Li, Jiaming Liu, Nuowei Han, Liang Heng, Yandong Guo, Hao Dong, and Yang Liu. 3dwg: 3d weakly supervised visual grounding via category and instance-level alignment, 2025. URL https://arxiv.org/abs/2505.01809.

[29] Xiaoqi Li, Jingyun Xu, Mingxu Zhang, Jiaming Liu, Yan Shen, Iaroslav Ponomarenko, Jiahui Xu, Liang Heng, Siyuan Huang, Shanghang Zhang, and Hao Dong. Object-centric prompt-driven vision-language-action model for robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27638–27648, June 2025.

[30] Toru Lin, Yu Zhang, Qiyang Li, Haozhi Qi, Brent Yi, Sergey Levine, and Jitendra Malik. Learning visuotactile skills with two multifingered hands. *IEEE International Conference on Robotics & Automation (ICRA)*, 2025.

[31] Jason Jingzhou Liu, Yulong Li, Kenneth Shaw, Tony Tao, Ruslan Salakhutdinov, and Deepak Pathak. Factr: Force-attending curriculum training for contact-rich policy learning, 2025. URL https://arxiv.org/abs/2502.17432.

[32] Rundong Luo*, Haoran Geng*, Congyue Deng, Puhao Li, Zan Wang, Baoxiong Jia, Leonidas Guidbas, and Siyuan Huang. Physpart: Physically plausible part completion for interactable objects. *International Conference on Robotics and Automation (ICRA)*, 2025. URL https://arxiv.org/abs/2408.13724.

[33] Haozhi Qi, Ashish Kumar, Roberto Calandra, Yi Ma, and Jitendra Malik. In-Hand Object Rotation via Rapid Motor Adaptation. In *Conference on Robot Learning (CoRL)*, 2022.

[34] Haozhi Qi, Brent Yi, Sudharshan Suresh, Mike Lambeta, Yi Ma, Roberto Calandra, and Jitendra Malik. General in-hand object rotation with vision and touch. In *Conference on Robot Learning*, pages 2549–2564. PMLR, 2023.

[35] Younggyo Seo, Carmelo Sferrazza, Haoran Geng, Michal Nauman, Zhao-Heng Yin, and Pieter Abbeel. Fasttd3: Simple, fast, and capable reinforcement learning for humanoid control, 2025. URL https://arxiv.org/abs/2505.22642.

[36] 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)*, pages 9698–9705. IEEE, 2024.

[37] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport:

What and where pathways for robotic manipulation. In *Conference on robot learning (CoRL)*, pages 894–906. PMLR, 2022.

[38] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning (CoRL)*, pages 785–799. PMLR, 2023.

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[40] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Charles Xu, Jianlan Luo, et al. Octo: An open-source generalist robot policy, 2023.

[41] Weikang Wan, Haoran Geng, Yun Liu, Zikang Shan, Yaodong Yang, Li Yi, and He Wang. Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. *arXiv preprint arXiv:2304.00464*, 2023.

[42] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. *arXiv preprint arXiv:2210.02697*, 2022.

[43] Yuran Wang, Ruihai Wu, Yue Chen, Jiarui Wang, Jiaqi Liang, Ziyu Zhu, Haoran Geng, Jitendra Malik, Pieter Abbeel, and Hao Dong. Dexgarmentlab: Dexterous garment manipulation environment with generalizable policy, 2025. URL https://arxiv.org/abs/2505.11032.

[44] 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, 2024. URL https://arxiv.org/abs/2409.17549.

[45] Yilin Wu, Wilson Yan, Thanard Kurutach, Lerrel Pinto, and Pieter Abbeel. Learning to manipulate deformable objects without demonstrations, 2020. URL https://arxiv.org/abs/1910.13439.

[46] Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, et al. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. *arXiv preprint arXiv:2303.00938*, 2023.

[47] Zhengtong Xu, Raghava Uppuluri, Xinwei Zhang, Cael Fitch, Philip Glen Crandall, Wan Shou, Dongyi Wang, and Yu She. Unit: Unified tactile representation for robot learning. *arXiv preprint arXiv:2408.06481*, 2024.

[48] Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, et al. Binding touch to everything: Learning unified multimodal tactile representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26340–26353, 2024.

[49] Zhao-Heng Yin, Binghao Huang, Yuzhe Qin, Qifeng Chen, and Xiaolong Wang. Rotating without seeing: Towards in-hand dexterity through touch. *arXiv preprint arXiv:2303.10880*, 2023.

[50] Kelin Yu, Yunhai Han, Qixian Wang, Vaibhav Saxena, Danfei Xu, and Ye Zhao. Mimictouch: Leveraging multi-modal human tactile demonstrations for contact-rich manipulation. *arXiv preprint arXiv:2310.16917*, 2023.

[51] 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. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6558–6565. IEEE, 2024.

[52] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy. *arXiv preprint arXiv:2403.03954*, 2024.

[53] Hui Zhang, Sammy Christen, Zicong Fan, Luocheng Zheng, Jemin Hwangbo, Jie Song, and Otmar Hilliges. Artigrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation. In *2024 International Conference on 3D Vision (3DV)*, pages 235–246, 2024. doi: 10.1109/3DV62453.2024.00016.

[54] Jialiang Zhang, Haoran Liu, Danshi Li, XinQiang Yu, Haoran Geng, Yufei Ding, Jiayi Chen, and He Wang. Dexgraspnet 2.0: Learning generative dexterous grasping in large-scale synthetic cluttered scenes. In *8th Annual Conference on Robot Learning*.

[55] Jialiang Zhang, Haoran Geng, Yang You, Congyue Deng, Pieter Abbeel, Jitendra Malik, and Leonidas Guibas. Rodrigues network for learning robot actions, 2025. URL https://arxiv.org/abs/2506.02618.

[56] Rongyu Zhang, Menghang Dong, Yuan Zhang, Liang Heng, Xiaowei Chi, Gaole Dai, Li Du, Yuan Du, and Shanghang Zhang. Mole-vla: Dynamic layer-skipping vision language action model via mixture-of-layers for efficient robot manipulation, 2025. URL https://arxiv.org/abs/2503.20384.

[57] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

[58] Tony Z Zhao, Jonathan Tompson, Danny Driess, Pete Florence, Kamyar Ghasemipour, Chelsea Finn, and Ayzaan Wahid. Aloha unleashed: A simple recipe for robot dexterity. *arXiv preprint arXiv:2410.13126*, 2024.

[59] Sun Zhaole, Jihong Zhu, and Robert B. Fisher. Dexdlo: Learning goal-conditioned dexterous policy for dynamic manipulation of deformable linear objects. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16009–16015, 2024. doi: 10.1109/ICRA57147.2024.10610754.