DeCAL: Towards Physically-Grounded Dexterous Vision-Language-Action Models via Contact-Aware Latent Co-Imagination

Yankai Fu[1,2]*, Ning Chen[1,2]*, Junkai Zhao[2]†, Heng Zhang[1], Guocai Yao[2], Pengwei Wang[2], Zhongyuan Wang[2], Shanghang Zhang[1,2]✉

[1]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University; [2]Beijing Academy of Artificial Intelligence

*Equal contribution, †Project leader, ✉Corresponding author

Project Webpage: https://aureleopku.github.io/DeCAL

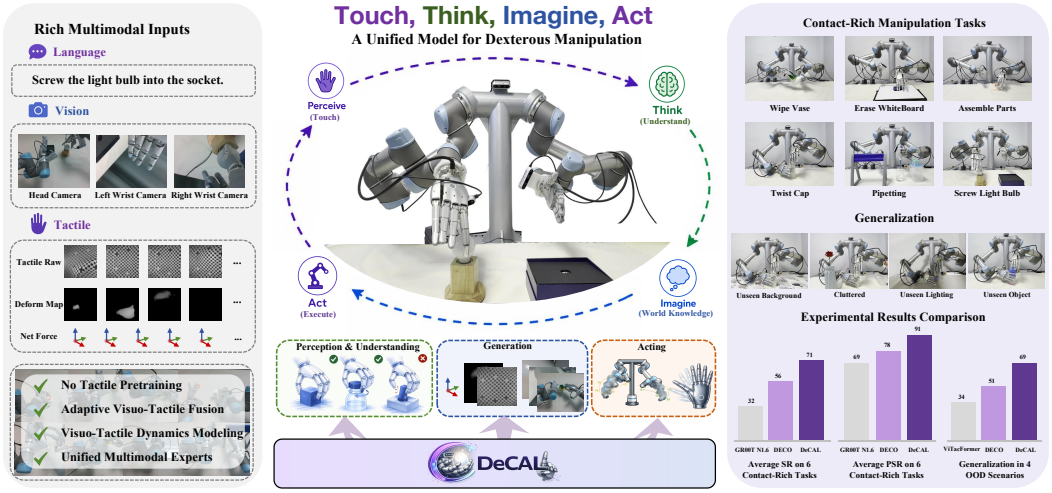

Figure 1: We present DeCAL, a physically-grounded dexterous VLA model that unifies perception, understanding, imagination, and action within a framework. Equipped with rich multimodal inputs, DeCAL achieves strong performance across diverse contact-rich dexterous manipulation tasks and demonstrates robust generalization to unseen scenarios.

Abstract: Dexterous manipulation involves contact-rich and fine-grained interactions with the physical world, posing significant challenges for existing vision-language-action (VLA) models due to severe visual occlusions and complex contact dynamics. While recent works have incorporated tactile sensing into robotic manipulation, most approaches still rely on homogeneous multimodal fusion, lacking adaptive tactile integration and explicit modeling of physical dynamics. In this work, we present DeCAL, a physically-grounded dexterous vision-language-action model that unifies understanding, imagination and action generation for contact-rich dexterous manipulation. Built upon a Mixture-of-Transformers (MoT) architecture, DeCAL leverages specialized experts for each capability while enabling efficient information flow among them. To effectively leverage tactile information, we introduce Adaptive Visuo-Tactile Fusion that dynamically regulates tactile interactions via a contact-aware gating strategy. Furthermore, we propose Visuo-Tactile Latent Co-Imagination to jointly model visual and tactile dynamics, equipping the policy with implicit physical world knowledge. Experimental results show that DeCAL consistently achieves state-of-the-art performance across all tasks, attaining a 71% average success rate and an 83.4% progress success rate, while also demonstrating strong generalization to unseen scenarios.

Keywords: Visuo-Tactile Learning, VLA, Dexterous Manipulation

1 Introduction

Dexterous manipulation plays a fundamental role in human daily life [1, 2], enabling fine-grained and contact-rich interactions with the physical world. Motivated by recent advances in vision-language models [3, 4, 5, 6], vision-language-action (VLA) models have achieved remarkable progress in robotic manipulation [7, 8, 9, 10], exhibiting strong capabilities in visual perception and semantic understanding. Building upon large-scale egocentric human demonstrations [11, 12, 13], recent works [14, 15, 16, 17] have extended VLA models to dexterous manipulation and achieved promising results on multi-fingered manipulation tasks. However, these approaches still struggle with contact-rich and fine-grained interactions due to severe visual occlusions, complex contact dynamics, and limited physical observability from visual inputs alone.

Tactile feedback [18, 19, 20, 21, 22] provides direct access to physical interaction signals, including contact state, force variation, slip, and object deformation, which is essential for dexterous manipulation [23, 24, 25, 26, 27]. Some studies [28, 29, 30] have incorporated tactile sensing into robotic manipulation systems to improve contact awareness and manipulation robustness [31, 32, 33]. However, most existing approaches primarily treat tactile inputs as auxiliary sensory signals and perform homogeneous multimodal fusion, lacking adaptive tactile interaction modeling. Moreover, existing methods are largely reactive and fail to explicitly model future interactions, limiting their ability to reason about evolving contact transitions during manipulation.

In this work, we investigate these problems by introducing DeCAL, a physically-grounded dexterous vision-language-action model that enables effective utilization of tactile signals for contact-rich manipulation. First, DeCAL is built upon a Mixture-of-Transformers architecture composed of three collaborative experts for understanding, imagination, and action generation. The three experts perform directional knowledge sharing through joint attention, providing implicit guidance for coherent and fine-grained dexterous manipulation. Second, we introduce an adaptive visuo-tactile fusion mechanism that dynamically injects tactile information via a contact-aware gating strategy, allowing the model to selectively emphasize tactile cues during physical interactions. Finally, we propose visuo-tactile latent co-imagination, which jointly models visuo-tactile dynamics through future-oriented latent imagination, equipping the policy with implicit world knowledge.

To comprehensively evaluate DeCAL, we conduct extensive real-world experiments on diverse contact-rich dexterous manipulation tasks. The comparative results demonstrate that DeCAL consistently achieves state-of-the-art performance, outperforming the strongest baseline by more than 15% average success rate. Meanwhile, DeCAL maintains efficient real-time performance, achieving an average inference latency of 0.27s per action chunk. Furthermore, comprehensive ablation studies verify the effectiveness of each proposed component, highlighting the promise of effectively leveraging tactile information for dexterous manipulation. In summary, our contributions are as follows:

- We present DeCAL, a dexterous Vision-Tactile-Language-Action framework that unifies understanding, generation, and action for effective visuo-tactile representation learning.
- We introduce Adaptive Visuo-Tactile Fusion and Visuo-Tactile Latent Co-Imagination to facilitate contact-rich and fine-grained dexterous manipulation.
- We demonstrate the effectiveness and generalization of our method through a range of real-world experiments.

2 Related Work

2.1 Vision-Language-Action Model

Vision-Language-Action (VLA) models have emerged as a powerful paradigm for robotic control [7, 34, 9, 17, 15, 35], leveraging the broad semantic and world knowledge embedded in large-scale pre-trained foundation models [4, 5, 6]. By fine-tuning these models on diverse robotic datasets [36, 37, 38, 39], researchers have demonstrated successful cross-domain capability transfer from general

intelligence to specific manipulation tasks. Representative examples include $\pi_{0.5}$ [7], and GR00T N1.6 [40]. More recently, a growing line of work, including InternVLA-A1 [41], MoTus [42], and LAST_0 [43], has advanced VLA research by adopting Mixture-of-Transformer (MoT) architectures. These approaches seamlessly integrate perception (understanding), generation, and action within a unified framework, achieving strong performance across a variety of complex manipulation tasks. Despite these advances, existing VLA models remain predominantly vision-centric and lack explicit modeling of physical contact dynamics, limiting their performance in tactile-intensive scenarios involving severe visual occlusions and fine-grained physical interactions. To address this, we integrate tactile sensing into VLA world modeling by jointly learning visual evolution and contact dynamics, enabling more fine-grained and physically grounded manipulation.

2.2 Tactile for Dexterous Manipulation

Dexterous manipulation poses a fundamental challenge for vision-only policies, as the intricate physical structure of multi-fingered hands often leads to severe self-occlusion and ambiguous contact observations [2, 44]. To mitigate this, recent studies have explored the integration of haptic feedback to maintain operational continuity under visual impairment [45, 46, 47, 48, 23, 24, 49]. For instance, some studies [30, 50, 51, 52] learn rich tactile representations through self-supervised objectives, capturing contact geometry and force-related cues to facilitate efficient downstream policy learning. Another line of work [53, 28, 29] predicts future tactile signals to guide action generation, showing the potential of tactile foresight for improving visuo-tactile coordination and contact-aware control. Despite these advances, how to effectively and adaptively couple visual and tactile signals still remains underexplored [46, 32]. Distinct from these approaches, our work integrates tactile sensing into VLA policy learning, leveraging the model's multimodal understanding of language, vision, and physical interactions [54]. By introducing a unified dynamics modeling framework and an adaptive learning mechanism, our model enables more robust visuo-tactile fusion, allowing the agent to dynamically prioritize visual and tactile cues according to task demands and environmental constraints.

3 Robot System Setup

Our system consists of a pair of 6-DoF UR5 robotic arms and two 22-DoF SharpaWave five-fingered dexterous hands. Visual observations are captured from two wrist-mounted cameras and one egocentric camera, all using Intel RealSense D435. Each fingertip is equipped with a high-resolution (320×240) vision-based tactile sensor developed by Sharpa, enabling fine-grained perception of contact dynamics during manipulation. Specifically, a built-in camera inside each sensor captures the deformation of the elastic surface during physical interaction. Based on the raw tactile images and visuo-tactile processing algorithms, we represent tactile signals in three complementary forms:

- **Raw Image (R).** Raw tactile observations are directly captured by the built-in camera inside each visuo-tactile sensor, recording the contact patterns and elastic surface deformations.
- **Net Force (F).** Each fingertip outputs a 6-DoF force representation consisting of 3-axis forces and torques. The force signals are estimated from raw tactile observations using a pretrained regression model, enabling force-aware tactile perception.
- **Deform Map (M).** Each fingertip outputs a deformation depth map represented as a 2D image through a translation model [55], where each pixel indicates the local deformation depth on the tactile surface, providing fine-grained geometric contact information.

4 Method

4.1 Preliminaries

The goal of our robot policy is to predict an action chunk from multimodal observations. Most existing VLA models lack explicit modeling of future interaction dynamics, limiting their ability to

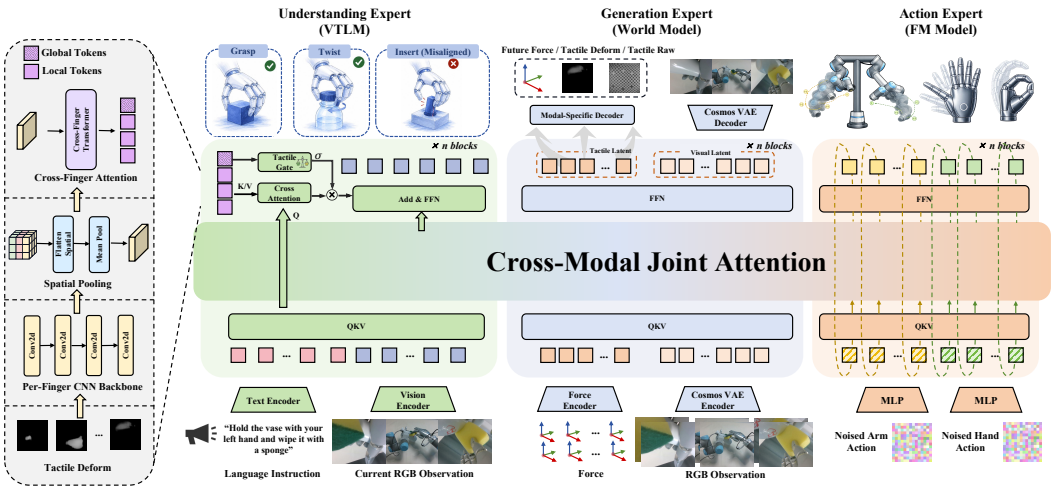

Figure 2: DeCAL is built upon a MoT architecture that unifies scene understanding, visuo-tactile dynamics foresight, and action generation. The Action Expert employs Factorized Flow Matching to decouple arm and hand motion, enabling better coordination and dexterous manipulation.

reason about fine-grained physical interactions. To address this, we propose DeCAL π_θ, a unified VLA framework that jointly performs understanding, imagination, and action generation. Given the current observation $o_t = (I_t, H_t, s_t)$, and the language instruction l, where I_t denotes the visual observations, H_t denotes the tactile observations, and s_t represents the robot state. DeCAL is trained to jointly maximize the likelihood of future visuo-tactile latent z_{t+H} and future actions $a_{t+1:t+H}$:

$$\max_{\theta} \mathbb{E}_{(o_t, l, a_{t+1:t+H}, z_{t+H}) \sim \mathcal{D}} \left[\log \pi_\theta(a_{t+1:t+H}, z_{t+H} \mid o_t, l) \right]. \tag{1}$$

4.2 Model Architecture

DeCAL adopts a Mixture-of-Transformers (MOT) architecture that seamlessly integrates scene understanding, visuo-tactile dynamics foresight, and action generation within a unified framework. As illustrated in Figure 2, the framework consists of three specialized transformer experts.

Understanding Expert. The understanding expert is built upon Qwen3-VL [3] for its strong multimodal understanding capability. Given language instructions and multi-view visual observations, the inputs are first encoded into text and visual tokens, which are then processed by transformer blocks to produce contextual embeddings shared with downstream experts through masked self-attention. We further extend the multimodal inputs to tactile observations through a dedicated cross-attention mechanism, where tactile features serve as keys and values, and visual-language features act as queries. The resulting representations are then residually fused with the self-attention features, enabling joint reasoning over visual, linguistic, and tactile observations within a unified latent space.

Generation Expert. The Generation Expert explicitly models visuo-tactile dynamics by predicting future visual and tactile observations from historical interactions, forming physically grounded multimodal world representations to guide downstream action generation. Despite the remarkable progress of recent video generation models in visual prediction, directly applying generative modeling to contact-rich manipulation remains challenging due to the strict real-time requirements of high-frequency robot control. Following Cai et al. [41], we adopt a parallel decoding strategy for future visuo-tactile generation. The proposed non-autoregressive paradigm significantly improves computational efficiency while remaining effective for modeling future visuo-tactile interactions.

Action Expert. Conditioned on the latent features from the understanding expert and generation expert, together with the robot proprioceptive states q_t, the action expert predicts an action chunk $a_{t:t+H}$ using a flow matching objective. Since arm and hand motions exhibit substantially distinct dynamics and control granularity, jointly modeling them through a shared denoising process may obscure fine-grained dexterous behaviors and weaken arm-hand coordination in high-DoF settings.

To address this, we propose **Factorized Flow Matching**, which explicitly decouples arm and hand motion generation. Specifically, arm and hand actions are initialized from independent noise distributions and projected into separate token sequences through MLP layers. The tokens are jointly processed by transformer blocks for coordinated motion generation before being decoded into clean action trajectories. Both the generation expert and action expert adopt Qwen3 [56] as their backbone.

Cross-Modal Joint Attention. We implement a blockwise attention mask to control the information flow across experts, as shown in Figure 3. Information is propagated unidirectionally from the understanding expert to the generation expert, and then to the action expert. Tokens in later experts can attend to tokens from preceding experts, while the reverse attention is prohibited. Within both the understanding expert and generation expert, tokens are bidirectionally attended for more effective multimodal reasoning and representation learning. In the action expert, state tokens can attend to themselves and all earlier blocks, while action tokens can attend to themselves, state tokens, and all preceding blocks. This asymmetric attention design enables action generation to fully condition on semantic understanding, imagined future interactions, and robot states.

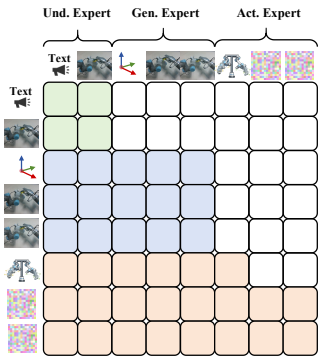

Figure 3: **Attention mask.**

4.3 Adaptive Visuo-Tactile Fusion

Tactile Feature Extraction. We use multi-finger deform maps as tactile inputs due to their rich geometric contact information. A shared ResNet-based [57] tactile encoder extracts tactile features from all fingertips. The extracted features are spatially pooled and projected into tactile tokens, which are subsequently fused through transformer layers to capture cross-finger interactions and global contact patterns. The tactile encoder finally produces two complementary representations: local tactile tokens that preserve fine-grained contact details for visuo-tactile interaction, and a global tactile token that summarizes the overall contact state for contact-aware tactile gating.

Adaptive Visuo-Tactile Cross-Attention. To incorporate tactile feedback into visual-language embeddings, we introduce a cross-attention mechanism at each transformer block. The local tactile tokens serve as keys and values, while the prefix projection of the transformer input provides the queries. The resulting fused output is then added to the embeddings from the block's joint self-attention via a residual connection, allowing the model to integrate tactile cues while preserving the original multimodal context.

Contact-Aware Gating Mechanism. In dexterous manipulation, physical contact is often intermittent and only arises during specific phases of a task, rather than being continuously present. Directly injecting tactile signals into the policy at all times can introduce biases in action generation and perturb the distribution of visual attention. To address this, we introduce a Contact-Aware Gating Mechanism that allows the policy to adaptively modulate the weights of tactile inputs based on the relevance of contact information at each timestep. Formally, the gated output is computed as:

$$\begin{aligned} \tilde{X} &= X + \sigma \odot \mathrm{CrossAttn}(Q_{vl}, K_{local}, V_{local}), \\ \sigma &= \mathrm{Gate}(z_{\mathrm{global}}), \end{aligned} \tag{2}$$

where X represents the original input embeddings to the transformer layer, Q_{vl} is the query from visual-language embeddings, while K_{local} and V_{local} correspond to the local tactile tokens serving as keys and values. The gating weight σ is derived from the global tactile token z_{global} through a lightweight projection function $\mathrm{Gate}(\cdot)$. This mechanism allows the model to selectively incorporate tactile feedback based on the overall contact state across all fingertips.

4.4 Visuo-Tactile Latent Co-Imagination

The generation expert performs future-oriented latent imagination to jointly model visuo-tactile dynamics. Given historical and current visuo-tactile observations, the model produces predictive latent representations that capture the underlying physical dynamics of the interaction.

Tactile Latent Prediction. 6-DoF contact force provides a compact and physically meaningful representation of contact dynamics, capturing both force magnitude and torque information. For tactile latent prediction, the forces from each fingertip are first mapped into embeddings using a dedicated force encoder, which are then processed by the policy to produce a tactile latent. The resulting latent can subsequently be decoded into representations of different granularities, including raw images, deform maps, and 6-DoF force vectors. These representations capture tactile dynamics from multiple perspectives, supporting physically grounded scene understanding and action generation.

Visual Latent Prediction. Inspired by recent world action models [58, 59, 60], we equip DeCAL with world knowledge by generating future visual latents from historical and current observations. This forward-looking representation provides temporally informed context that enhances planning and decision-making in contact-rich manipulation tasks. We encode the multi-view visual images using the Cosmos VAE tokenizer [61], which provides expressive visual latents while preserving fine-grained details. To efficiently handle long visual sequences, we further apply a token compression mechanism following Cai et al. [41]. Specifically, the 32×32 latent feature grid is downsampled to 4×4 using a convolutional layer with an 8×8 kernel, reducing sequence length while retaining essential spatial information. The compressed tokens are then decoded in parallel to generate future visual frames, leveraging the KV cache from the understanding expert.

During training, the predicted visual latents are directly regressed toward the Cosmos VAE targets encoded from future visual observations. While tactile latents, conditioned on current 6-DoF forces, are indirectly supervised by reconstructing future raw tactile images, deform maps, and 6-DoF forces. At inference time, the generation expert jointly predicts future visual and tactile latents to provide forward-looking context for action generation, while the reconstruction decoders are omitted, avoiding unnecessary pixel-space reconstruction.

5 Experiment

In this section, we evaluate DeCAL through extensive experiments addressing two key questions: (1) How does DeCAL perform against state-of-the-art policies and generalize to OOD scenarios (Section 5.2, 5.3, 5.5)? (2) How does each component contribute to the overall performance (Section 5.4)?

5.1 Experiment Setup

Tasks. We evaluate DeCAL on six contact-rich and fine-grained dexterous manipulation tasks, as illustrated in Figure 4: (1) *Wipe Vase*, (2) *Erase Whiteboard*, (3) *Assemble Parts*, (4) *Twist Cap*, (5) *Pipetting*, (6) *Screw Light Bulb*. Each task is collected with 100 high-quality demonstrations and evaluated with 20 trials by default. We collect expert demonstrations through a teleoperation system with MetaGlove Pro and VIVE Trackers, following Fu et al. [17], where the gloves retarget human hand motion to dexterous hands and the trackers control the robotic arm end-effectors. More details about the teleoperation setup and task specifications are provided in the Appendix.

Baselines and Evaluation Metrics. We compare DeCAL with two state-of-the-art VLA models, GR00T N1.6 [40] and InternVLA-A1 [41], as well as two tactile-based specialist policies, ViTacFormer [28] and DECO [32], which are built upon different architectures. To enable a fairer comparison in terms of tactile modality usage, we further reproduce InternVLA-A1^t by naively incorporating tactile deform maps and net-force signals into the VLM backbone. We use two metrics to evaluate model performance: Success Rate (SR), indicating the entire task is successfully completed, and Progress Rate (PSR), capturing the average completion ratio across all task stages.

5.2 Results and Analysis

Results on Real-World Experiments. As shown in Table 1, DeCAL achieves the highest success rate across all tasks, outperforming all baselines, while also attaining the best PSR on the majority of tasks. Vision-based policies perform reasonably well on simpler tasks, such as Wipe Vase and Pipetting. However, they struggle with contact-rich dexterous manipulation requiring precise physical interaction

Table 1: **Main results of six real-world tasks.** Each experiment is evaluated with 20 trials.

Method	Wipe Vase		Erase Whiteboard		Assemble Parts		Twist Cap		Pipetting		Screw Light Bulb	
	SR	PSR	SR	PSR	SR	PSR	SR	PSR	SR	PSR	SR	PSR
GR00T N1.6	30.0%	78.8%	50.0%	55.0%	30.0%	60.0%	10.0%	46.7%	**60.0%**	85.0%	10.0%	45.0%
InternVLA-A1	65.0%	85.0%	50.0%	58.3%	15.0%	43.3%	5.0%	36.7%	35.0%	78.8%	30.0%	**53.8%**
ViTacFormer	80.0%	88.8%	45.0%	50.0%	15.0%	71.7%	65.0%	81.7%	55.0%	**86.3%**	25.0%	46.3%
DECO	90.0%	95.0%	60.0%	70.0%	35.0%	61.7%	70.0%	83.3%	45.0%	71.3%	35.0%	41.3%
InternVLA-A1^t	75.0%	87.5%	45.0%	51.7%	45.0%	65.0%	25.0%	55.0%	20.0%	58.8%	25.0%	31.3%
DeCAL (Ours)	**100.0%**	**100.0%**	**80.0%**	**86.7%**	**65.0%**	**85.0%**	**80.0%**	**93.3%**	**60.0%**	**86.3%**	**40.0%**	48.8%

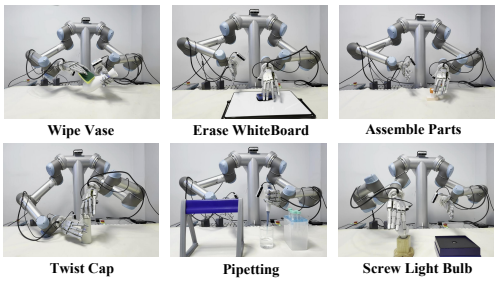

Figure 4: **Visualization of dexterous tasks.**

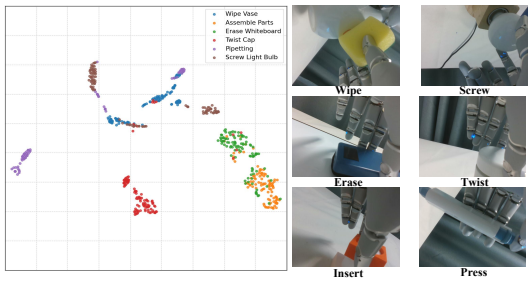

Figure 5: **Visualize t-SNE of global tactile tokens.**

and fine-grained contact understanding. Policies equipped with tactile sensing can incorporate richer environmental feedback, improving physical interaction awareness. But naively injecting tactile signals may also disturb the distribution of visual representations, leading to reduced accuracy in visual grounding and grasping. This issue becomes particularly evident in the Assemble Parts task, where effective coordination between vision and tactile is crucial. Benefiting from the adaptive visuo-tactile fusion mechanism and the joint modeling of visuo-tactile dynamics, DeCAL enhances physical interaction awareness while preserving robust visual grounding capabilities, thereby leading to improved performance on fine-grained dexterous manipulation tasks.

Tactile Latent Analysis. We analyze the global tokens extracted by the tactile encoder using t-SNE [62]. For each task, we sample frames from different contact phases, which correspond to diverse interaction patterns such as twisting, insertion, and wiping. As illustrated in Figure 5, the learned global tactile representations exhibit clear clustering behavior under different contact modes, indicating that the encoder is able to capture meaningful and structured physical interaction patterns.

Adaptive Tactile Gate. Next, we evaluate the effectiveness of the adaptive tactile gating mechanism. Taking the Assemble Parts and Twist Cap tasks as examples, Figure 6 visualizes the tactile gate outputs across different frames within a testing episode. When the hand is not interacting with the environment, the tactile gate values remain consistently low, indicating that the policy is primarily dominated by visual signals. In contrast, once meaningful physical contact occurs between the fingers and the objects, the gate values increase significantly, allowing tactile information to play a more prominent role in the policy and influence action generation.

Tactile Force Prediction. We further compare the predicted tactile forces with the ground-truth values in the validation data. As shown in Figure 7, the predicted 6D forces highly align with the ground-truth variations across different contact stages. This demonstrates that the proposed visuo-tactile latent co-imagination can effectively model underlying contact dynamics from multimodal

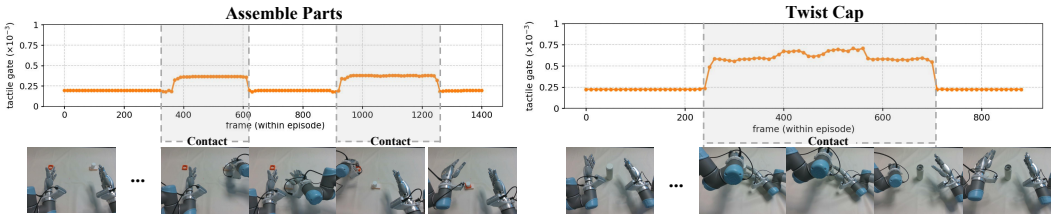

Figure 6: **Tactie gate.** The tactile gate value σ increases during contact phases.

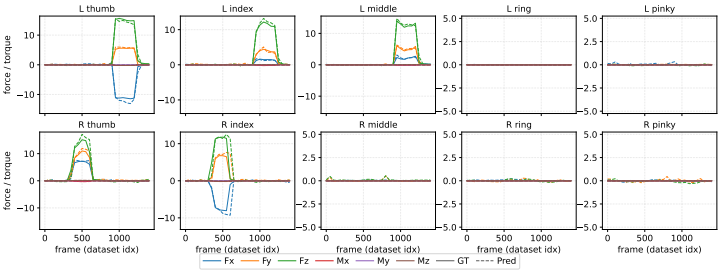

Figure 7: **Force prediction.**

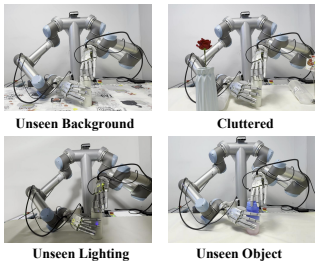

Figure 8: **OOD scenarios.**

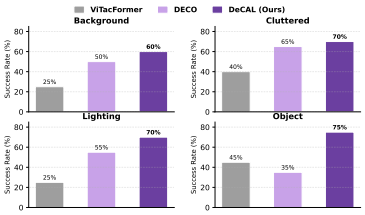

Figure 9: **Generalization results.**

Table 2: **Ablation study of each component.**

Factorized FM	Tac Gating	Vis Gen	Tac Gen	Assemble Parts	Twist Cap
✗	✓	✓	✓	25.0%	35.0%
✓	✓	✗	✗	20.0%	30.0%
✓	✓	✗	✓	35.0%	45.0%
✓	✓	✓	✗	55.0%	60.0%
✓	✗	✓	✓	50.0%	70.0%
✓	✓	✓	✓	65.0%	80.0%

observations. Moreover, the learned dynamics are further propagated to the action expert through joint attention, enabling more stable and precise action generation for contact-rich dexterous manipulation.

5.3 Generalization

Besides the remarkable effectiveness, DeCAL also demonstrates promising generalization capabilities under four *out-of-distribution* scenarios: (1) **Unseen Background**, where a tablecloth perturbs the visual distribution. (2) **Cluttered Environment**, where random objects introduce visual distractions. (3) **Unseen Lighting**, where we use dimmer lighting conditions to simulate different illumination settings. (4) **Unseen Object**, where a novel cup with different shape, diameter, and height is introduced for manipulation, as shown in Figure 8. Taking Twist Cap as an example, we report the success rate of DeCAL, DECO, and ViTacFormer under four OOD scenarios in Figure 9. The results demonstrate that DeCAL generalizes effectively across various distribution shifts, achieving a 75.0% success rate under the unseen object setting and substantially outperforming all baselines.

5.4 Ablation Studies

In this section, we analyze the contribution of each component in DeCAL to the overall policy performance. We conduct ablation studies focusing on four aspects: (1) Factorized Flow Matching; (2) Tactile Gating Mechanism; (3) Visual Latent Generation; (4) Tactile Latent Generation. As shown in Table 2, Factorized Flow Matching contributes significantly to the overall task performance. Without this mechanism, we observe noticeable inconsistency and poor coordination between the arms and dexterous hands during manipulation. In addition, visual and tactile latent generation effectively models motion and contact dynamics, leading to more accurate action generation. Finally, the adaptive tactile gating mechanism enables the policy to dynamically adjust the importance of visual and tactile cues across different interaction stages, further improving the success rate on contact-rich manipulation tasks. More quantitative and qualitative results are provided in the Appendix.

Table 3: **Visual Generation Results.**

Method	Assemble Parts		Twist Cap	
	Cos ↑	LPIPS ↓	Cos ↑	LPIPS ↓
InternVLA-A1	0.913	0.243	0.908	0.257
DeCAL (Ours)	**0.946**	**0.222**	**0.927**	**0.245**

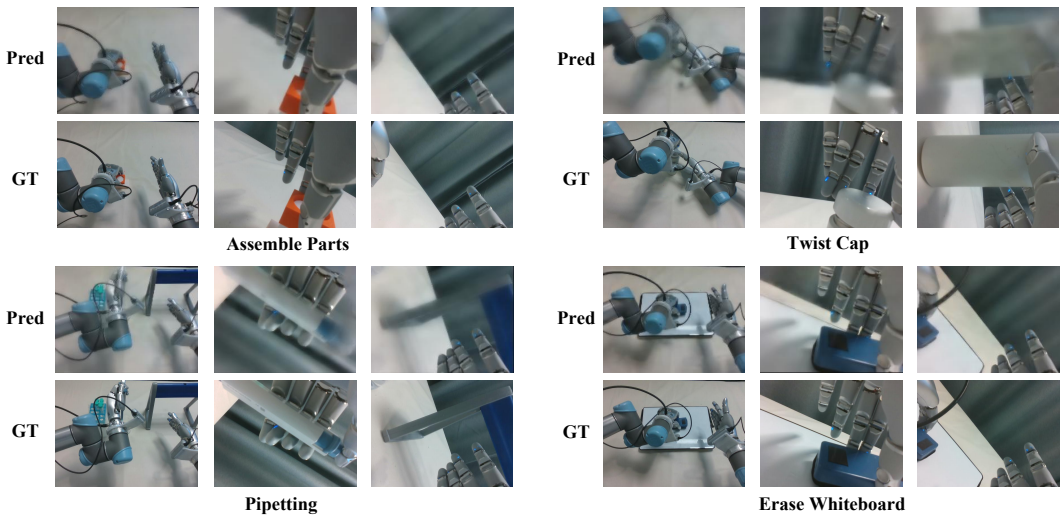

Figure 10: **Qualitative Results of Future Visual Generation.**

5.5 Generation Results

We evaluate future visual generation by comparing DeCAL with InternVLA-A1 [41], as both adopt a MoT architecture for future-frame prediction. Experiments are conducted on Assemble Parts and Twist Cap tasks, where models are trained on the training set and evaluated on the validation set by comparing the predicted frames against the ground-truth future observations. We assess generation quality using Cosmos feature similarity (Cos ↑) and LPIPS (↓), which measure semantic consistency and perceptual similarity, respectively. As shown in Table 3, DeCAL consistently achieves better visual generation quality than InternVLA-A1. By incorporating tactile feedback, the model gains additional information about physical interactions and contact dynamics, enabling a more comprehensive understanding of environmental changes. Furthermore, the joint modeling of visuo-tactile dynamics helps the model better capture the evolution of future states under physical contact, resulting in more accurate and realistic future-frame predictions. More qualitative examples are presented in Figure 10 and Appendix E.1.

6 Conclusions

In this paper, we present DeCAL, a physically-grounded dexterous vision-language-action model that effectively leverages tactile feedback for contact-rich manipulation. By unifying understanding, generation, and action within a collaborative framework, DeCAL enables coherent multimodal reasoning and fine-grained dexterous control. Furthermore, DeCAL adaptively fuses visuo-tactile information through contact-aware tactile gating and models visuo-tactile dynamics via future-oriented latent imagination, endowing the policy with implicit world knowledge. Extensive experiments on real-world dexterous tasks demonstrate the effectiveness and generalization of our approach.

7 Limitations

First, our method relies on accurate tactile perception, which can be affected by sensor noise, calibration errors, and model drift, potentially leading to performance degradation during long-term and high-load operation. Second, the teleoperation system does not provide fingertip force feedback to the operator, which may result in delayed contact adjustment, suboptimal contact regulation and limit the quality of tactile-aware demonstrations. Moreover, our current framework does not leverage large-scale visuo-tactile pretraining. We believe that scaling up diverse tactile dexterous data for pretraining could further improve the model's capability, which we consider an important direction for future research.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62476011), the Beijing Natural Science Foundation (L252060). We would like to express our sincere gratitude to Yifan Ye and Yunfan Lou for their insightful discussions and valuable feedback on the methodological design. We are also grateful to Xiansheng Chen for his support with the hardware setup.

References

[1] Y. Chen, T. Wu, S. Wang, X. Feng, J. Jiang, Z. Lu, S. McAleer, H. Dong, S.-C. Zhu, and Y. Yang. Towards human-level bimanual dexterous manipulation with reinforcement learning. *Advances in Neural Information Processing Systems*, 35:5150–5163, 2022.

[2] Y. Fu, Q. Feng, N. Chen, Z. Zhou, M. Liu, M. Wu, T. Chen, S. Rong, J. Liu, H. Dong, et al. Cordvip: Correspondence-based visuomotor policy for dexterous manipulation in real-world. *arXiv preprint arXiv:2502.08449*, 2025.

[3] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[4] B. R. Team, M. Cao, H. Tan, Y. Ji, X. Chen, M. Lin, Z. Li, Z. Cao, P. Wang, E. Zhou, et al. Robobrain 2.0 technical report. *arXiv preprint arXiv:2507.02029*, 2025.

[5] G. R. Team, S. Abeyruwan, J. Ainslie, J.-B. Alayrac, M. G. Arenas, T. Armstrong, A. Balakrishna, R. Baruch, M. Bauza, M. Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

[6] J. Yang, R. Tan, Q. Wu, R. Zheng, B. Peng, Y. Liang, Y. Gu, M. Cai, S. Ye, J. Jang, et al. Magma: A foundation model for multimodal ai agents. In *Proceedings of the computer vision and pattern recognition conference*, pages 14203–14214, 2025.

[7] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[8] M. J. Kim, C. Finn, and P. Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025.

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

[10] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

[11] R. Hoque, P. Huang, D. J. Yoon, M. Sivapurapu, and J. Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025.

[12] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024.

[13] R. Punamiya, S. Kareer, Z. Liu, J. Citron, R.-Z. Qiu, X. Cai, A. Gavryushin, J. Chen, D. Liconti, L. Y. Zhu, et al. Egoverse: An egocentric human dataset for robot learning from around the world. *arXiv preprint arXiv:2604.07607*, 2026.

[14] R. Yang, Q. Yu, Y. Wu, R. Yan, B. Li, A.-C. Cheng, X. Zou, Y. Fang, X. Cheng, R.-Z. Qiu, et al. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025.

[15] H. Luo, Y. Wang, W. Zhang, S. Zheng, Z. Xi, C. Xu, H. Xu, H. Yuan, C. Zhang, Y. Wang, et al. Being-h0. 5: Scaling human-centric robot learning for cross-embodiment generalization. *arXiv preprint arXiv:2601.12993*, 2026.

[16] X. Cai, R.-Z. Qiu, G. Chen, L. Wei, I. Liu, T. Huang, X. Cheng, and X. Wang. In-n-on: Scaling egocentric manipulation with in-the-wild and on-task data. *arXiv preprint arXiv:2511.15704*, 2025.

[17] Y. Fu, N. Chen, J. Zhao, S. Shan, G. Yao, P. Wang, Z. Wang, and S. Zhang. Metis: Multi-source egocentric training for integrated dexterous vision-language-action model. *arXiv preprint arXiv:2511.17366*, 2025.

[18] R. Patel, R. Ouyang, B. Romero, and E. Adelson. Digger finger: Gelsight tactile sensor for object identification inside granular media. In *International Symposium on Experimental Robotics*, pages 105–115. Springer, 2020.

[19] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020.

[20] T. P. Tomo, A. Schmitz, W. K. Wong, H. Kristanto, S. Somlor, J. Hwang, L. Jamone, and S. Sugano. Covering a robot fingertip with uskin: A soft electronic skin with distributed 3-axis force sensitive elements for robot hands. *IEEE Robotics and Automation Letters*, 3(1):124–131, 2017.

[21] L. Zhang, Y. Wang, and Y. Jiang. Tac3d: A novel vision-based tactile sensor for measuring forces distribution and estimating friction coefficient distribution. *arXiv preprint arXiv:2202.06211*, 2022.

[22] Y. Ye, Y. Fu, Y. Lv, B. Hou, J. Cen, L. Kong, D. Zheng, T. Chen, J. Liu, Z. Cao, et al. Data pyramid for embodied manipulation. *arXiv preprint arXiv:2607.24744*, 2026.

[23] I. Guzey, Y. Dai, B. Evans, S. Chintala, and L. Pinto. See to touch: Learning tactile dexterity through visual incentives. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13825–13832. IEEE, 2024.

[24] Y. Yuan, H. Che, Y. Qin, B. Huang, Z.-H. Yin, K.-W. Lee, Y. Wu, S.-C. Lim, and X. Wang. Robot synesthesia: In-hand manipulation with visuotactile sensing. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6558–6565. IEEE, 2024.

[25] T. Lin, Y. Zhang, Q. Li, H. Qi, B. Yi, S. Levine, and J. Malik. Learning visuotactile skills with two multifingered hands. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5637–5643. IEEE, 2025.

[26] Q. Liu, Q. Ye, Z. Sun, Y. Cui, G. Li, and J. Chen. Masked visual-tactile pre-training for robot manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13859–13875. IEEE, 2024.

[27] V. Dave, F. Lygerakis, and E. Rueckert. Multimodal visual-tactile representation learning through self-supervised contrastive pre-training. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8013–8020. IEEE, 2024.

[28] L. Heng, H. Geng, K. Zhang, P. Abbeel, and J. Malik. Vitacformer: Learning cross-modal representation for visuo-tactile dexterous manipulation. *arXiv preprint arXiv:2506.15953*, 2025.

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

[30] T. Wu, J. Li, J. Zhang, M. Wu, and H. Dong. Canonical representation and force-based pretraining of 3d tactile for dexterous visuo-tactile policy learning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6786–6792. IEEE, 2025.

[31] J. Huang, Y. Ye, Y. Gong, X. Zhu, Y. Gao, and K. Zhang. Spatially anchored tactile awareness for robust dexterous manipulation. *arXiv preprint arXiv:2510.14647*, 2025.

[32] X. Li, Y. Sun, L. Zhang, B. Huang, Y. Peng, Y. Meng, H. Jiang, S. Xie, G. Yao, A. Knoll, et al. Deco: Decoupled multimodal diffusion transformer for bimanual dexterous manipulation with a plugin tactile adapter. *arXiv preprint arXiv:2602.05513*, 2026.

[33] Z. He, H. Fang, J. Chen, H.-S. Fang, and C. Lu. Foar: Force-aware reactive policy for contact-rich robotic manipulation. *IEEE Robotics and Automation Letters*, 2025.

[34] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[35] Y. Ye, J. Cen, J. Chen, and Z. Lu. Self-evolved imitation learning in simulated world. *IEEE Robotics and Automation Letters*, 2026.

[36] O. X.-E. Collaboration, A. O'Neill, A. Rehman, A. Gupta, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, et al. Open X-Embodiment: Robotic learning datasets and RT-X models. https://arxiv.org/abs/2310.08864, 2023.

[37] K. Wu, C. Hou, J. Liu, Z. Che, X. Ju, Z. Yang, M. Li, Y. Zhao, Z. Xu, G. Yang, et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. *arXiv preprint arXiv:2412.13877*, 2024.

[38] S. Wu, X. Liu, S. Xie, P. Wang, X. Li, B. Yang, Z. Li, K. Zhu, H. Wu, Y. Liu, et al. Robocoin: An open-sourced bimanual robotic data collection for integrated manipulation. *arXiv preprint arXiv:2511.17441*, 2025.

[39] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[40] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

[41] J. Cai, Z. Cai, J. Cao, Y. Chen, Z. He, L. Jiang, H. Li, H. Li, Y. Li, Y. Liu, et al. Internvla-a1: Unifying understanding, generation and action for robotic manipulation. *arXiv preprint arXiv:2601.02456*, 2026.

[42] H. Bi, H. Tan, S. Xie, Z. Wang, S. Huang, H. Liu, R. Zhao, Y. Feng, C. Xiang, Y. Rong, et al. Motus: A unified latent action world model. *arXiv preprint arXiv:2512.13030*, 2025.

[43] Z. Liu, J. Liu, H. Chen, J. Yu, Z. Guo, C. Hou, C. Gu, X. Mi, R. Zhang, K. Wu, et al. Last_0: Latent spatio-temporal chain-of-thought for robotic vision-language-action model. *arXiv preprint arXiv:2601.05248*, 2026.

[44] Z. Zhang, W. Wang, H. Sun, Q. Ding, X. Chu, G. Fang, and K. Au. Fingervip: Learning real-world dexterous manipulation with fingertip visual perception. *arXiv preprint arXiv:2604.21331*, 2026.

[45] H. Xue, J. Ren, W. Chen, G. Zhang, Y. Fang, G. Gu, H. Xu, and C. Lu. Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation. *arXiv preprint arXiv:2503.02881*, 2025.

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

[47] I. Guzey, B. Evans, S. Chintala, and L. Pinto. Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play. *arXiv preprint arXiv:2303.12076*, 2023.

[48] Z.-H. Yin, B. Huang, Y. Qin, Q. Chen, and X. Wang. Rotating without seeing: Towards in-hand dexterity through touch. *arXiv preprint arXiv:2303.10880*, 2023.

[49] C. Yuan, Z. Zhang, M. Zhou, W. Chen, Y. Wang, Z. Liu, D. Niu, S. Wang, H. Zhang, W. Zhang, et al. Ftp-1: A generalist foundation tactile policy across tactile sensors for contact-rich manipulation. *arXiv preprint arXiv:2606.13102*, 2026.

[50] S. Funabashi, T. Isobe, F. Hongyi, A. Hiramoto, A. Schmitz, S. Sugano, and T. Ogata. Multi-fingered in-hand manipulation with various object properties using graph convolutional networks and distributed tactile sensors. *IEEE Robotics and Automation Letters*, 7(2):2102–2109, 2022.

[51] Y. Zheng, S. Gu, W. Li, Y. Zheng, Y. Zang, S. Tian, X. Li, C. Hao, C. Gao, S. Liu, et al. Omnivta: Visuo-tactile world modeling for contact-rich robotic manipulation. *arXiv preprint arXiv:2603.19201*, 2026.

[52] F. Yang, C. Feng, Z. Chen, H. Park, D. Wang, Y. Dou, Z. Zeng, X. Chen, R. Gangopadhyay, A. 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.

[53] Y. Lou, Y. Ye, Y. Fu, J. Cen, X. Chi, Y. Lyu, P. Jia, S. Han, Z. Lu, and S. Zhang. Dreamtac: A unified tactile world action model for contact-rich robot manipulation. *arXiv preprint arXiv:2606.08737*, 2026.

[54] S. Bai, W. Song, J. Chen, Y. Ji, Z. Zhong, J. Yang, H. Zhao, W. Zhou, W. Zhao, Z. Li, et al. Towards a unified understanding of robot manipulation: A comprehensive survey. *arXiv preprint arXiv:2510.10903*, 2025.

[55] L. Su, Z. Peng, R. Ren, S. Mao, J. Du, K. Zhang, and X. Zhu. Tacmap: Bridging the tactile sim-to-real gap via geometry-consistent penetration depth map. *arXiv preprint arXiv:2602.21625*, 2026.

[56] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[57] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[58] M. J. Kim, Y. Gao, T.-Y. Lin, Y.-C. Lin, Y. Ge, G. Lam, P. Liang, S. Song, M.-Y. Liu, C. Finn, et al. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*, 2026.

[59] Q. Feng, J. Yu, J. Liu, Y. Jia, Z. Wu, H. Chen, Z. Qian, S. Gu, P. Jia, S. Ma, et al. Harmowam: Harmonizing generalizable and precise manipulation via adaptive world action models. *arXiv preprint arXiv:2605.10942*, 2026.

[60] T. Yuan, Z. Dong, Y. Liu, and H. Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026.

[61] N. Agarwal, A. Ali, M. Bala, Y. Balaji, E. Barker, T. Cai, P. Chattopadhyay, Y. Chen, Y. Cui, Y. Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[62] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[63] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[64] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788*, 2024.

Appendix

A Additional Method Detail

A.1 Training objective

DeCAL is trained with three complementary objectives: visual foresight generation, tactile foresight generation, and action prediction. For visual foresight, we supervise the predicted latent representations using the COSMOS latent features, encouraging the model to focus on task-relevant regions. For tactile foresight, the model predicts multiple forms of tactile representations (e.g., raw images, deformation maps, and 6-DoF force vectors), each supervised with a corresponding reconstruction loss to capture fine-grained and global contact dynamics. Action prediction is trained under the flow matching paradigm [63], which models the transport velocity between noisy and target action distributions. The final training objective is a weighted sum of the three losses: $\mathcal{L}_{\text{total}} = \lambda_v \mathcal{L}_{\text{visual}} + \lambda_t \mathcal{L}_{\text{tactile}} + \mathcal{L}_{\text{action}}$.

DeCAL is trained with three complementary objectives: visual foresight generation, tactile foresight generation, and action prediction. For visual foresight, the model predicts future visual latent representations $\hat{z_v}$ conditioned on historical visual frames. These predictions are supervised using the pretrained COSMOS latent features $\hat{z_v}$, with a reconstruction loss that encourages the model to attend to task-relevant regions:

$$\mathcal{L}_{visual} = \frac{1}{N_v} \sum_{i=1}^{N_v} ||\hat{z_v}^{(i)} - z_v^{(i)}||_2^2, \tag{3}$$

where N_v denotes the number of visual tokens.

For tactile foresight, the model predicts multiple complementary future tactile representations: raw tactile images $\hat{R}$, deformation maps $\hat{M}$, and 6-DoF net force/torque vectors $\hat{F}$. Each is supervised with a corresponding reconstruction loss to capture both fine-grained local contact dynamics and global tactile interactions:

$$L_{tactile} = \lambda_R \mathcal{L}_{img}(\hat{R}, R) + \lambda_M \mathcal{L}_{def}(\hat{M}, M) + \lambda_F \mathcal{L}_{force}(\hat{F}, F), \tag{4}$$

where λ_R, λ_M, λ_F balance the contributions of each tactile modality.

For action prediction, we adopt the flow matching paradigm [63], which learns a transport vector field v_θ that maps a noisy action to the target action. Formally, let s_t denote the proprioceptive state at time t, and let $\hat{a}^\tau_{t:t+k}$ be an interpolated noisy action chunk constructed as:

$$\hat{a}^\tau_{t:t+k} = (1-\tau)\epsilon + \tau a_{t:t+k}, \quad \epsilon \sim \mathcal{N}(0, I), \quad \tau \sim Beta(1.5, 1.0), \tag{5}$$

where $a_{t:t+k}$ is the expert action chunk, and τ progresses from 0 to 1 over K steps. Given the dataset of expert demonstrations $\mathcal{D}_2$, The model learns a velocity field $v_\theta(s_t, \hat{a}^\tau_{t:t+k}, h_{und}, h_{gen})$ that transports noisy actions toward the target, conditioned on contextual features from the understanding expert h_{und} and generation expert h_{gen}:

$$\mathcal{L}_{action} = \mathbb{E}_{\xi_2 \sim \mathcal{D}_2}[||v_\theta(q_t, \hat{a}^\tau_{t:t+k}, h_{und}, h_{gen}) - (a_{t:t+k} - \epsilon)||_2^2]. \tag{6}$$

The overall training objective is a weighted sum of the three losses:

$$\mathcal{L}_{total} = \lambda_v \mathcal{L}_{visual} + \lambda_t \mathcal{L}_{tactile} + \mathcal{L}_{action} \tag{7}$$

where λ_v and λ_t control the relative importance of visual, tactile objectives. This formulation encourages the model to jointly reason over multi-modal signals and generate physically grounded, contact-aware actions.

A.2 Inference Pipeline

During inference, DeCAL embeds multi-view visual observations, language instructions, and tactile inputs into prefix and middle tokens, which are processed by the understanding expert and generation expert. The prefix tokens produce contextual representations stored in KV caches, allowing subsequent computations to efficiently attend to all prior context. Subsequently, middle embeddings, including predicted visuo-tactile foresight, are integrated with these cached states to form latent representations that capture both task understanding and visuo-tactile dynamics. Finally, actions are generated iteratively via factorized flow matching, conditioned on the current state and the cached context. Running on a single NVIDIA RTX 4090 GPU, DeCAL achieves an average inference latency of 0.27 s per action chunk, demonstrating efficient inference for real-time dexterous manipulation.

B Policy Implementation Details

For training DeCAL, we initialize the policy from Internvla-A1-3B [41] pretrained weights, which have been extensively trained on large-scale vision-language-action data and exhibit strong generalization across diverse tasks. Training is conducted on 8 NVIDIA H100 GPUs, with a per-device batch size of 4. We optimize the model using AdamW with the learning rate of $5.0e^{-5}$, along with a linear warmup (2,000 steps) and decay schedule down to $5.0e^{-6}$ over 100,000 steps. During training, the generation expert receives a temporal observation history consisting of the previous 15 frames (approximately 0.5 seconds of context at 30 Hz), together with the current observation, enabling the model to capture short-term visuo-tactile dynamics for action prediction. The policy predicts action chunks of length 50 during both training and inference.

C Real-Robot System

As illustrated in Figure 11, our real-robot platform consists of a pair of 6-DoF UR5 robotic arms and two 22-DoF SharpaWave five-fingered dexterous hands. Visual observations are captured using two wrist-mounted cameras and one egocentric camera, all based on Intel RealSense D435 sensors, providing both local manipulation views and global scene perception. Each fingertip is further equipped with a high-resolution (320×240) vision-based tactile sensor developed by Sharpa, enabling fine-grained perception of contact geometry and interaction dynamics during dexterous manipulation. The entire system operates at 30 Hz for real-time visuo-tactile control.

Self-collected Robot Data. We collect robot demonstrations through human teleoperation with a glove-tracker system [64]. Three VIVE trackers are placed on the operator's chest and wrists to record spatial positions of these key points. The wrist poses are then computed relative to the chest coordinate frame through coordinate transformation, and subsequently mapped to the arm joint configurations via inverse kinematics (IK) solvers. Simultaneously, Manus Metagloves Pro are employed to retarget human hand motions onto the dexterous hand, enabling accurate transfer of finger and joint movements. During demonstration, we capture visual observations from an egocentric view as well from the left and right wrist cameras. Each fingertip provides rich tactile signals, including three modalities: raw tactile images, deformation maps, and 6-DoF net forces. The entire system operates at 30Hz, balancing motion fidelity with reliable multi-sensor synchronization.

D Task Details

Wipe Vase. In this task, the robot must remove graffiti marks from a vase through coordinated bimanual manipulation. The robot first grasps and lifts the vase with its left hand, then picks up a sponge with its right hand. It continuously adjusts the vase orientation while maintaining stable contact between the sponge and the curved surface to remove the graffiti. After cleaning, the robot places both the vase and the sponge back onto the table. Success is achieved if all visible graffiti marks are removed and both objects are returned to their original locations.

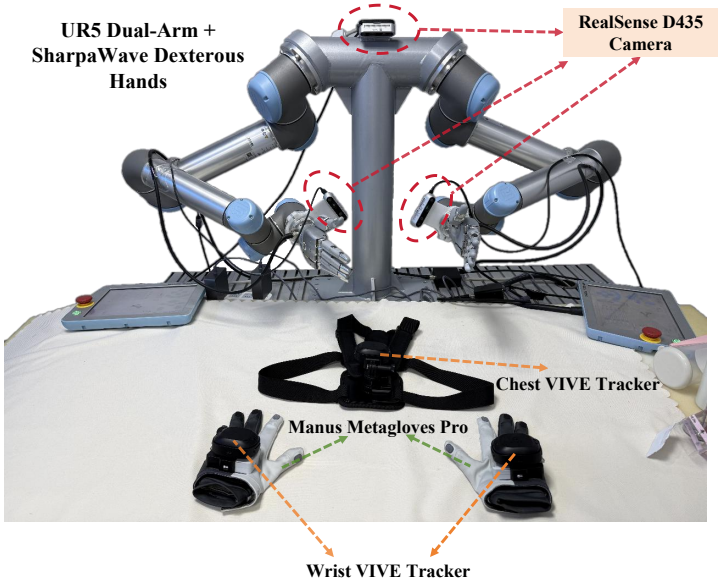

Figure 11: **Real Robot Platform.** Our setup consists of a dual UR5 robotic arm and a pair of SharpaWave five-finger dexterous hands. Visual observations are captured with one head-mounted and two wrist-mounted cameras, while high-resolution tactile data are collected from each fingertip. Expert demonstrations are acquired via a wearable glove-tracker teleoperation system.

Erase Whiteboard. The robot is required to erase handwritten markings from a whiteboard. It first grasps an eraser with its left hand and establishes contact with the whiteboard surface. The robot then performs a controlled top-to-bottom wiping motion to remove the markings while maintaining stable contact throughout the cleaning process. Success is achieved if all visible markings are removed from the whiteboard.

Assemble Parts. The robot is required to assemble a plug-and-socket pair through precise dexterous manipulation. It first grasps and positions the white socket with its right hand, then picks up the orange plug with its left hand and aligns it above the insertion point. The robot must continuously adjust the plug pose based on contact feedback while performing the insertion. This contact-rich task requires accurate alignment and continuous adaptation to contact feedback during insertion. Success is achieved if the plug is fully inserted into the socket.

Twist Cap. The robot is required to unscrew a bottle cap through precise dexterous manipulation. It first stabilizes the bottle with its right hand, then grasps the cap using the thumb, index finger, and middle finger of its left hand. The robot gradually rotates the cap while maintaining a stable grasp until it is fully unscrewed, and subsequently places the cap on the table. Success is achieved if the cap is completely removed from the bottle and placed on the table.

Pipetting. The robot is required to dispense liquid using a pipette and subsequently return it to a holder. It first grasps the pipette with its left hand and moves it above a target beaker. The robot then presses the plunger with its thumb to release the liquid. After dispensing, the pipette is transferred to the right hand and placed onto a designated pipette holder. This task requires accurate thumb-actuated control, stable object handover, and precise placement. Success is achieved if the liquid is dispensed and the pipette is correctly returned to the holder.

Screw Light Bulb. The robot first grasps a light bulb from the workbench with its left hand and places it into the socket. It then uses the thumb and index finger of its right hand to rotate the bulb clockwise until it is fully tightened. This task requires coordinated finger movements and continuous contact feedback to maintain stable grasping and accuratlely regulate the screwing motion. success is achieved if the bulb is fully tightened in the socket and illuminates.

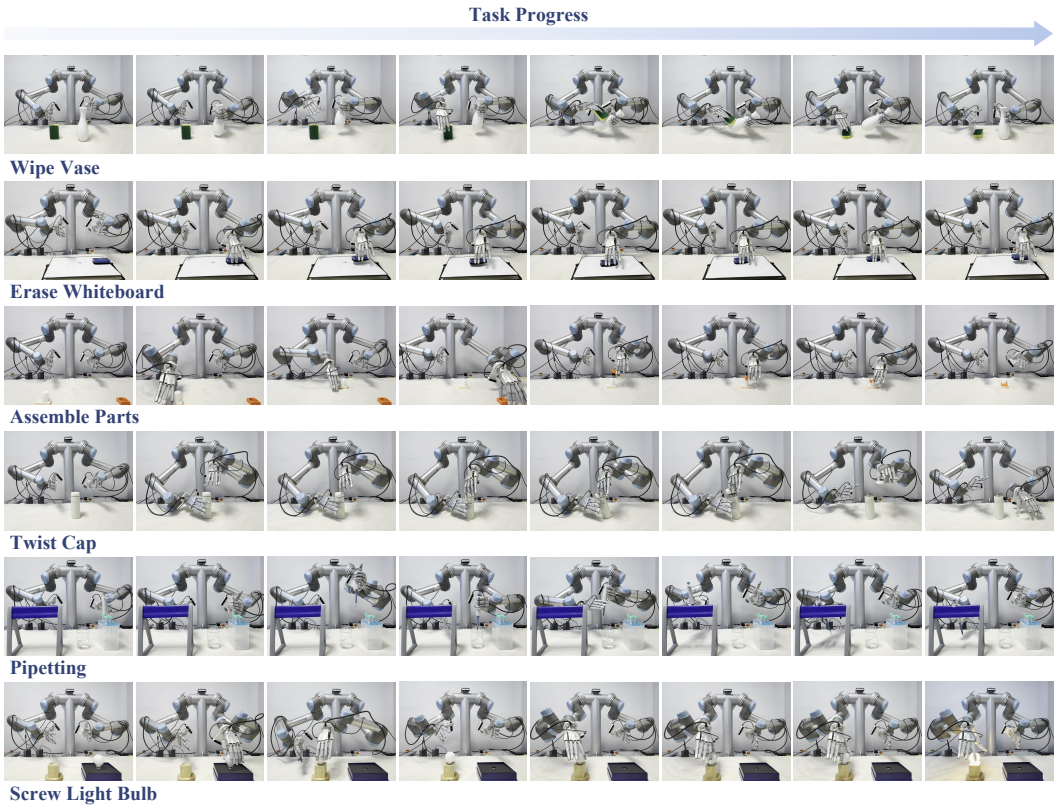

Figure 12: **Visualization of task progress.**

Table 4: **Task descriptions**

Task	Key Manipulation Steps
Wipe Vase	**S1**: Grasp vase → **S2**: Grasp sponge → **S3**: Wipe vase surface → **S4**: Place vase
Erase Whiteboard	**S1**: Pick up eraser → **S2**: Erase whiteboard → **S3**: Place eraser back
Assemble Parts	**S1**: Pick place socket → **S2**: Pick plug → **S3**: Insert plug into socket
Twist Cap	**S1**: Grasp bottle → **S2**: Unscrew cap → **S3**: Place cap
Pipetting	**S1**: Grasp pipette → **S2**: Dispense liquid → **S3**: Hand over pipette → **S4**: Hang pipette
Screw Light Bulb	**S1**: Grasp bulb → **S2**: Insert bulb into socket → **S3**: Screw bulb → **S4**: Illuminate bulb

The key manipulation steps for each task are summarized in Table 4. Additional visualizations of the dexterous manipulation tasks can be found in Figure 12.

E Additional Quantitative and Qualitative Results

E.1 Supplementary Generation Results.

We further visualize the generated future tactile deformation maps in Figure 13. Although the generated tactile observations do not perfectly reconstruct fine-grained deformation details and exhibit some loss of local structure, they successfully preserve contact-related information. In particular, the high-response regions consistently correspond to fingertip-object contact areas, providing informative representations of future contact states. This suggests that the model is able to capture the underlying evolution of tactile interactions, which is sufficient for guiding downstream action generation.

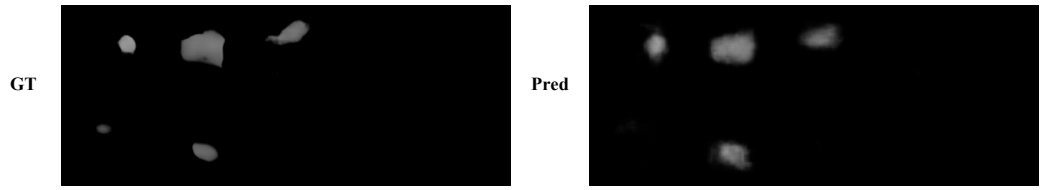

Figure 13: **Qualitative Results of Future tactile deform Generation.**

Table 5: **Stage-wise success rates for ID scenarios**.

Method	Wipe Vase				Erase Whiteboard			Assemble Parts			Twist Cap			Pipetting				Screw Bulb				Avg
	S1→S2→S3→S4				S1→S2→S3			S1→S2→S3			S1→S2→S3			S1→S2→S3→S4				S1→S2→S3→S4				
GR00T N1.6	**1.00**	0.95	0.90	0.30	0.60	0.55	0.50	0.80	0.70	0.30	0.85	0.45	0.10	**1.00**	**1.00**	0.80	**0.60**	**0.90**	0.60	0.20	0.10	0.69
InternVLA-A1	**1.00**	**1.00**	0.75	0.65	0.65	0.60	0.50	0.65	0.50	0.15	0.95	0.10	0.05	**1.00**	**1.00**	0.80	0.35	0.75	**0.70**	**0.40**	0.30	0.68
ViTacFormer	**1.00**	0.95	0.80	0.80	0.55	0.50	0.45	**1.00**	**1.00**	0.15	**1.00**	0.80	0.65	**1.00**	0.95	**0.95**	0.55	0.70	0.65	0.25	0.25	0.79
DECO	**1.00**	**1.00**	0.90	0.90	0.75	0.75	0.60	0.80	0.70	0.35	**1.00**	0.80	0.70	**1.00**	0.75	0.65	0.45	0.45	0.45	0.40	0.35	0.78
InternVLA-A1^t	0.95	0.95	0.85	0.75	0.60	0.50	0.45	0.80	0.70	0.45	0.90	0.50	0.25	0.95	0.90	0.30	0.20	0.35	0.35	0.30	0.25	0.64
Ours	**1.00**	**1.00**	**1.00**	**1.00**	**0.95**	**0.85**	**0.80**	**1.00**	0.90	**0.65**	**1.00**	**1.00**	**0.80**	**1.00**	0.95	0.90	**0.60**	0.65	0.50	**0.40**	**0.40**	**0.91**

E.2 In-Domain (ID) Stage-Wise Results.

Table 5 reports the success rates of individual task stages under in-domain settings. Overall, DeCAL achieves the highest average stage success rate of 0.91, substantially outperforming all baselines. We observe that vision-based VLAs often perform well in early stages that primarily rely on visual grounding, such as object localization and grasping, but their performance degrades significantly in later stages involving sustained contact-rich interactions. Tactile specialist policies, such as ViTacFormer and DECO, generally achieve better performance during contact-intensive stages by leveraging direct physical feedback. However, the absence of explicit visuo-tactile dynamics modeling limits their understanding of physical interactions and future state evolution, leading to weaker long-horizon planning capabilities. In contrast, DeCAL consistently maintains strong performance across all stages by combining adaptive visuo-tactile fusion with explicit visuo-tactile dynamics modeling. For example, in Assemble Parts, DeCAL achieves substantially higher success rates in the final insertion stage, where precise alignment and contact-aware adjustment are essential. Similarly, in Twist Cap, DeCAL maintains a 100% success rate during the cap-twisting stage, demonstrating its ability to reason about sustained physical interactions.

E.3 Out-of-Domain (OOD) Stage-Wise Results

We further report the stage-wise success rates on the Twist Cap task under four challenging generalization settings in Table 6. While all methods experience performance degradation under distribution shifts, DeCAL exhibits better robustness across diverse scenarios, including changes in background, lighting conditions, scene clutter, and object geometry. ViTacFormer exhibits noticeably less stable hand motions under unseen environments, often resulting in object slippage during the cap-twisting stage. DECO demonstrates stronger robustness to environmental perturbations, but suffers a more pronounced performance degradation when manipulating unseen objects. In contrast, DeCAL consistently achieves superior performance across all generalization settings. Notably, under the *Unseen Object* setting, where the test cups differ significantly in height, diameter, and shape, DeCAL still achieves a final-stage success rate of 0.75, exceeding the best baseline by 30% points. We attribute this improvement to the joint modeling of visuo-tactile dynamics, which enables the policy to capture transferable interaction patterns beyond appearance-specific cues and therefore generalize more effectively to unseen scenarios.

Table 6: **Stage-wise success rates for OOD scenarios on Twist Cap.**

Method	Unseen Background			Cluttered			Unseen Lighting			Unseen Object			Avg
	S1→S2→S3			S1→S2→S3			S1→S2→S3			S1→S2→S3			
ViTacFormer	0.80	0.60	0.25	**1.00**	**1.00**	0.40	0.90	0.85	0.25	0.90	**0.90**	0.45	0.69
DECO	**0.90**	**0.75**	0.50	**1.00**	0.85	0.65	**0.95**	0.60	0.55	0.70	0.60	0.35	0.70
Ours	0.75	**0.75**	**0.60**	**1.00**	0.80	**0.70**	**0.95**	**0.90**	**0.65**	**0.95**	**0.90**	**0.75**	**0.81**

E.4 Supplementary Generalization Analysis

In the generalization experiments, we evaluate the models under four types of unseen scenarios, including unseen background, unseen lighting condition, cluttered scene, and unseen object. The first three settings mainly introduce environment-level perturbations, while the unseen-object setting introduces an object-level distribution shift. DECO and ViTacFormer benefit from random image augmentations during training, which improves their robustness to visual changes such as background variation, illumination shifts, and moderate scene clutter. However, the unseen-object setting is substantially more challenging. As shown in Figure 14, we replace the training object with a novel cup that differs in color, diameter, shape, and height. For the Twist Cap task, such object-level changes require the policy to adjust not only the visual localization of the object, but also the grasping height, contact position, and wrist pose in real time. Both baselines show a clear performance drop under this object-level shift, suggesting that visual augmentation alone is insufficient for handling changes in contact dynamics. In contrast, DeCAL maintains a success rate of 75%, demonstrating its stronger ability to generalize to the novel object by leveraging tactile feedback and future-oriented latent imagination.

Figure 14: **Novel Object.**

E.5 Supplementary Ablation Study Analysis

We further provide a detailed analysis of the ablation results. (1) Removing Factorized Flow Matching leads to a noticeable performance drop. Without separating arm and hand motion during denoising, the policy models the full action space with a shared denoising process, which weakens arm-hand coordination. This often leads to inconsistent motions between arm-level reaching and fine-grained finger movements, especially in grasping and bimanual tasks where small pose or timing errors can cause unstable contact or failed execution. (2) Removing visual and tactile latent generation weakens performance in contact-sensitive tasks. Without this future-oriented visuo-tactile prediction, the policy struggles to anticipate how observations and contact states evolve after actions. In the Assemble Parts task, the policy may hesitate or oscillate near the socket during the "insert plug into socket" step, failing to make a consistent progress toward insertion. This indicates that latent generation provides useful predictive representations for modeling both motion and contact dynamics. (3) Removing the tactile gating mechanism reduces the policy's ability to adaptively balance visual-language features. Naively fusing tactile features with visual-language embeddings may introduce noise or interfere with visual grounding, especially when contact are weak. Introducing the tactile gating mechanism mitigates this issue by modulating the contribution of tactile features according to the current interaction state.

F Baseline Settings

For GR00T N1.6. [40] All experiments are conducted on a single NVIDIA H100 GPU. We use the full joint configuration of the dual-arm dual-hand system, with two 6-DoF robot arms and two 22-DoF dexterous hands, resulting in a 56-DoF action space for representing the complete joint action. The model is initialized from the pretrained GR00T-N1.6-3B checkpoint and fine-tuned for

50K steps with a batch size of 32. We optimize the model using AdamW with an initial learning rate of 1.0×10^{-4}. The learning rate is warmed up during the first 5% of the training steps and then decayed using a cosine learning rate schedule.

For InternVLA-A1 [41]. All experiments are conducted on 8 NVIDIA H100 GPUs. We use the full joint configuration of the dual-arm dual-hand system, consisting of two 6-DoF robot arms and two 22-DoF dexterous hands (a total of 56-DoF). We extend the maximum action dimensionality to support this full joint action representation and initialize the model from the pretrained InternVLA-A1-3B checkpoint, with newly introduced parameters randomly initialized. The model is trained with mixed precision for 100K steps, using a per-GPU batch size of 4, an action chunk size of 50, and AdamW with a learning rate of 5.0×10^{-5}. The learning rate is warmed up for the first 2K steps and then linearly decayed over 200K steps.

For ViTacFormer [28]. All experiments are conducted on a single NVIDIA H100 GPU. To adapt ViTacFormer to our visuo-tactile dexterous manipulation setting, we use the 6-D net force from each fingertip as the tactile input and the supervision signal, since the original model does not support tactile image inputs. We use ResNet-18 as the visual encoder and optimize it with a smaller learning rate of 1.0×10^{-5}, while the remaining modules, including the Transformer backbone, action prediction head, and tactile fusion layers, are randomly initialized and trained with a learning rate of 1.0×10^{-4}. We set the KL weight to 10, the action chunk size to 100, the hidden dimension to 512, and the batch size to 128. The model is trained for 2,000 epochs.

For DECO [32]. All experiments are conducted on 4 NVIDIA H100 GPUs. We adapt DECO to our visuo-tactile dexterous manipulation setting by using the 6-D net force from each fingertip as the tactile input. Following the original design, we adopt a pretrained ResNet-34 image encoder as the visual backbone and fine-tune it in a full-parameter manner. The model is optimized with SGD for 500 epochs, using a batch size of 1024. We set the initial learning rate to 1.0×10^{-4} and use a cosine annealing schedule with a minimum learning rate of 5.0×10^{-6}. The first epoch is used for learning-rate warmup.

For InternVLA-A1^t. For a fair comparison, we construct a tactile-augmented variant of InternVLA-A1, denoted as InternVLA-A1^t, by incorporating both tactile deforms and 6-D net force signals as additional VLM inputs, matching the input modalities used by DeCAL. Specifically, the current-frame deformation maps from 10 fingertips are encoded by a Multi-Finger Tactile Encoder with a shared CNN and an inter-finger Transformer, yielding compact tactile features. Meanwhile, each fingertip's 6-D force is projected into the hidden space by a lightweight MLP to obtain force features. These tactile and force features are concatenated and injected into each understanding Transformer block through an additional 8-head cross-attention layer, using the prefix hidden states as queries and the tactile features as keys and values, followed by a residual connection. All other hyperparameters, pretrained weights, camera inputs, and action settings are kept the same as InternVLA-A1.

G Failure Case Analysis

Through extensive real-world experiments, we identify two primary categories of failure modes that can adversely affect the performance of DeCAL: The first category is related to **bimanual coordination**. For tasks that require precise synchronization between two arms, such as transferring a pipette from one hand to the other in the Pipetting task, small discrepancies in arm motion, hand alignment, or grasp timing may lead to unstable contact and eventually cause the object to slip or deviate from the desired pose. The second category arises from the **limited field of view** of the head-mounted camera. In tasks involving large-range motions, the manipulated object may occasionally move outside the visible region, making it difficult for the policy to maintain reliable visual tracking and state estimation. A potential direction for mitigating this limitation is to incorporate wider-angle or fisheye cameras, which can provide more comprehensive visual coverage of the workspace. Another promising direction is to introduce active perception, where the robot dynamically adjusts its viewpoint or selects informative observations during execution to better track task-relevant objects and reduce visual uncertainty.