ImplicitRDP: An End-to-End Visual-Force Diffusion Policy with Structural Slow-Fast Learning

Anonymous CVPR submission

Paper ID *****

Abstract

Contact-rich manipulation requires combining global visual context with local force feedback. We propose ImplicitRDP, an end-to-end visual-force diffusion policy that unifies visual planning and reactive force control in a single network. Our Structural Slow-Fast Learning uses causal attention to process low-frequency visual tokens and high-frequency force tokens for closed-loop control within an action chunk. We further introduce Virtual-target-based Representation Regularization, which predicts a virtual target in the action space to encourage effective use of force feedback and to avoid modality collapse. Experiments on real-world contact-rich tasks show that ImplicitRDP outperforms both vision-only and hierarchical visual-force baselines with a simpler training pipeline.

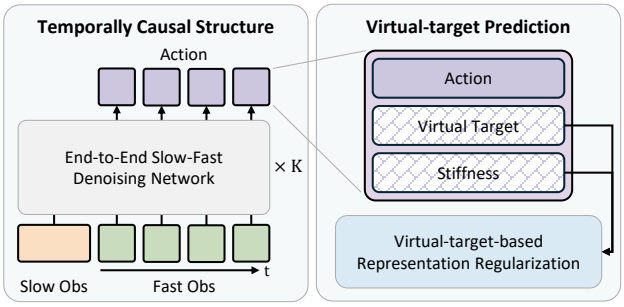

Figure 1. **Structural Slow-Fast Learning:** we leverage a temporally causal structure to enable end-to-end closed-loop force-based control within an action chunk. **Virtual-target-based Representation Regularization**: we further incorporate virtual target prediction as an auxiliary task to prevent modality collapse.

1. Introduction

Contact-rich manipulation requires both global visual context and local force feedback. Although imitation learning has achieved strong performance in robotic manipulation [2, 5, 11, 18, 29], integrating high-frequency force signals into action-chunked policies remains difficult. Recent approaches such as Reactive Diffusion Policy (RDP) [24] address this challenge with hierarchical slow-fast architectures, but the explicit separation between visual planning and force reaction introduces an information bottleneck and brittle hand-designed coordination between modules.

We propose **ImplicitRDP** (Fig. 1), an end-to-end visual-force diffusion policy built on a unified Transformer. Our *Structural Slow-Fast Learning* mechanism concatenates low-frequency visual tokens and high-frequency force tokens under temporal causality, allowing action tokens to attend directly to both modalities and enabling closed-loop force control within each chunk. To prevent modality collapse [13], we further introduce *Virtual-Target-based Representation Regularization*, which predicts a virtual target in the action space instead of raw force and encourages the policy to use force feedback in a more actionable way.

We evaluate ImplicitRDP on two contact-rich real-world tasks, box flipping and switch toggling. The end-to-end formulation outperforms both visual-only and hierarchical visual-force baselines while using a simpler unified pipeline.

Overall, this paper makes the following contributions:

- We propose ImplicitRDP, an end-to-end visual-force policy with structural slow-fast learning that simultaneously processes slow and fast observations while realizing closed-loop force control.
- We introduce an auxiliary task based on virtual-target prediction that encourages the policy to adaptively adjust weights of different modalities. It maps desired forces into the Cartesian action space and appropriately weights losses according to force magnitudes, leading to more effective guidance than conventional force prediction.
- Extensive experiments on two representative contact-rich tasks demonstrate that ImplicitRDP achieves higher performance than baseline methods and provides a more streamlined and unified training framework as well.

2. Methodology

We briefly introduce the standard Diffusion Policy, and then present *structural slow-fast learning (SSL)* and *virtual-target-based representation regularization (VRR)*. Additional implementation details are provided in Supp. Sec. 5.

2.1. Preliminary: Diffusion Policy

Let $\mathbf{O}_t \triangleq \{o_{t-h_o+1}, \ldots, o_t\}$ denote the observation window of length h_o at time t, and let $\mathbf{A}_t \triangleq \{a_{t-h_o+1}, \ldots, a_{t-h_o+h_a}\}$ denote the action chunk with horizon h_a. Following Diffusion Policy (DP) [5], the clean action chunk $\mathbf{A}_t^0$ is corrupted into a noisy chunk $\mathbf{A}_t^k$ with Gaussian noise ϵ^k at diffusion step k, and the original DDPM objective is:

$$\mathcal{L}_\epsilon = \mathbb{E}_{k,\epsilon^k,(\mathbf{O}_t,\mathbf{A}_t^0)}[\|\epsilon^k - \epsilon_\theta(\mathbf{O}_t, \mathbf{A}_t^k, k)\|^2]. \quad (1)$$

DP typically executes it with receding-horizon control. As a result, it remains open-loop within each chunk, which motivates our method.

2.2. Structural Slow-Fast Learning

To enable high-frequency closed-loop control within the action chunking, we introduce *structural slow-fast learning*. Unlike RDP [24], which employs a two-stage slow-fast architecture, structural slow-fast learning realizes end-to-end action modeling with variant-frequency observations through temporally causal structure and consistent inference mechanism.

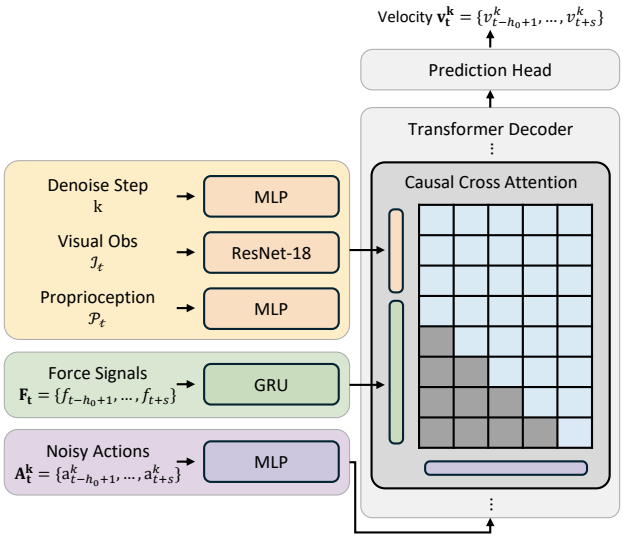

Figure 2. **Network Architecture of ImplicitRDP.** We enforce a temporally causal structure using a GRU for force signal encoding and a causal attention mask for action-force interaction, which enables structural slow-fast learning.

2.2.1. Temporally Causal Structure

As shown in Fig. 2, we design ImplicitRDP based on the standard Transformer-based [20] DP with ResNet-18 [8] visual encoder. We separate the observations into the "slow" part (visual observations $\mathcal{I}_t$ and proprioception $\mathcal{P}_t$) and the "fast" part (force signals $\mathbf{F}_t \triangleq \{f_{t-h_o+1}, \ldots, f_{t-h_o+h_a}\}$). Unlike standard Transformer-based DP which directly encodes all observations and applies full cross-attention between the action tokens and observation tokens, we treat the "fast" force signals as a temporal sequence aligned with the action chunk which are used for closed-loop control. To prevent future information leakage, we modify the model structure to meet the temporal causality requirement. First, we use GRU [6] for force encoding to preserve causality constraints. Second, we employ a causal attention mask for the force tokens, ensuring that the prediction of action $a_{t-h_o+s}(1 \leq s \leq h_a)$ can attend to force tokens $\{f_{t-h_o+1}, \ldots, f_{t-h_o+s}\}$ but not future forces. These structural constraints enable ImplicitRDP to achieve parallel training efficiency comparable to standard DP, while also enabling temporally causal processing of dense force signals.

2.2.2. Consistent Inference Mechanism

The causal structure of ImplicitRDP naturally supports variable-length action tokens, which enables closed-loop force control within an action chunk. A direct way to exploit this property is to resample a slightly longer action sequence at every control step. However, because diffusion models are stochastic, independently sampled trajectories can become inconsistent across consecutive steps and lead to unstable execution. To avoid this issue, we use deterministic DDIM [17] with $\eta = 0$ and sample the initial noise only once per chunk. Slow observation encoding and noise are cached at the start of the chunk, while each control step updates only the fast observations, reruns DDIM with the cached context, and executes the last action. Full pseudocode is provided in Supp. Alg. 1.

2.3. Virtual-target-based Representation Regularization

To prevent the end-to-end policy from relying solely on a single modality, we introduce a novel representation regularization method. Inspired by [28], which predicts future torque to enforce physical understanding, we propose predicting the *virtual target*. Unlike raw force, the virtual target resides in the same Cartesian space as the robot's action, facilitating a more unified representation learning process.

2.3.1. Virtual Target Formulation

The concept of the virtual target is derived from compliance control theory [15]. A standard compliance system can be

modeled as a spring-mass-damper system:

$$f_{ext} = M\ddot{x}_{vt} + D\dot{x}_{vt} + K(x_{vt} - x_{real}), \tag{2}$$

where f_{ext} denotes the external wrench, and M, D, and K represent the inertia, damping, and stiffness matrices. x_{real} is the robot's actual pose, and x_{vt} is the virtual target. In the context of quasi-static manipulation, we can ignore the inertia and damping terms. Consequently, the virtual target can be derived given the current stiffness and measured force:

$$x_{vt} = x_{real} + K^{-1} f_{ext}. \tag{3}$$

2.3.2. Adaptive Stiffness Assignment

A manipulation task may consist of multiple phases, such as approaching the object, making contact, and manipulating it. Thus, it is inappropriate to apply the same force-based representation regularization uniformly across all phases. Instead, we adopt the heuristic strategy from ACP [9] to assign an adaptive stiffness matrix.

We decompose the stiffness into a generalized force direction and its orthogonal subspace. Specifically, we assign a high stiffness k_{high} to the directions that are orthogonal to the force. For the force direction, we define an adaptive stiffness scalar k_{adp} that varies with the force magnitude $\|f_{ext}\|$:

$$k_{adp} = \begin{cases} k_{max}, & \|f_{ext}\| < f_{min} \\ k_{max} - \frac{k_{max}-k_{min}}{f_{max}-f_{min}}(\|f_{ext}\| - f_{min}), & \text{otherwise} \\ k_{min}, & \|f_{ext}\| > f_{max} \end{cases} \tag{4}$$

where f_{min} and f_{max} are thresholds determining the sensitivity to contact, and k_{max}, k_{min} define the stiffness range.

2.3.3. Unified Training Objective

To incorporate this regularization into the diffusion framework, we use a unified prediction space. We construct an augmented action vector $a_{aug,t}$ by concatenating the original action a_t, the calculated virtual target x_{vt}, and the stiffness magnitude k_{adp}:

$$a_{aug,t} = \text{concat}([a_t, x_{vt}, k_{adp}]). \tag{5}$$

The diffusion policy is then trained to denoise the sequence $\mathbf{A}_t^0 \triangleq \{a_{aug,t-h_o+1}, \ldots, a_{aug,t-h_o+h_a}\}$ of the augmented vectors. During inference, we discard the auxiliary components and execute only $\hat{a}_t$.

2.3.4. Advantages over Force Prediction

While equation 3 implies that predicting x_{vt} is mathematically equivalent to predicting f_{ext} (given x_{real} and K), the virtual target offers two significant advantages for representation learning.

The first one is **objective alignment.** Force sensors typically measure force signals in the TCP frame, whereas actions are trajectories in the robot base or world frame. In contrast, the virtual target x_{vt} is also a motion trajectory computed in the same coordinate system as the action space. This alignment helps the network to learn a consistent representation for both motion planning and force understanding.

The second one is **adaptive importance weighting.** The use of adaptive stiffness acts as a dynamic weighting mechanism. Consider the deviation $\Delta x \triangleq x_{vt} - x_{real} = K^{-1} f_{ext}$.

- In free motion, $\|f_{ext}\|$ is small (mostly sensor noise). According to Eq. 4, K becomes large (k_{max}), making K^{-1} small. Consequently, $\Delta x \to 0$, and $x_{vt} \approx x_{real}$.
- During contact, $\|f_{ext}\|$ is large. K becomes small (k_{min}), making K^{-1} large. This amplifies Δx, causing x_{vt} to deviate significantly from x_{real}.

This mechanism effectively assigns higher loss weights to high-force contact events, forcing the network to pay attention to critical force feedback while ignoring noise during free motion.

3. Experiments

We evaluate ImplicitRDP on real-world contact-rich manipulation tasks against visual-only and hierarchical visual-force baselines, and further analyze the roles of closed-loop control and auxiliary regularization. For more details and results, please refer to Supp. Sec. 6.

3.1. Experimental Setup

We design two representative contact-rich manipulation tasks:

1. **Box Flipping:** The robot must push a thin phone box against a fixture to flip it from a flat position to an upright one. This task requires sustained contact and continuous force adjustment under a narrow operating regime, making open-loop execution particularly brittle.
2. **Switch Toggling:** The robot needs to toggle a circuit breaker switch. The challenge of this task is that the switch requires a relatively large force to actuate, while vision-only policy cannot determine whether the triggering threshold has been reached. Unlike box flipping, switch toggling requires a short-duration force burst, which is common in tasks like vegetable chopping.

For each task, we collected 40 demonstrations.

3.2. Results and Analysis

3.2.1. Comparison with Baselines

As illustrated in Tab. 1, the end-to-end ImplicitRDP consistently achieves the best performance compared to both the vision-only DP and the hierarchical visual-force policy RDP.

In the box flipping task, the vision-only DP often applies force far exceeding normal levels, resulting in dangerous

crushing of the phone box, as shown in Supp. Fig. 5. This failure stems from the inability of visual observations to determine whether the applied force is appropriate. Similarly, in the switch toggling task, DP tends to start the upward toggling motion before the required triggering force is reached, as shown in Supp. Fig. 6, because the visual difference between the triggered and un-triggered states is negligible.

Regarding RDP, while it performs adequately in box flipping, it struggles with switch toggling. RDP frequently contacts the wrong location during the approach phase, as illustrated in Supp. Fig. 6. We hypothesize that this is because the fast policy in RDP compresses raw actions into a latent space, leading to precision loss during free-space motion.

In contrast, ImplicitRDP realizes closed-loop control based on force signals and performs end-to-end denoising directly in the original action space. This allows it to adaptively weigh different modalities, ensuring both accurate reactivity during contact and precise movement during free motion.

Table 1. Success Rate Compared with Baseline Methods

Method	Box Flipping	Switch Toggling
DP	0/20	8/20
RDP	16/20	10/20
ImplicitRDP (Ours)	**18/20**	**18/20**

3.2.2. Effectiveness of Closed-Loop Control

To validate the importance of the closed-loop mechanism provided by Structural Slow-Fast Learning (SSL), we compare the full model against open-loop variants. As shown in Tab. 2, when both SSL and Virtual-target-based Representation Regularization (VRR) are removed, and the network relies only on low-frequency visual and force signals for open-loop control, performance drops significantly across both tasks. Even when VRR is reintroduced, the open-loop variant still suffers from a performance decline compared to the complete ImplicitRDP.

Notably, the performance drop is much more pronounced in the box flipping task. According to Supp. Fig. 5, the primary cause of failure in the open-loop setting is excessive force application. This is probably because box flipping requires the sustained application of a constant force, while an open-loop network cannot adjust its actions in real-time within a chunk based on force feedback. Consequently, the applied force deviates from the target, pushing the state into an out-of-distribution region and leading to task failure. These results demonstrate that the closed-loop force control realized by SSL in ImplicitRDP is critical for improving performance in contact-rich tasks, particularly those requiring sustained force maintenance.

Table 2. Comparison between Open-loop and Closed-loop Control

Method	Box Flipping	Switch Toggling
ImplicitRDP w.o. SSL and VRR	6/20	5/20
ImplicitRDP w.o. SSL	4/20	15/20
ImplicitRDP (Ours)	**18/20**	**18/20**

3.2.3. Auxiliary Task Analysis

We further analyze the impact of different auxiliary tasks in Tab. 3. We find that using the virtual target as the prediction objective yields the best performance on both tasks. While standard force prediction provides some improvement over using no auxiliary task, it remains inferior to VRR. Qualitative failures in Supp. Fig. 5 and Supp. Fig. 6 indicate that policies trained with other auxiliary strategies tend to lose contact prematurely during both box pushing and switch toggling. The attention visualization in Supp. Fig. 4 shows that without the auxiliary task, the model fails to learn the importance relationships between different modalities. These results confirm that these networks fail to fully utilize high-frequency force inputs, resulting in modality collapse. In comparison, VRR in ImplicitRDP resides in the same space as the action and employs adaptive weighting, which helps regularize the model representation, encouraging the network to focus on critical force data and thereby enhancing performance.

Table 3. Comparison of Different Auxiliary Tasks

Auxiliary Task	Box Flipping	Switch Toggling
None	6/20	6/20
Force Prediction	8/20	10/20
Virtual Target Prediction (Ours)	**18/20**	**18/20**

4. Conclusion

In this paper, we present ImplicitRDP, a novel end-to-end framework that reconciles low-frequency visual planning and high-frequency force control. By embedding structural slow-fast learning directly into the diffusion process, we eliminate the need for separate policy hierarchies, allowing a single network to dynamically attend to different modalities with variant frequencies. Additionally, our proposed virtual-target auxiliary task effectively regularizes the representation space, ensuring the policy adaptively leverages physical feedback rather than over-relying on one single modality. Experimental results confirm that this unified approach not only simplifies the training pipeline but also achieves superior performance in contact-rich manipulation compared with baselines. Future work will investigate extending this unified framework to Vision-Language-Action (VLA) models, as well as integrating other high-frequency modalities such as tactile sensing.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 1, 3

[3] Claire Chen, Zhongchun Yu, Hojung Choi, Mark Cutkosky, and Jeannette Bohg. Dexforce: Extracting force-informed actions from kinesthetic demonstrations for dexterous manipulation. *IEEE Robotics and Automation Letters*, 2025. 3

[4] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024. 1

[5] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44 (10-11):1684–1704, 2025. 1, 2, 3

[6] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014. 2

[7] Flexiv Rizon 4 robotic arm. https://www.flexiv.com/products/rizon, 2024. 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[9] Yifan Hou, Zeyi Liu, Cheng Chi, Eric Cousineau, Naveen Kuppuswamy, Siyuan Feng, Benjamin Burchfiel, and Shuran Song. Adaptive compliance policy: Learning approximate compliance for diffusion guided control. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4829–4836. IEEE, 2025. 3, 4

[10] Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. *arXiv preprint arXiv:2412.14803*, 2024. 4

[11] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. 1, 3

[12] Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025. 4

[13] Jason Jingzhou Liu, Yulong Li, Kenneth Shaw, Tony Tao, Ruslan Salakhutdinov, and Deepak Pathak. Factr: Force-attending curriculum training for contact-rich policy learning. *arXiv preprint arXiv:2502.17432*, 2025. 1, 4

[14] Wenhai Liu, Junbo Wang, Yiming Wang, Weiming Wang, and Cewu Lu. Forcemimic: Force-centric imitation learning with force-motion capture system for contact-rich manipulation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1105–1112. IEEE, 2025. 3

[15] Matthew T Mason. Compliance and force control for computer controlled manipulators. *IEEE Transactions on Systems, Man, and Cybernetics*, 11(6):418–432, 2007. 2

[16] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 1

[17] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[18] Generalist AI Team. Gen-0: Embodied foundation models that scale with physical interaction. *Generalist AI Blog*, 2025. https://generalistai.com/blog/preview-uqlxvb-bb.html. 1, 3

[19] Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. *arXiv preprint arXiv:2412.15109*, 2024. 4

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[21] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. *arXiv preprint arXiv:2312.13139*, 2023. 4

[22] Yansong Wu, Zongxie Chen, Fan Wu, Lingyun Chen, Liding Zhang, Zhenshan Bing, Abdalla Swikir, Sami Haddadin, and Alois Knoll. Tacdiffusion: Force-domain diffusion policy for precise tactile manipulation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11831–11837. IEEE, 2025. 3

[23] Xiaomeng Xu, Yifan Hou, Zeyi Liu, and Shuran Song. Compliant residual dagger: Improving real-world contact-rich manipulation with human corrections. *arXiv preprint arXiv:2506.16685*, 2025. 3

[24] Han Xue, Jieji Ren, Wendi Chen, Gu Zhang, Yuan Fang, Guoying Gu, Huazhe Xu, and Cewu Lu. Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation. *arXiv preprint arXiv:2503.02881*, 2025. 1, 2, 3

[25] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 3

[26] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22963–22974, 2025. 3

[27] Jiawen Yu, Hairuo Liu, Qiaojun Yu, Jieji Ren, Ce Hao, Haitong Ding, Guangyu Huang, Guofan Huang, Yan Song, Panpan Cai, et al. Forcevla: Enhancing vla models with a force-aware moe for contact-rich manipulation. *arXiv preprint arXiv:2505.22159*, 2025. 3

[28] Zongzheng Zhang, Haobo Xu, Zhuo Yang, Chenghao Yue, Zehao Lin, Huan-ang Gao, Ziwei Wang, and Hao Zhao. Ta-vla: Elucidating the design space of torque-aware vision-language-action models. *arXiv preprint arXiv:2509.07962*, 2025. 2, 4

[29] 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. 1, 3

[30] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019. 1

[31] Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792*, 2025. 4