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

Anonymous Author(s)

***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.**

I. Introduction

Contact-rich manipulation requires both global visual context and local force feedback. Although imitation learning has achieved strong performance in robotic manipulation [1, 2, 3, 4, 5], integrating high-frequency force signals into action-chunked policies remains difficult. Recent approaches such as Reactive Diffusion Policy (RDP) [6] 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 [7], 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.

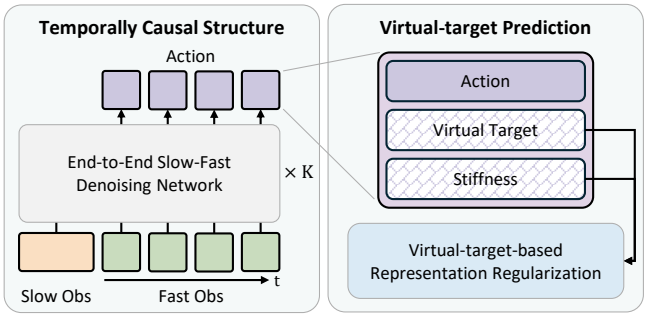

Fig. 1. **Structural Slow-Fast Learning:** we leverage a temporally causal structure to enable end-to-end closed-loop force-based control within an action chunk. **Virtual-target-based Representation Regularization**: we further incorporate virtual target prediction as an auxiliary task to prevent modality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

II. 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. V.

A. 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) [1], 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.

B. Structural Slow-Fast Learning

To enable high-frequency closed-loop control within the action chunking, we introduce *structural slow-fast learning*. Unlike RDP [6], 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.

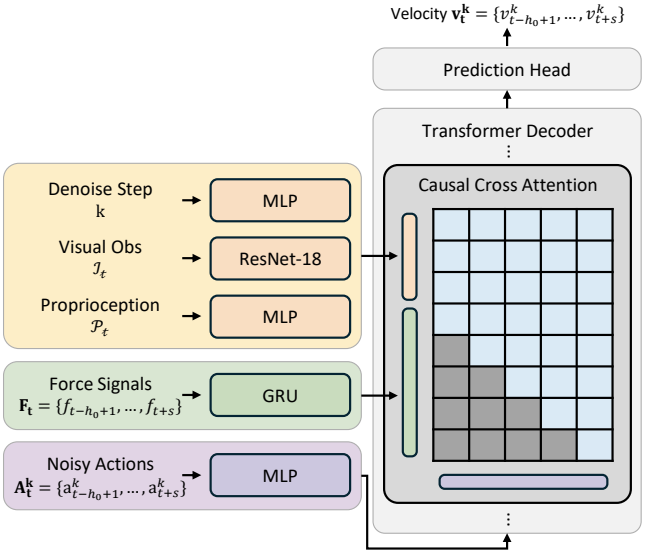

Fig. 2. **Network Architecture of ImplicitRDP.** We enforce a temporally causal structure using a GRU for force signal encoding and a causal attention mask for action-force interaction, which enables structural slow-fast learning.

1) Temporally Causal Structure: As shown in Fig. 2, we design ImplicitRDP based on the standard Transformer-based [8] DP with ResNet-18 [9] 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 $\mathbf{F}_t \triangleq \{f_{t-h_o+1}, \dots, f_{t-h_o+h_a}\}$). 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 [10] 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 $\{f_{t-h_o+1}, \dots, f_{t-h_o+s}\}$ 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) 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 [11] 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.

C. 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 [12], 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.

1) Virtual Target Formulation: The concept of the virtual target is derived from compliance control theory [13]. 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) 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 [14] 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.

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 $\hat{a}_t$.

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.

III. 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. VI.

A. Experimental Setup

Our hardware setup (see Supp. Fig. 4) is built on a Flexiv Rizon 4s [15] robot arm, which is equipped with both joint torque sensors and a 6-axis force/torque sensor at the end effector. This allows us to use the joint torque sensors for kinematic teaching during data collection, while obtaining contact forces directly from the 6-axis F/T sensor at the same time. A joystick and our custom compliant fingertips are mounted on the robot's end effector. We use a webcam as a wrist camera to record visual observations. All data is recorded at 10 Hz.

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.

We compare ImplicitRDP against the following baselines:

- **Diffusion Policy (DP):** Standard CNN-based DP with vision-based open-loop control.
- **Reactive Diffusion Policy (RDP):** The state-of-the-art hierarchical slow-fast visual-force learning method.
- **ImplicitRDP w.o. SSL and VRR:** Similar to standard transformer-based DP, but augmented with open-loop force inputs and techniques that improve training stability in Sec. V-A.1.
- **ImplicitRDP w.o. SSL:** Similar to the previous one, but with VRR used.
- **ImplicitRDP w. Different Auxiliary Tasks:** Use no other task or use force prediction as the auxiliary task.
- **ImplicitRDP w. Different Training Choices:** Use alternative implementation choices discussed in Sec. V-A.1.

B. Results and Analysis

1) Comparison with Baselines: As illustrated in Tab. I, 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 I

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**

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. II, 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 II

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) Auxiliary Task Analysis: We further analyze the impact of different auxiliary tasks in Tab. III. 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 Fig. 3 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.

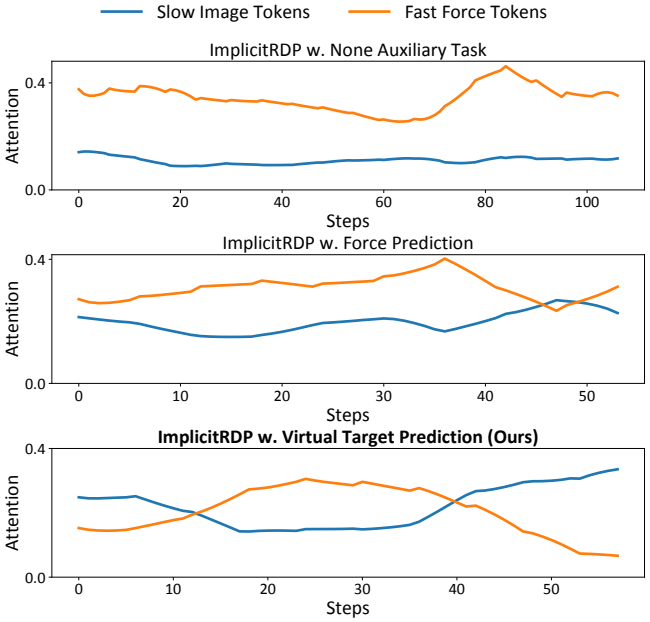

Fig. 3. **Attention Weight Visualization.** We visualize the summed attention weights of visual tokens and force tokens from the first transformer layer in the switch toggling task. The weights are averaged over all heads and all queries. A sliding window of size 10 is applied to smooth the curves.

TABLE III

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**

IV. 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] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, vol. 44, no. 10-11, pp. 1684–1704, 2025. 1, 2

[2] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023. 1, 2

[3] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024. 1, 2

[4] 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. 1, 2

[5] G. A. Team, "Gen-0: Embodied foundation models that scale with physical interaction," *Generalist AI Blog*, 2025, https://generalistai.com/blog/preview-uqlxvb-bb.html. 1, 2

[6] 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. 1, 2

[7] J. J. Liu, Y. Li, K. Shaw, T. Tao, R. Salakhutdinov, and D. Pathak, "Factr: Force-attending curriculum training for contact-rich policy learning," *arXiv preprint arXiv:2502.17432*, 2025. 1, 3

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. 2

[9] 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*, 2016, pp. 770–778. 2

[10] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014. 2

[11] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020. 2

[12] Z. Zhang, H. Xu, Z. Yang, C. Yue, Z. Lin, H.-a. Gao, Z. Wang, and H. Zhao, "Ta-vla: Elucidating the design space of torque-aware vision-language-action models," *arXiv preprint arXiv:2509.07962*, 2025. 2, 3

[13] M. T. Mason, "Compliance and force control for computer controlled manipulators," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 11, no. 6, pp. 418–432, 2007. 2

[14] Y. Hou, Z. Liu, C. Chi, E. Cousineau, N. Kuppuswamy, S. Feng, B. Burchfiel, and S. Song, "Adaptive compliance policy: Learning approximate compliance for diffusion guided control," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 4829–4836. 2, 3

[15] https://www.flexiv.com/products/rizon, 2024. 3

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

[17] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5745–5753. 1

[18] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," *arXiv preprint arXiv:2402.10329*, 2024. 1

[19] Y. Wu, Z. Chen, F. Wu, L. Chen, L. Zhang, Z. Bing, A. Swikir, S. Haddadin, and A. Knoll, "Tacdiffusion: Force-domain diffusion policy for precise tactile manipulation," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 11 831–11 837. 2

[20] W. Liu, J. Wang, Y. Wang, W. Wang, and C. 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)*. IEEE, 2025, pp. 1105–1112. 2

[21] C. Chen, Z. Yu, H. Choi, M. Cutkosky, and J. Bohg, "Dexforce: Extracting force-informed actions from kinesthetic demonstrations for dexterous manipulation," *IEEE Robotics and Automation Letters*, 2025. 2

[22] X. Xu, Y. Hou, Z. Liu, and S. Song, "Compliant residual dagger: Improving real-world contact-rich manipulation with human corrections," *arXiv preprint arXiv:2506.16685*, 2025. 2

[23] J. Yu, H. Liu, Q. Yu, J. Ren, C. Hao, H. Ding, G. Huang, G. Huang, Y. Song, P. Cai, *et al.*, "Forcevla: Enhancing vla models with a force-aware moe for contact-rich manipulation," *arXiv preprint arXiv:2505.22159*, 2025. 2

[24] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023. 3

[25] 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. 3

[26] T. Yin, Q. Zhang, R. Zhang, W. T. Freeman, F. Durand, E. Shechtman, and X. Huang, "From slow bidirectional to fast autoregressive video diffusion models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 22 963–22 974. 3

[27] H. Wu, Y. Jing, C. Cheang, G. Chen, J. Xu, X. Li, M. Liu, H. Li, and T. Kong, "Unleashing large-scale video generative pre-training for visual robot manipulation," *arXiv preprint arXiv:2312.13139*, 2023. 3

[28] Y. Tian, S. Yang, J. Zeng, P. Wang, D. Lin, H. Dong, and J. Pang, "Predictive inverse dynamics models are scalable learners for robotic manipulation," *arXiv preprint arXiv:2412.15109*, 2024. 3

[29] Y. Hu, Y. Guo, P. Wang, X. Chen, Y.-J. Wang, J. Zhang, K. Sreenath, C. Lu, and J. Chen, "Video prediction policy: A generalist robot policy with predictive visual representations," *arXiv preprint arXiv:2412.14803*, 2024. 3

[30] S. Li, Y. Gao, D. Sadigh, and S. Song, "Unified video action model," *arXiv preprint arXiv:2503.00200*, 2025. 3

[31] C. Zhu, R. Yu, S. Feng, B. Burchfiel, P. Shah, and A. Gupta, "Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets," *arXiv preprint arXiv:2504.02792*, 2025. 3

V. Methodology Details

This section provides additional implementation details.

Algorithm 1 Consistent Inference in ImplicitRDP

Require: Policy Network π_θ, Slow Observation Encoder $\mathcal{E}_{slow}$, Fast Observation Encoder $\mathcal{E}_{fast}$, Slow Observation Horizon h_o, Execution Horizon h_e, Latency Steps l

Initialize step $t \leftarrow 0$
// **slow loop**: action chunk modeling
while Task not done **do**
  // cache slow context and noise
  $\mathbf{O}_{slow} \leftarrow \text{GetSlowObservation}(len = h_o)$
  $\mathbf{Z}_{slow} \leftarrow \mathcal{E}_{slow}(\mathbf{O}_{slow})$
  $\mathbf{A}^K \leftarrow \mathcal{N}(0, \mathbf{I})$
  // **fast loop**: closed-loop control within the horizon
  for $i \leftarrow 0$ **to** $h_e - 1$ **do**
    // update fast context
    $\mathbf{O}_{fast} \leftarrow \text{GetFastObservation}(len = h_o + i + l)$
    $\mathbf{Z}_{fast} \leftarrow \mathcal{E}_{fast}(\mathbf{O}_{fast})$
    // get a noisy action sequence of a certain length
    $\mathbf{A}_t^K \leftarrow \mathbf{A}^K[: i + h_o + l]$
    // consistent denoising with cache
    $\hat{\mathbf{A}}_t^0 \leftarrow \text{DDIM}(\pi_\theta, \mathbf{Z}_{slow}, \mathbf{Z}_{fast}, \mathbf{A}_t^K, \eta = 0)$
    // execute the current step action
    $a_t \leftarrow \hat{\mathbf{A}}_t^0[-1]$
    Execute a_t
    $t \leftarrow t + 1$
    if Task done **then**
      break
    end if
  end for
end while

A. Implementation Details

Performing contact-rich tasks with force control imposes stringent requirements on the execution precision of the entire system. To ensure better performance, we adjust various system components, ranging from learning objectives to hardware and low-level controller.

1) Learning Stability: Directly utilizing force signals in an end-to-end network can lead to instability. We observe that the powerful fitting capability of Transformer-based DP often results in overfitting to high-frequency noise within force signals, causing action jitter during inference. We address this through two key modifications.

First, we replace the standard ϵ-prediction parameterization with velocity-prediction. While ϵ-prediction and sample-prediction are common, we found that velocity-prediction strikes a better balance between inference stability and adherence to conditional information. Following the formulation in [16], the relationship between velocity $\mathbf{v}_k$, noise ϵ, and the original sample $\mathbf{A}_t^0$ is defined as

$$\mathbf{v}_t^k \triangleq \sqrt{\bar{\alpha}_k}\epsilon - \sqrt{1-\bar{\alpha}_k}\mathbf{A}_t^0. \tag{6}$$

The corresponding training loss is formulated in Supp. Eq. 7.

$$\mathcal{L}_{\mathbf{v}} = \mathbb{E}_{k,\epsilon^k,(\mathbf{O}_t,\mathbf{A}_t^0)}[\|\mathbf{v}_t^k - \mathbf{v}_\theta(\mathbf{O}_t, \mathbf{A}_t^k, k)\|^2]. \tag{7}$$

Second, we adopt Euler angles for rotation representation instead of 6D rotation [17] or quaternions. Since the three dimensions of Euler angles are independent, this representation reduces the coupling in rotation regression, thereby further enhancing action stability. Notably, because our policy predicts relative actions [18], the discontinuities and Gimbal lock issues are naturally avoided.

2) Hardware Design: Effective force-based learning requires distinctive physical signals. When both the end-effector and the manipulated object are rigid, the variations in action adjustments resulting from force feedback are often subtle and easily drowned out by noise, significantly increasing the difficulty of policy learning. To mitigate this, we design a custom compliant fingertip. This hardware compliance ensures that contact with objects of any stiffness always produces distinctive reactivity signals, providing the network with clear, high-quality pairs of force feedback and action adjustments to learn from.

3) Controller Tuning: Since ImplicitRDP relies on the policy to learn reactive behaviors based on force, the low-level controller must provide precise position tracking rather than inherent compliance. Therefore, we modified the robot's default impedance controller, specifically tuning the integral gain parameters (k_i) of the PI controller in the Cartesian space. This adjustment ensures that the robot faithfully tracks the high-frequency adjustments commanded by the policy.

VI. Experiments Details

This section provides additional experimental details, quantitative results and qualitative analyses.

A. Experimental Setup Details

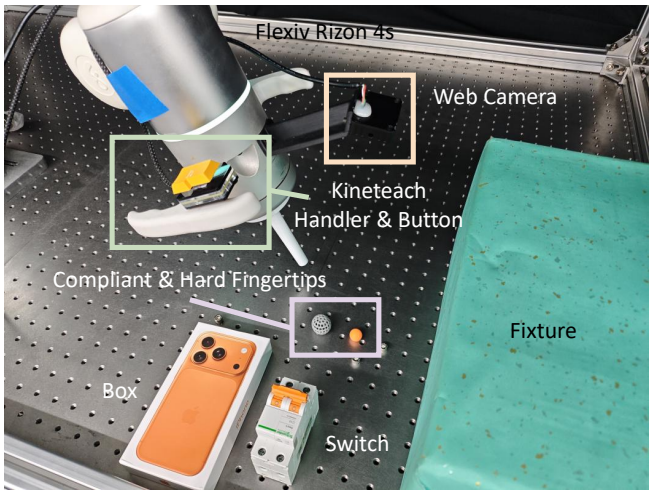

Fig. 4. **Hardware Setup.** The system utilizes a Flexiv Rizon 4s robot arm. A handle and button are mounted between the seventh joint and 6-axis F/T sensor for kinematic teaching. We also design a custom compliant fingertip to ensure distinctive reactivity signals during contact-rich interactions.

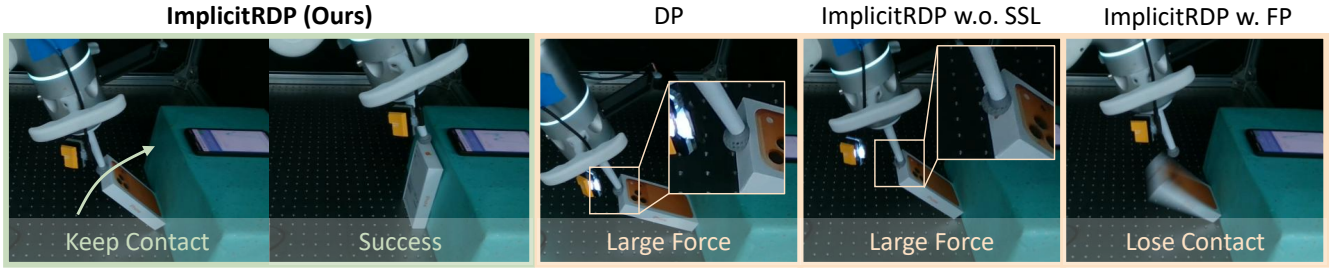

Fig. 5. **Box Flipping Task and Failure Cases.** The goal is to push a thin phone box against a fixture to flip it upright while maintaining a delicate force limit ($< 14N$). Vision-only or open-loop baselines lack closed-loop, force-based adjustment and apply excessive force, resulting in squeezing the fingertip. ImplicitRDP successfully utilizes the force feedback to complete the task safely.

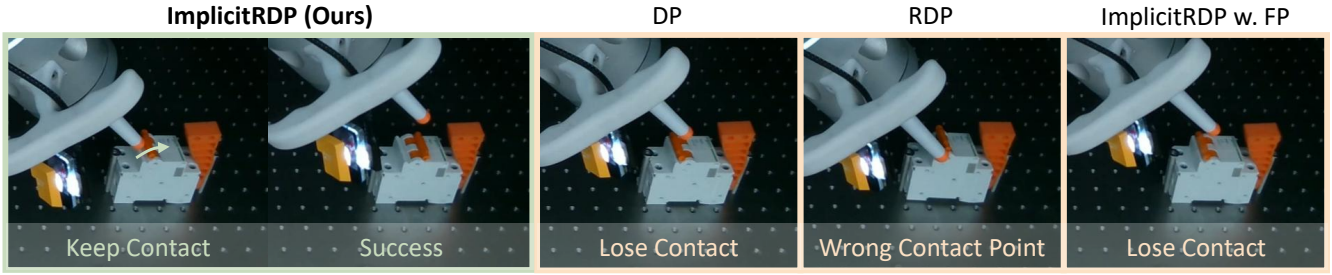

Fig. 6. **Switch Toggling Task and Failure Cases.** The robot has to locate and apply a specific force to toggle a circuit breaker switch. DP tends to initiate the toggling motion prematurely before the triggering force threshold is reached, while RDP often misses the precise contact location due to latent compression errors. ImplicitRDP accurately approaches the switch and perceives force to toggle the switch successfully.

For box flipping, we intentionally applied a relatively small contact force ($\sim 8N$) during demonstration collection. During evaluation, any case where the applied force exceeds $14N$ is considered a failure, which prevents the policy from completing the task through brute-force pushing.

B. Learning Stability Analysis

Supp. Tab. IV shows the impact of prediction parameterization and rotation representation. Results indicate that our choice of velocity-prediction consistently outperforms ϵ-prediction and sample-prediction, particularly in the box flipping task which requires continuous force application. Furthermore, using Euler angles proves superior to 6D rotation. The latter struggles with unstable actions in switch toggling due to worse noise tolerance caused by non-independent representation. In general, the combination of velocity-prediction and Euler angles achieves the highest stability and success rates across both tasks.

TABLE IV
ABLATION STUDY ON LEARNING STABILITY

Method	Box Flipping	Switch Toggling
ImplicitRDP (ϵ-prediction)	9/20	18/20
ImplicitRDP (sample-prediction)	7/20	14/20
ImplicitRDP (6D Rotation)	16/20	12/20
ImplicitRDP (Ours)	**18/20**	**18/20**

C. Failure Cases

In the box flipping task, the main failure mode of visual-only or open-loop methods is excessive force application. Without high-frequency force-based correction, the policy cannot maintain the narrow force range required to keep the box in contact while avoiding collapse of the compliant fingertip.

In the switch toggling task, DP often begins the upward motion before sufficient contact force has been accumulated, while RDP is more prone to spatial errors during the approach stage. ImplicitRDP benefits from direct access to both geometric context and high-frequency force feedback, allowing it to localize the contact point accurately and trigger the switch with the required force burst.

VII. Related Works

A. Imitation Learning with Force Input

Imitation Learning (IL) has emerged as a dominant paradigm in robotic manipulation [1, 2]. Notable works [3, 4, 5] have demonstrated exceptional scalability when utilizing visual observations as input, successfully tackling complex challenges ranging from deformable object manipulation to long-horizon assembly.

Recent research [14, 19, 20, 21, 22, 23] has begun to integrate force and torque measurements as additional modalities within the IL framework, which aims to enhance the model's understanding of contact states and exerted forces, and improve performance in contact-rich scenarios. Most of these works incorporate force/torque signals from the robot's Tool Center Point (TCP) directly into the policy input. However, there is a critical limitation in these approaches. Due to action chunking [1, 2], the control within each chunk remains effectively open-loop. This prevents the system from reacting to high-frequency force feedback in real-time, which consequently constrains performance in contact-rich tasks. To address this, Reactive Diffusion Policy (RDP) [6] proposed a hierarchical slow-fast architecture to achieve closed-loop control based on force signals. It utilizes a slow network

to predict latent action and a fast network to decode the latent combined with the latest force signals into executable actions.

Despite its effectiveness, the two-stage design of RDP introduces complexity in training and hyperparameter tuning and limits potential scalability. In this work, we propose ImplicitRDP, which achieves similar force-based closed-loop control but within a unified framework. Inspired by the success of causal modeling in domains such as large language models [24, 25] and video generation[26], we implement a structural slow-fast learning mechanism via causal attention. This allows for high-frequency force injection and end-to-end training as well.

B. Mitigate Modality Collapse

While RDP enforces attention to different modalities through its hierarchical architecture, standard end-to-end networks often struggle with modality collapse where models are unable to flexibly adjust the weights across multiple modalities.

To mitigate this, FACTR [7] introduced a curriculum learning strategy that blurs visual inputs during the early stages of training, guiding the network to autonomously learn how to weight different modalities. However, this approach adds training complexity and reduces generalizability across different tasks.

Alternatively, introducing future prediction as representation regularization has shown promise in robotic manipulation [27, 28, 29, 30, 31]. These works demonstrate that using future observation prediction as an auxiliary task significantly enhances policy robustness and representation quality. Building on this insight, recent work has applied future prediction to policies with force input. TA-VLA [12] employs future torque prediction as an additional objective, finding that it encourages the model to learn physically grounded internal representations and improves manipulation performance.

In this work, we also leverage the paradigm of future prediction but propose a novel target. Inspired by classical compliance controllers, we utilize the virtual target, which is calculated via adaptive compliance parameters [14], as the prediction objective. Unlike raw force/torque, the virtual target resides in the same space as the action and assigns varying weights to force signals based on their magnitude. We demonstrate that this approach provides tighter representation regularization, facilitating more effective utilization of different modalities.