ForceFlow: Learning to Feel and Act via Contact-Driven Flow Matching

Shuoheng Zhang[♦1], **Yifu Yuan**[♦,✉1], **Hongyao Tang**[1], **Yan Zheng**[1], **Qiaojun Yu**[3], **Pengyi Li**[1], **Guowei Huang**[2], **Helong Huang**[2], **Xingyue Quan**[2], **Jianye Hao**[✉1]

[1]Tianjin University, [2]Huawei Noah's Ark Lab, [3]Shanghai AI Lab
[♦]Equal contribution, [✉]Corresponding Author (yuanyf@tju.edu.cn, jianye.hao@tju.edu.cn)

Existing imitation learning methods enable robots to interact autonomously with the physical environment. However, contact-rich manipulation tasks remain a significant challenge due to complex contact dynamics that demand high-precision force feedback and control. Although recent efforts have attempted to integrate force/torque sensing into policies, how to build a simple yet effective framework that achieves robust generalization under multimodal observations remains an open question. In this paper, we propose **ForceFlow**, a force-aware reactive framework built upon flow matching. For contact-stage policy design, we investigate force signal fusion mechanisms and adopt an asymmetric multimodal fusion architecture that treats force as a global regulatory signal, combined with a joint prediction paradigm that enhances the policy's understanding of instantaneous force and historical information, thereby achieving deep coupling between force and motion. For task-level hierarchical decomposition, we divide manipulation into a vision-dominant approach stage (VLM-based pointing for target localization) and a touch-dominant interaction stage (force-driven contact execution), with a Vision-to-Force (V2F) handover mechanism that explicitly decouples spatial generalization from contact regulation. Experimental results across six real-world contact-rich tasks demonstrate that ForceFlow achieves a 37% success rate improvement over the strong baseline ForceVLA while maintaining significantly lower cost. Moreover, ForceFlow exhibits accurate force signal prediction and demonstrates superior performance in contact force self-regulation and zero-shot out-of-distribution (OOD) generalization.

Project: https://jokeresc.github.io/ForceFlow-page
Code: https://github.com/JokerESC/ForceFlow
Datasets: https://huggingface.co/datasets/JokerESC/ForceFlow
Contact: zsh_2024244119@tju.edu.cn

1 Introduction

In contact-rich manipulation, humans follow a natural staged perception process: vision is first used to localize the target and guide the hand toward it, while perceptual dominance shifts from vision to force and touch once physical contact is established Zhao et al. (2025). For tasks such as precision assembly, surface wiping, and peg-in-hole insertion He et al. (2025), successful execution requires not only reaching the correct interaction region, but also regulating contact forces through rapid local adjustment. Enabling robots to achieve a similar seeing-to-feeling modal transition is a core problem in general-purpose embodied manipulation Zhang et al. (2025); Yu et al. (2025).

Recent progress in imitation learning has enabled robots to acquire diverse manipulation skills from expert demonstrations Zhao et al. (2023); Chi et al. (2023); Kim et al. (2024). However, these methods remain fundamentally vision-centric. Although visual observations are effective for macro-level positioning, they

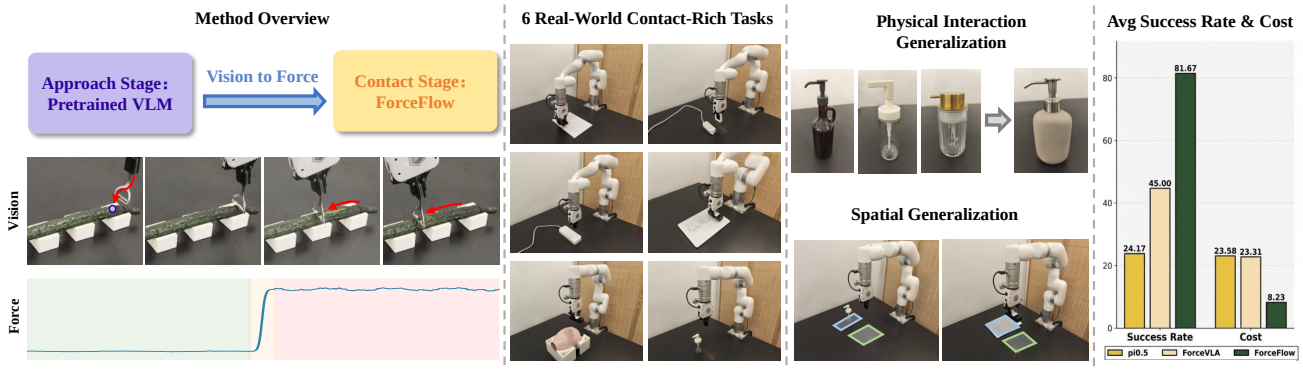


Figure 1 ForceFlow integrates V2F handover mechanism with flow matching to achieve precise robotic manipulation in contact-rich tasks, demonstrating significantly higher success rates and superior OOD generalization compared to baselines.

are often insufficient to independently support fine contact regulation due to occlusion, limited spatial resolution, and weak sensitivity to subtle contact events at the interaction interface Liu et al. (2025). Consequently, many vision-dominant policies can move the end-effector near the target, yet still struggle to reliably detect misalignment, regulate compliance, or adapt to changes in stiffness and friction during physical interaction.

Recent efforts have begun to incorporate force/torque signals into imitation learning frameworks Zhang et al. (2025); Liu et al. (2025); Zhao et al. (2025), but two key challenges remain. The first is multi-modal fusion under contact dynamics: force signals are low-dimensional, high-frequency, and strongly temporally structured, making them easily overshadowed by high-dimensional visual features during end-to-end training, which prevents policies from fully exploiting force information. The second is the dual generalization requirement: contact-rich manipulation simultaneously demands *physical interaction generalization* (stable execution under unseen stiffness, friction, and force responses) and *spatial generalization* (reaching the correct interaction region despite significant visual or positional shifts). Existing methods typically do not explicitly separate these two difficulties, leading to brittle performance when either physical properties or workspace layout changes.

To address these challenges, we propose **ForceFlow**, a force-aware reactive control framework built on Flow Matching (Lipman et al., 2022). We choose Flow Matching as the policy backbone because its deterministic ODE path generation provides lower inference latency and more stable trajectory output compared with diffusion sampling, which is critical for closed-loop contact control that must respond to high-frequency force feedback in real time. At the policy level, ForceFlow integrates temporal force history as a global regulatory signal into the network through asymmetric multimodal fusion, while retaining vision as a selective spatial reference, thereby effectively preventing force signals from being overshadowed by visual features. The policy jointly predicts actions and next-step contact force, encouraging the model to internalize the coupling between force and motion and enabling proactive force regulation. At the task level, we further introduce a **Vision-to-Force (V2F)** handover mechanism that divides fine manipulation into a vision-dominant approach stage and a force-dominant interaction stage. In the former, a Vision-Language Model localizes the target through a pointing mechanism; in the latter, once the end-effector reaches the VLM-derived 3D approach waypoint, the ForceFlow policy takes over for contact-rich execution. This design explicitly decouples spatial generalization from contact regulation, reflecting the natural transition from seeing to feeling in contact-rich manipulation. We systematically evaluate ForceFlow on six real-world contact-rich tasks spanning short-horizon contact establishment and continuous contact regulation. ForceFlow achieves an average success rate of 81.67%, outperforming the strong baseline ForceVLA by 37% while maintaining significantly lower contact force cost. Furthermore, ForceFlow exhibits stronger robustness under both unseen physical property variations and spatially out-of-distribution settings.

Our contributions are threefold: (1) We propose ForceFlow, a flow-matching policy for contact-rich

manipulation that integrates temporal force history through asymmetric multimodal fusion and jointly predicts motion and next-step contact force. (2) We introduce the V2F handover mechanism that separates vision-guided spatial localization from force-driven contact execution, following a staged transition from vision dominance to force dominance. (3) We validate the framework on six real-world benchmark tasks, demonstrating clear gains over strong baselines including ForceVLA and $\pi_{0.5}$ in task success rate, contact force fidelity, and robustness to both physical and spatial distribution shifts.

2 Related Work

Imitation Learning for Robotics Imitation Learning (IL) (Ross et al., 2011) acquires manipulation skills from expert demonstrations. Early Behavioral Cloning (BC) (Osa et al., 2018) struggles in complex long-horizon tasks, prompting the adoption of generative models. Methods like ACT (Zhao et al., 2023) and Diffusion Policy Chi et al. (2023) enable long-horizon smooth trajectory generation. More recently, Vision-Language-Action (VLA) models (e.g., OpenVLA (Kim et al., 2024), $\pi_{0.5}$ (Black et al., 2025), and GROOT N1 (Bjorck et al., 2025)) leverage large-scale pre-training for open-world generalization. Despite these advances, these state-of-the-art methods remain inherently vision-centric, treating manipulation primarily as a kinematic trajectory generation problem. By neglecting high-frequency force-torque dynamics, they often exhibit brittleness in contact-rich scenarios where visual feedback becomes unreliable, and success is dictated by the precise regulation of contact forces. To address this issue, we propose ForceFlow, which employs a flow matching architecture supplemented by historical force feedback and contact force prediction to ensure accurate force regulation in contact-rich tasks.

Contact-Rich Manipulation Contact-rich tasks require continuous force regulation (Zhao et al., 2026; Li et al., 2026), motivating the integration of force or tactile signals into policy learning. Existing approaches typically introduce force as additional observations (Adeniji et al., 2025; Wu et al., 2025) or infer it from vision and proprioception (Liu et al., 2025). Other works focus on multimodal fusion via contact-aware gating (He et al., 2025), residual refinement (Zhao et al., 2025), or hierarchical modeling (Xue et al., 2025). Several recent models (Yu et al., 2025; Cheng et al., 2025; Hao et al., 2025; Huang et al., 2025; Zhang et al., 2025) further incorporate force/torque information into VLA-style frameworks. However, a common bottleneck persists: high-dimensional visual features tend to dominate end-to-end optimization, causing low-dimensional yet temporally rich force signals to be underutilized, a phenomenon known as *modal masking* (Wang et al., 2020; Wu et al., 2022; Dong et al., 2025). Moreover, the diversity of sensor configurations across prior work complicates fair comparison. We therefore adopt the same observation setting as ForceVLA (multi-view RGB, end-effector pose, and 6D force/torque) and focus on suppressing modal masking through asymmetric fusion.

3 The ForceFlow Framework

To mitigate the critical issue where high-dimensional visual features readily overshadow subtle, sparse force signals, we propose **ForceFlow**, a force-aware reactive framework for contact-rich manipulation.

To further achieve robust spatial generalization, we integrate ForceFlow with a hierarchical **Vision-to-Force (V2F) handover mechanism**. The V2F mechanism decouples manipulation into a macro-level *Approach Stage* and a micro-level *Interaction Stage*. During the Approach Stage, a VLM handles target localization and guides the end-effector into the local interaction workspace; once the designated waypoint is reached, V2F executes the handover and ForceFlow takes over as a local expert for closed-loop contact control.

3.1 Preliminary

We consider the task of training a contact-rich policy $\pi_\theta(\mathbf{a}_t|\mathcal{O}_t)$ from a set of expert demonstrations $\mathcal{D}$. The multimodal observation at time t is defined as $\mathcal{O}_t = \{I_{\text{arm},t}, I_{\text{fix},t}, \mathbf{q}_t, \mathbf{F}_t^{\text{hist}}\}$, comprising dual-view RGB images, proprioception $\mathbf{q}_t \in \mathbb{R}^{d_q}$, and a force-torque history window $\mathbf{F}_t^{\text{hist}} \in \mathbb{R}^{H \times d_f}$. Here, H denotes the temporal history horizon and d_f the dimension of force feedback. Based on $\mathcal{O}_t$, the policy π_θ predicts an action chunk $\mathbf{a}_t$ of length H^a.

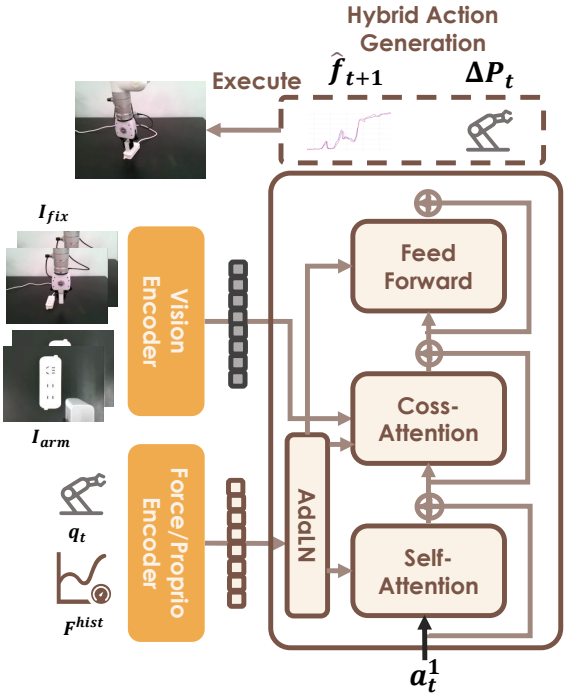


Figure 2 The ForceFlow Architecture.

3.2 Approach Stage: Pointing Mechanism and V2F Handover

As discussed in Section 1, contact-rich manipulation requires both spatial and physical generalization, yet these two capabilities demand fundamentally different mechanisms: the former relies on semantic reasoning, while the latter depends on high-frequency force feedback. Coupling both within a single end-to-end policy causes mutual performance degradation. We therefore propose the **Vision-to-Force (V2F) handover mechanism**, which explicitly divides the manipulation task into a vision-dominant Approach Stage and a force-dominant Interaction Stage, with stage switching triggered by a spatial arrival condition. This design delegates spatial generalization entirely to the VLM's zero-shot reasoning capability, and reserves physical generalization for ForceFlow's force-aware closed-loop regulation.

VLM Pointing Mechanism. During the Approach Stage, we leverage the VLM's open-world spatial reasoning to perform semantic target localization. Given a natural language instruction and a global camera view I_{fix}, the VLM predicts the pixel coordinates (u, v) of the target contact keypoint, isolating the spatial generalization problem as a pure semantic localization task. To equip the VLM with precise localization ability, we construct a visually grounded dataset by manually annotating the 2D coordinates $(u_{\text{gt}}, v_{\text{gt}})$ of target interaction regions on the initial frames of expert demonstrations, formatting them as VQA conversational pairs (image + language instruction $\rightarrow$ coordinates), and fine-tuning the VLM accordingly.

V2F Handover and Inference. At inference time, the VLM predicts the target pixel $(\hat{u}, \hat{v})$, which is deprojected into a 3D approach waypoint in the robot's base frame using depth information and camera intrinsics. A motion planner navigates the end-effector to this region. Upon arrival, the V2F handover is triggered (strictly based on positional criteria), transferring control to the ForceFlow policy for force-aware closed-loop regulation. Since global spatial reasoning has already been completed by the VLM during the Approach Stage, ForceFlow operates solely as a local expert focused on contact dynamics, without the need to model global visual features.

3.3 Contact Stage: Contact-Driven Flow Matching Policy

Once the VLM guides the end-effector to the 3D approach waypoint within the local interaction window, the system transitions to the ForceFlow policy for the Interaction Stage. ForceFlow focuses exclusively on

high-frequency contact dynamics without bearing the burden of global spatial reasoning. To this end, we design a policy combining an asymmetric multimodal fusion mechanism with a flow matching generative backbone. An overview of the ForceFlow architecture is shown in Figure 2.

Asymmetric Multimodal Fusion. In contact-rich manipulation, high-dimensional visual features can easily overshadow subtle, low-dimensional force signals during end-to-end training. To prevent this visual dominance, we partition the observation set $\mathcal{O}_t$ into two specialized pathways:

- **Force-Centric Vector Condition (c_{vec}):** To ensure that force signals persistently influence the generation process across all network layers, we encode the temporal force history $\mathbf{F}_t^{\text{hist}}$ (of length H) together with proprioception $\mathbf{q}_t$ into a unified global representation c_{vec}. This vector is injected into the Diffusion Transformer (DiT) via Adaptive Layer Normalization (AdaLN) (Peebles and Xie, 2023). By modulating feature statistics globally at every network layer, force signals act as a persistent regulatory constraint rather than being marginalized.
- **Visual Sequence Condition (c_{seq}):** Multi-view RGB observations are processed into spatial features. Rather than pooling them into a single vector, we preserve their temporal order to form a sequence condition c_{seq}, integrated via Cross-Attention. This allows the model to selectively attend to relevant spatio-temporal visual cues while remaining consistent with the global force state.

Hybrid Action Generation via Flow Matching. We employ Flow Matching (Lipman et al., 2022; Dong et al., 2024) to learn a deterministic velocity field over a **hybrid action space** $\mathbf{a}_t = \left[\Delta\mathbf{p}_t, \hat{\mathbf{f}}_{t+1}\right]$. By jointly predicting the motion command $\Delta\mathbf{p}_t$ and the expected contact force $\hat{\mathbf{f}}_{t+1}$, the policy is encouraged to internalize the correlation between robot motion and contact feedback.

We construct a linear probability path interpolating between expert hybrid actions $\mathbf{a}_t^0 \sim p_{\text{data}}(\mathbf{a})$ and a standard Gaussian prior $\mathbf{a}_t^1 \sim \mathcal{N}(0, \mathbf{I})$. For a flow time step $k \in [0, 1]$, the intermediate state is $\mathbf{a}_t^k = (1-k)\mathbf{a}_t^0 + k\mathbf{a}_t^1$. We parameterize the neural velocity field $v_\theta(\mathbf{a}_t^k, k, c_{\text{vec}}, c_{\text{seq}})$ to approximate the constant target drift $\mathbf{u}_t^k = \mathbf{a}_t^1 - \mathbf{a}_t^0$ using the following objective:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{k,\mathbf{a}_t^0,\mathbf{a}_t^1} \left\| v_\theta(\mathbf{a}_t^k, k, c_{\text{vec}}, c_{\text{seq}}) - \mathbf{u}_t^k \right\|^2 \tag{1}$$

During inference, we recover the noise-free hybrid action $\mathbf{a}_t^0$ by solving the ODE $d\mathbf{a}_t^k = v_\theta(\mathbf{a}_t^k, k, c_{\text{vec}}, c_{\text{seq}})dk$ from $k = 1$ to $k = 0$ using a deterministic numerical solver. At execution time, only the motion command $\Delta\mathbf{p}_t$ is sent to the robot controller; the force prediction $\hat{\mathbf{f}}_{t+1}$ serves as a joint training objective that encourages the network to internalize the coupling between force and motion, rather than directly participating in low-level force control. This ensures physically consistent trajectory generation while enhancing the policy's awareness of contact states through the joint prediction mechanism.

4 Experiments

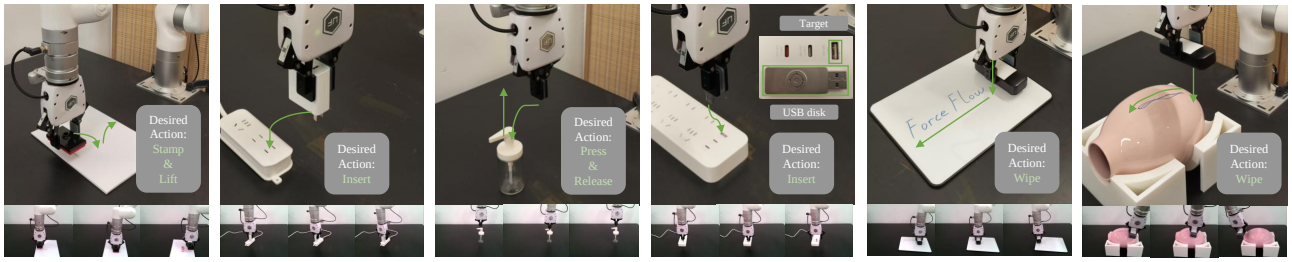
We evaluate ForceFlow on six contact-rich manipulation tasks to answer four key research questions:

RQ1 (Effectiveness): How does ForceFlow compare against state-of-the-art vision-centric and force-aware baselines (e.g. ForceVLA)?

RQ2 (Force Fidelity): Can the framework learn to regulate interaction forces and maintain consistency due to the deterministic nature of flow matching?

RQ3 (Generalization): Does the V2F handover mechanism enhance robustness in OOD scenarios?

RQ4 (Ablation): What are the individual contributions of force history and active force prediction?



(a) Stamp (b) Plug (c) Press (d) Insert (e) Clean WB (f) Clean Vase

Figure 3 Contact-Rich Task Suite. We evaluate ForceFlow on six diverse manipulation tasks categorized into: (a)-(d) **Short-Horizon Contact Tasks** requiring precise contact establishment; and (e)-(f) **Continuous Contact Tasks** demanding consistent normal force tracking along trajectories.

Table 1 Quantitative Comparison of Task Success Rates (SR). We report the success rate (%) over 20 trials for each task. **Bold** indicates the best performance.

Method	Stamp	Plug	Press	Insert	Clean WB	Clean Vase	Avg.
$\pi_{0.5}$	0%	60%	30%	45%	10%	0%	24.17%
ACT	0%	30%	5%	0%	15%	0%	8.33%
Diffusion Policy	0%	40%	20%	50%	75%	0%	30.83%
ForceVLA	20%	70%	65%	15%	100%	0%	45%
ForceFlow (w/o Force)	20%	75%	0%	40%	100%	30%	44.17%
ForceFlow (Ours)	**85%**	**90%**	**90%**	**60%**	**100%**	**65%**	**81.67%**

Additionally, we qualitatively demonstrate ForceFlow's capability in continuous force regulation through a cucumber peeling task.

4.1 Experimental Setup

Task Suite (Figure 3). We select six tasks categorized into: (1) short-horizon contact tasks that require precise modulation to establish contact (Stamping, Plug/USB Insertion, Press Button); and (2) continuous contact tasks which demand consistent normal force (Clean Whiteboard, Clean Vase). Detailed task settings are provided in Appendix A.

We employ two complementary metrics:

Success Rate (SR): The percentage of successful trials ($N = 20$) based on task-specific completion criteria (e.g., successful insertion or clean wiping).

Force Fidelity (MAE Cost): To quantify physical compliance and stability, we measure the deviation between the policy's interaction force and the expert's reference. We define the metric $\mathcal{J}_{\text{force}}$ as:

$$\mathcal{J}_{\text{force}} = \frac{1}{N} \sum_{i=1}^{N} \left| \hat{F}^{(i)}_{\text{policy}} - F_{\text{expert}} \right| \tag{2}$$

where $N = 20$ is the number of trials. The force statistic $\hat{F}$ is task-dependent: for **Short-Horizon Contact Tasks**, it denotes the *peak contact force* ($\max_t \|\mathbf{f}_t\|$) to capture impact intensity; for **Continuous Contact Tasks**, it represents the *average effective force* calculated only during the contact phase (where $\|\mathbf{f}_t\| > 5\text{N}$) to evaluate tracking consistency.

Table 2 Force Fidelity Analysis (MAE Cost). The table reports the MAE between the executed force and the expert demonstration force. The unit is Newton (N). **Bold** indicates the lowest cost (highest fidelity). **Lower is better.**

Method	Stamp	Plug	Press	Insert	Clean WB	Clean Vase	Avg.
$\pi_{0.5}$	31.99	21.41	17.39	50.89	11.93	7.87	23.58
ACT	31.86	25.54	31.81	38.71	11.91	30.45	28.38
Diffusion Policy	32.26	15.79	24.86	23.85	8.22	23.56	21.42
ForceVLA	30.03	9.59	30.94	37.82	20.16	11.29	23.31
ForceFlow (w/o Force)	30.03	13.36	37.50	34.75	7.16	13.24	22.67
ForceFlow (Ours)	**10.61**	**3.58**	**5.03**	**21.79**	**4.59**	**3.76**	**8.23**

4.2 RQ1: Effectiveness in Contact-Rich Tasks

We compare ForceFlow against $\pi_{0.5}$ (Black et al., 2025), ACT (Zhao et al., 2023), Diffusion Policy (Chi et al., 2023), ForceVLA (Yu et al., 2025), and the ablation baseline ForceFlow (w/o Force). As presented in Table 1, ForceFlow achieves an average success rate of **81.67%**, significantly outperforming the best baseline (ForceVLA, 45%). We analyze this performance gain through three functional paradigms of force perception:

Resolving Visual Ambiguity (Force as Primary Modality). In tasks like *Stamping* and *Press Button*, vision-centric baselines collapse (0–30% SR) because they cannot perceive environmental properties such as paper stack thickness (ranging from 1 to 50 sheets) or varying spring constants. They tend to converge to a mean terminal height, leading to either insufficient pressure or excessive collision. In contrast, ForceFlow exploits the 10-step force history to detect the exact onset of resistance. By treating force-torque profiles as the ground truth for state transitions, it achieves high success rates (85% and 90%) regardless of visual uncertainties in object height or stiffness.

Navigating Geometric Constraints (Force as Auxiliary Modality). For *Plug* and *USB Insertion* involving sub-millimeter tolerances, visual feedback often erroneously indicates completion when the connector is actually stalled by friction. ForceFlow achieves 60% SR on the challenging USB task (where baselines largely fail) by perceiving the reactive torque τ generated by misalignment. Combined with active force prediction, the policy generates subtle sliding and wiggling motions to feel the opening and dynamically align the components, overcoming geometric jamming that vision-only policies cannot resolve.

Stability in Continuous Interaction (Force as Regulatory Modality). The advantage of ForceFlow is most pronounced in tracking non-linear geometries. While ForceVLA performs well on the planar *Whiteboard* (100%), it fails completely on the curved *Clean Vase* (0%) due to the continuously changing surface normals. ForceFlow utilizes active force prediction as a regulatory mechanism to proactively compliance-match the surface curvature. This allows it to maintain a stable interaction envelope on irregular 3D surfaces, achieving a 65% success rate.

4.3 RQ2: Force Regulation and Fidelity (Cost Analysis)

To answer RQ2, we conduct an in-depth analysis of the model's force control behavior from three perspectives: statistical force deviation (Force Cost), stability across trials, and the consistency between predicted and real contact forces during single execution. ForceFlow significantly reduces Force Cost across all tasks. The average Force Cost drops from the 20–30 N range of vision-dominant models to 8.23 N. The reduction is particularly notable ($> 50\%$) in instantaneous or semi-instantaneous contact tasks such as Stamp, Plug, and Press. Although ForceFlow (w/o Force) can still complete operations in some tasks, its Force Cost is significantly higher than the full model, highlighting that Success Rate alone is an

incomplete proxy for force quality, as geometrically correct actions can still deviate severely from expert force distributions.

Stability Across Trials (20 Test Statistics) Figure 5 illustrates the Maximum Contact Force and Average Effective Force for ForceFlow and ForceFlow (w/o Force) across 20 independent trials. ForceFlow's force output is highly consistent across trials. Fluctuations are minimal, indicating that the model has learned a *task-dependent force regulation strategy* rather than memorizing a single trajectory. Conversely, the model without force feedback exhibits obvious instability. ForceFlow (w/o Force) shows distinct issues: ForceFlow (w/o Force) exhibits severe instability (e.g., violent oscillations and abnormal spikes) in contact establishment, stably excessive forces in continuous tracking, and clear open-loop characteristics (e.g., insensitivity to paper thickness in Stamping). This demonstrates that relying solely on vision and proprioception prevents the model from forming a stable closed-loop mechanism for contact force regulation.

Consistency of Force Prediction in Single Execution Figure 6 displays the alignment between ForceFlow's predicted contact force curve (red) and the real measured force (blue) across six representative tasks. In short-contact tasks, predicted forces track real forces synchronously during both contact establishment and release. This suggests that the model can infer the impending contact state via force history, vision, and proprioception. In continuous tasks (Clean Whiteboard/Clean Vase), the predicted force smoothly adjusts following surface changes, without high-frequency jitter, proving the model learns state-based continuous force regulation.

4.4 RQ3: Generalization to OOD Environments

This section evaluates the robustness of ForceFlow across complex real-world variations by decoupling visual localization from force regulation. We analyze two critical dimensions of generalization: (1) **Physical Interaction Generalization**, where contact objects or tools change but spatial locations remain fixed, and (2) **Spatial Generalization**, where target objects are placed in OOD spatial locations.

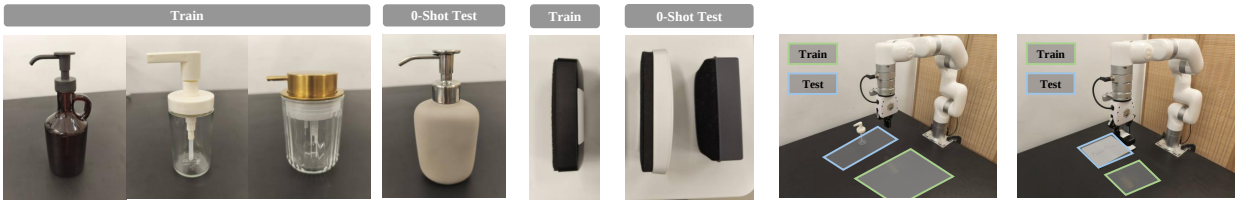


Figure 4 OOD Evaluation Setup. (a) **Physical Interaction Generalization:** Unseen objects (e.g., distinct bottles, different erasers) used for zero-shot testing. (b) **Spatial Generalization:** The testing workspace (marked "Test") is spatially disjoint from the training distribution (marked "Train"), requiring V2F for localization.

Physical Interaction Generalization: Zero-Shot Adaptability. We evaluate physical adaptability by replacing training tools with unseen variants without fine-tuning (see Figure 4a). We conducted 10 independent trials per task using objects with distinct physical properties: for **Clean Vase** and **Whiteboard**, we introduced two novel erasers with significantly different stiffness and thickness; for **Press Button**, we utilized a new bottle with a distinct height and spring trigger threshold.

As shown in Table 3, vision-centric baselines fail entirely when tool friction or object elasticity changes, as they cannot perceive high-frequency physical feedback. In contrast, ForceFlow maintains high success rates by adaptively adjusting contact forces through its temporal force perception and active prediction mechanisms. By encoding F_{hist}, the model perceives contact states and achieves proactive compliance even when the physical interaction parameters deviate from the training distribution.

Spatial Generalization: Hierarchical Grounding. To evaluate robustness against spatial shifts, we configure a testing workspace that is spatially disjoint from the training distribution, as illustrated in Figure 4(b). To tackle large spatial variations, we implement V2F, specifically utilizing Embodied-R1 Yuan et al. (2025) as the high-level planner. This hierarchical strategy allows the system to localize targets in OOD positions significantly outside the training distribution.

Results in Table 4 highlight the necessity of V2F integration. Without semantic guidance, low-level policies (including standalone baselines) fail to locate objects in OOD regions. By utilizing V2F, the system predicts semantic pixel coordinates from the global view I_{fix} to guide initial movement. Once the end-effector enters the local interaction window, ForceFlow takes over for high-frequency regulation. This separation of concerns ensures that the model handles diverse backgrounds and spatial shifts while maintaining the determinism required for contact-intensive tasks.

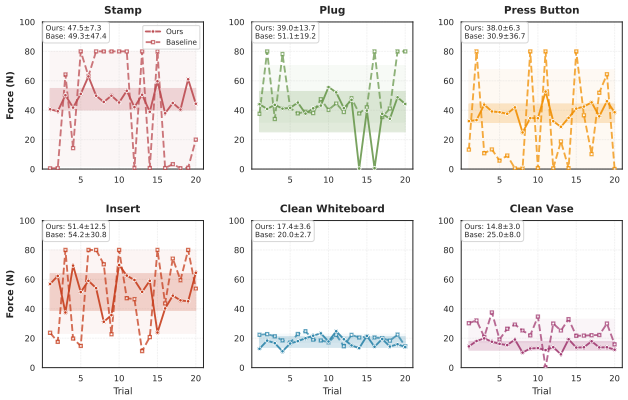


Figure 5 Statistical Force Distribution. Maximum or average interaction forces for six representative tasks across 20 independent trials. Error bars/shading indicate standard deviation, highlighting the stability of the learned force regulation strategy.

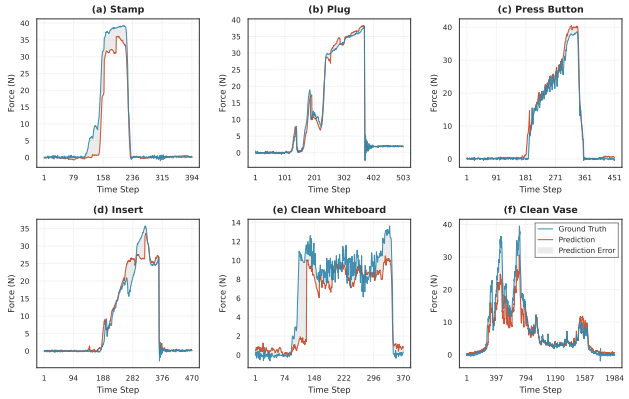


Figure 6 Comparison of Predicted and Measured Forces. Alignment between the predicted (red) and measured ground truth (blue) contact forces. Shaded areas represent prediction error, demonstrating high temporal fidelity during interaction.

Table 3 Physical Interaction Generalization (SR %).

Method	Press	Clean WB	Clean Vase
$\pi_{0.5}$	0%	0%	0%
ACT	0%	0%	0%
Diffusion Policy	0%	0%	0%
ForceVLA	40%	90%	0%
ForceFlow(Ours)	**80%**	**100%**	**60%**

Table 4 Spatial Generalization (SR %).

Method	Press	Plug	Clean WB
$\pi_{0.5}$	0%	0%	0%
ACT	0%	0%	0%
Diffusion Policy	0%	0%	0%
ForceVLA	0%	0%	0%
ForceFlow	0%	0%	0%
ForceFlow + V2F	**40%**	**10%**	**50%**

Table 5 Ablation Study on the Stamp Task.

Variant	SR (%)	Cost (N)
w/o Force History(1-step)	55%	15.50
w/o Force Prediction	80%	12.52
w/o Both(1-step + No Pred)	40%	18.21
ForceFlow(Full)	**85%**	**10.61**

4.5 RQ4: Ablation Study

This section investigates the individual contributions of temporal force history and active force prediction to the overall performance and stability of ForceFlow. We evaluate three ablation variants across our task suite to quantify the impact of our force-centric design choices.

As detailed in Table 5, the results reveal different roles for each component. Temporal force history proves decisive for task success. Replacing the 10-step window with a single reading causes a drastic SR drop (85% to 55%). This performance degradation is due to the fact that instantaneous force measurement is inherently susceptible to interference from contact instability or noise from the sensor itself. This instability is visually corroborated in Figure 7. The 1-step variant exhibits significant high-frequency

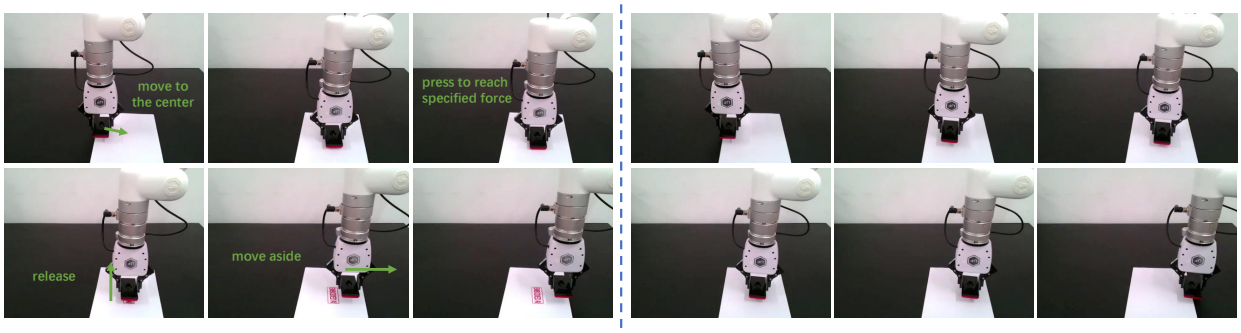


Figure 7 Qualitative comparison of force regulation in the Stamping task. The upper part illustrates the stable execution of ForceFlow (Full), while the lower part shows the w/o Force History (1-step) variant. Without temporal context, the single-step policy suffers from severe oscillations and force spikes due to sensor noise, whereas ForceFlow maintains a stable and smooth contact profile.

jitter and abrupt force spikes during contact, indicating an inability to distinguish between signal noise and true physical feedback. In contrast, ForceFlow leverages the historical window to smooth out these fluctuations, resulting in a stable interaction. Without temporal context to filter these fluctuations, the policy fails to reliably estimate the true interaction state. In contrast, active force prediction acts as a compliance regularizer. Removing the prediction head results in a marginal SR decline (80%) but a notable increase in Force Cost (10.61N to 12.52N), suggesting that it smooths the interaction rather than dictating feasibility. The significant degradation when both are removed (40% SR) confirms their synergistic necessity: history ensures correct decision-making, while prediction refines execution quality.

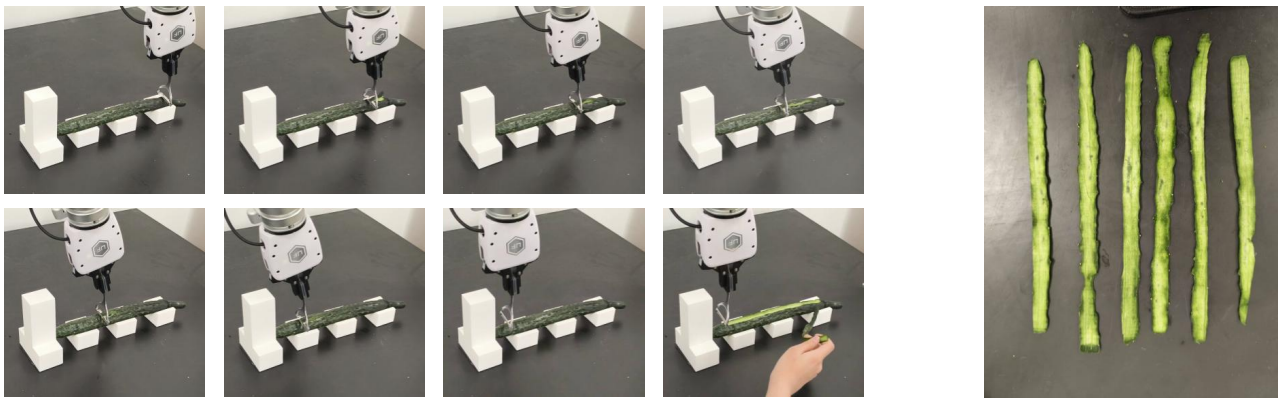


Figure 8 Qualitative results of cucumber peeling. (Left) ForceFlow proactively regulates force to track the contour. (Right) Uniform, unbroken peeled strips evidence a consistent cutting depth.

4.6 Qualitative Analysis: Continuous Force Regulation in Cucumber Peeling

We qualitatively evaluate ForceFlow on a cucumber peeling task (Fig. 8, Left), which requires maintaining precise normal force along a varying-stiffness surface stabilized on V-groove supports.

Visual ambiguity often causes vision-centric baselines to miss contact or apply excessive force, jamming the tool. Conversely, ForceFlow utilizes temporal force history to accurately detect the onset of resistance upon initial contact, establishing optimal cutting depth without over-pressing.

During sliding, the model leverages active force prediction to dynamically adjust downward pressure, seamlessly adapting to geometric variations (e.g., local bumps or tapering ends) to track the contour.

The resulting uniform, unbroken peeled strips (Fig. 8, Right) evidence ForceFlow's ability to maintain a stable interaction envelope and consistent cutting depth.

5 Conclusion

In this paper, we propose ForceFlow, a force-aware reactive framework based on the Flow Matching method for contact-rich manipulation tasks. Through the temporal force history and a joint prediction mechanism, ForceFlow enables robots to achieve deep synergy between visual perception and physical contact dynamics. The Vision-to-Force (V2F) mechanism effectively decouples fine manipulation into a VLM-guided approach stage and a force-dominant interaction stage, ensuring the system exhibits excellent robustness against spatial and physical distribution shifts. Experimental results on six real-world tasks demonstrate that ForceFlow improves the success rate by 37% compared with state-of-the-art baselines, while demonstrating exceptional zero-shot OOD generalization capability. Despite these promising results, ForceFlow still has certain limitations. For instance, the current framework relies on high-fidelity force/torque sensors, which may restrict its deployment on low-cost robotic platforms. Furthermore, future research can develop an adaptive V2F switching framework to further enhance the overall manipulative dexterity of embodied robots in unstructured environments.

References

Ademi Adeniji, Zhuoran Chen, Vincent Liu, Venkatesh Pattabiraman, Raunaq Bhirangi, Siddhant Haldar, Pieter Abbeel, and Lerrel Pinto. Feel the force: Contact-driven learning from humans. *arXiv preprint arXiv:2506.01944*, 2025.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, brian ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *9th Annual Conference on Robot Learning*, 2025. URL https://openreview.net/forum?id=vlhoswksBO.

Zhengxue Cheng, Yiqian Zhang, Wenkang Zhang, Haoyu Li, Keyu Wang, Li Song, and Hengdi Zhang. Omnivtla: Vision-tactile-language-action model with semantic-aligned tactile sensing. *arXiv preprint arXiv:2508.08706*, 2025.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems*, 2023. URL https://doi.org/10.15607/RSS.2023.XIX.026.

Zibin Dong, Yifu Yuan, Jianye Hao, Fei Ni, Yi Ma, Pengyi Li, and Yan Zheng. Cleandiffuser: An easy-to-use modularized library for diffusion models in decision making. In *NeurIPS*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/9e08a1db869a9646418e3371b24c6ae6-Abstract-Datasets_and_Benchmarks_Track.html.

Zibin Dong, Yicheng Liu, Yinchuan Li, Hang Zhao, and Jianye HAO. Conditioning matters: Training diffusion policies is faster than you think. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=pKQcmLHoGG.

Peng Hao, Chaofan Zhang, Dingzhe Li, Xiaoge Cao, Xiaoshuai Hao, Shaowei Cui, and Shuo Wang. Tla: Tactile-language-action model for contact-rich manipulation. *arXiv preprint arXiv:2503.08548*, 2025.

Zihao He, Hongjie Fang, Jingjing Chen, Hao-Shu Fang, and Cewu Lu. FoAR: Force-aware reactive policy for contact-rich robotic manipulation. In *ICRA 2025 Workshop: Beyond Pick and Place*, 2025. URL https://openreview.net/forum?id=cbjluXVaJz.

Jialei Huang, Shuo Wang, Fanqi Lin, Yihang Hu, Chuan Wen, and Yang Gao. Tactile-vla: unlocking vision-language-action model's physical knowledge for tactile generalization. *arXiv preprint arXiv:2507.09160*, 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=ZMnD6QZAE6.

Yang Li, Hongru Jiang, Junjie Xia, Hongquan Zhang, Jinda Du, Yunsong Zhou, Jia Zeng, Ce Hao, Jieji Ren, Qiaojun Yu, et al. Forcevla2: Unleashing hybrid force-position control with force awareness for contact-rich manipulation. *arXiv preprint arXiv:2603.15169*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *CoRR*, abs/2210.02747, 2022. URL https://doi.org/10.48550/arXiv.2210.02747.

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.

Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J. Andrew Bagnell, Pieter Abbeel, and Jan Peters. An algorithmic perspective on imitation learning. *Found. Trends Robotics*, 7(1-2):1–179, 2018. URL https://doi.org/10.1561/2300000053.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

Stéphane Ross, Geoffrey J. Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*, pages 627–635, 2011. URL http://proceedings.mlr.press/v15/ross11a/ross11a.pdf.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, pages 12692–12702, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_What_Makes_Training_Multi-Modal_Classification_Networks_Hard_CVPR_2020_paper.html.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J. Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *ICML*, pages 24043–24055, 2022. URL https://proceedings.mlr.press/v162/wu22d.html.

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.

Han Xue, Jieji Ren, Wendi Chen, Gu Zhang, Fang Yuan, Guoying Gu, Huazhe Xu, and Cewu Lu. Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation. In *ICRA 2025 Workshop: Beyond Pick and Place*, 2025. URL https://openreview.net/forum?id=zRhjjLGUAp.

Jiawen Yu, Hairuo Liu, Qiaojun Yu, Jieji Ren, Ce Hao, Haitong Ding, Guangyu Huang, Guofan Huang, Yan Song, Panpan Cai, Wenqiang Zhang, and Cewu Lu. ForceVLA: Enhancing VLA models with a force-aware moe for contact-rich manipulation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=2845H8Ua5D.

Yifu Yuan, Haiqin Cui, Yaoting Huang, Yibin Chen, Fei Ni, Zibin Dong, Pengyi Li, Yan Zheng, and Jianye Hao. Embodied-r1: Reinforced embodied reasoning for general robotic manipulation. *arXiv preprint arXiv:2508.13998*, 2025.

Zongzheng Zhang, Haobo Xu, Zhuo Yang, Chenghao Yue, Zehao Lin, Huan ang Gao, Ziwei Wang, and Hao Zhao.

Elucidating the design space of torque-aware vision-language-action models. In *9th Annual Conference on Robot Learning*, 2025. URL https://openreview.net/forum?id=HAmi1X11BO.

Ruiteng Zhao, Wenshuo Wang, Yicheng Ma, Xiaocong Li, Francis EH Tay, Marcelo H Ang Jr, and Haiyue Zhu. Fd-vla: Force-distilled vision-language-action model for contact-rich manipulation. *arXiv preprint arXiv:2602.02142*, 2026.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023. URL https://openreview.net/forum?id=e8Eu1lqLaf.

Zifan Zhao, Siddhant Haldar, Jinda Cui, and Lerrel Pinto. Touch begins where vision ends: Generalizable policies for contact-rich manipulation. In *Second Workshop on Out-of-Distribution Generalization in Robotics at RSS 2025*, 2025. URL https://openreview.net/forum?id=vbW7BVKAeb.

A Detailed Experimental Setup

Table 6 Detailed Training Hyperparameters and Model Architecture

Category	Hyperparameter	Value	Description
Training	Batch Size	64	Per-device batch size
	Max Steps	100,000	Total training steps
	Precision	bf16-mixed	Mixed precision training
	Gradient Accumulation	1	Accumulation batches
	Checkpoint Interval	5,000	Steps between checkpoints
	Random Seed	0	Reproducibility seed
Data	Image Resolution	320×240	RGB image size
	Observation Horizon (H_o)	2	Number of observation frames
	Action Horizon (H_a)	64	Action sequence length
	Force History (H_{force})	10	Historical force steps
	State Dimension	7	6D pose + 1D gripper
	Force Dimension	6	3D force + 3D torque
	Action Dimension	13	6D pose + gripper + 6D force
	Data Workers	8	Parallel data loading threads
DiT-1D	Model Dimension	384	Base feature dimension
	Attention Heads	6	Multi-head attention
	Transformer Depth	12	Number of transformer blocks
	Vector Embedding	256	Low-dim + force embedding
	Sequence Embedding	512	Dual-view image embedding
	Head Type	MLP	Output head architecture
	Cross-Attention	Yes	Visual-to-action attention
	AdaLN	Yes	Adaptive Layer Normalization
	Timestep Embedding	Fourier	Scale: 0.2 (untrainable)
	Activation	SiLU	Swish activation function
Condition Encoder	Visual Backbone	ResNet-18	Pre-trained, dual views
	Image Embedding	2×256	Arm view + fixed view
	Low-dim Encoder	2-layer MLP	Hidden: $64 \to 64$
	Force Encoder	2-layer MLP	$(H_{\text{force}} \times 6) \to 128$
	Total Vector Dim	256	$64 \times H_o + 128$
	Dropout	0.0	No dropout in encoder
Diffusion	Algorithm	Flow Matching	Continuous-time diffusion
	Sampling Steps	Variable	Inference-time adjustable
	Normalization	MinMax	State and action normalization
	Image Normalization	$[-1, 1]$	Mean: 0.5, Std: 0.5

A.1 Task Description and Challenges

We evaluate our framework on six real-world contact-rich manipulation tasks, categorized into short-horizon contact establishment and continuous contact maintenance.

Short-Horizon Contact Tasks. These tasks require establishing contact with appropriate target forces; insufficient or excessive force leads to failure.

Stamping (Visual Ambiguity): The robot stamps stacks of paper of varying thicknesses (1 sheet to $\sim$5 cm). Since stack height is visually indistinguishable, the policy must rely on real-time force feedback to regulate vertical motion and apply optimal stamping pressure.

Plug Insertion (Spatial Randomness): The robot inserts a power plug into a socket. Initial poses are randomized, and the socket is subject to random deflections. Success requires coarse visual alignment followed by force-guided insertion to overcome contact friction.

USB Insertion (Tight Tolerance): Similar to the plug task but with sub-millimeter geometric tolerances. The policy must utilize subtle torque cues to "feel" the narrow opening of the USB-A port and complete insertion without jamming.

Press Button (Cross-Object): The robot interacts with three distinct spring-loaded sanitizer bottles. Due to significant variations in height and design, the policy must adaptively regulate contact force to complete a full press-and-release cycle.

Continuous Contact Tasks. These tasks require maintaining stable interaction forces over a longer time horizon.

Clean Whiteboard (Planar Constant Force): The robot wipes markings off a fixed whiteboard surface. This requires maintaining a continuous normal force along a 2D planar trajectory to ensure effective cleaning.

Clean Vase (Geometric Complexity): The robot cleans marks off a curved 3D surface. Although the mount is fixed, the vase may be subject to perturbations, leading to variations in contact normals. This task evaluates force regulation against non-linear, unknown geometries.

A.2 Data Collection and Training Setup

The hardware setup consists of a 6-Dof UFactory xArm6 robotic arm with a 1-Dof gripper, an Intel RealSense L515 camera for the global view, and a D435 camera for the wrist view. Expert demonstrations were collected using two methods: a 3Dconnexion SpaceMouse and a Meta Quest Pro VR headset. The hardware platform is illustrated in Fig. 9.

For each task, we collected 50–100 expert demonstrations via teleoperation at a control frequency of 30 Hz. The dataset includes synchronized dual-view 320×240 RGB observations, 7D proprioceptive states (6D pose and 1D gripper), and 10-step force-torque histories. Our policy is trained using a high-capacity Diffusion Transformer (DiT) backbone to regress the continuous rectified velocity field.

The models were trained on a compute node equipped with 4$\times$ NVIDIA GeForce RTX 4090 GPUs (24 GB VRAM each), 48 physical CPU cores, and 283 GB system RAM. We employed the AdamW optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$, weight decay = 0.01) with a cosine learning rate schedule starting from 1×10^{-4}. Training utilized a batch size of 64 with a gradient accumulation of 1 (effective batch size 64), completing 100,000 steps in approximately 8–10 hours per task. To ensure training stability, we employed bfloat16 mixed precision with gradient clipping ($\|\nabla\| = 1.0$). The model architecture and key hyperparameters are detailed in Table 6.

A.3 Inference and Evaluation Protocol

During inference, actions are generated by deterministically integrating the learned velocity field using an Ordinary Differential Equation (ODE) solver. The policy predicts a 64-step action chunk, executing the first 32 steps before replanning to ensure temporal smoothness. Each task is evaluated over 20 independent trials with randomized initial conditions. Task success is determined by task-specific completion criteria (e.g., full insertion or a clear imprint), while force fidelity is assessed by measuring deviations from expert force profiles.

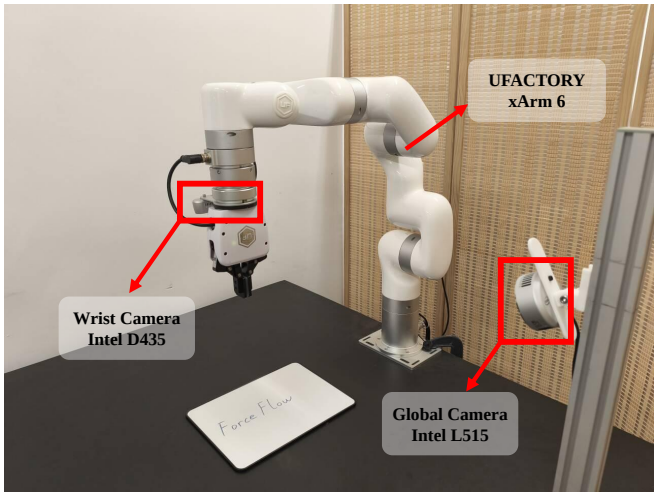


Figure 9 Hardware System

B Mathematical Formulation

This section provides the formal mathematical grounding for the **ForceFlow** framework, detailing the construction of the probability path, the asymmetric conditioning mechanism, and the joint optimization for active compliance.

B.1 Linear Probability Path and Velocity Field

To ensure the trajectory smoothness and determinism required for stable physical contact, ForceFlow employs Continuous Rectified Flow (CRF). Consistent with Section 3.1, we define the hybrid action space at task time step t as $\mathbf{a}_t = [\Delta\mathbf{p}_t, \hat{\mathbf{f}}_{t+1}] \in \mathcal{A}$, where $\Delta\mathbf{p}_t \in \mathbb{R}^{d_p}$ represents the motion command (pose increments and gripper status) and $\hat{\mathbf{f}}_{t+1} \in \mathbb{R}^{d_f}$ represents the target contact force for active compliance.

We construct a linear probability path between the expert action distribution $\mathbf{a}_t^0 \sim p_{\text{data}}$ and a standard Gaussian prior $\mathbf{a}_t^1 \sim \mathcal{N}(0, \mathbf{I})$. For a flow time step $k \in [0, 1]$ (superscript), the intermediate state $\mathbf{a}_t^k$ is defined as:

$$\mathbf{a}_t^k = (1-k)\mathbf{a}_t^0 + k\mathbf{a}_t^1 \tag{3}$$

This interpolation defines a rectified flow with a constant ground-truth velocity field $\mathbf{u}_t^k$:

$$\mathbf{u}_t^k(\mathbf{a}_t^k, k) = \frac{d\mathbf{a}_t^k}{dk} = \mathbf{a}_t^1 - \mathbf{a}_t^0 \tag{4}$$

B.2 Asymmetric Multimodal Fusion Architecture

To prevent high-dimensional visual features from drowning out sparse force signals, we implement an asymmetric fusion strategy within the DiT blocks. The conditioning comprises a vector condition c_{vec} (derived from proprioception $\mathbf{q}_t$ and force history $\mathbf{F}_t^{\text{hist}}$) and a sequence condition c_{seq} (derived from multi-view visual observations).

Global Regulation via AdaLN. The vector condition c_{vec} modulates the feature statistics of each Transformer block via AdaLN:

$$\text{AdaLN}(h, c_{\text{vec}}) = \gamma(c_{\text{vec}}) \odot \text{LayerNorm}(h) + \beta(c_{\text{vec}}) \tag{5}$$

where h is the hidden state, and γ, β are scale and shift parameters regressed from c_{vec}. This ensures haptic feedback provides a global constraint on the generation process, preventing the "modality ignoring" issue common in naive fusion.

Spatio-Temporal Grounding via Cross-Attention. Visual sequences c_{seq} are integrated via Multi-Head Cross-Attention (MHCA) to provide spatial grounding. Let h be the intermediate action tokens acting as queries; the keys K and values V are projected from c_{seq}:

$$Q = hW_Q, \quad K = c_{\text{seq}}W_K, \quad V = c_{\text{seq}}W_V \tag{6}$$

$$\text{MHCA}(h, c_{\text{seq}}) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{7}$$

where d_k is the dimension per attention head and W_Q, W_K, W_V are learnable projection matrices.

B.3 Training and Inference

The neural velocity field v_θ is trained to jointly predict the motion and force components while optimizing the flow-matching objective.

Training Objective with Active Compliance. The DiT decoding head outputs a hybrid velocity prediction $v_\theta(\mathbf{a}_t^k, k, c_{\text{vec}}, c_{\text{seq}})$. The training objective minimizes the mean squared error against the target velocity $\mathbf{u}_t^k$:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{k\sim[0,1],\mathbf{a}_t^0,\mathbf{a}_t^1}\left[\|v_\theta(\mathbf{a}_t^k, k, c_{\text{vec}}, c_{\text{seq}}) - (\mathbf{a}_t^1 - \mathbf{a}_t^0)\|^2\right] \tag{8}$$

This joint prediction compels the model to internalize the causal relationship between movement and physical feedback, enabling proactive compliance during contact.

Deterministic Inference via ODE Solving. During inference, actions are generated by solving the Ordinary Differential Equation (ODE) from noise ($k = 1$) to data ($k = 0$):

$$d\mathbf{a}_t^k = v_\theta(\mathbf{a}_t^k, k, \mathcal{O}_t)dk \tag{9}$$

where $\mathcal{O}_t = \{c_{\text{vec}}, c_{\text{seq}}\}$. By integrating the predicted velocity field using an Euler solver, ForceFlow produces smooth, deterministic trajectories that eliminate the high-frequency jitter common in stochastic diffusion models.

C Real-world Experiments Visualization

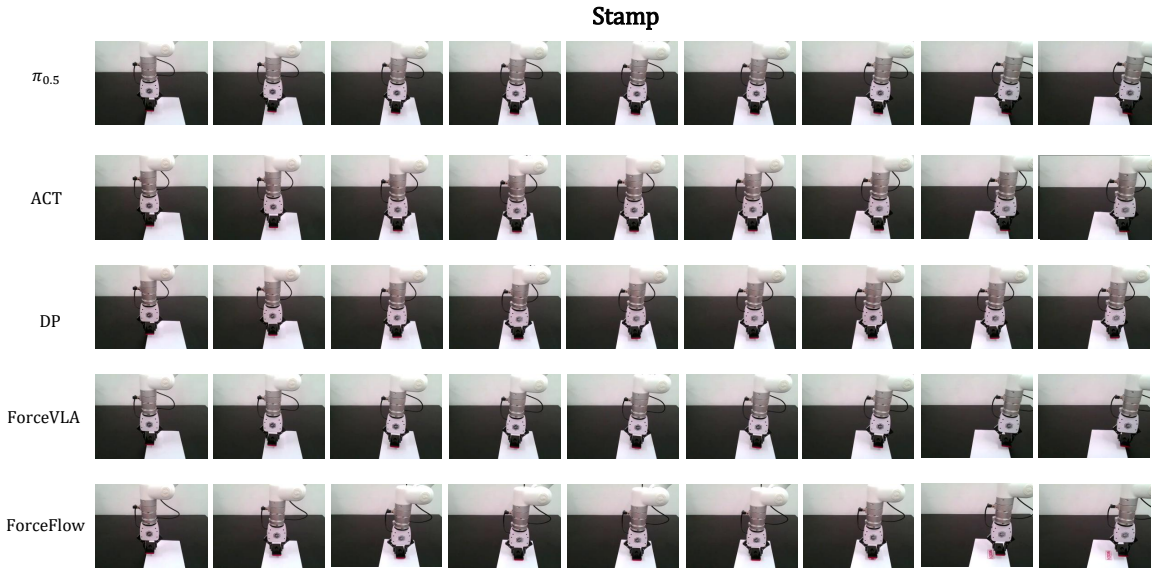


Figure 10 Stamp

Figure 11 Plug

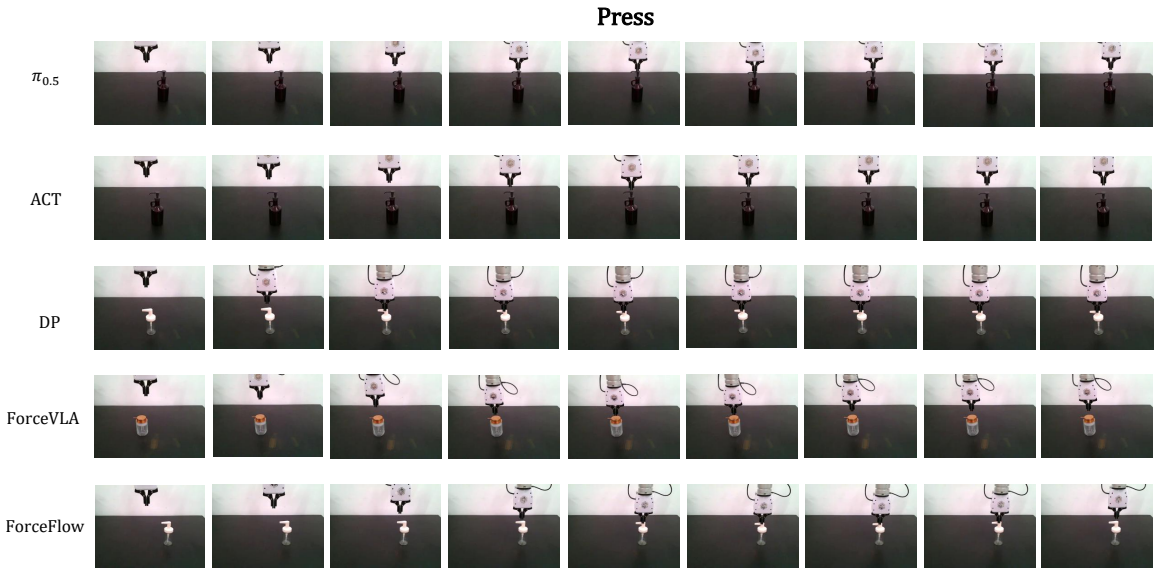


Figure 12 Press

Figure 13 Insert

Figure 14 Clean Whiteboard

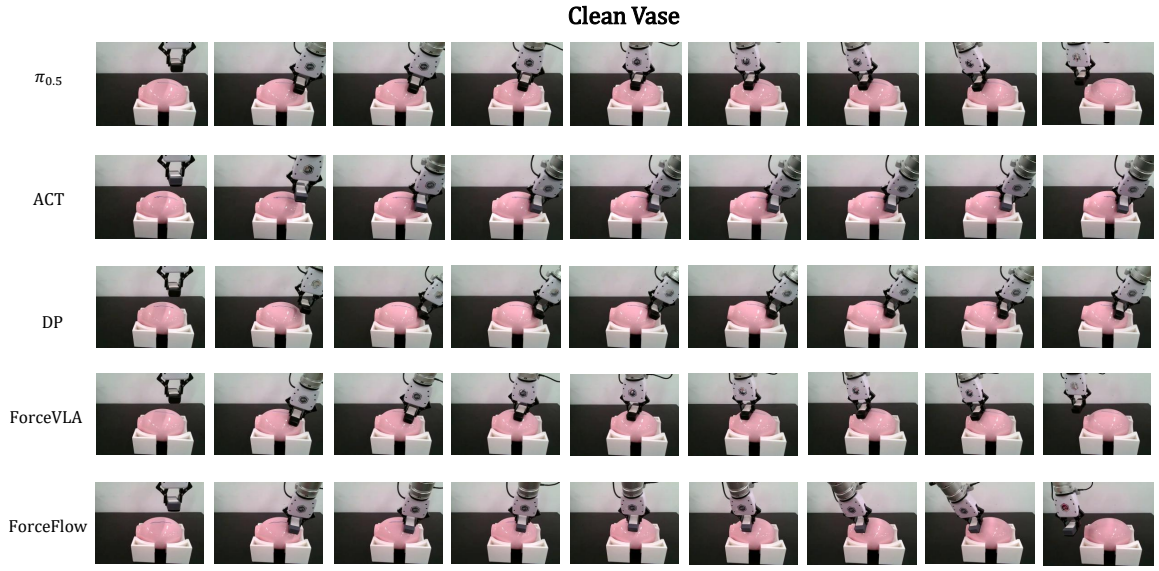


Figure 15 Clean vase

D In-depth Analysis of Force Modalities in Contact-Rich Tasks

To further elucidate why **ForceFlow** significantly outperforms vision-centric baselines, we categorize the role of force perception in our task suite into three distinct functional paradigms based on the interaction physics.

D.1 Force as the Primary Modality: Resolving Visual Ambiguity

In scenarios where visual cues are insufficient to determine the state of completion or environmental properties, force perception acts as the primary modality for action grounding.

In *Stamping Task*, the thickness of a paper stack (ranging from 1 to 50 sheets) is visually negligible from a top-down or wrist camera. A vision-only policy tends to converge to a mean terminal height, leading to insufficient pressure on thin stacks or excessive collision on thick ones. ForceFlow utilizes the 10-step force history to detect the exact moment of contact, using the force-torque as stopping signal to trigger the stamp action regardless of the absolute height. In *Press Button Task*, different sanitizer bottles possess varying spring constants and trigger depths. Because the model must reach a specific internal pressure to complete the press, ForceFlow treats the force-torque (F/T) profile as the ground truth for state transition, allowing it to generalize to unseen bottles where pixel-level movement does not directly correlate with successful activation.

D.2 Force as an Auxiliary Modality: Navigating Geometric Constraints

In tasks requiring sub-millimeter precision, such as *Plug* and *USB Insertion*, visual alignment provides a coarse initialization, but the successful completion relies on force-guided searching to overcome geometric jamming.

When a USB-A male connector hits the edge of a port, visual feedback often shows the target is reached, but the state is actually a high-friction stall. ForceFlow perceives the reactive torque τ generated by misalignment. By predicting the target force $\mathbf{F}_{t+1}^{\text{pred}}$, the model generates small-scale lateral sliding and wiggling motions. This force-aware behavior allows the robot to feel the opening and align the components dynamically, a capability entirely absent in vision-only models.

D.3 Force as a Regulatory Modality: Stable Continuous Interaction

For tasks involving continuous surface contact, force perception is critical for maintaining a stable normal force (F_n), ensuring the end-effector neither loses contact nor damages the substrate.

In the *Clean Whiteboard* task, the normal vector is constant, yet ForceFlow maintains a consistent $5N$ pressure to ensure cleaning efficacy despite potential arm vibration. The *Clean Vase* task introduces non-planar geometry where the normal vector changes continuously. As the arm moves along the curve, ForceFlow adaptively adjusts its joint torques by sensing the change in the F/T vector. The active force prediction mechanism allows the policy to proactively compliance-match the surface curvature, maintaining a stable interaction envelope that vision-centric models fail to track.

E Ablation on Visual Modality in Contact Stage

While our primary contribution focuses on alleviating the modality masking issue and proving the necessity of force feedback in contact-rich tasks (as demonstrated by the *w/o Force* baseline in Section 4.2), it is equally important to understand the role of visual input during the interaction stage. The current robotic imitation learning paradigm is inherently vision-centric; the visual modality serves as the absolute foundation for macro-level localization and motion planning.

To investigate the coupling between vision and force during fine-grained manipulation, we conduct an additional ablation study targeting the visual modality. We design a *w/o vis* variant: the policy is allowed to use visual observations during the VLM-guided *Approach Stage* for spatial localization. However, upon triggering the V2F handover and entering the *Interaction Stage*, we completely mask the visual input, forcing the ForceFlow policy to rely solely on closed-loop force perception and proprioception. We evaluate this variant across our task suite with 10 trials per task.

As shown in Table 7, completely ablating visual input during contact leads to distinct behavioral outcomes depending on the task geometry. For tasks requiring primarily single-axis pressure regulation (e.g., *Stamp*, *Clean Whiteboard*), high-frequency force feedback alone is sufficient to maintain a high success rate (80% and 90%, respectively). However, for tasks involving complex spatial trajectories, narrow tolerances, or non-planar geometries (*Plug*, *Insert*, *Clean Vase*), removing visual input leads to complete failure (0% success rate). This validates our Asymmetric Fusion architecture: both modalities are indispensable. Vision provides continuous spatial grounding to prevent the end-effector from drifting, while force feedback regulates the physical compliance.

Table 7 Ablation on Visual Modality in Contact Stage. We report the success rate (%) over 10 trials for each task. The *w/o vis* variant masks visual input entirely during the interaction stage.

Method	Stamp	Plug	Press	Insert	Clean WB	Clean Vase	Avg.
ForceFlow (Full)	85%	90%	90%	60%	100%	65%	**81.6%**
w/o vis	80%	0%	15%	0%	90%	0%	30.8%

F Discussion on Baseline Selection and Observation Modalities

In Section 4, we primarily compared ForceFlow against the state-of-the-art force-aware model, ForceVLA (Yu et al., 2025). This selection was deliberate to ensure a rigorous, controlled evaluation. The current landscape of force-aware imitation learning lacks a unified standard for modality representations, with different frameworks employing vastly different sensor hardware and data structures.

Table 8 summarizes the observation modalities utilized by recent related works. Models such as OmniVTLA (Cheng et al., 2025), RDP (Xue et al., 2025), and ViTaL (Zhao et al., 2025) rely on high-dimensional tactile arrays (e.g., GelSight sensors), which capture fine-grained local geometry but differ significantly in physical properties and data dimensionality from global force-torque measurements. Conversely, TA-VLA (Zhang et al., 2025) utilizes internal joint angles and joint torques, while methods like FeelTheForce (Adeniji et al., 2025) and FoAR (He et al., 2025) process 3D point clouds rather than 2D RGB images.

ForceVLA is the only state-of-the-art model that shares our exact modality alignment (Multi-view RGB + EEF Pose + 6D EEF Wrench). Comparing models with mismatched observation spaces or underlying hardware platforms confounds the evaluation, making it impossible to determine whether performance gaps stem from the architectural design of the multimodal fusion module or simply from differences in feature extraction and sensor density. By benchmarking against ForceVLA, we strictly control for environmental variables, demonstrating that the significant 37% performance improvement achieved by ForceFlow is entirely attributable to our proposed flow-matching formulation and the Asymmetric Fusion architecture, which effectively prevents the modality masking issue prevalent in standard MoE or concatenation-based fusion strategies.

G Additional Experiments and System Details

In this section, we provide supplementary experiments and implementation details to further evaluate the fairness, efficiency, and hardware reliability of the ForceFlow framework.

Table 8 Comparison of Observation Modalities in Force-Aware Imitation Learning Methods. ForceFlow strictly aligns with ForceVLA's input space to ensure a fair architectural comparison.

Method	Visual Modality	Proprioception	Force/Tactile Modality
ForceMimic (Liu et al., 2025)	3D Point Clouds	EEF Pose	-
OmniVTLA (Cheng et al., 2025)	Multi-view RGB	EEF Pose	High-dim Tactile Arrays
RDP (Xue et al., 2025)	RGB	EEF Pose	3D Tactile Deformation Field
ViTaL (Zhao et al., 2025)	RGB	EEF Pose	High-dim Tactile Arrays
FeelTheForce (Adeniji et al., 2025)	3D Key points	3D Key points	1D Norm Force
FoAR (He et al., 2025)	3D Point Clouds	EEF Pose	6D EEF Wrench
TA-VLA (Zhang et al., 2025)	Multi-view RGB	Joint	Joint Torques
ForceVLA (Yu et al., 2025)	Multi-view RGB	EEF Pose	6D EEF Wrench
ForceFlow (Ours)	Multi-view RGB	EEF Pose	6D EEF Wrench (w/ History)

G.1 Fairness of the V2F Mechanism in OOD Scenarios

To further validate the superiority of our Hierarchically Decoupled architecture (V2F) in handling spatial Out-of-Distribution (OOD) scenarios, we conducted an additional evaluation equipping all baseline methods with the identical V2F upper-level module.

As shown in Table 9, even with the V2F module solving the coarse navigation and spatial alignment problem, the end-to-end baseline policies still fail drastically in contact-rich execution. This demonstrates that once spatial alignment is achieved, the physical interaction and force regulation become the true bottleneck. The significant performance gap confirms that the V2F mechanism does not merely provide an "unfair structural advantage" in navigation, but rather enables the lower-level ForceFlow policy to fundamentally resolve complex contact dynamics.

Table 9 Spatial OOD Generalization with V2F Equipped Baselines (SR %). We conducted 10 independent trials per task to ensure a fair comparison.

Method	Press	Plug	Clean WB	Avg. Success
$\pi_{0.5}$ + V2F	0%	0%	0%	0%
ACT + V2F	0%	0%	0%	0%
Diffusion Policy + V2F	0%	0%	0%	0%
ForceVLA + V2F	20%	0%	20%	13.33%
ForceFlow + V2F (Ours)	**40%**	**10%**	**50%**	**33.33%**

G.2 System Latency and Inference Frequency

To ensure a rigorous comparison of system efficiency, we evaluated the inference latency and execution details across all methods. The underlying robot control frequency is uniformly fixed at 30 Hz across all baselines to maintain physical execution consistency.

For the proposed framework, the VLM in the V2F module runs only once to provide a coarse target. ForceFlow then takes over for real-time continuous control. At each timestep, the policy predicts a 64-step action chunk. By executing only the first 32 steps ($\sim$ 1s) and immediately updating all sensory inputs for the next inference, the system forms a continuous closed loop without excessive delay. Detailed latency metrics are provided in Table 10.

Table 10 Comparison of System Latency and Execution Details.

Method	Inference Latency	Action Horizon	Executed Steps	Replanning Cycle
ACT	~ 6.4 ms	100	100	~ 3.340 s
$\pi_{0.5}$	~ 81.3 ms	50	25	~ 0.915 s
Diffusion Policy	~ 109.9 ms	16	8	~ 0.377 s
ForceVLA	~ 84.1 ms	50	25	~ 0.917 s
V2F (Upper-level)	~ 1805 ms	-	-	Runs only once
ForceFlow (Lower-level)	~ 83.3 ms	64	32	~ 1.150 s

G.3 Data Collection Interface and Hardware Setup

Our data collection pipeline is designed to ensure high-quality, physically grounded expert demonstrations. To avoid the prohibitive hardware costs associated with active haptic feedback systems, we developed a Custom Real-time Force Visualization UI (see Figure 16).

During teleoperation, operators monitor precise force/torque numerical panels and dynamic curves alongside the multi-view visual feeds. This explicit feedback loop allows operators to perform precise visuo-motor compensation during complex contact phases, ensuring the recorded demonstrations are deliberate, safe, and of high quality. Furthermore, to capture reliable and high-fidelity interaction forces, our hardware system utilizes the official UFACTORY xArm 6-axis force torque sensor equipped on the robot's end-effector.

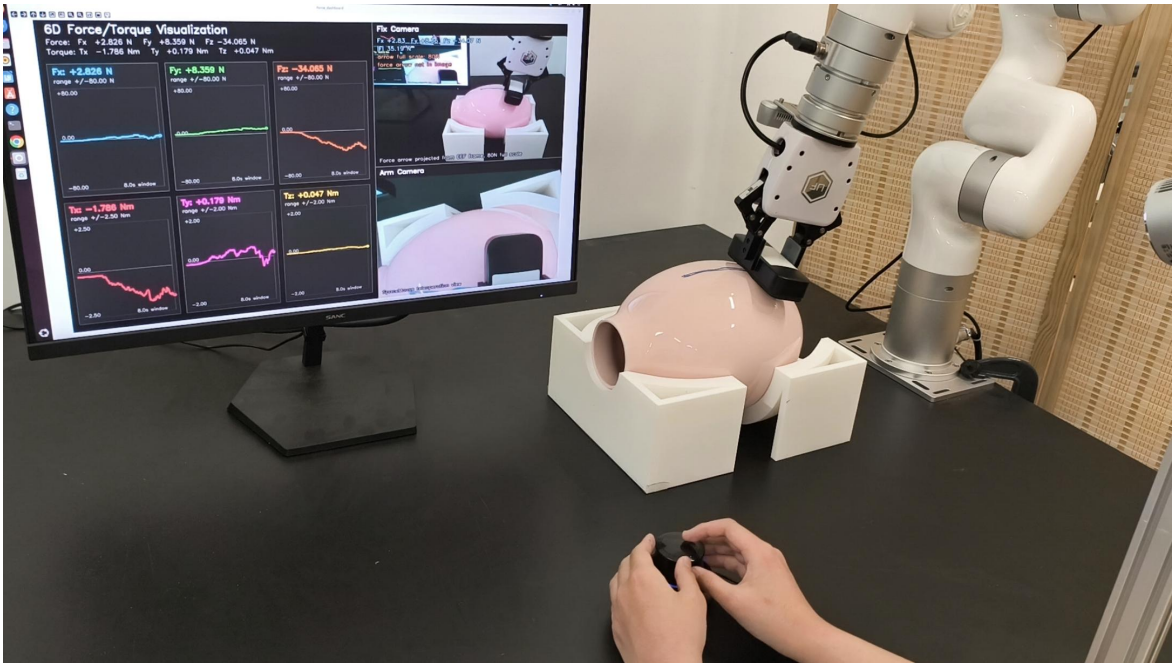


Figure 16 Custom Real-time Force Visualization UI. The interface displays synchronized multi-view visual feeds (Arm Camera and Fix Camera) alongside real-time 6D force (F_x, F_y, F_z) and torque (T_x, T_y, T_z) dynamic curves, enabling operators to achieve precise visuo-motor compensation during teleoperation.