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

Shuoheng Zhang[1*] **Yifu Yuan**[1*†] **Hongyao Tang**[1] **Yan Zheng**[1,3†] **Qiaojun Yu**[4]
Pengyi Li[1] **Guowei Huang**[2] **Helong Huang**[2] **Xingyue Quan**[2] **Jianye Hao**[1,3†]
[1]Tianjin University [2]Huawei Noah's Ark Lab [3]MemoraX AI [4]Shanghai AI Lab

Abstract

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 page: https://jokeresc.github.io/ForceFlow-page.

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 [34]. For tasks such as precision assembly, surface wiping, and peg-in-hole insertion [9], 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 [31, 26, 16].

Recent progress in imitation learning has enabled robots to acquire diverse manipulation skills from expert demonstrations [33, 5, 12]. However, these methods remain fundamentally vision-centric. Although visual observations are effective for macro-level positioning, they are often insufficient to independently support fine contact regulation due to occlusion, limited spatial resolution, and

*Equal contribution.
†Corresponding authors: yuanyf@tju.edu.cn, yanzheng@tju.edu.cn, jianye.hao@tju.edu.cn.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

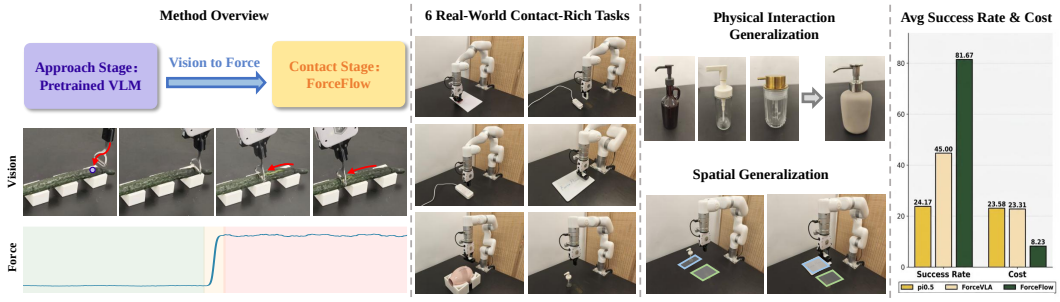


Figure 1: **Overview of ForceFlow.** ForceFlow integrates V2F handover mechanism with flow matching to achieve precise robotic manipulation in contact-rich tasks, demonstrating significantly higher success rates and superior OOD generalization compared to baselines.

weak sensitivity to subtle contact events at the interaction interface [15]. 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 [31, 15, 34], but two key challenges remain. The first is multimodal 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 [14]. 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) [20] acquires manipulation skills from expert demonstrations. Early Behavioral Cloning (BC) [18] struggles in complex long-horizon tasks, prompting the adoption of generative models. Methods like ACT [33] and Diffusion Policy [5] enable long-horizon smooth trajectory generation. Beyond imitation learning, MODULI [28] studies preference-conditioned diffusion planning for offline multi-objective reinforcement learning, demonstrating generalization to unseen preferences. More recently, Vision-Language-Action (VLA) models (e.g., OpenVLA [12], $\pi_{0.5}$ [3], and GROOT N1 [2]) leverage large-scale pre-training for open-world generalization. Complementary advances span spatially grounded reasoning [29, 11], embodied foundation models [30, 22], and unified capability evaluation [17]. Despite these advances, vision-centric action policies still need explicit mechanisms for contact-force regulation. 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 [32, 13], motivating the integration of force or tactile signals into policy learning. Existing approaches typically introduce force as additional observations [1, 24] or infer it from vision and proprioception [15]. Other works focus on multimodal fusion via contact-aware gating [9], residual refinement [34], or hierarchical modeling [25]. Several recent models [26, 4, 8, 10, 31] 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* [21, 23, 7]. 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.

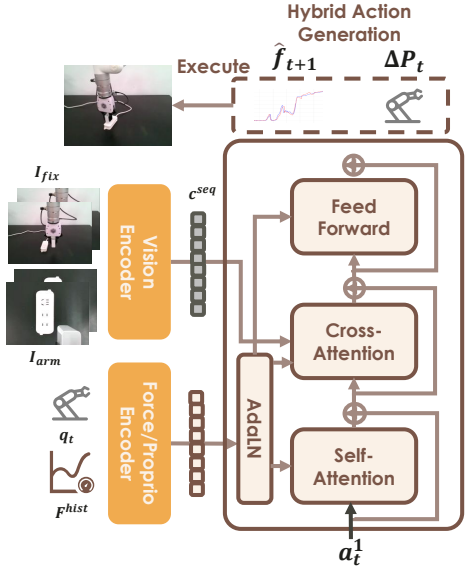


Figure 2: **The ForceFlow Architecture.**

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.

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

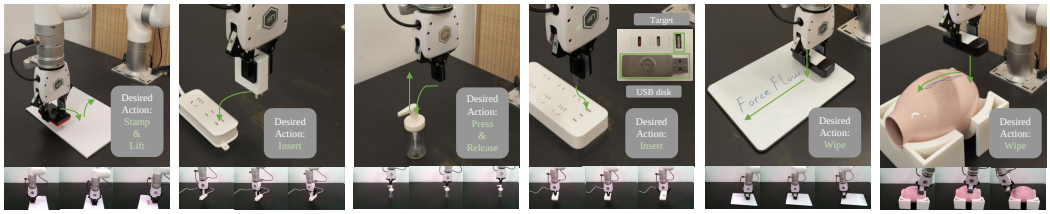
Once the VLM guides the end-effector to the 3D approach waypoint within the local interaction window, the system transitions to the ForceFlow policy for the Interaction Stage. ForceFlow focuses exclusively on high-frequency contact dynamics without bearing the burden of global spatial reasoning. To this 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) [19]. 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 [14, 6] 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



(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%**

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?

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.

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**

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:

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

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 $\|\mathbf{f}_t\| > 5$N) to evaluate tracking consistency.

4.2 RQ1: Effectiveness in Contact-Rich Tasks

We compare ForceFlow against $\pi_{0.5}$ [3], ACT [33], Diffusion Policy [5], ForceVLA [26], 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 4 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 5 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.

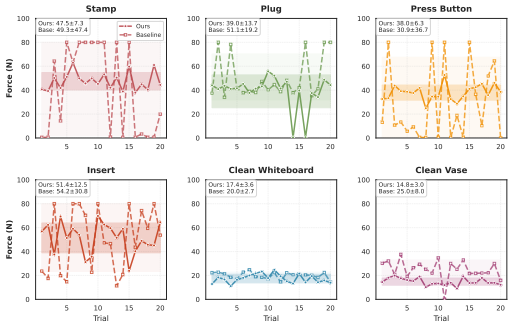


Figure 4: **Statistical Force Distribution**. Maximum or average interaction forces for six representative tasks across 20 independent trials. Error bars/shading indicate standard deviation, highlighting the stability of the learned force regulation strategy.

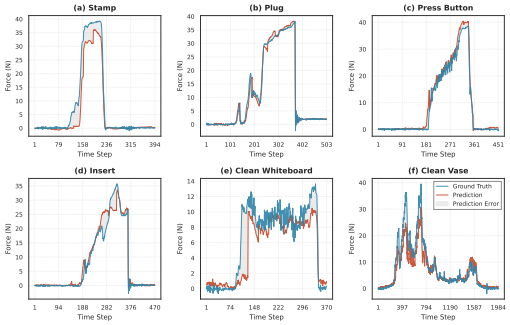


Figure 5: Comparison of Predicted and Measured Forces. Alignment between the predicted (red) and measured ground truth (blue) contact forces. Shaded areas represent prediction error, demonstrating high temporal fidelity during interaction.

Table 3: **Physical 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
All Policies (w/o V2F)	0%	0%	0%
$\pi_{0.5}$ + V2F	0%	0%	0%
ACT + V2F	0%	0%	0%
Diffusion Policy + V2F	0%	0%	0%
ForceVLA + V2F	20%	0%	20%
ForceFlow + V2F	**40%**	**10%**	**50%**

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.

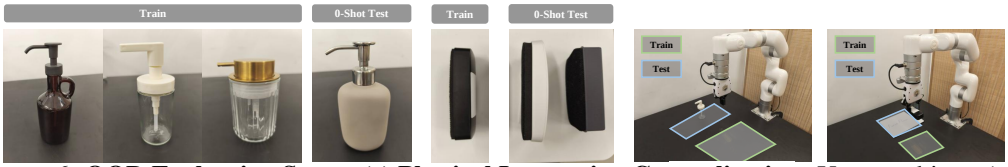


Figure 6: **OOD Evaluation Setup.** (a) **Physical Interaction Generalization:** Unseen objects (e.g., distinct bottles, different erasers) used for zero-shot testing. (b) **Spatial Generalization:** The testing workspace (marked "Test") is spatially disjoint from the training distribution (marked "Train"), 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 6a). 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 large workspace shifts, we configure a testing workspace that is spatially disjoint from the training distribution (Figure 6b) and conduct 10 independent trials per task across **Press**, **Plug**, and **Clean Whiteboard**. Within the V2F handover mechanism, we utilize Embodied-R1 [27] during the vision-dominant Approach Stage to predict target pixel coordinates from the global view I_{fix} and guide the end-effector to the 3D approach waypoint before handing over control to the contact policy. We evaluate all methods under two settings in Table 4: standalone end-to-end execution (*w/o V2F*) and hierarchical execution where all policies are equipped with the identical V2F approach module (*+ V2F*).

Results in Table 4 demonstrate the necessity of coupling V2F localization with force-aware contact regulation. Without V2F handover (*w/o V2F*), all standalone policies fail to locate targets in OOD regions (0% SR). Furthermore, even when V2F resolves coarse spatial navigation and guides the end-effector to the 3D approach waypoint (*+ V2F*), vision-centric baselines ($\pi_{0.5}$ + V2F, ACT + V2F, and Diffusion Policy + V2F) still fail completely (0% SR), while ForceVLA + V2F achieves only 13.33% average SR (20% on Press, 0% on Plug, and 20% on Clean Whiteboard). In contrast, **ForceFlow + V2F** achieves the highest performance across all tasks (40%, 10%, and 50%, averaging 33.33% SR). This confirms that once spatial localization is decoupled via V2F, closed-loop force regulation during the force-dominant Interaction Stage becomes the decisive factor for compensating for local positioning residuals and completing contact-rich execution.

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.

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.

Contributions of Temporal Force History and Joint Force Prediction. 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 susceptible to interference from contact instability or noise from the sensor itself. This instability is corroborated in Figure 7. The 1-step variant exhibits significant high-frequency 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.

Table 5: **Ablation Study on the Stamp Task.** Effects of temporal force history and active force prediction on success rate (SR) and force cost.

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**

Role of Selective Visual Reference in the Interaction Stage. While the *ForceFlow (w/o Force)* baseline in Table 1 and Table 2 validates the necessity of treating temporal force history as a global regulatory signal (c_{vec}), we further examine the role of retaining vision as a selective spatial reference (c_{seq}) within our asymmetric multimodal fusion architecture. We evaluate a **ForceFlow (w/o Vision)** variant over 20 trials per task: the policy utilizes visual observations during the VLM-guided Approach Stage for target localization, but completely masks the visual inputs upon triggering the V2F handover, relying solely on temporal force history $\mathbf{F}_t^{\text{hist}}$ and proprioception $\mathbf{q}_t$ during the force-dominant Interaction Stage. As shown in Table 6, ablating visual reference during contact execution yields distinct outcomes depending on task geometry. For tasks dominated by single-axis normal pressure regulation on planar surfaces (*Stamp* and *Clean Whiteboard*), closed-loop force feedback alone suffices to maintain high success rates (80% and 90%, respectively). However, for tasks involving sub-millimeter geometric tolerances (*Plug* and *Insert*) or non-planar 3D surface trajectories (*Press* and *Clean Vase*), masking visual input leads to severe degradation (0% to 15% SR), reducing the average success rate from 81.67% to 30.83%. This confirms that both pathways in our asymmetric fusion design are indispensable: global force modulation governs physical compliance, while selective visual attention provides continuous spatial grounding to prevent end-effector drift during fine contact interaction.

Table 6: **Ablation on Visual Modality in the Interaction Stage.** We report the success rate (%) over 20 trials for each task. *ForceFlow (w/o Vision)* retains VLM pointing in the Approach Stage but masks the visual sequence condition (c_{seq}) entirely during the Interaction Stage.

Method	Stamp	Plug	Press	Insert	Clean WB	Clean Vase	Avg.
ForceFlow (w/o Vision)	80%	0%	15%	0%	90%	0%	30.83%
ForceFlow (Full)	**85%**	**90%**	**90%**	**60%**	**100%**	**65%**	**81.67%**

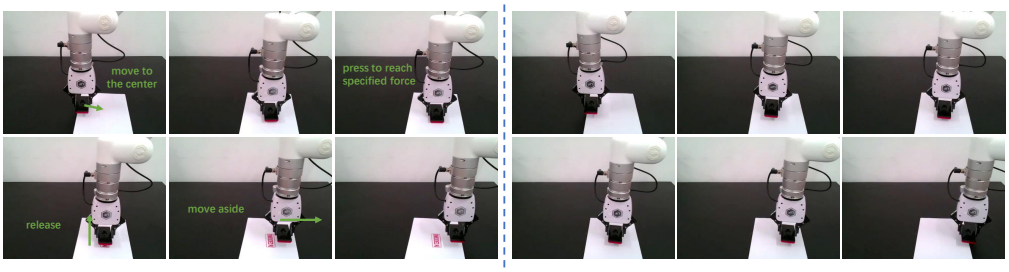


Figure 7: Qualitative comparison of force regulation in the Stamping task. The upper part illustrates the stable execution of ForceFlow (Full), while the lower part shows the w/o Force History (1-step) variant. Without temporal context, the single-step policy suffers from severe oscillations and force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

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.

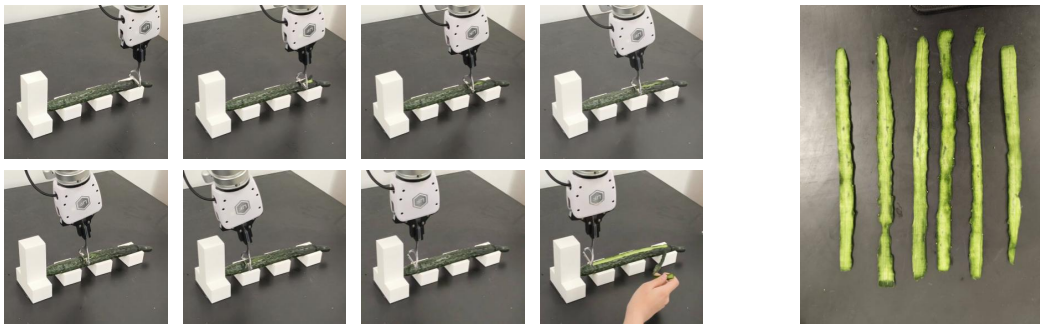


Figure 8: **Qualitative results of cucumber peeling.** (Left) ForceFlow proactively regulates force to track the contour. (Right) Uniform, unbroken peeled strips evidence a consistent cutting depth.

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.

Acknowledgement

This work is supported by the National Natural Science Foundation of China Youth Student Basic Research Project (Grant No. 625B2128). This work is also supported by the National Natural Science Foundation of China (Grant Nos. 62533021, 62422605, 92370132) and the National Key Research and Development Program of China (Grant No. 2026ZD0126104). Yifu Yuan is also supported by the Young Science and Technology Scientists Sponsorship Program by CAST - Doctoral Student Special Plan. Qiaojun Yu is supported by the State Key Laboratory of Industrial Control Technology, Zhejiang University (Grant No. ICT2026B50). We would also like to thank our collaborators Guowei Huang, Helong Huang, and Xingyue Quan at Huawei Noah's Ark Lab for their insightful discussions and strong support in this joint project.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] Yaoting Huang, Yifu Yuan, Linqi Han, Chengwen Li, Shuoheng Zhang, Xianze Yao, Hongyao Tang, Yan Zheng, and Jianye Hao. RoboPIN: Grounded embodied reasoning via pinned chain-of-thought. In *Proceedings of the ACM International Conference on Multimedia*, 2026. URL https://arxiv.org/abs/2606.15753.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Shaojin Ma, Min Zhang, Hongyao Tang, Ruitao Wang, Jianye Hao, and Yan Zheng. DDCG: Decoupled dual-critic guidance for embodied agents. In Hung-Yi Lee, Sunipa Dev, Jessy Li, and Desmond Elliott, editors, *Findings of the Association for Computational Linguistics: EMNLP 2026*, pages 9339–9360, Budapest, Hungary, October 2026. Association for Computational Linguistics. ISBN 979-8-89176-462-0. URL https://aclanthology.org/2026.findings-emnlp.445/.

[17] Fei Ni, Min Zhang, Pengyi Li, Yifu Yuan, Lingfeng Zhang, Yuecheng Liu, Peilong Han, Longxin Kou, Shaojin Ma, Jinbin Qiao, David Gamaliel Arcos Bravo, Yuening Wang, Xiao Hu, Zhanguang Zhang, Xianze Yao, Yutong Li, Zhao Zhang, Ying Wen, Ying-Cong Chen, Xiaodan Liang, Liang Lin, Bin He, Haitham Bou-Ammar, He Wang, Huazhe Xu, Jiankang Deng, Shan Luo, Shuqiang Jiang, Wei Pan, Yang Gao, Stefanos Zafeiriou, Jan Peters, Yuzheng Zhuang, Yingxue Zhang, Yan Zheng, Hongyao Tang, and Jianye Hao. Embodied Arena: A comprehensive, unified, and evolving evaluation platform for embodied AI. *arXiv preprint arXiv:2509.15273*, 2025. URL https://arxiv.org/abs/2509.15273.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Ziyi Wang, Xumin Yu, Yongming Rao, Yonggen Ling, Yunheng Li, Oran Wang, Mingqi Gao, Yuchen Zhou, Yves Liang, Zuyan Liu, Yani Zhang, Rui Huang, Xiaoran Xu, Bowen Yuan, Yifu Yuan, Xu Tan, He Zhang, Yufei Huang, Shenghao Zhang, Hongsheng Wu, Han Hu, and Zhengyou Zhang. Hy-Embodied-VLM-1.0: Efficient physical-world agents. *arXiv preprint arXiv:2607.12894*, 2026. URL https://arxiv.org/abs/2607.12894.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] Yifu Yuan, Zhenrui Zheng, Zibin Dong, and Jianye Hao. MODULI: Unlocking preference generalization via diffusion models for offline multi-objective reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 73743–73766, 2025. URL https://proceedings.mlr.press/v267/yuan25n.html.

[29] Yifu Yuan, Haiqin Cui, Yibin Chen, Zibin Dong, Fei Ni, Longxin Kou, Jinyi Liu, Pengyi Li, Yan Zheng, and Jianye Hao. From seeing to doing: Bridging reasoning and decision for robotic manipulation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2505.08548.

[30] Yifu Yuan, Yaoting Huang, Xianze Yao, Yutong Li, Shuoheng Zhang, Linqi Han, Pengyi Li, Jiangeng Sun, Wenting Jia, Zhao Zhang, Yuhao Liu, Ruihao Liao, Yucheng Hu, Qiyu Wu, Yuxiao Li, Zibin Dong, Fei Ni, Yan Zheng, Shuyang Gu, Yi Ma, Hongyao Tang, Han Hu, and Jianye Hao. Embodied-R1.5: Evolving physical intelligence via embodied foundation models. In *Advances in Neural Information Processing Systems*, 2026. URL https://arxiv.org/abs/2606.11324.

[31] 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.

[32] 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.

[33] 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.

[34] 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 7: 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 7.

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.

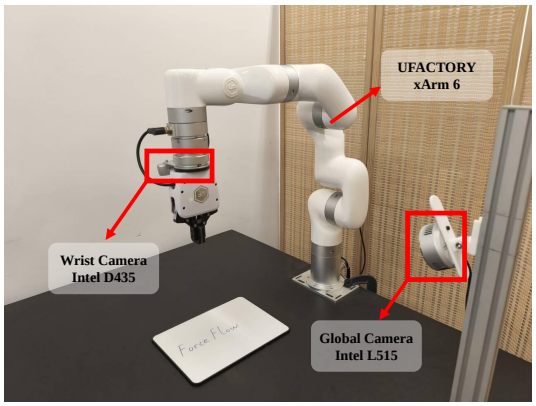


Figure 9: Hardware System

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 $k \in [0, 1]$, 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

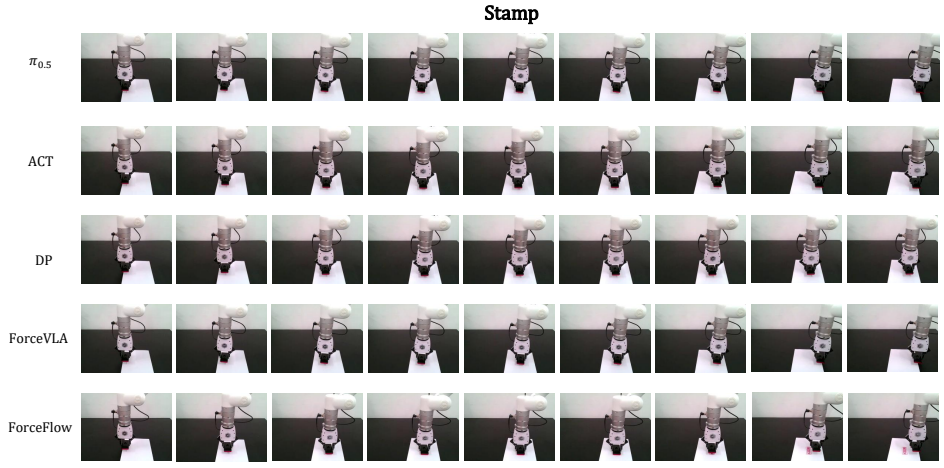


Figure 10: **Stamp**



Figure 11: **Plug**

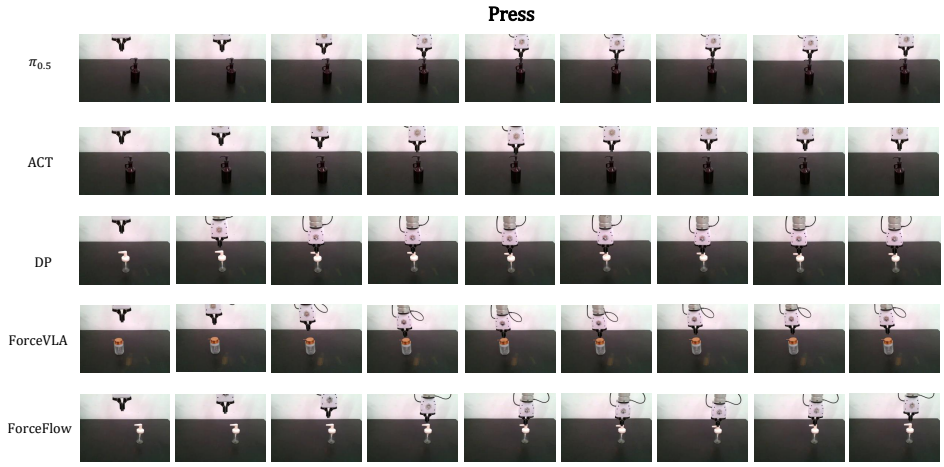


Figure 12: **Press**



Figure 13: **Insert**



Figure 14: **Clean Whiteboard**

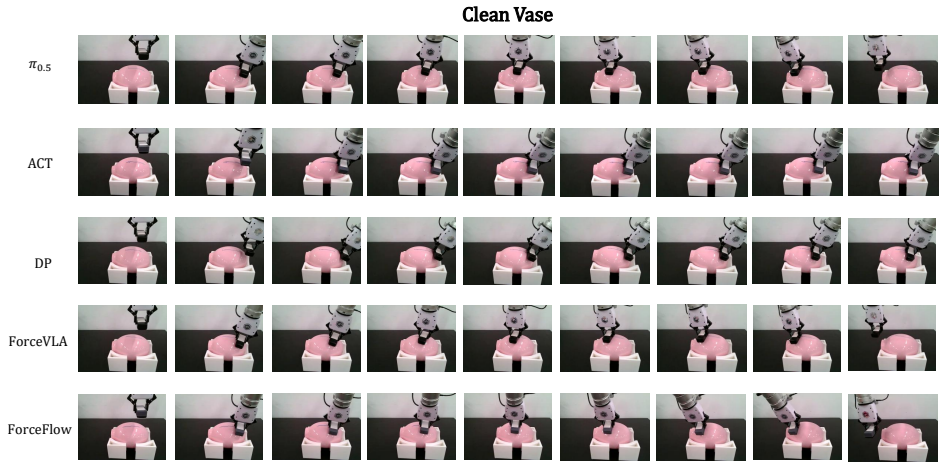


Figure 15: **Clean vase**

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 Discussion on Baseline Selection and Observation Modalities

In Section 4, we primarily compared ForceFlow against the state-of-the-art force-aware model, ForceVLA [26]. 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 [4], RDP [25], and ViTaL [34] 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 [31] utilizes internal joint angles and joint torques, while methods like FeelTheForce [1] and FoAR [9] 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.

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 [15]	3D Point Clouds	EEF Pose	-
OmniVTLA [4]	Multi-view RGB	EEF Pose	High-dim Tactile Arrays
RDP [25]	RGB	EEF Pose	3D Tactile Deformation Field
ViTaL [34]	RGB	EEF Pose	High-dim Tactile Arrays
FeelTheForce [1]	3D Key points	3D Key points	1D Norm Force
FoAR [9]	3D Point Clouds	EEF Pose	6D EEF Wrench
TA-VLA [31]	Multi-view RGB	Joint	Joint Torques
ForceVLA [26]	Multi-view RGB	EEF Pose	6D EEF Wrench
ForceFlow (Ours)	Multi-view RGB	EEF Pose	6D EEF Wrench (w/ History)

F 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 9.

Table 9: 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

F.1 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.

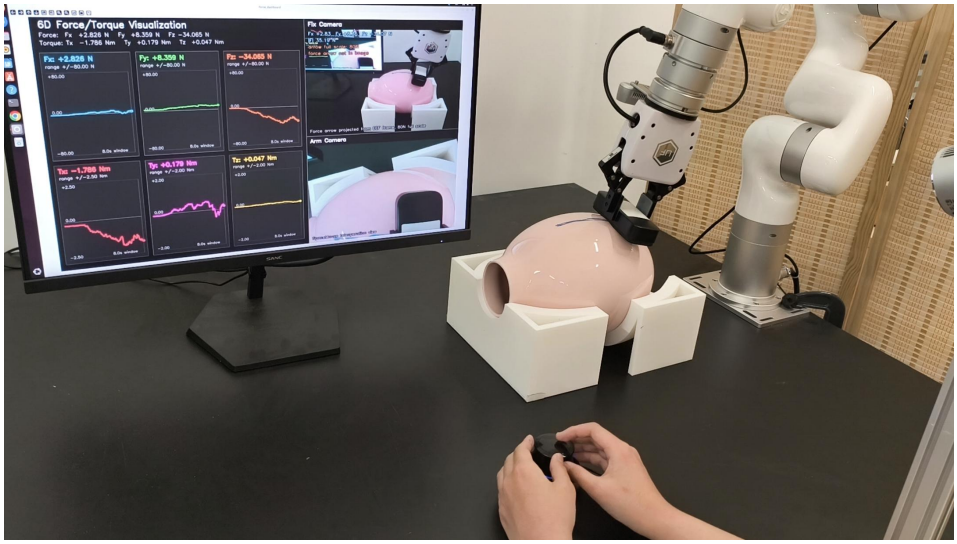


Figure 16: **Custom Real-time Force Visualization UI.** The interface displays synchronized multi-view visual feeds (Arm Camera and Fix Camera) alongside real-time 6D force (F_x, F_y, F_z) and torque (T_x, T_y, T_z) dynamic curves, enabling operators to achieve precise visuo-motor compensation during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

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 (Introduction) explicitly state our main contributions, including the ForceFlow framework, the asymmetric multimodal fusion architecture, and the Vision-to-Force (V2F) handover mechanism. These claims accurately match the scope of our evaluations and the experimental results detailed in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of our work are explicitly discussed in the Conclusion section, where we acknowledge the framework's reliance on high-fidelity force/torque sensors, which may restrict its deployment on low-cost robotic platforms.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper is primarily methodological and empirical. While we provide the formal mathematical formulation of our framework in Section 3 and Appendix B, the work does not introduce formal theoretical results (e.g., new mathematical theorems or lemmas) that require formal proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We fully disclose the information required to reproduce our experiments in Section 4 and Appendix A. This includes detailed descriptions of the hardware setup, data collection pipeline, task definitions, evaluation protocols, and a comprehensive table of model architecture and training hyperparameters (Table 7).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have open-sourced all training and evaluation code, pre-trained checkpoints, and real-world demonstration datasets. All resources and documentation are publicly available at our project page: https://jokeresc.github.io/ForceFlow-page.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We detail the experimental setup in Section 4.1 and provide a comprehensive breakdown of all training and test details—including optimizer type, learning rate schedules, and specific hyperparameters—in Appendix A (particularly Table 7).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: We evaluate the statistical significance and variance of our framework by conducting 20 independent trials for each task. Specifically, Figure 4 illustrates the mean and standard deviation (indicated by error bars/shading) of the interaction forces to demonstrate stability across different runs.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We explicitly detail the compute resources in Appendix A.2, specifying the hardware configuration (4× NVIDIA RTX 4090 GPUs with 24 GB VRAM, 48 CPU cores, and 283 GB RAM) and reporting the estimated training time of approximately 8-10 hours per task.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics. Our work focuses on foundational robotic manipulation skills in a safe laboratory environment and does not involve applications, datasets, or methods that pose societal harm or ethical risks.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper presents foundational algorithmic research in robotic imitation learning and low-level control. As it focuses on basic physical interaction capabilities (e.g., plugging, wiping surfaces) in controlled settings, it does not present a direct path to negative societal applications or harms.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our models and datasets do not pose a high risk for misuse. The collected data and the trained ForceFlow policy are strictly specialized for low-level robotic manipulation tasks (e.g., plugging, wiping surfaces) in controlled physical environments and do not involve large-scale internet scraping or high-risk generative capabilities.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We utilize open-source pre-trained visual encoders and Vision-Language Models (e.g., ResNet-18 and Embodied-R1). We properly credit the original creators in the manuscript and adhere to their respective open-source licenses and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce a new contact-rich manipulation dataset collected via robot teleoperation. Detailed documentation regarding the hardware setup and data collection pipeline is provided in Appendix A.2 and F.1. The dataset, along with full training and evaluation code, is publicly available at `https://jokeresc.github.io/ForceFlow-page`.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This research does not involve crowdsourcing or research with human subjects. All expert demonstrations were collected internally via teleoperation.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work focuses strictly on robotic manipulation in a controlled laboratory environment and does not involve human subjects or biological data; therefore, Institutional Review Board (IRB) approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We explicitly describe the integration of a pre-trained Vision-Language Model (Embodied-R1) as a semantic spatial planner during the Approach Stage within our V2F framework. Additionally, Large Language Models were solely used for text editing and polishing (e.g., grammar, spelling, word choice) to improve the readability of the manuscript.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.