ViTaR: Visuo-Tactile Residual Adaptation for Foundation VLA Manipulation

Anonymous submission

Abstract

As Vision-Language-Action (VLA) models scale toward real-world deployment, contact-rich manipulation exposes a critical blind spot: these policies encode broad visual-semantic priors yet remain unaware of local contact events, producing identical actions whether contact is established, lost, or destabilized. Existing remedies either modify VLA internals, risking catastrophic forgetting, or demand online reinforcement under near-failure contact conditions. Both grant tactile unbounded influence over action generation, conflicting with the priors that make VLAs generalizable. We introduce ViTaR, which reframes tactile feedback from an action-generating perceptual input to an execution modulator that selects and scales bounded residual corrections atop a frozen VLA, preserving pretrained capabilities by construction. ViTaR decomposes adaptation into two stages: Effect-Guided Modeling determines *whether* and *which* correction is locally justified via outcome-grounded preference evidence, and Residual Action Modulation converts this evidence into a residual choice with continuously scaled gain from real-time visuotactile observations. On the UniVTAC benchmark spanning seven contact-rich tasks, ViTaR achieves 61.3% average success, a 30.6 percentage-point improvement over its frozen VLA base that also surpasses purpose-built tactile baselines. Physical-robot experiments confirm that bounded tactile modulation transfers to real sensor noise and dynamics.

Introduction

Vision-Language-Action (VLA) models encode broad visual-semantic priors from large-scale pretraining (Brohan et al. 2023b,a; Kim et al. 2024; Black et al. 2024; Chi et al. 2023; Physical Intelligence et al. 2025; Ghosh et al. 2024; Wang et al. 2024, 2026), yet contact-rich tasks such as insertion, connector mating, and force-sensitive grasping depend on resolving local physical interactions that external cameras cannot observe (Cao et al. 2026; Huang et al. 2025). As these models are deployed to increasingly contact-intensive settings, this gap becomes a critical bottleneck. Once deployed, a frozen VLA is blind to contact dynamics: it commits to the same action regardless of whether contact is established, slipping, or lost.

Existing approaches to incorporating tactile feedback share a common structural commitment: they grant touch *unbounded* influence over action generation. Methods that fuse tactile into VLA internals (Huang et al. 2025; Li et al. 2026; Morissette et al. 2026) risk catastrophic forgetting, while reinforcement-based adaptation (Ma et al. 2026; Zheng et al. 2026) demands exploratory rollouts at near-failure contact states where safety margins are thinnest. In both paradigms, tactile can override any dimension of the output action, directly conflicting with the visual-semantic priors that make VLAs generalizable.

Our key observation is that a frozen VLA already provides the correct semantic action *direction*; what it lacks is calibration of execution under contact conditions absent during pretraining. This asymmetry motivates a paradigm shift: **tactile feedback should serve as an execution modulator rather than participating in unconstrained action generation** (Fig. 1). Realizing this paradigm demands two ingredients: a *correction mechanism* preserving the base action space, and a *learning signal* that ranks corrections by local effect. Residual policies (Johannink et al. 2019; Hoque et al. 2022) supply the first, but existing formulations lack contact feedback and assume state-independent corrections. For the learning signal, global reward regression is ill-suited because absolute scores are not comparable across heterogeneous contact states. Within-state preference comparison (Bradley and Terry 1952) resolves this: ranking corrections at the *same* restored contact state recovers relative local effects without cross-state calibration, providing exactly the signal that contact-conditioned residual selection requires.

We introduce **ViTaR** (**Vi**suo-**Ta**ctile **R**esidual Adaptation), which operationalizes this paradigm through two coupled stages (Fig. 2). *Effect-Guided Modeling* (EGM) assesses whether the current contact state is improvable and ranks available residuals by expected local effect via pairwise outcome comparisons at restored decision points. *Residual Action Modulation* (RAM) converts this evidence into a discrete action choice—retain the base action or select a residual—and predicts a tactile-conditioned gain. The frozen VLA action is preserved whenever contact evidence does not justify intervention.

On the UniVTAC benchmark (Chen et al. 2026), ViTaR achieves 61.3% average success across seven contact-rich tasks, a 30.6pp improvement over its frozen OpenVLA-OFT (Qiu et al. 2023) base, and consistent gains over purpose-built tactile baselines. Physical-robot experiments validate transfer across diverse contact modalities. Our contributions are threefold:

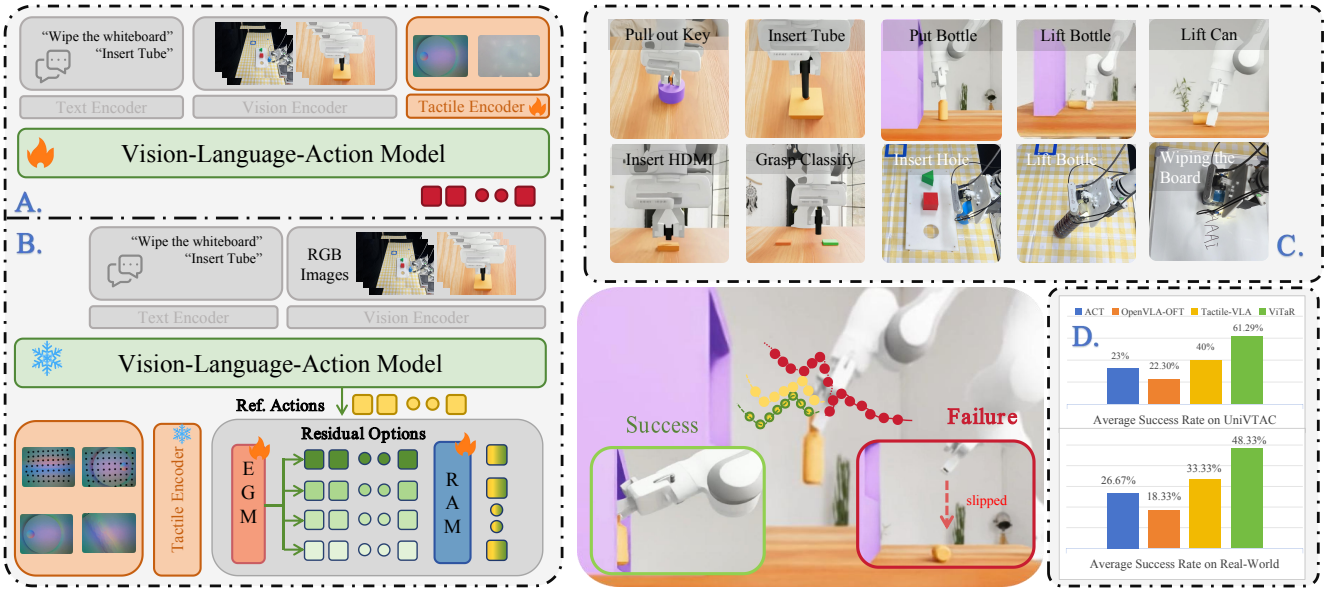

Figure 1: ViTaR at a glance. (A) A paradigm for introducing tactile modality into a pretrained VLA. (B) The tactile-modality integration paradigm proposed by ViTaR. (C) The seven contact-rich UniVTAC tasks and three physical-robot contact-rich tasks; task names shown in white indicate physical-robot tasks. (D) Task success rates of baselines and ViTaR across UniVTAC and physical-robot tasks; grouped bars follow the fixed left-to-right order ACT, OpenVLA-OFT, Tactile-VLA, and ViTaR.

- We propose ViTaR, a *tactile-as-execution-modulation* framework that augments a frozen VLA with bounded residual corrections selected and scaled by tactile evidence, without generating new action directions or modifying pretrained representations.
- We introduce Effect-Guided Modeling (EGM), which ranks candidate corrections by observed outcome differences within each contact state via pairwise preference learning, eliminating the need for globally calibrated rewards.
- We validate ViTaR on the UniVTAC benchmark and physical-robot tasks, demonstrating a 30.6pp gain over the frozen VLA base and successful transfer across insertion, sliding, and grasping.

Related Work

Tactile integration into VLA and visuomotor policies. Tactile sensing captures proximal contact evidence that external cameras cannot resolve (Cao et al. 2026). Recent methods fuse touch into VLA policies through diverse mechanisms: deep token fusion (Huang et al. 2025), temporally decoupled injection (Li et al. 2026), feature-level modulation (Morissette et al. 2026), and residual tactile representations (Zhang et al. 2026). Beyond fusion, TacCoRL (Ma et al. 2026) and TORL-VLA (Zheng et al. 2026) use online RL at contact bottlenecks; Tactile-WAM (Wu et al. 2026) routes tactile within a world-action model; HTD (Niu et al. 2026) predicts future tactile states; and TactiDex (Ni et al. 2026) establishes a tactile-guided dexterous benchmark. Foundation tactile encoders (Yuan et al. 2026a; Higuera et al. 2024) and large-scale visuotactile datasets (Hua et al. 2026) further support these efforts. Despite their diversity, all above methods treat tactile as a perceptual modality participating in action generation with unbounded influence. ViTaR departs from this design: it constrains tactile to a bounded execution-modulation role, selecting and scaling structured residuals atop the frozen base policy without altering its pretrained representations.

Residual and intervention policy learning. Residual policy learning augments a fixed base policy with a learned correction term, preserving base competence while enabling task-specific adaptation (Silver et al. 2019; Johannink et al. 2019). Intervention learning (Hoque et al. 2022; Mandlekar et al. 2020) and shared-autonomy frameworks (Haldar et al. 2023) similarly overlay corrective actions under predefined trigger conditions. These methods validate the value of structured corrections but operate without contact feedback and assume correction directions are independent of the local physical state. ViTaR extends this paradigm along two axes: it conditions both correction selection and magnitude on real-time contact evidence, and grounds the selection in outcome-based preference supervision rather than hand-designed triggers.

Preference-based learning for embodied agents. Preference-based reward learning acquires relative quality judgments without globally calibrated scalar rewards (Christiano et al. 2017; Bradley and Terry 1952). Extensions to embodied settings address online feedback (Lee, Smith, and Abbeel 2021), demonstration-based extrapolation (Brown, Goo, and Niekum 2019), unified multi-feedback platforms (Yuan et al. 2024), annotator heterogeneity (Yuan et al. 2026b), and risk sensitivity (Tung et al. 2026). However, existing applications uniformly learn global reward

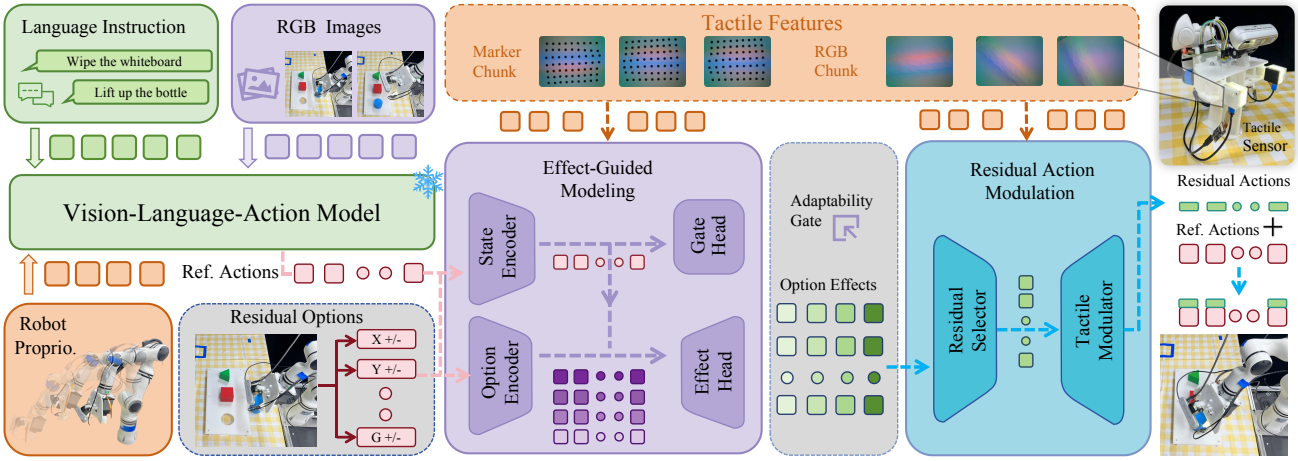

Figure 2: Overview of ViTaR. A frozen VLA maps language, RGB observations, and proprioception to a base action. Marker and tactile-image chunks provide local contact evidence. EGM estimates state adaptability and relative effects for the base action and residual options; RAM then selects an action choice and predicts a continuous scale for the selected residual before execution.

models from cross-episode trajectory comparisons. ViTaR operates at a fundamentally different granularity: it learns *within-state* effect rankings from local branch comparisons at restored decision points, recovering which correction is locally superior without requiring a reward function that generalizes across heterogeneous contact configurations.

Method

Problem Formulation

We consider contact-rich manipulation augmented by a frozen pretrained Vision-Language-Action (VLA) policy π_{VLA}. At each decision step t, the frozen VLA maps observation o_t and language instruction ℓ to a base action chunk,

$$A_t^{\mathrm{ref}} = \pi_{\mathrm{VLA}}(o_t, \ell), \qquad A_t^{\mathrm{exec}} = A_t^{\mathrm{ref}} + \Delta A_t, \quad (1)$$

where A_t^{ref} denotes the frozen VLA output and ΔA_t is a contact-aware residual correction. An *action choice* either retains the base action ($\Delta A_t = 0$) or selects a structured 7D delta $d(u)$ from a small task-aligned candidate set $\mathcal{R}$, whose directions are derived from the predominant signed components of frozen VLA action chunks. For an H-step action chunk, the same scaled delta is applied uniformly: $\Delta A_t = (\alpha\, d(u), \ldots, \alpha\, d(u)) \in \mathbb{R}^{H\times 7}$.

Overview

ViTaR augments the frozen base policy with two learned components. *Effect-Guided Modeling* (EGM) maps the local state and available action choices to an *adaptability score* $g_i \in \mathbb{R}$ and relative *effect scores* $\{e_i(c)\}_{c\in\mathcal{C}_i}$; *Residual Action Modulation* (RAM) uses them to retain the base action or select and continuously scale a residual. Throughout EGM and RAM, i indexes a local decision point where residuals are evaluated, whereas t indexes a control step in the associated H-step action chunk. Thus, ΔA_i denotes the residual selected at decision point i, which instantiates ΔA_t in Eq. (1) at each step of its chunk.

Both modules are supervised by short-horizon branch rollouts restored to common decision points (detailed in §). At deployment, only current observations are used: a marker-derived contact descriptor m_i conditions residual selection (Fig. 3), while a visuotactile summary z_i determines the scaling gain (Fig. 2).

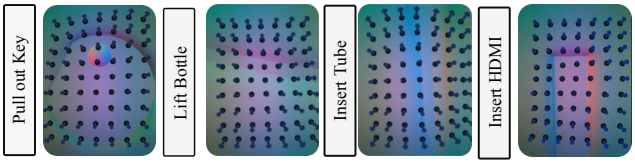

Figure 3: Marker motion on raw simulated GelSight Mini images from UniVTAC. Arrows indicate the displacement directions of markers between pre-contact and post-contact observations. These local motions provide evidence of contact deformation and stability for the marker-derived contact descriptor m_i.

Effect-Guided Modeling

A residual that improves one contact state may degrade another; EGM therefore learns state-conditioned effect evidence by comparing action choices within the same local state. It serves as a supervision source for RAM rather than a deployed action policy.

Branch-Outcome Preference Learning A *branch* is a short-horizon rollout executed from a restored local state under a single action choice. Let s_i and x_i denote the decision point and its representation, and $c \in \mathcal{C}_i \equiv \{\varnothing\} \cup \mathcal{R}_i$ an action choice. The base choice retains A_t^{ref}; $u \in \mathcal{R}_i$ ap-

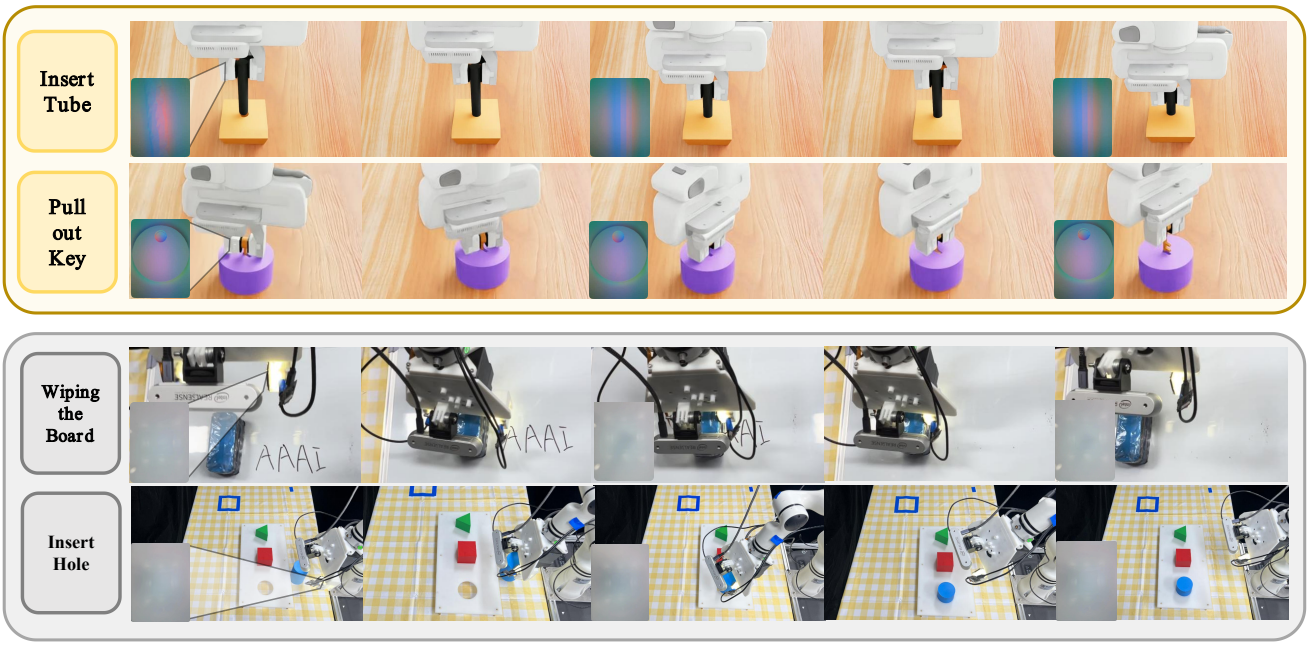

Figure 4: Representative contact-rich task sequences. The upper rows illustrate UniVTAC scenarios spanning insertion, placement, and pulling; the lower rows show the physical *Wiping the Board* and *Insert Hole* tasks. Insets show tactile observations where they are available.

plies its fixed $d(u) \in \mathbb{R}^7$ at every action step. Approximate restoration is handled via reliability weighting; see Eq. (5).

For each branch, we form a local outcome score from task and contact measurements,

$$q_i(c) = \mathbf{w}_q^\top \boldsymbol{\phi}_i(c), \tag{2}$$

where $\boldsymbol{\phi}_i(c)$ contains task outcome, marker-based contact retention, and restoration quality with fixed weights $\mathbf{w}_q$. Because $q_i(c)$ is meaningful only within a single decision group, we compare each residual with the base action at the same decision point,

$$\Delta q_i(u) = q_i(u) - q_i(\varnothing). \tag{3}$$

Using a threshold $\delta = 0.05$, we partition the residual options into positive, neutral, and negative sets,

$$\begin{aligned} \mathcal{U}_i^+ &= \{u \in \mathcal{R}_i : \Delta q_i(u) > \delta\}, \\ \mathcal{U}_i^0 &= \{u \in \mathcal{R}_i : |\Delta q_i(u)| \le \delta\}, \\ \mathcal{U}_i^- &= \{u \in \mathcal{R}_i : \Delta q_i(u) < -\delta\}. \end{aligned} \tag{4}$$

This induces preference relations: positive $\succ$ base $\succ$ negative, and positive $\succ$ neutral; neutral options are not directly compared with the base action.

We weight each preference by both its score gap and the quality of its restored branches,

$$\begin{aligned} w_{i,c_i^+,c_i^-} &= \max\{\delta, q_i(c_i^+) - q_i(c_i^-)\}\kappa_i(c_i^+)\kappa_i(c_i^-), \\ \kappa_i(c) &= \begin{cases} 1, & c = \varnothing \text{ or restoration checks pass,} \\ 0.25, & \text{otherwise.} \end{cases} \end{aligned} \tag{5}$$

This weighting favors large outcome gaps from reliably restored branches without treating approximate restoration as exact counterfactual evidence. We fit the scalar comparator $e_i(c)$ with a weighted Bradley–Terry objective (Bradley and Terry 1952),

$$\mathcal{L}_{\text{rank}} = \mathbb{E}_i \, \mathbb{E}_{(c_i^+, c_i^-)} \left[-w_{i,c_i^+,c_i^-} \log \sigma\big(e_i(c_i^+) - e_i(c_i^-)\big) \right]. \tag{6}$$

The resulting objective learns local effect orderings without requiring a globally consistent branch-score scale; the post-rollout scores $q_i(c)$ serve only as offline supervision targets.

Adaptability and Effect Scoring The state representation x_i concatenates the marker-derived contact descriptor m_i, visual-proprioceptive observations, language embedding, execution history, and a base-action encoding $\phi_A(A_i^{\text{ref}})$. Crucially, all components of x_i are available at inference time, unlike the post-rollout terms in $\boldsymbol{\phi}_i(c)$ that are used only for offline target construction. EGM predicts

$$g_i = G_\theta(x_i), \qquad e_i(c) = E_\theta(x_i, c). \tag{7}$$

g_i estimates the potential for a useful residual; $e_i(u) - e_i(\varnothing)$ compares a residual with the base action under the current contact state.

The positive set in Eq. (4) directly defines the binary adaptability target,

$$y_i^{\text{adapt}} = \begin{cases} 1, & \mathcal{U}_i^+ \neq \varnothing, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

We optimize the per-decision binary cross-entropy

$$\mathcal{L}_{\text{adapt}} = -y_i^{\text{adapt}} \log \sigma(g_i) - (1 - y_i^{\text{adapt}}) \log(1 - \sigma(g_i)),$$

where σ is the logistic sigmoid. The EGM objective is

$$\mathcal{L}_{\mathrm{EGM}} = \mathcal{L}_{\mathrm{rank}} + \lambda_{\mathrm{adapt}} \mathcal{L}_{\mathrm{adapt}}. \tag{9}$$

λ_{adapt} balances the two objectives. EGM is trained to convergence and frozen before its outputs are used to construct RAM supervision targets.

Residual Action Modulation

RAM converts EGM's within-state preference evidence into two deployment-time decisions: which action choice to execute, and at what tactile-conditioned gain to apply the selected residual.

Preference-Guided Residual Selection Because EGM is trained to rank choices within a single branch group, its raw scores and score gaps are ordinal within each state and should not be compared across states. Let $r_i^{\mathrm{EGM}}(c) \in [0,1]$ denote the normalized ordinal rank of c in the ordering induced by $\{e_i(c) : c \in \mathcal{C}_i\}$, and let $\mathbf{r}_i^{\mathrm{EGM}}$ collect these ranks. This encoding preserves only the within-state order of the current option set.

RAM action-choice labels are constructed only within the corresponding branch group:

$$c_i^{\mathrm{tar}} = \begin{cases} \varnothing, & \mathcal{U}_i^+ = \varnothing, \\ \arg\max\limits_{u \in \mathcal{U}_i^+} \Delta q_i(u), & \text{otherwise.} \end{cases} \tag{10}$$

Thus, the target identifies whether to retain the base action or, when a positive residual exists, which observed residual performed best, without comparing outcome margins across states. The selector maps the current state, option set, and EGM evidence to

$$p_i(a) = \Pi_\psi\left(a \mid x_i, \mathcal{R}_i, g_i, \mathbf{r}_i^{\mathrm{EGM}}\right), \qquad a \in \{\varnothing\} \cup \mathcal{R}_i. \tag{11}$$

We train Π_ψ via cross-entropy against c_i^{tar}; notably, the base action is treated as an ordinary class label rather than a post-hoc fallback gate. At deployment, $u_i^* = \arg\max_a p_i(a)$ uses only current observations and EGM evidence.

Tactile-Conditioned Residual Scaling Given the selected residual direction u_i^*, a scaling network predicts the execution gain from x_i and the bilateral tactile summary z_i:

$$\alpha_i = \sigma(M_\eta(x_i, u_i^*, z_i)), \qquad \Delta A_i = \begin{cases} 0, & u_i^* = \varnothing, \\ \alpha_i d(u_i^*), & \text{otherwise.} \end{cases} \tag{12}$$

z_i encodes bilateral tactile images via per-sensor spatial statistics and temporal differences. By design, z_i modulates only the gain α_i; it cannot alter the base action, residual direction, or candidate set.

Because branch data is collected at a single residual amplitude, there is no empirical ground truth for the optimal scale. We therefore repurpose the ordinal rank evidence as an auxiliary gain target, acknowledging that it serves as a within-state proxy rather than a physical magnitude:

$$\alpha_i^{\mathrm{tar}} = \begin{cases} 0, & c_i^{\mathrm{tar}} = \varnothing, \\ r_i^{\mathrm{EGM}}(c_i^{\mathrm{tar}}), & \text{otherwise.} \end{cases} \tag{13}$$

The combined RAM objective is $\mathcal{L}_{\mathrm{RAM}} = \mathcal{L}_{\mathrm{selector}} + \lambda_\alpha \mathcal{L}_{\mathrm{scale}}$, where $\mathcal{L}_{\mathrm{selector}}$ is cross-entropy for action choice and $\mathcal{L}_{\mathrm{scale}}$ is weighted binary cross-entropy for the auxiliary gain target. Supplementary multiscale rollouts directly evaluate whether the deployed α_i produces favorable outcomes from an evaluated scale set on held-out states.

Experiments

Having formulated tactile feedback as evidence for selecting and scaling residual corrections around a frozen VLA policy, we now evaluate whether this formulation translates into consistent task-level gains and separately examine the quality of its learned components. The experiments address three questions:
(1) Does ViTaR improve task success over its frozen VLA base policy across simulated and physical contact-rich tasks?
(2) Does EGM's within-state preference learning yield reliable effect evidence for selecting residual corrections?
(3) Do RAM's effect-guided residual selection and tactile-conditioned scaling yield better residual control than direct residual RL?

Experimental Setup

Benchmarks and Tasks UniVTAC (Chen et al. 2026) provides synchronized visual, proprioceptive, and tactile observations for seven contact-rich tasks: *Lift Bottle*, *Pull-out Key*, *Lift Can*, *Put Bottle*, *Insert HDMI*, *Insert Tube*, and *Grasp Classify*. We additionally evaluate *Insert Hole*, *Lift Bottle*, and *Wiping the Board* on a physical robot. The first two mirror their UniVTAC counterparts in contact requirements, while *Wiping the Board* introduces sustained sliding contact not present in the simulated suite. Representative sequences appear in Fig. 4.

Hardware Setup Physical demonstrations are collected through ExUMI, a VR-operated UMI-style interface (Chi et al. 2024), with 9DTact vision-based tactile sensors (Lin et al. 2024). The physical system comprises a 6-DoF RealMan RM-65B, a custom tactile gripper driven by a DH Robotics PGIA-series motor, and wrist and third-person RealSense D455 cameras. Supplementary material details the platform and sensing configuration.

Baselines For ACT (Zhao et al. 2023), VITaL (George et al. 2024), and ACT+UniVTAC, we report the task-level results provided by the UniVTAC benchmark (Chen et al. 2026). For Tactile-VLA, we adapt its tactile-fusion architecture (Huang et al. 2025) to the UniVTAC observation and control interface, testing direct tactile fusion as a baseline. We also evaluate $\pi_{0.5}$ (Physical Intelligence et al. 2025), a strong open-source VLA model without tactile input, under the same seven-task protocol. OpenVLA-OFT (Qiu et al. 2023) is our frozen base-policy reference, fine-tuned with OFT on 50 trajectories per task. We do not reproduce AT-VLA, TacFiLM, ResTacVLA, TacCoRL, or TORL-VLA (Li et al. 2026; Morissette et al. 2026; Zhang et al. 2026; Ma et al. 2026; Zheng et al. 2026): at the time of evaluation, we could not obtain public implementations and task-specific checkpoints that were reproducible under the shared UniVTAC

Method	Lift Bottle	Pull-out Key	Lift Can	Put Bottle	Insert HDMI	Insert Tube	Grasp Classify	Avg.
ACT	42.0%	28.0%	20.0%	28.0%	15.0%	45.0%	50.0%	32.6%
VITaL	72.0%	47.0%	8.0%	32.0%	6.0%	34.0%	**100.0%**	42.7%
ACT + UniVTAC	71.0%	46.0%	29.0%	31.0%	28.0%	56.0%	99.0%	51.4%
Tactile-VLA	**97.0%**	32.0%	15.0%	10.0%	12.0%	56.0%	58.0%	40.0%
$\pi_{0.5}$	92.0%	38.0%	**70.0%**	14.0%	18.0%	38.0%	68.0%	48.3%
OpenVLA-OFT (ref.)	45.0%	33.0%	21.0%	8.0%	9.0%	55.0%	44.0%	30.7%
ViTaR (ours)	88.0%	**55.0%**	44.0%	**35.0%**	**38.0%**	**69.0%**	**100.0%**	**61.3%**

Table 1: Success rate (%) on seven contact-rich UniVTAC tasks. ViTaR operates on top of the frozen OpenVLA-OFT base policy. Best reported result in each column is in bold.

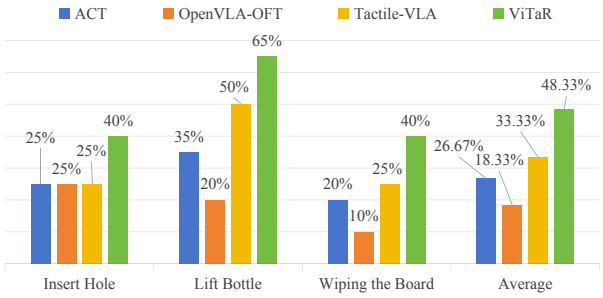

Figure 5: End-to-end success rates on physical-robot contact-rich tasks. Within each group, bars follow the fixed left-to-right order ACT, OpenVLA-OFT, Tactile-VLA, and ViTaR; the rightmost group reports average success.

control interface and physical-robot protocol. We therefore restrict end-to-end comparison to accessible baselines for which a common evaluation interface can be established; the recent methods above are discussed in Related Work rather than treated as directly comparable reproduced systems.

Evaluation Protocol and Metrics We evaluate each UniVTAC method–task pair over 100 episodes and each physical task over 20 trials, constrained by physical execution cost, using end-to-end binary success as the primary metric. The physical protocol tests consistency across insertion, sustained sliding, and grasping rather than precisely estimating any single-task rate. ViTaR uses 20 restored decision groups per task for short-horizon branch supervision, whose aggregate execution horizon is shorter than that of the 50 expert demonstrations used to fine-tune OpenVLA-OFT. For ablations, Rec., Pres., and Unsafe. denote paired recovery, paired preservation, and safety/early termination, respectively; P>B and B>N are held-out positive-over-base and base-over-negative ordering accuracies; FPR is measured at 80% recall; and Off. T1 is held-out agreement with the best observed positive residual. Supplementary multiscale and option-set controls report scale-bin accuracy, local outcome regret, the fraction of groups with a positive option, and mean best branch gain. All non-success metrics are post-hoc diagnostics and never inputs to the deployed policy.

Main Results

UniVTAC Success Rates Table 1 compares ViTaR with its OpenVLA-OFT base policy and representative imitation, tactile, and foundation-policy baselines on seven contact-rich UniVTAC tasks. ViTaR attains the highest average success rate (61.3%), doubling the frozen base policy's performance (30.7% $\rightarrow$ 61.3%, a 30.6pp gain). Because ViTaR retains the base policy's semantic action direction rather than replacing it, its achievable absolute performance remains constrained by the frozen base's semantic action quality. This improvement is achieved without modifying the VLA's pretrained representations, indicating that contact-conditioned residual corrections can substantially improve a frozen policy while preserving its visual-language action prior intact.

Relative to OpenVLA-OFT, ViTaR improves success on all seven UniVTAC tasks (Table 1), with particularly large gains on Lift Bottle, Put Bottle, Insert HDMI, and Grasp Classify. ViTaR does not dominate every task individually: $\pi_{0.5}$ is stronger on Lift Can (70% vs. 44%), and Tactile-VLA achieves the highest Lift Bottle result (97% vs. 88%). We therefore interpret the comparison as broad improvement around the frozen base policy rather than uniform task-level dominance.

Physical-Robot Success Rates Figure 5 reports end-to-end success on the three physical contact-rich tasks. ViTaR achieves a 48.3% average success rate, improving on OpenVLA-OFT by 30.0 percentage points, on Tactile-VLA by 15.0 percentage points, and on ACT by 21.7 percentage points. The improvement appears in all three tasks: ViTaR reaches 40.0% on *Insert Hole*, 40.0% on *Wiping the Board*, and 65.0% on *Lift Bottle*. These tasks require qualitatively different contact behaviors (precision alignment, sustained sliding, and stable grasping), so the physical gains reflect the generality of bounded tactile modulation across contact modalities rather than a single task-specific effect. The absolute success level (48.3%) is bounded by the base policy's semantic action quality; improvements over the base are consistent with the simulated margins. Given the three-task, 20-trial physical protocol, these results provide limited transfer evidence rather than resolving the sim-to-real gap or establishing broad real-world generalization.

Ablation Studies

Comparison with direct residual RL. Table 2 compares learning formulations around the same frozen OpenVLA-OFT reference, residual coordinates, decision points, and safety limits. PPO-option selects from the same discrete base-action/residual interface, whereas SAC-residual learns a bounded continuous residual without EGM or RAM supervision. ViTaR reaches 61.3% average success, versus 36.0% and 32.0%, and has lower unsafe/early-termination rates (0.04 versus 0.25 and 0.31). Within the shared bounded-residual protocol, these results favor effect-guided residual selection over the tested direct RL formulations. We emphasize that this is a comparison of learning formulations under matched conditions, not a general claim about RL for robotic manipulation.

Method	Rec. ↑	Pres. ↑	Unsafe. ↓	Avg. (%) ↑
PPO-option	.55	.66	.25	36.0%
SAC-residual	.45	.74	.31	32.0%
Full ViTaR	**.79**	**.92**	**.04**	**61.3%**

Table 2: Comparison with direct residual RL. Recovery and preservation are post-hoc matched-start measures; Unsafe. is the safety/early-termination rate.

Variant	Rec. ↑	Unsafe. ↓	Avg. ↑
w/o State	.46	.16	41.0%
w/o Tac.	.61	.13	46.0%
w/o Both	.40	.19	36.0%
Fixed Scale	.57	.11	45.0%
Full ViTaR	**.79**	**.04**	**61.3%**

Table 3: Contact information and continuous scaling ablation. "w/o State" removes the marker-derived contact descriptor, "w/o Tac." removes the tactile summary, and "w/o Both" removes both. Rec. is paired recovery under matched-start evaluation.

Contact information and continuous scaling. Table 3 separates the marker-derived contact descriptor m_i, used for residual selection, from the tactile summary z_i, used for continuous scaling. Fixing the scale reduces success from 61.3% to 45.0%. Removing m_i lowers paired recovery from 0.79 to 0.46, while removing z_i lowers success to 46.0%; removing both yields 36.0% success and 0.19 unsafe/early termination. These controls support complementary and non-redundant roles: m_i is critical for *whether* to intervene (Rec. drops from 0.79 to 0.46 without it), while z_i is critical for *how much* to apply (success drops from 61.3% to 46.0% without it). Supplementary multiscale rollouts directly assess the deployed scale choices against observed outcomes.

Effect-guided modeling. Table 4 isolates the contribution of each EGM design choice by progressively adding components to a direct score-regression baseline. Pairwise ranking replaces pointwise regression; outcome-gap weighting emphasizes informative pairs; and restoration-aware weighting down-weights unreliably restored branches. The cumulative effect improves P>B / B>N ordering from 0.63 / 0.60 to 0.83 / 0.88 and reduces FPR@80% from 0.40 to 0.12, translating to a 23.3pp end-to-end success gain (38.0% → 61.3%).

Variant	P>B ↑	B>N ↑	FPR ↓	Avg. (%) ↑
RawScore	.63	.60	.40	38.0%
PairRank	.73	.72	.25	40.0%
+ Gap	.71	.76	.29	44.0%
+ Restore	**.83**	**.88**	**.12**	**61.3%**

Table 4: EGM ablation on held-out branch groups. FPR is measured at 80% recall.

Residual action modulation. Table 5 evaluates how EGM's effect evidence is best converted into a deployment-time action choice. Using EGM scores directly (EGM-Greedy) or restricting supervision to a single evidence source (Outcome-only, Effect-only) each degrades performance. The full joint selector, conditioned on both EGM rankings and branch-outcome evidence, attains the strongest held-out top-1 agreement (0.79), paired recovery (0.79), and average success (61.3%), indicating that both evidence sources contribute non-redundantly.

Variant	Off. T1 ↑	Rec. ↑	Avg. (%) ↑
EGM-Greedy	.45	.41	36.0%
Outcome-only	.64	.56	42.0%
Effect-only	.51	.50	42.0%
Unweighted	.68	.61	47.0%
Two-stage	.62	.57	44.0%
Full RAM	**.79**	**.79**	**61.3%**

Table 5: RAM ablation. Off. T1 is held-out oracle-positive top-1 agreement; Rec. uses matched-start rollouts.

Conclusion

We presented ViTaR, a framework that reframes tactile feedback from an action-generating perceptual input to a bounded execution modulator for frozen VLA policies. By decomposing contact-aware adaptation into effect-guided residual ranking and tactile-conditioned gain scaling, ViTaR improves the frozen VLA baseline by 30.6 percentage points on the UniVTAC benchmark and demonstrates consistent gains on physical-robot tasks spanning insertion, sliding, and grasping, all without modifying the base policy's pretrained representations. Future work includes extending the paradigm to multi-finger dexterous manipulation with heterogeneous tactile sensors, scaling the branch-comparison protocol to longer-horizon tasks.

References

Black, K.; Brown, N.; Driess, D.; Esmail, A.; Equi, M.; Finn, C.; Fusai, N.; Groom, L.; Hausman, K.; Ichter, B.; Levine, S.; Li, Y.; Liang, J.; Pertsch, K.; Ramos, F.; Sanketi, P.; Vuong, Q.; Xia, F.; Xiao, T.; Xu, P.; and Zeng, A. 2024. π_0: A Vision-Language-Action Flow Model for General Robot Control. arXiv:2410.24164.

Bradley, R. A.; and Terry, M. E. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4): 324–345.

Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Chen, X.; Choromanski, K.; Ding, T.; Driess, D.; Dubey, A.; Finn, C.; Florence, P.; Fu, C.; Arenas, M. G.; Gopalakrishnan, K. H.; Hausman, K.; Herzog, A.; Hsu, J.; Ichter, B.; Irpan, A.; Joshi, N.; Julian, R.; Kalashnikov, D.; Kuang, Y.; Leal, I.; Lee, K.-H.; Lee, T.-W. E.; Levine, S.; Lu, Y.; Michalewski, H.; Mordatch, I.; Pertsch, K.; Rao, K.; Reymann, K.; Ryoo, M. S.; Salazar, G.; Sanketi, P. R.; Sermanet, P.; Singh, J.; Singh, A.; Soricut, R.; Tran, H. T.; Vanhoucke, V.; Vuong, Q.; Wahid, A.; Welker, S.; Wohlhart, P.; Xia, F.; Xiao, T.; Xu, P.; Xu, S.; Yu, T.; and Zitkovich, B. 2023a. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *Proceedings of the Conference on Robot Learning*.

Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Dabis, J.; Finn, C.; Gopalakrishnan, K.; Hausman, K.; Herzog, A.; Hsu, J.; Ibarz, J.; Ichter, B.; Irpan, A.; Jackson, T.; Jesmonth, S.; Joshi, N. J.; Julian, R.; Kalashnikov, D.; Kuang, Y.; Leal, I.; Lee, K.-H.; Levine, S.; Lu, Y.; Malla, U.; Manjunath, D.; Mordatch, I.; Nachum, O.; Parada, C.; Peralta, J.; Perez, E.; Pertsch, K.; Quiambao, J.; Rao, K.; Ryoo, M.; Salazar, G.; Sanketi, P.; Sayed, K.; Singh, J.; Sontakke, S.; Stone, A.; Tan, C.; Tran, H.; Vanhoucke, V.; Vega, S.; Vuong, Q.; Xia, F.; Xiao, T.; Xu, P.; Xu, S.; Yu, T.; and Zitkovich, B. 2023b. RT-1: Robotics Transformer for Real-World Control at Scale. arXiv:2212.06817.

Brown, D. S.; Goo, W.; and Niekum, S. 2019. Better-than-Demonstrator Imitation Learning via Automatically-Ranked Demonstrations. arXiv:1907.03976.

Cao, Z.; Tian, D.; Guan, R.; Mu, Y.; Sun, X.; Liang, S.; Liu, D.; Huang, T.; Yue, Y.; Ding, H.; Fang, B.; Zhou, A.; Han, Q.-L.; and Xiong, H. 2026. Tactile-Based Multimodal Fusion in Embodied Intelligence: A Survey of Vision, Language, and Contact-Driven Paradigms. arXiv:2605.17336.

Chen, B.; Wan, W.; Chen, T.; Guo, X.; Xu, C.; Qi, Y.; Zhang, H.; Wu, L.; Xu, T.; Li, Z.; Wu, Y.; Li, R.; Yang, X.; Luo, P.; Sui, W.; and Mu, Y. 2026. UniVTAC: A Unified Simulation Platform for Visuo-Tactile Manipulation Data Generation, Learning, and Benchmarking. arXiv:2602.10093.

Chi, C.; Feng, S.; Du, Y.; Xu, Z.; Cousineau, E.; Burchfiel, B.; and Song, S. 2023. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Proceedings of Robotics: Science and Systems*.

Chi, C.; Xu, Z.; Feng, S.; Cousineau, E.; Du, Y.; Burchfiel, B.; Tedrake, R.; and Song, S. 2024. Universal Manipulation Interface: In-the-Wild Robot Teaching Without In-the-Wild Robots. In *Robotics: Science and Systems*.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

George, A.; Gano, S.; Katragadda, P.; and Farimani, A. B. 2024. VITaL Pretraining: Visuo-Tactile Pretraining for Tactile and Non-Tactile Manipulation Policies. arXiv:2403.11898.

Ghosh, D.; Walke, H.; Pertsch, K.; Black, K.; Mees, O.; Dasari, S.; Hejna, J.; Xu, C.; Luo, J.; Kreiman, T.; Tan, Y. L.; Chen, L. Y.; Sanketi, P.; Vuong, Q.; Xiao, T.; Finn, C.; Levine, S.; and Zhao, T. Z. 2024. Octo: An Open-Source Generalist Robot Policy. arXiv:2405.12213.

Haldar, S.; Pari, J.; Rai, A.; and Pinto, L. 2023. Teach a Robot to FISH: Versatile Imitation from One Minute of Demonstrations. arXiv:2303.01497.

Higuera, C.; Sharma, A.; Bodduluri, C. K.; Fan, T.; Lancaster, P.; Kalakrishnan, M.; Kaess, M.; Boots, B.; Lambeta, M.; Wu, T.; and Mukadam, M. 2024. Sparsh: Self-Supervised Touch Representations for Vision-Based Tactile Sensing. In *Proceedings of the Conference on Robot Learning*.

Hoque, R.; Balakrishna, A.; Novoseller, E. R.; Wilcox, A.; Brown, D. S.; and Goldberg, K. 2022. ThriftyDAgger: Budget-Aware Novelty and Risk Gating for Interactive Imitation Learning. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, 598–608.

Hua, Q.; Li, X.; Yan, Z.; Li, Y.; Zhang, C.; Li, Y.; and Liu, Y. 2026. VTouch++: A Multimodal Dataset with Vision-Based Tactile Enhancement for Bimanual Manipulation. arXiv:2604.20444.

Huang, J.; Wang, S.; Lin, F.; Hu, Y.; Wen, C.; and Gao, Y. 2025. Tactile-VLA: Unlocking Vision-Language-Action Model's Physical Knowledge for Tactile Generalization. arXiv:2507.09160.

Johannink, T.; Bahl, S.; Nair, A.; Luo, J.; Kumar, A.; Loskyll, M.; Ojea, J. A.; Solowjow, E.; and Levine, S. 2019. Residual Reinforcement Learning for Robot Control. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 6023–6029.

Kim, M. J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E.; Lam, G.; Sanketi, P.; Vuong, Q.; Kollar, T.; Burchfiel, B.; Tedrake, R.; Sadigh, D.; Levine, S.; Liang, P.; and Finn, C. 2024. OpenVLA: An Open-Source Vision-Language-Action Model. arXiv:2406.09246.

Lee, K.; Smith, L.; and Abbeel, P. 2021. PEBBLE: Feedback-Efficient Interactive Reinforcement Learning via Relabeling Experience and Unsupervised Pre-training. In *International Conference on Machine Learning*.

Li, X.; Cai, M.; Xu, J.; Zhu, J.; Fan, H.; Shen, Y.; Ren, G.; and Dong, H. 2026. AT-VLA: Adaptive Tactile Injection for Enhanced Feedback Reaction in Vision-Language-Action Models. arXiv:2605.07308.

Lin, C.; Zhang, H.; Xu, J.; Wu, L.; and Xu, H. 2024. 9DTact: A Compact Vision-Based Tactile Sensor for Accurate 3D

Shape Reconstruction and Generalizable 6D Force Estimation. *IEEE Robotics and Automation Letters*, 9(2): 923–930.

Ma, S.; Liang, Y.; Yu, C.; Chen, Y.; Su, H.; Zhu, Y.; Yang, Y.; and Jiang, C. 2026. TacCoRL: Integrating Tactile Feedback into VLA via Simulation. arXiv:2606.11743.

Mandlekar, A.; Xu, D.; Martín-Martín, R.; Zhu, Y.; Fei-Fei, L.; and Savarese, S. 2020. Human-in-the-Loop Imitation Learning using Remote Teleoperation. arXiv:2012.06733.

Morissette, C.; Abyaneh, A.; Chang, W.-D.; Houssaini, A.; Meger, D.; Lin, H.-C.; Tremblay, J.; and Dudek, G. 2026. Tactile Modality Fusion for Vision-Language-Action Models. arXiv:2603.14604.

Ni, S.; Zhang, H.; Wei, Z.; Chen, G.; Zhang, C.; Shi, Y.; and Wang, J. 2026. TactiDex: A Real-World Tactile-Guided Benchmark for Human-Like Dexterous Manipulation. arXiv:2607.09190.

Niu, Y.; Fang, Z.; Chen, B.; Zhou, S.; Senthilkumaran, R. K.; Zhang, H.; Chen, B.; Qiu, C.; Tseng, H. E.; Francis, J.; and Zhao, D. 2026. Learning Versatile Humanoid Manipulation with Touch Dreaming. arXiv:2604.13015.

Physical Intelligence; Black, K.; Brown, N.; Darpinian, J.; Dhabalia, K.; Driess, D.; Esmail, A.; Equi, M.; Finn, C.; Fusai, N.; Galliker, M. Y.; Ghosh, D.; Groom, L.; Hausman, K.; Ichter, B.; Jakubczak, S.; Jones, T.; Ke, L.; LeBlanc, D.; Levine, S.; Li-Bell, A.; Mothukuri, M.; Nair, S.; Pertsch, K.; Ren, A. Z.; Shi, L. X.; Smith, L.; Springenberg, J. T.; Stachowicz, K.; Tanner, J.; Vuong, Q.; Walke, H.; Walling, A.; Wang, H.; Yu, L.; and Zhilinsky, U. 2025. $\pi_{0.5}$: A Vision-Language-Action Model with Open-World Generalization. arXiv:2504.16054.

Qiu, Z.; Liu, W.; Feng, H.; Xue, Y.; Feng, Y.; Liu, Z.; Zhang, D.; Weller, A.; and Schölkopf, B. 2023. Controlling text-to-image diffusion by orthogonal finetuning. *Advances in Neural Information Processing Systems*, 36: 79320–79362.

Silver, T.; Allen, K.; Tenenbaum, J.; and Kaelbling, L. 2019. Residual Policy Learning. arXiv:1812.06298.

Tung, Y.-S.; Wu, Y.; Jiang, W.; Roncone, A.; and Hayes, B. 2026. Risk-Aware Preference Learning for Stochastic Outcomes. arXiv:2607.15483.

Wang, L.; Chen, X.; Zhao, J.; and He, K. 2024. Scaling Proprioceptive-Visual Learning with Heterogeneous Pre-trained Transformers. arXiv:2409.20537.

Wang, Y.; Ding, P.; Li, L.; Cui, C.; Ge, Z.; Tong, X.; Song, W.; Zhao, H.; Zhao, W.; Hou, P.; et al. 2026. Vla-adapter: An effective paradigm for tiny-scale vision-language-action model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 40, 18638–18646.

Wu, S.; You, L.; Zhu, J.; Liu, Y.; Zhang, C.; Liu, J.; Wang, W.; Li, Q.; Li, J.; and Zhao, H. 2026. Tactile-WAM: Touch-Aware World Action Model with Tactile Asymmetric Attention. arXiv:2606.26663.

Yuan, C.; Zhang, Z.; Zhou, M.; Chen, W.; Wang, Y.; Liu, Z.; Niu, D.; Wang, S.; Zhang, H.; Zhang, W.; Hu, Y.; Gong, Y.; Xing, W.; Wen, C.; Lu, C.; Zhang, K.; and Gao, Y. 2026a. FTP-1: A Generalist Foundation Tactile Policy Across Tactile Sensors for Contact-Rich Manipulation. arXiv:2606.13102.

Yuan, Y.; Hao, J.; Ma, Y.; Dong, Z.; Liang, H.; Liu, J.; Feng, Z.; Zhao, K.; and Zheng, Y. 2024. Uni-RLHF: Universal Platform and Benchmark Suite for Reinforcement Learning with Diverse Human Feedback. arXiv:2402.02423.

Yuan, Z.; Wang, R.; Zhao, D.; Yang, B.; and Min, B.-C. 2026b. PrefMoE: Robust Preference Modeling with Mixture-of-Experts Reward Learning. arXiv:2605.00384.

Zhang, P.; Xie, B.; Meng, X.; Guo, X.; Hao, C.; Deng, F.; Cheng, L.; and Wang, T. 2026. Feeling the Unexpected: ResTacVLA for Contact-Rich Manipulation via Residual Tactile Representation. arXiv:2607.03387.

Zhao, T. Z.; Kumar, V.; Levine, S.; and Finn, C. 2023. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. In *Proceedings of Robotics: Science and Systems*.

Zheng, H.; Yang, Y.; Ma, K.; Xu, S.; Xie, T.; Li, G.; Wang, X.; Ma, Y.; Liu, S.; Mao, Y.; and Liu, B. 2026. TORL-VLA: Tactile Guided Online Reinforcement Learning for Contact-Rich Manipulation. arXiv:2606.09337.