PiVLA: Physical-Interaction-aware Vision-Language-Action Policy for Force-Sensitive Manipulation

Anonymous authors
Paper under double-blind review

Abstract

Vision-Language-Action (VLA) models produce semantically and geometrically correct motions, yet the same motion can crush or drop an object depending on grasp force. Recent force-aware VLAs treat force as an input or control target but do not represent what interaction a given context calls for. We present PiVLA, a VLA that learns an intermediate *physical interaction representation*: a dedicated token fusing vision, language, and contact observations, supervised by per-group safe-force distributions and recurring interaction prototypes, and used to directly guide a flow-matching action expert. We validate on two real-robot tasks transferring i) damage-prone produce and ii) appearance-identical cartons with disjoint safe grasp intervals, where the required grasp can only be resolved through contact. PiVLA outperforms force-agnostic and equal-input force-aware baselines, and ablations confirm the contribution of each component. Code and data will be released.

1 Introduction

Vision-Language-Action models (VLAs) extend the semantic understanding of large Vision-Language Models (Driess et al., 2023) into real-world action. By mapping visual observations and instructions into robot commands, and trained on large multi-robot demonstration corpora (O'Neill et al., 2024; Khazatsky et al., 2024), models such as RT-2 (Brohan et al., 2023), OpenVLA (Kim et al., 2024), π_0 (Black et al., 2024), and $\pi_{0.5}$ (Intelligence et al., 2025) have made rapid progress across diverse manipulation scenarios. Pretrained vision-language knowledge gives a robot an increasingly powerful prior for deciding *what to do* in the physical world.

However, correct actions do not necessarily imply safe physical interaction (Brunke et al., 2022). The same actions may lead to substantially different physical outcomes depending on the object, manipulation stage, and interaction dynamics. Consider two visually identical closed containers that differ only in fill level: the grasp that secures the full one crushes the empty one, and the grasp the empty one tolerates lets the full one slip (Figure 1). Which case is at hand is unrecoverable from vision alone, a limit that also bounds the physical reasoning of vision-language models (Gao et al., 2024)—it becomes available only at contact, after an excessive grasp has already caused deformation. Human grasping resolves the same ambiguity by pairing a predictive grip plan with tactile signals that correct it after contact onset (Johansson & Flanagan, 2009). Safe interaction thus reduces to neither recognition nor reactive force control (Hogan, 1984; Raibert & Craig, 1981): recognition supplies the regime a category typically requires but cannot resolve the instance; contact resolves the instance only after the policy has committed. A policy must hold both, and must hold the first before contact begins.

Force-aware VLAs already address part of this, admitting force as an additional modality (Yu et al., 2026) or making its regulation context-dependent through stage-dependent force prompts (Li et al., 2026) or tactile-feedback targets (Huang et al., 2025; Bi et al., 2026). In each case, however, the appropriate interaction is carried by a single force value and never by a description of what the context admits. The policy is never asked the prior question: **what physical interaction should the robot maintain to accomplish the task safely?**

We argue that a VLA for safe physical interaction should learn an intermediate physical interaction representation, bridging semantic understanding and force-aware action generation. A fixed fragility

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

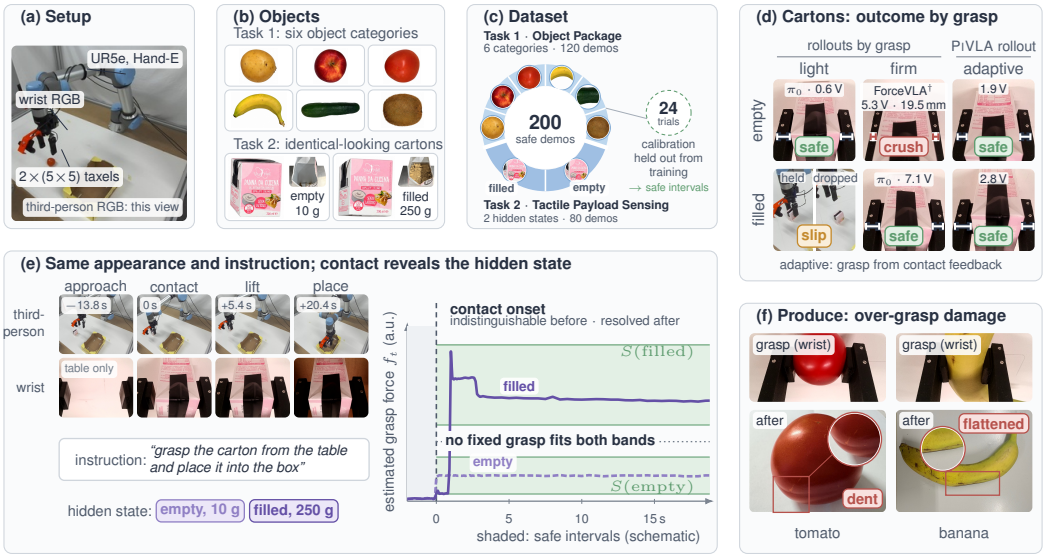

Figure 1: **Force-sensitive transfer with appearance-identical objects.** (a) Setup. (b) *Object Package* (Task 1) and *Tactile Payload Sensing* (Task 2). (c) 120 *Object Package* and 80 *Tactile Payload Sensing* safe demonstrations; 24 calibration trials held out. (d) *Tactile Payload Sensing* rollouts by applied grasp (wrist view, peak grasp): a light grasp drops the filled carton and a firm one crushes the empty carton, its fingertips sinking in until their visible width halves (arrows); PiVLA succeeds on both. Each cell is one rollout, from different policies. (e) Grasp force is uninformative before contact and separates after. (f) Over-grasp damage: the grasp and the instance afterwards; the tomato dent accumulated over 15 grasps by all three policies.

label cannot express this: a single object category may admit two disjoint safe-force intervals, so no constant target is correct for all instances. We instead model appropriate physical behavior as a distribution over interaction forces, whose center describes the regime a context calls for and whose spread reflects variation among safe demonstrations. Recurring across objects and stages, these distributions span a space of physical interaction regimes that can be learned from force together with vision-language context and used to guide action generation.

Studying this requires data that existing datasets do not provide: appearance and instruction held fixed while physical state changes, paired with calibrated slip and damage boundaries. We collect a real-robot dataset over two task families with these properties and build PiVLA, in which a contextualized physical interaction token (fusing vision, language, and contact) predicts the stage-group force distribution and its interaction prototype, and directly guides action generation. Our main contributions are:

- PiVLA-data: a real-world dataset that decouples visual appearance from physical state by construction and pairs demonstrations with calibrated safe-force intervals.
- PiVLA: an end-to-end VLA framework that learns a physical interaction representation from fused vision–language–contact context, supervises it with demonstrated force distributions and interaction prototypes, and uses it to guide safe, context-aware action generation.
- Evaluation against force-agnostic and equal-input force-aware baselines on two real-robot tasks and unseen conditions, with ablations isolating the contributions of physical awareness, direct action guidance, and contact-derived representations.

2 Related Work

Vision–language–action models. Building on language-conditioned robot transformers (Brohan et al., 2022) and embodied multimodal language models (Driess et al., 2023), RT-2 (Brohan et al., 2023) represents actions as tokens within a vision–language model; OpenVLA (Kim et al., 2024)

provides an open fine-tunable alternative; Octo (Team et al., 2024) and RDT-1B (Liu et al., 2025) pair pretrained backbones with diffusion heads (Chi et al., 2025; Zhao et al., 2023). π_0 (Black et al., 2024) couples a pretrained VLM with a flow-matching (Lipman et al., 2022) action expert; $\pi_{0.5}$ (Intelligence et al., 2025) extends this to open-world generalization. Later work improves action tokenization (Pertsch et al., 2025), fine-tuning (Kim et al., 2025), and backbone insulation (Driess et al., 2026). Our work builds on π_0 and studies how an intermediate physical interaction representation complements semantic grounding in force-sensitive manipulation.

Force-aware VLA policies. Classical manipulation regulates contact through impedance (Hogan, 1984) or hybrid force–position control (Raibert & Craig, 1981); learned policies predict such parameters from observation (Hou et al., 2025). ForceVLA (Yu et al., 2026) integrates six-axis wrench feedback through a force-aware MoE (Shazeer et al., 2017); FD-VLA (Zhao et al., 2026) achieves force awareness without physical sensors via force token distillation; ForceVLA2 (Li et al., 2026) connects force awareness to hybrid force–position control via stage-dependent prompts. These methods represent the appropriate interaction as a single value. PIVLA instead supervises a shared physical token with empirical force distributions and soft prototypes, organizing interaction regimes rather than tracking a target value.

Datasets for contact and force. Large-scale datasets focus on visual and linguistic generality: Open X-Embodiment (O'Neill et al., 2024), BridgeData V2 (Walke et al., 2023), and DROID (Khazatsky et al., 2024) use vision and proprioception; RH20T (Fang et al., 2024) adds force–torque and tactile signals. Force-aware VLA datasets record demonstrated force (Yu et al., 2026; Li et al., 2026); visuo-tactile corpora align touch with vision and language (Fu et al., 2024). None reports safe force deviation bounds, and none holds appearance fixed while physical state changes; our dataset provides both.

Tactile feedback for planning and control. Touch has been used to predict grasp outcomes (Calandra et al., 2017; 2018), learn multimodal representations (Lee et al., 2019; 2020; Li et al., 2022), and act on contact directly (Lin et al., 2025; Huang et al., 2024; Xue et al., 2025), supported by vision-based tactile sensors (Yuan et al., 2017; Lambeta et al., 2020). Tactile–language models reason about object properties from touch (Yu et al., 2024; Yang et al., 2024), and generalist policies have been fine-tuned with tactile inputs grounded through language (Jones et al., 2025). VLA-Touch (Bi et al., 2026) augments a base VLA via a tactile–language planner and a separate interpolant controller; Tactile-VLA (Huang et al., 2025) predicts force and position targets for a hybrid controller. Our distinction is that the same token predicts stage-group force statistics and prototype probabilities (Snell et al., 2017; Caron et al., 2020) while directly guiding action generation, rather than post-processing chunks or supplying setpoints to a separate controller.

3 PIVLA

We introduce PIVLA, a framework that explicitly models physical interaction constraints rather than treating force sensing as a reactive auxiliary input. Safe manipulation depends not only on visual semantics but on how the robot interacts with an object at each task stage; PIVLA learns a compact physical interaction representation that bridges multimodal perception and action generation. An overview of our framework is shown in Fig. 2.

3.1 VISION-LANGUAGE-TACTILE MODALITIES INTEGRATION

In contact-rich manipulation, raw tactile measurements are inherently ambiguous without visual and semantic context. To address this, we encode tactile measurements as explicit multimodal tokens, allowing them to interact directly with vision-language representations prior to physical reasoning. Inspired by π_0 (Black et al., 2024), the PIVLAarchitecture integrates visual, text, and tactile modalities to condition a flow-matching action expert. At time step t, the policy receives an observation tuple $\boldsymbol{o}_t = (\mathbf{I}_t^{\text{ext}}, \mathbf{I}_t^{\text{wrist}}, \ell, \boldsymbol{s}_t, T_t)$, where I_t^{ext} and I_t^{wrist} denote the wrist-mounted and third-person RGB observations, respectively. l is the language task instruction. $\boldsymbol{s}_t \in \mathbb{R}^7$ contains the tool position, rotation vector, and gripper width. A SigLIP-based VLM with a PaliGemma-style architecture processes I_t^{ext}, I_t^{wrist}, and ℓ into a sequence of vision-language contextual embeddings $E_{VL} \in \mathbb{R}^{N_{VL} \times D_m}$. Different from conventional force-aware VLA formulations that directly use a 6D force-torque vector, we retain the spatially distributed fingertip measurements captured by the

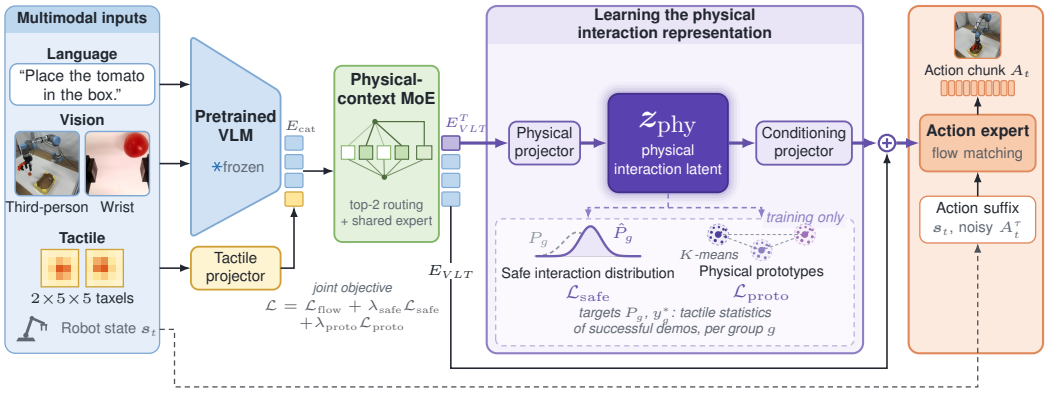

Figure 2: **Overview of PiVLA.** Our model encodes language, RGB, and tactile inputs into a physical interaction latent $\boldsymbol{z}_{\text{phy}}$, which conditions a flow-matching action expert. A frozen VLM encodes language and two RGB views; a tactile projector produces token E_T. A physical-context MoE fuses both into E_{VLT}; the contextualized tactile token E^T_{VLT} is projected to $\boldsymbol{z}_{\text{phy}}$. During training, two auxiliary heads supervise $\boldsymbol{z}_{\text{phy}}$: $\mathcal{L}_{\text{safe}}$ aligns it with the group's successful-demo tactile distribution P_g; $\mathcal{L}_{\text{proto}}$ assigns it to K physical prototypes via K-means. At inference, $\boldsymbol{z}_{\text{phy}}$ is added to E_{VLT} and passed to the action expert. Joint objective: $\mathcal{L} = \mathcal{L}_{\text{flow}} + \lambda_{\text{safe}}\mathcal{L}_{\text{safe}} + \lambda_{\text{proto}}\mathcal{L}_{\text{proto}}$.

tactile sensors. Each fingertip provides a 5×5 taxel array of voltage measurements, which is flattened into a 25-dimensional vector. Thus, the left $\tilde{\boldsymbol{x}}^L_t \in \mathbb{R}^{25}$ and right $\tilde{\boldsymbol{x}}^R_t \in \mathbb{R}^{25}$ fingertip measurements are concatenated. The finger tactile observation is denoted as:

$$T_t = [\tilde{\boldsymbol{x}}^L_t; \tilde{\boldsymbol{x}}^R_t] \in \mathbb{R}^{50}. \tag{1}$$

The tactile observation is projected into a tactile embedding $E_T \in \mathbb{R}^{D_m}$ by a multi-layer perceptron (MLP) layer. Subsequently, the high-dimensional tactile token E_T is appended to the vision-language feature sequence to form the unified multi-modal representation $E_{cat} = [E_{VL}; E_T]$, which serves as the input to the downstream physical-context MoE.

3.2 Physical-context Mixture-of-Experts

While global multimodal self-attention enables broad feature interactions, contact-rich physical manipulation exhibits heterogeneous information processing requirements across different modalities and execution phases. To effectively model these disparate dynamics, we introduce a Physical-Context MoE module. This architecture systematically integrates high-level contextual understanding from visual and linguistic cues with the immediate, fine-grained tactile dynamics captured at the end-effector.

The combined multimodal sequence $E_{cat} \in \mathbb{R}^{(N_{VL}+1)\times D_m}$ is first refined by a Transformer encoder layer with multi-head self-attention (N_{head} heads) and an FFN, yielding $E_{enc} \in \mathbb{R}^{(N_{VL}+1)\times D_m}$. This is then processed by an MoE comprising a shared expert and $E = 4$ routed experts. The shared expert (a standard MLP) is applied uniformly to all tokens; in parallel, a gating network dispatches each token to the top-$k = 2$ most relevant routed experts. Outputs from shared and routed experts are aggregated, merged via a residual connection, and projected to yield the fused sequence E_{VLT}.

3.3 Physical-reasoning

Through the Physical-context MoE, raw tactile tokens are deeply contextualized by visual and semantic constraints. From the fused sequence E_{VLT}, we extract the contextualized tactile token $E^T_{VLT} \in \mathbb{R}^{D_m}$ and map it into a physical interaction latent space:

$$z_{phy} = \phi_{phy}(E^F_{VLT}) \tag{2}$$

where z_{phy} denotes the continuous physical latent embedding. Rather than treating tactile feedback as a passive sensory signal for action regression, we utilize z_{phy} to infer explicit physical constraints

that govern safe and effective contact. We model physical interaction at the level of group $g = (task, object, stage)$, where each group corresponds to a particular manipulation context. Successful expert demonstrations belonging to a given group g provide a statistical foundation for the required physical interaction profile at that specific phase. To translate latent physical features into explicit constraints, our physical reasoning module incorporates two complementary heads: a continuous Safe Interaction Distribution Head and a discrete Physical Interaction Prototype Head.

Safe Interaction Distribution Prediction. To enforce safe contact boundaries, we represent physical interaction constraints as a continuous distribution rather than a deterministic target. For each context group g, we precompute the empirical stage-level tactile statistics from successful demonstrations:

$$q_g = [\mu_g, \sigma_g] \in \mathbb{R}^{50} \tag{3}$$

where, $\mu_g \in \mathbb{R}^{25}$ is the mean and $\sigma_g \in \mathbb{R}^{25}$ is the standard deviation of tactile signals, respectively. Given the latent physical representation z_{phy}, we use a MLP head to predict the safe interaction distribution, and the predicted interaction distribution is consequently modeled as:

$$[\hat{\mu}, \hat{\sigma}] = H_{safe}(z_{phy}), \quad \hat{P}_g = N(\hat{\mu}, \hat{\sigma}^2)$$

Unlike predicting a physical signal independently for every action step, our formulation explicitly models the interaction constraint at the manipulation-stage level. The network is supervised by minimizing the Kullback-Leibler (KL) divergence between the empirical demonstration distribution P_g and the predicted distribution $\hat{P}_g$:

$$L_{safe} = D_{KL}(P_g || \hat{P}_g) \tag{4}$$

Physical Interaction Prototypes. While the safe interaction distribution models continuous physical constraints, distinct $(\text{task}, \text{object}, \text{stage})$ triplets often share underlying physical interaction behaviors. To capture such reusable interaction patterns for generalization, we further construct a compact physical interaction prototype space in the dataset. Given the group-level interaction statistics $\delta = \{q_g\}_{g=1}^{G}$, we perform K-Means clustering in the physical interaction statistics space:

$$\{c_k\}_{k=1}^{K} = KMeans(\delta) \tag{5}$$

where $c_k \in \mathbb{R}^{50}$ represents a physical interaction prototype. Rather than semantic object identities, these prototypes are defined from physical interaction statistics. Therefore, different objects or tasks can be assigned to the same prototype if they exhibit similar physical interaction characteristics. For each group g, we measure the Euclidean distance $d_{g,k} = ||q_g, c_k||^2$ to each prototype center c_k. We then apply a temperature-scaled softmax transformation to obtain soft ground-truth prototype assignments:

$$y^*_{g,k} = \frac{exp(-d_{g,k}/\tau_q)}{\sum_{j=1}^{K} exp(-d_{g,j}/\tau_q)} \tag{6}$$

where $\tau_q > 0$ is a temperature hyperparameter. Given the latent physical representation z_{phy}, the prototype prediction head H_{proto} (an MLP) predicts the soft assignment distribution $\hat{y}$. We design the interaction prototype supervision by using soft-label cross entropy:

$$\hat{y} = \text{softmax}(H_{proto}(z_{phy})), \quad L_{proto} = -\sum_{k=1}^{K} y^*_{g,k} log \hat{y}_k \tag{7}$$

By unifying continuous distribution prediction with discrete prototype classification, the physical reasoning supervision is composed by safe physical interaction supervision and the interaction prototype supervision, enabling the latent physical representation z_{phy} to capture physical interaction patterns that generalize beyond individual task-object combinations.

3.4 Physical-Interaction-Conditioned Action Generation

To ensure generated actions adhere to physical interaction constraints, we condition the flow-matching action expert on both E_{VLT} and $\mathbf{z}_{phy}$. The physical latent $\mathbf{z}_{phy}$ is projected into the token embedding space via a dedicated MLP, yielding $E_{phy} \in \mathbb{R}^{(N_{VL}+1)\times D_m}$, and integrated additively:

$$E_{\text{pref}} = E_{VLT} + E_{phy}. \tag{8}$$

E_{pref} encodes visual semantics, language instructions, tactile measurements, and inferred physical constraints, and serves as the observation context for action generation. The noisy action trajectory $A_t^\tau \in \mathbb{R}^{T\times 7}$ and robot state s_t are iteratively refined at each denoising step τ:

$$A_t^{\tau+1} = A_t^\tau + \Delta\tau \cdot F_\theta(A_t^\tau, s_t, E_{\text{pref}}, \tau), \tag{9}$$

starting from $A_0 \sim \mathcal{N}(0, I)$ and producing the executed action chunk A_t. The flow-matching supervision is:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}\left[\|F_\theta(A_t^\tau) - U_\tau\|^2\right], \tag{10}$$

where U_τ is the target flow interpolated between A_0 and the ground-truth action chunk at $\tau \in [0, 1]$. The complete framework is trained end-to-end with a joint objective:

$$\mathcal{L} = \mathcal{L}_{\text{flow}} + \lambda_{\text{safe}}\mathcal{L}_{\text{safe}} + \lambda_{\text{proto}}\mathcal{L}_{\text{proto}}, \tag{11}$$

where λ_{safe} and λ_{proto} balance action generation with physical awareness.

4 BENCHMARK

We train and evaluate PIVLA on PiVLA-data to answer three questions: **Q1:** When does contact feedback improve safe manipulation, and when is it necessary? **Q2:** Does a physical interaction representation improve on equal-input force fusion? **Q3:** What are the contributions of physical awareness, direct guidance, and contact-derived tokens?

Hardware. A UR5e arm with a Robotiq Hand-E gripper (Figure 1a), bilateral 5×5 fingertip taxel arrays, third-person and wrist-mounted RGB cameras, continuous gripper width, and 6-DoF end-effector poses.

Tasks. All tasks follow a pick–transfer–place procedure but differ in how the interaction regime must be inferred: a) **Object Package.** The robot picks a specified produce item and places it in a target container. The 6 categories (potato, apple, tomato, banana, cucumber, kiwifruit) differ in mass, compliance, and damage threshold. b) **Tactile Payload Sensing.** Same procedure, but items are lidded opaque paper cartons, either empty or sand-filled. Since they are visually indistinguishable, the robot must infer the required interaction from contact.

Data Collection. Demonstrations are collected via teleoperation. For Task 2, a second experimenter randomly interleaves empty and filled cartons without notifying the teleoperator, keeping grasp height and trajectory criteria standardized. Safe force intervals are measured per condition using separate scripted trials: $f_{\min}$ prevents slip during transfer; $f_{\max}$ is the deformation threshold. Calibration trials are excluded from training and used solely for evaluation.

Dataset statistics. The dataset comprises teleoperated demos plus calibration trials establishing $[f_{\min}, f_{\max}]$ per condition. Demonstrations are segmented into contact stages (grasp, lift, translate, place) and zero-corrected from initial static windows. 15–20% of episodes are stratified into a validation split by instance and session; all normalizers, bounds, and prototype targets are fit on training data only. Interaction profiles are described by a 5D descriptor: mean grasp force (μ), log force variance ($\log\sigma$), contact area (A), center of pressure (C), and closure stiffness (κ). $K = 6$ K-means clusters on training descriptors yield soft prototype targets separating distinct physical regimes.

5 EXPERIMENT

5.1 EVALUATION PROTOCOL AND METRICS

Evaluation protocol. Each policy runs 10 trials per condition on instances not in the demonstrations, randomized and interleaved across policies and conditions. Damage raters are blind to policies.

Evaluation metrics. We leverage three primary evaluation metrics to evaluate our framework: a) *Task Success.* The percentage of test trials where a robotic model correctly and completely finishes the given task out of the total number of attempts. b) *Safe Success.* The percentage of test trials with no slip and no permanent deformation. c) *Distribution Error* is the KL divergence between predicted

Table 1: **Task 1: macro-average across 6 object categories** (per-category breakdown in Table 8). Rates are trial-level (%); v^+/v^-: fraction of trials whose hold-stage grasp exceeds/falls below the demonstrated range $\mu \pm 2\sigma$ (Eq. 12).

Model	Task ↑	Safe ↑	Damage ↓	Slip ↓	v^+ ↓	v^- ↓
No-force π_0	97.2	80.6	16.7	2.8	66.7	2.8
ForceVLA†	74.9	53.7	28.8	8.3	38.1	4.2
PIVLA	88.2	**84.9**	**3.3**	6.9	**16.9**	15.8

and training-group Gaussian targets. d) *Slip and Damage Rates.* The percentage of test trials with the contact tactile in a demonstrated safe range:

$$v_s^+(c) = \Pr\big[\bar{T}_s(c) > \mu_s(c) + 2\sigma_s(c)\big], \qquad v_s^-(c) = \Pr\big[\bar{T}_s(c) < \mu_s(c) - 2\sigma_s(c)\big], \tag{12}$$

where $\Pr[\cdot]$ is the probability operator. $v_s^+(c)$ denotes the upper-bound violation rate for the damage cases. $v_s^-(c)$ denotes lower-bound violation rate for the slip cases. T_t is formed from episode-zeroed taxel measurements as in training; $\bar{T}_s(c)$ is its mean over contact frames of stage s in condition c. $\mu_s(c)$ is the mean contact tactile. $\sigma_s(c)$ is the standard deviation of the contact tactile.

5.2 BASELINES

All learned policies are initialized from π_0 (Black et al., 2024), fine-tuned on the same demonstrations with the same action horizon and training budget, and keep the vision–language weights frozen.

No-force π_0 receives images, the instruction, and the robot state, including the measured gripper aperture, but no contact-load measurements. This baseline evaluates performance without explicit load feedback while retaining the same proprioceptive observations.

ForceVLA† adapts the FVLMoE late-fusion module of ForceVLA (Yu et al., 2026) to the same backbone and contact input as PIVLA. Matching sensory inputs controls for sensor availability when evaluating the physical interaction token, its supervision, and its action guidance.

Ablation variants. Three ablations assess the roles of physical awareness, direct action guidance, and contact information in PIVLA (Q3). *Architecture only* retains all physical modules and the guidance term $\boldsymbol{G}_{\text{phy}}$ but sets $\lambda_{\text{dist}} = \lambda_{\text{proto}} = 0$, evaluating auxiliary supervision at fixed model capacity. *No extra guidance* removes $\boldsymbol{G}_{\text{phy}}$ while retaining both auxiliary losses; their influence on the action expert through the shared FVLMoE and $\boldsymbol{G}_{\text{fvl}}$ remains intact. *Vision–language token* computes $\boldsymbol{z}_{\text{phy}}$ from a prefix token before contact fusion, evaluating the role of contact information in the physical representation.

5.3 MAIN RESULTS

Task 1: Object Package (Q1–Q2). Table 1 and Figure 3 report performance across 6 object categories (120 trials). Macro-averaged safe-success rates are 80.6%, 53.7%, and 84.9% for No-force π_0, ForceVLA†, and PIVLA. PIVLA gains $+31.2$ pp in safe success over ForceVLA†; the gap over No-force π_0 ($+4.3$ pp) is not significant at this sample size. The continuous grasp signal is more discriminative: PIVLA holds least firmly in five of six categories (by factors of 1.9–$23.6\times$ over No-force π_0), and uniquely relaxes its grasp through the carry on banana. Referred to the demonstrated range ($\mu \pm 2\sigma$ per group), No-force π_0 over-grasps on 66.7% of trials and rarely under-grasps (2.8%); ForceVLA† reduces this to 38.1%; PIVLA reaches 16.9% over and 15.8% under—a symmetric spread rather than a bias—and lands inside the demonstrated range most often (67.2% vs. 57.8% and 30.6%).

Task 2: Tactile Payload Sensing (Q1–Q2). Table 2 and Figure 3 report outcomes and grasp force. ForceVLA† fails both conditions: it crushes empty cartons (20% safe success) and is stopped by the force limiter on all filled-carton trials. π_0 and PIVLA both achieve 100% safe success on both conditions, so outcomes do not separate them. Grasp force partially does: all policies hold far above the demonstrated range (saturating $v^\pm$), so Table 2 reports standardized distance d instead. On filled cartons PIVLA is closest to demonstrated interaction ($+3.3\sigma$ vs. $+5.0\sigma$ and $+7.9\sigma$); on empty

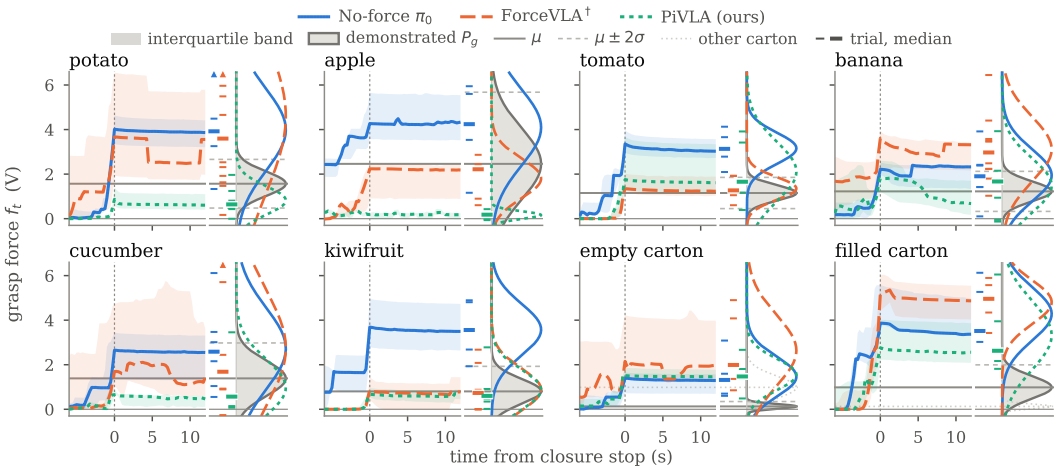

Figure 3: **Grasp force over time and hold-force distribution, per condition.** Traces: median $\pm$ interquartile range; gray fill: demonstrated P_g; left-edge ticks: per-trial hold medians; densities independently normalized.

Table 2: **Task 2: carton transfer.** Trial-level rates (%). d: hold-stage grasp of the median trial in units of σ from the demonstrated mean; used in place of $v^{\pm}$ because all policies exceed the demonstrated range ($v^{+} = 100\%$) on both conditions—note that exceeding the demonstrated range does not imply failure on these objects, as No-force π_0 demonstrates. All ForceVLA† filled-carton failures are force-limiter stops.

	Empty carton				Filled carton			
Model	Task ↑	Safe ↑	Damage ↓	d	Task ↑	Safe ↑	Damage ↓	d
No-force π_0	100.0	100.0	0.0	$+10.9\sigma$	100.0	100.0	0.0	$+5.0\sigma$
ForceVLA†	100.0	20.0	80.0	$+17.0\sigma$	0.0	0.0	0.0	$+7.9\sigma$
PiVLA	100.0	100.0	0.0	$+12.3\sigma$	100.0	100.0	0.0	$+3.3\sigma$

cartons it is not ($+12.3\sigma$ vs. $+10.9\sigma$). All three modulate between conditions, but that π_0—with no contact input—modulates at all is the more salient finding.

Grasp-force regulation and failure modes. Force and outcome metrics do not rank policies identically, and the disagreement is informative. On Task 1, No-force π_0 deviates most from demonstrated interaction (66.7% over-grasp vs. 38.1% and 16.9%) yet outperforms ForceVLA† in safe success (80.6% vs. 53.7%): excess grasp alone does not lose trials on these objects. What costs ForceVLA† trials is inconsistency (Figure 4): wide interquartile range on cucumber and potato, 80% damage on empty cartons, and force-limiter stops on every filled-carton trial. PiVLA is the only policy that is both close to and consistent with the demonstrated interaction; neither property alone accounts for its performance. Per-stage rates are given in the appendix. (Table 7).

5.4 Ablation Studies

Table 3 reports Task 2 ablation results. All variants ran until five successes (biasing rates upward), on the same cartons after the main campaign in arm-blocked order. Carton wear over the empty-carton block (-0.70 mm/trial closing aperture) means the *Vision–language token* variant, which ran last, met the most worn carton. Safe success on empty cartons uses the aperture rule (< 21.1 mm) rather than damage ratings, which were uninformative by this point. Six *Vision–language token* and five *Architecture only* filled-carton trials during a bulged-carton state are excluded.

Effect of physical awareness (Q3). *Architecture only* drops -16.7 and -20.8 pp versus the full model. All failures share one mechanism: the arm presses onto the carton without lifting it—the same failure ForceVLA† shows on every filled-carton trial. This joint ablation does not isolate either auxiliary loss.



Figure 4: **Qualitative rollouts.** PiVLA successes (left) vs. ForceVLA† failures (right) on the same objects; insets: wrist camera frame. ForceVLA† fails at both ends: over-grasping crushes the empty carton and stalls on the potato, while under-grasping drops the kiwifruit.

Table 3: **PiVLA ablations on Task 2.** Aw.: physical awareness (both losses); $\boldsymbol{G}_{\text{phy}}$: direct guidance pathway; $\boldsymbol{h}_f$: token from contact fusion (vs. vision–language prefix). Rates (%) run until 5 successes. Empty-carton Safe uses gripper aperture ($< \mu - 2\sigma$) as proxy for damage due to carton wear.

Component				Task 2, empty		Task 2, filled	
Aw.	$\boldsymbol{G}_{\text{phy}}$	$\boldsymbol{h}_f$	Variant	Task ↑	Safe ↑	Task ↑	Safe ↑
—	✓	✓	Architecture only	83.3 (5/6)	83.3	62.5 (5/8)	62.5
✓	—	✓	No extra guidance	100.0 (5/5)	100.0	100.0 (5/5)	100.0
✓	✓	—	Vision–language token	71.4 (5/7)	28.6	83.3 (5/6)	83.3
✓	✓	✓	PiVLA (full)	100.0 (5/5)	100.0	83.3 (5/6)	83.3

Effect of direct action guidance (Q3). *No extra guidance* succeeds on all ten trials ($5/5$ each condition), matching or exceeding the full model. At this sample size $\boldsymbol{G}_{\text{phy}}$ adds nothing measurable; the physical representation may still be used via the shared pathway.

Effect of contact-conditioned representations (Q3). *Vision–language token* reaches $71.4\%/83.3\%$ task success and 28.6% safe success on empty cartons, with three successes closing below the aperture threshold. Two caveats apply: it ran on the most worn carton, and both prints appear in training, so vision alone partially separates conditions before contact.

6 Conclusion

We presented PiVLA, a VLA policy that learns an intermediate physical interaction representation: a token fusing vision, language, and contact, supervised with per-group safe-force distributions and interaction prototypes, and used to directly guide flow-based action generation. PiVLA achieves 84.9% safe success on Object Package and 100% on both carton conditions, outperforming baselines in safe success and grasp consistency, and holding closest to the demonstrated interaction distribution (16.9% over-grasp vs. 38.1% and 66.7%). Ablations identify physical awareness as the critical component: removing it causes systematic pressing failures, while direct action guidance adds no measurable gain at this sample size.

Limitations: interaction targets are scalar Gaussians per group, capturing neither instance-level variation nor temporal contact structure; σ_g reflects demonstrated variation rather than physical tolerance; and all experiments use a single robot. Richer distributions and broader embodiments are left for future work.

AI USE STATEMENT

Generative AI tools assisted with manuscript drafting and editing, reference formatting, software development, and experimental feedback. All experimental measurements, citations, analyses, and results were verified manually by the authors, who take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

The model and training details are given in Section 3 and Appendix C, data collection and preprocessing in Section 4 and Appendix A, and the evaluation protocol in Section 5. Code and artifacts will be provided in case of acceptance.

REFERENCES

Jianxin Bi, Kevin Yuchen Ma, Ce Hao, Mike Shou Zheng, and Harold Soh. Vla-touch: Enhancing vision-language-action model with dual-level tactile feedback. *IEEE Robotics and Automation Letters*, 2026.

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

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. 2022.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Lukas Brunke, Melissa Greeff, Adam W Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5(1):411–444, 2022.

Roberto Calandra, Andrew Owens, Manu Upadhyaya, Wenzhen Yuan, Justin Lin, Edward H Adelson, and Sergey Levine. The feeling of success: Does touch sensing help predict grasp outcomes? 2017.

Roberto Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, Edward H Adelson, and Sergey Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3(4):3300–3307, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. volume 33, pp. 9912–9924, 2020.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. volume 44, pp. 1684–1704. Sage Publications Sage UK: London, England, 2025.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. 2023.

Danny Driess, Jost Springenberg, Brian Ichter, Lili Yu, Adrian Li-Bell, Karl Pertsch, Allen Ren, Homer Walke, Quan Vuong, Lucy Xiaoyang Shi, et al. Knowledge insulating vision-language-action models: Train fast, run fast, generalize better. *Advances in Neural Information Processing Systems*, 38:102867–102888, 2026.

Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 653–660. IEEE, 2024.

Letian Fu, Gaurav Datta, Huang Huang, William Chung-Ho Panitch, Jaimyn Drake, Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, and Ken Goldberg. A touch, vision, and language dataset for multimodal alignment. 2024.

Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 12462–12469. IEEE, 2024.

Neville Hogan. Impedance control: An approach to manipulation. pp. 304–313, 1984.

Yifan Hou, Zeyi Liu, Cheng Chi, Eric Cousineau, Naveen Kuppuswamy, Siyuan Feng, Benjamin Burchfiel, and Shuran Song. Adaptive compliance policy: Learning approximate compliance for diffusion guided control. pp. 4829–4836, 2025.

Binghao Huang, Yixuan Wang, Xinyi Yang, Yiyue Luo, and Yunzhu Li. 3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing. 2024.

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, 2025.

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

Roland S Johansson and J Randall Flanagan. Coding and use of tactile signals from the fingertips in object manipulation tasks. *Nature Reviews Neuroscience*, 10(5):345–359, 2009.

Joshua Jones, Oier Mees, Carmelo Sferrazza, Kyle Stachowicz, Pieter Abbeel, and Sergey Levine. Beyond sight: Finetuning generalist robot policies with heterogeneous sensors via language grounding. pp. 5961–5968, 2025.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset, 2024.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. 2024.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. 2025.

Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020.

Michelle A Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 International conference on robotics and automation (ICRA)*, pp. 8943–8950. IEEE, 2019.

Michelle A Lee, Brent Yi, Roberto Martín-Martín, Silvio Savarese, and Jeannette Bohg. Multimodal sensor fusion with differentiable filters. pp. 10444–10451, 2020.

Hao Li, Yizhi Zhang, Junzhe Zhu, Shaoxiong Wang, Michelle A Lee, Huazhe Xu, Edward Adelson, Li Fei-Fei, Ruohan Gao, and Jiajun Wu. See, hear, and feel: Smart sensory fusion for robotic manipulation. 2022.

Yang Li, Zhaxizhuoma Zhaxizhuoma, Hongru Jiang, Junjie Xia, Hongquan Zhang, Jinda Du, Yunsong Zhou, Jia Zeng, Ce Hao, Jieji Ren, et al. Forcevla2: Unleashing hybrid force-position control with force awareness for contact-rich manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8911–8920, 2026.

Toru Lin, Yu Zhang, Qiyang Li, Haozhi Qi, Brent Yi, Sergey Levine, and Jitendra Malik. Learning visuotactile skills with two multifingered hands. pp. 5637–5643, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. 2022.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. In *International Conference on Learning Representations*, volume 2025, pp. 29982–30009, 2025.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. 2025.

Marc H Raibert and John J Craig. Hybrid position/force control of manipulators. 1981.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. 2017.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. volume 30, 2017.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on robot learning*, pp. 1723–1736. PMLR, 2023.

Han Xue, Jieji Ren, Wendi Chen, Gu Zhang, Yuan Fang, Guoying Gu, Huazhe Xu, and Cewu Lu. Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation. 2025.

Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, et al. Binding touch to everything: Learning unified multimodal tactile representations. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26330–26343. IEEE, 2024.

Jiawen Yu, Hairuo Liu, Qiaojun Yu, Jieji Ren, Ce Hao, Haitong Ding, Guangyu Huang, Guofan Huang, Yan Song, Panpan Cai, et al. Forcevla: Enhancing vla models with a force-aware moe for contact-rich manipulation. volume 38, pp. 93409–93439, 2026.

Samson Yu, Kelvin Lin, Anxing Xiao, Jiafei Duan, and Harold Soh. Octopi: Object property reasoning with large tactile-language models. 2024.

Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017.

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. 2026.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. 2023.