When Looking Changes the Grasp: Recovery Interventions in Frozen VLA Manipulation

Anonymous Author(s)
Affiliation
Address
email

Abstract: External recovery supervisors for frozen vision-language-action (VLA) policies must verify whether a grasp succeeded before intervening, yet the robot's own gripper often occludes the object from overhead cameras. While moving the arm can restore visibility, the physical maneuver used to inspect the grasp risks altering contact and disrupting a manipulation that would otherwise succeed. To isolate this physical observer effect from control interruption, we evaluate a task-trained frozen SmolVLA policy on a simulated SO-101 arm across six manipulation scenarios (30 continuations) using a five-condition ablation that separates policy pause/reset, gripper opening, upward arm clearance, and their combination. Nominal execution, opening in place, and closed-gripper clearance each complete 5/6 tasks, and pause/reset completes 4/6, whereas combined opening and clearance collapses completion to 0/6 by leaving the released cube on the table as the arm withdraws. Because passive wrist-camera heuristics detect at most 2/6 historical failures without false alarms (0/8) in offline replays, reliable VLA recovery cannot rely on active viewing maneuvers that destroy the grasp state they are meant to assess. **Companion video:** https://vla-grasp-intervention.netlify.app/

Keywords: Manipulation recovery, active perception, frozen VLA

1 Introduction

When a robot closes its gripper around an object, an external recovery supervisor must verify whether the grasp succeeded before the policy proceeds to transport or placement. Missing a failed pickup leads to downstream task failure, whereas falsely declaring failure interrupts a trajectory that would otherwise succeed. Effective supervision therefore requires accurately diagnosing contact state while preserving nominal execution whenever a grasp is viable.

Obtaining clear visual evidence for this decision is difficult under self-occlusion. The gripper often hides the object from overhead cameras, and nearly identical finger apertures can accompany different contact configurations. Withdrawing the arm exposes the workspace to the camera, but the motion itself alters the arm's position and contact with the object; opening the gripper during withdrawal changes that interaction further. Consequently, the clearer image captures the physical aftermath of the sensing maneuver rather than the grasp state that prompted it.

This coupling between sensing actions and physical state is particularly critical when wrapping external supervisors around frozen vision-language-action (VLA) models such as RT-2 (Zitkovich et al., 2023) and SmolVLA (Shukor et al., 2025). While recent active-perception methods improve manipulation by selecting informative virtual views (Liu et al., 2026) or adjusting camera viewpoints under occlusion (Luo et al., 2026), they evaluate sensing actions primarily by visual observability. In single-arm manipulation, however, moving the arm to clear an occluded overhead view physically modifies the environment state from which the frozen policy must resume.

Submitted to the 10th Conference on Robot Learning (CoRL 2026). Do not distribute.

We investigate this physical observer effect using a task-trained frozen SmolVLA policy on a simulated SO-101 arm. Motivated by preliminary localization and recovery diagnostics, we design a five-condition ablation across six manipulation scenarios that decouples policy pause/reset from gripper opening, upward arm clearance, and their combination. Alongside an offline audit of passive wrist-camera signals during nominal execution, our study evaluates how the physical composition of a sensing maneuver impacts downstream task completion.

2 Experimental Setup and Intervention Study

Figure 1 illustrates our control and supervision pipeline for an SO-101 arm manipulating a 25 mm cube in MuJoCo (Todorov et al., 2012). In the nominal loop (top), a task-trained frozen SmolVLA policy maps overhead and wrist RGB, proprioception, and language instructions to 50-step chunks of six absolute joint-position targets. Downstream, an external alignment controller adds smoothed arm-target offsets based on joint feedback and a cached cube-position reference, executing one joint-limit-clipped command per step. When triggered, a separate recovery supervisor (bottom) monitoring only overhead RGB and proprioception overrides execution. It couples gripper opening with upward arm withdrawal to clear the overhead view before relocalizing the cube and deciding whether to retry. Because this physical maneuver (where opening's standalone visibility gain is unmeasured) precedes re-observation, the new camera frame captures an already altered manipulation state.

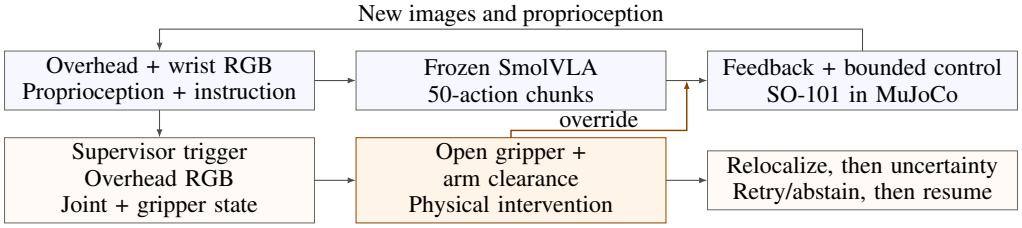


Figure 1: Policy and historical supervisor. Both cameras feed SmolVLA; the supervisor excludes wrist RGB. Its physical intervention precedes the new observation and uncertainty decision.

Our evaluation centers on six deterministic manipulation scenarios: three testing cube pickup and three adding target placement. Pickup success requires grasp acquisition and a *stable lift* (peak cube height increase ≥ 30 mm, without requiring a sustained hold), while placement additionally requires a final target distance ≤ 50 mm and cube-center height < 27.5 mm. These criteria evaluate end-of-episode task completion rather than contact retention during the intervention. Using these scenarios, we conduct a 30-run intervention ablation (6 scenarios $\times$ 5 conditions) alongside a 14-trajectory offline wrist-camera audit.

In the 30-run ablation, all five conditions for a given scenario initialize the alignment reference from the declared cube spawn and share matched seeds and nominal histories up to a closure-related trigger (with one recorded prestate differing slightly). At the trigger, each scenario branches into five continuations: *nominal* continues uninterrupted, while four conditions override control for 25 steps under a common pause/reset protocol that crosses gripper opening with arm withdrawal. Specifically, *reset only* holds the arm in place with closure commanded (bundling waiting, holding, clamping, and resets); *opening only* holds the arm with the gripper open; *clearance only* withdraws the arm with closure commanded; and *opening plus clearance* combines both. After 25 steps, the four interrupted branches discard pending policy actions, separately reset controller offsets and gripper open/close state, and update the external alignment reference to the simulator cube XY sampled at the trigger. These coordinates guide external controller offsets rather than direct SmolVLA inputs, without using post-clearance camera estimates, an uncertainty gate, or a scripted corrective reach.

In the complementary offline audit, we examine 14 recorded nominal trajectories (eight historical successes and six failures, sharing five scenarios with the ablation) to test whether passive observations could flag failures without moving the arm. We compare gripper-aperture and wrist-camera

red-area/centroid rules on images sampled 15 steps after closure detection (with a final-frame fallback). In the five shared scenarios, this sampling window falls after the ablation's intervention trigger, so the audit evaluates post-closure separability rather than pre-intervention online decisions.

3 Results

3.1 Why better localization did not resolve recovery

Preliminary post-contact diagnostics tested whether a second grasp was physically feasible. In one diagnostic state, feeding exact simulator cube coordinates to the external controller enabled a corrective retry that achieved a 119.1 mm lift, whereas relying on the existing overhead-camera estimate caused the supervisor to reject the estimate and decline the retry, reaching only 5.6 mm. Although this single-state test did not demonstrate autonomous recovery, it motivated clearing the arm to improve overhead visibility. In two development trials, upward arm clearance reduced localization error from 41.69 to 5.24 mm and from 52.74 to 4.39 mm. Despite this sharper view, earlier clearance-based procedures completed 0/6 tasks compared to 5/6 under nominal execution, with the uncertainty-gated variant declining every retry after opening and withdrawal. This mismatch raised the question of whether policy interruption, gripper opening, or arm motion caused the collapse.

3.2 Separating the components of the intervention

Figure 2 compares the five ablation conditions across our six scenarios to decouple the control pause from gripper opening and arm withdrawal. Nominal execution completes 5/6 tasks, reset only completes 4/6, opening only completes 5/6, clearance only completes 5/6, and combined opening and clearance completes 0/6. Pausing policy prediction alone does not explain the collapse, as reset only retains four successes (losing one placement trial). Likewise, neither physical action fails on its own, though identical 5/6 totals mask different scenario outcomes: clearance only repairs one placement failure but loses one nominal pickup success.

The intermediate trajectories in Figure 2a explain why only the combined maneuver fails across all five nominal passes. With closure commanded, upward clearance carries the grasped cube along with the gripper. Opening alone leaves the cube supported on the table during the intervention prefix (with post-prefix grasp events unlogged), but keeps the open fingers directly around it so the resumed policy still completes 5/6 tasks. By contrast, combining opening with withdrawal releases the cube and pulls the arm away; end-of-intervention telemetry confirms table support and zero gripper contact in all six combined runs, and resumed execution completes no tasks.

Limited passive evidence. Following the 14-replay offline setup in Section 2, we evaluated whether wrist-camera and aperture rules could detect failures without moving the arm. Against historical labels, wrist-camera rules flag 1/6 failures and 0/8 successes, while a combined wrist-and-aperture rule flags 2/6 failures and 0/8 successes (all in development examples). An aperture-only rule triggering outside $[0.242, 0.245]$ rad flags all 6/6 failures and all 8/8 successes, whereas overhead rules flag neither. Because these rules miss most failures and sample after the intervention trigger, whether passive observations can support timely online recovery decisions remains an open question.

4 Discussion and Limitations

Our ablation demonstrates that the performance drop during recovery is not an unavoidable byproduct of moving the arm for visibility, but rather depends on the intervention's composition and the state from which the frozen policy resumes. While closed-gripper clearance preserves task completion in most scenarios, combining gripper opening with arm withdrawal destroys the grasp state. This insight complements active-perception methods that optimize virtual views (Liu et al., 2026) or camera placement under occlusion (Luo et al., 2026), highlighting that observation quality must

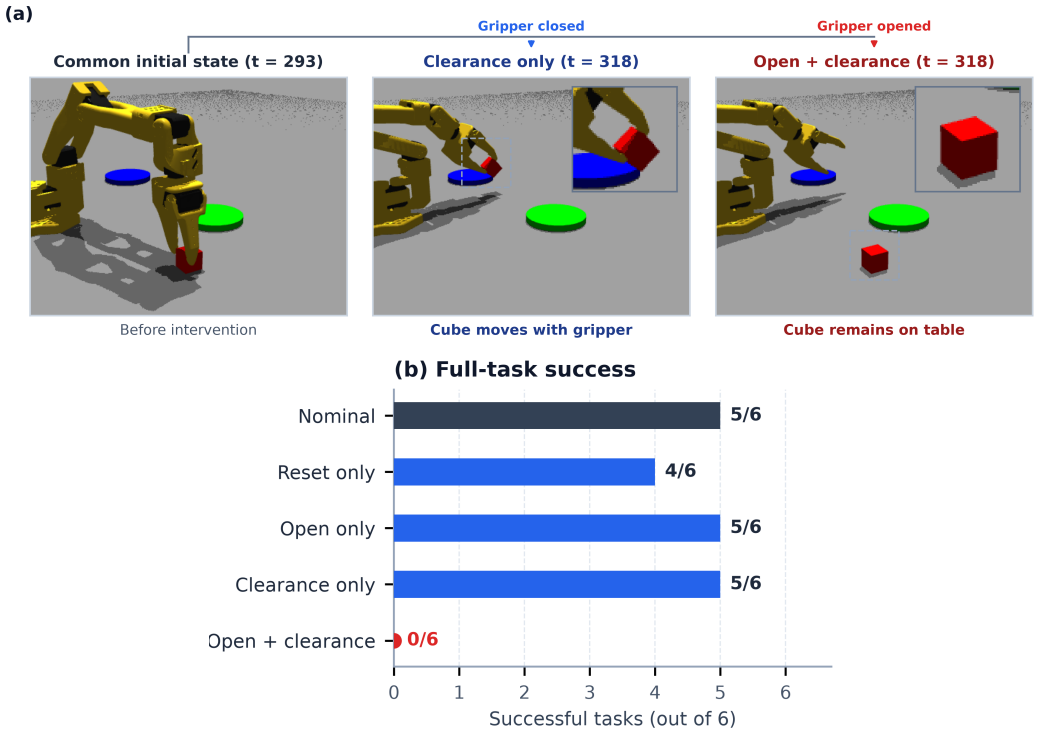


Figure 2: Physical intervention and task completion. (a) From a common illustrated state, closed-gripper clearance carries the cube; opening during withdrawal leaves it on the table. Frames are intermediate. (b) Full-task outcomes on six reused scenarios.

be weighed against its physical effect on manipulation. Furthermore, while contact-aware uncertainty calibration addresses contact-dependent prediction errors (Marques et al., 2026), a decision triggered after a disruptive physical action cannot prevent that action's consequences.

These findings are subject to several scope limitations. The study evaluates a single simulated SO-101 configuration, one frozen task-trained policy, and six deterministic scenarios without measuring generalization or hardware reliability. In one scenario, a clearance-only prestate differed slightly by 0.0022 rad in gripper angle and 0.1 mm in cube height; excluding this trial preserves the qualitative contrast ($4/5$ vs. $0/5$ successes). Additionally, privileged reset coordinates preclude interpreting the setup as fully autonomous recovery, and inherited labels with differing spawn settings mean our offline rule counts reflect exploratory label agreement rather than a verified online safety guarantee.

Looking forward, designing reliable recovery requires determining whether onboard observations can identify failed grasps prior to intervention. Prospective monitors must evaluate failure detection, false intervention rates, and intervention costs together, comparing contact-preserving observation strategies against release and withdrawal. Because simulated friction, compliance, and control delays may alter whether a physical SO-101 retains the cube during withdrawal, validating these relative outcomes on physical hardware remains an essential next step.

5 Conclusion

Physical interventions intended to clarify grasp state can themselves disrupt the manipulation task. In our simulated SO-101 setup, combining gripper opening with arm withdrawal is far more damaging than either action or control interruption alone, showing that active perception must be evaluated by its physical impact, not just observability gains. Reliable sim-to-real recovery thus calls for contact-preserving monitoring and robust hardware validation.

References

B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, Q. Vuong, V. Vanhoucke, H. Tran, R. Soricut, A. Singh, J. Singh, P. Sermanet, P. R. Sanketi, G. Salazar, M. S. Ryoo, K. Reymann, K. Rao, K. Pertsch, I. Mordatch, H. Michalewski, Y. Lu, S. Levine, L. Lee, T.-W. E. Lee, I. Leal, Y. Kuang, D. Kalashnikov, R. Julian, N. J. Joshi, A. Irpan, B. Ichter, J. Hsu, A. Herzog, K. Hausman, K. Gopalakrishnan, C. Fu, P. Florence, C. Finn, K. A. Dubey, D. Driess, T. Ding, K. M. Choromanski, X. Chen, Y. Chebotar, J. Carbajal, N. Brown, A. Brohan, M. G. Arenas, and K. Han. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR, 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

M. Shukor, D. Aubakirova, F. Capuano, P. Kooijmans, S. Palma, A. Zouitine, M. Aractingi, C. Pascal, M. Russi, A. Marafioti, S. Alibert, M. Cord, T. Wolf, and R. Cadene. SmolVLA: A vision-language-action model for affordable and efficient robotics, 2025. URL https://arxiv.org/abs/2506.01844. arXiv preprint.

Z. Liu, Y. Gu, Y. Wang, X. Xue, and Y. Fu. ActiveVLA: Injecting active perception into vision-language-action models for precise 3D robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8141–8151, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Liu_ActiveVLA_Injecting_Active_Perception_into_Vision-Language-Action_Models_for_Precise_3D_CVPR_2026_paper.html.

K. Luo, Y. Huang, Q. Zhong, X. Song, Y. Liu, and L. Lin. Recovering the view: Benchmarking physical active vision for occlusion recovery in robotic manipulation, 2026. URL https://arxiv.org/abs/2609.37292. arXiv preprint.

E. Todorov, T. Erez, and Y. Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012. doi:10.1109/IROS.2012.6386109. URL https://doi.org/10.1109/IROS.2012.6386109.

L. Marques, K. Popov, and D. Berenson. Particle-based conformal prediction for contact-aware uncertainty calibration in stratified configuration spaces. In *Proceedings of the Fifteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 329 of *Proceedings of Machine Learning Research*, pages 813–843. PMLR, 2026. URL https://proceedings.mlr.press/v329/marques26a.html.