SafeLab: An Interactive High-Fidelity Benchmark for Embodied Safety in Scientific Robotics

Anonymous Authors[1]

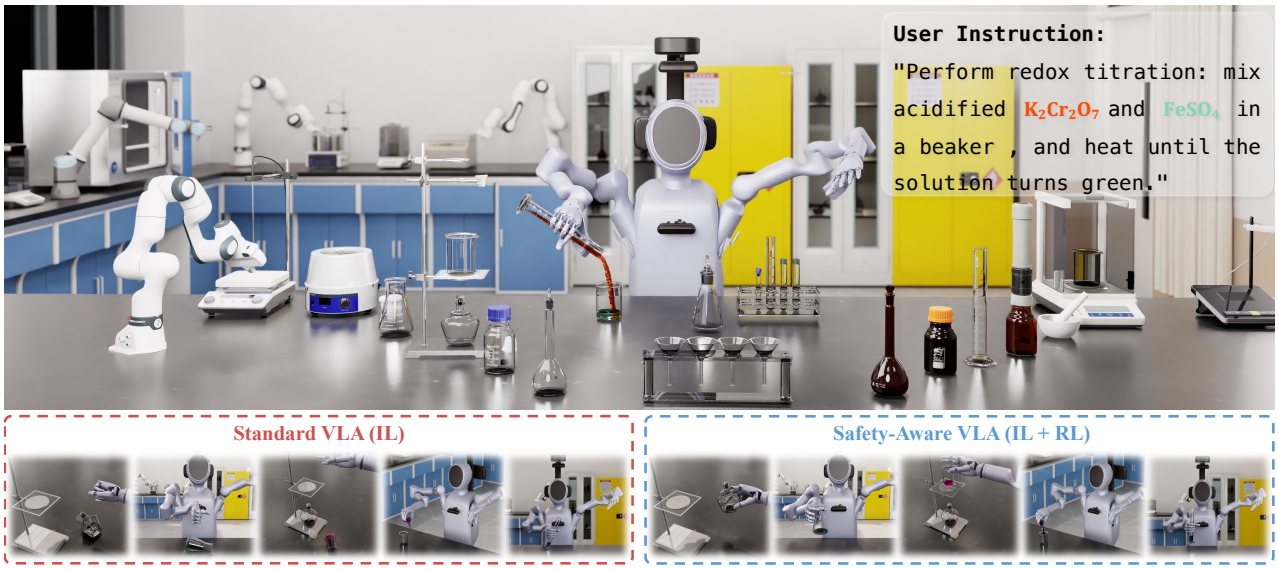


Figure 1. **Zero-Tolerance Manipulation: Benchmarking Irreversible Failures in Scientific Robotics.**

Abstract

Laboratory automation driven by scientific embodied agents represents a critical frontier in modern laboratories. Unlike conventional robotic domains, laboratory environments impose zero-tolerance constraints on manipulation precision and collision, as minor deviations can lead to irreversible chemical hazards or equipment damage. This naturally makes the automated laboratory an ideal testbed for advancing embodied safety. However, existing benchmarks predominantly feature high-tolerance manipulation tasks where intermediate failures are largely reversible. More critically, current Vision-Language-Action (VLA) models trained via static imitation learning cannot satisfy these strict constraints. Because they merely mimic successful demonstrations, they lack the ability to recover from execution drift, leading to catastrophic compounding errors in precision-critical domains. Overcoming this limitation requires transitioning from static datasets to interactive environments that support Reinforcement Learning (RL) for dynamic error recovery. To this end, we introduce SafeLab, a generative simulation benchmark designed for the full lifecycle of safe robot learning. Grounded in a high-fidelity chemistry lab, our framework integrates an **LLM engine** for procedural task synthesis, an **automated expert** for scalable demonstration collection, and an **interactive environment** for continuous RL refinement. Leveraging this infrastructure, we release a dataset of 6,000+ **complex trajectories** to evaluate state-of-the-art VLA models. Experiments reveal that current embodied agents fail significantly under these safety constraints. In contrast, our RL post-training pipeline enables agents to learn active error correction, mitigating hazardous failures and improving success rates by 37%, thereby establishing SafeLab as a critical platform for developing reliable and safe generalist agents.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

While scientific discovery drives technological progress, traditional experimental pipelines remain fundamentally time-consuming and labor-intensive. To accelerate this process, researchers are increasingly deploying scientific embodied agents to automate complex laboratory workflows. However, extending generalist robots into scientific domains introduces strict manipulation challenges distinct from conventional tasks. Laboratory environments frequently involve hazardous chemical reagents and fragile instrumentation, requiring high precision in interaction. In such settings, minor execution errors can trigger irreversible accidents, such as chemical spillage or equipment destruction. Consequently, realizing the full potential of scientific automation strictly hinges on solving the safety-reliability gap. This zero-tolerance nature transforms the automated laboratory from a mere application domain to the ultimate testbed for rigorously evaluating embodied safety (see Figure 1).

Despite this critical need for safety evaluation, existing robotic benchmarks fail to capture these strict constraints. Foundational simulation suites feature high-error-tolerance, rigid-body manipulation tasks where intermediate failures are easily reversible (Yu et al., 2020; Liu et al., 2023). Evaluated on these forgiving environments, current Vision-Language-Action (VLA) models exhibit artificially high performance (Zitkovich et al., 2023; Kim et al., 2024; Black et al., 2025). However, these models are trained primarily via static imitation learning. By passively cloning trajectories, they often succumb to causal confusion, relying on spurious visual correlations rather than understanding physical dynamics (de Haan et al., 2019). Furthermore, when deployed in precision-critical domains, they are inherently susceptible to covariate shift. Lacking interactive feedback, standard policies accumulate compounding errors (Ross & Bagnell, 2010), leading to unrecoverable physical disasters, such as knocking over glassware or spilling reagents.

Addressing this limitation requires shifting from static evaluation to interactive environments that support Reinforcement Learning (RL) for dynamic error recovery. To this end, we introduce SafeLab, a generative benchmark designed for the full lifecycle of safe robot learning (see Figure 2). Grounded in a high-fidelity chemistry lab, our framework integrates three core modules. First, an LLM-driven generative engine synthesizes physically valid environments with diverse task logic from natural language. Second, to address the lack of safe training data, we construct an automated expert system that generates scalable, teleoperation-free demonstrations serving as safe priors for policy initialization. Crucially, we provide a safety-aware RL interface. By incorporating penalties for irreversible failure modes including fluid spillage and glass breakage, this environment provides the dense feedback signals necessary for agents to evolve from blind imitation to active hazard avoidance.

Our contributions are threefold: (1) We introduce SafeLab, a high-fidelity generative benchmark that unites procedural task synthesis with rigorous physics simulation. Crucially, we validate the simulation's physical fidelity via real-world trajectory replay, establishing a verified and scalable testbed for embodied safety in scientific robotics. (2) Addressing the critical data scarcity in hazardous environments, we develop an automated expert system that synthesizes large-scale, collision-free demonstrations. These serve as risk-free safe priors for policy initialization, bypassing the dangers of real-world data collection. (3) Through systematic evaluation, we expose the vulnerability of current VLA models to compounding errors in zero-tolerance settings. We demonstrate that our safety-aware RL pipeline effectively mitigates these hazardous failures, achieving a 37% improvement in success rates by enabling active error correction.

2. Related Work

Automated Laboratories. Laboratory automation has evolved from stationary, script-based systems such as Opentrons (Opentrons Labworks Inc., 2024) and Chemspeed (Chemspeed Technologies AG, 2024) to more flexible robotic platforms. To overcome the spatial limitations of fixed hardware, prior work has introduced mobile manipulators that navigate shared laboratory environments to connect instruments (Burger et al., 2020; Dai et al., 2024). More recently, systems including A-Lab (Szymanski et al., 2023), Organa (Darvish et al., 2025), and CREStt (Zhang et al., 2025b) integrate Large Language Models (LLMs) to enable high-level reasoning for autonomous synthesis and materials discovery.

Despite these advances, most systems rely on classical Task and Motion Planning or low-level API execution under deterministic assumptions, limiting their ability to handle dynamic physical uncertainties such as fluid sloshing or glassware instability. Scaling laboratory automation to hazardous, zero-tolerance settings therefore remains risky without a safety-aware learning pipeline (Cooper et al., 2025). SafeLab addresses this gap by providing a high-fidelity benchmark for training embodied agents under safety constraints.

Simulation Benchmarks for Embodied Agents. Simulation benchmarks are essential for scaling robot learning. Early suites such as Meta-World (Yu et al., 2020) and RLBench (James et al., 2020) standardized rigid-body manipulation tasks, while newer platforms including ManiSkill3 (Tao et al., 2025) and RoboCasa (Nasiriany et al., 2024) improved visual fidelity and throughput in household settings. To further expand task diversity, GenSim (Wang et al., 2024a) and RoboGen (Wang et al., 2024b) employ LLMs to synthesize environments via code generation, though this often introduces physical inconsistencies; RoboTwin (Chen et al., 2025) mitigates this through

Table 1. **Systematic Comparison of Simulation Benchmarks.** While existing frameworks typically prioritize either broad semantic scalability or specific physical fidelity, SafeLab uniquely integrates high-fidelity fluid dynamics, physically grounded task verification, and massively parallel infrastructure to enable interactive, safety-aware policy learning. (✓: Supported; ✗: Not Supported; ●: Partial/Limited)

Features	LIBERO (Liu et al., 2023)	RoboCasa (Nasiriany et al., 2024)	GenSim (Wang et al., 2024a)	LeHome (Li et al., 2025b)	AutoBio (Lan et al., 2025)	LabUtopia (Li et al., 2025a)	SafeLab (Ours)
Physics & Simulation Fidelity							
High-Fidelity Fluid Dynamics	✗	✗	✗	✗	✓	✓	✓
Transparent Object Assets	✗	✗	✗	✗	✓	✓	✓
Irreversible Failure Modes	✗	✗	✗	✓	✗	✓	✓
Generative Task Logic							
Open-Ended Task Generation	✗	✓	✓	✗	✗	✗	✓
Sim-in-the-Loop Verification	●	✓	✓	✗	●	●	✓
Physically Grounded Logic	✓	✓	✗	✗	✓	✗	✓
Learning Infrastructure							
RL Training Interface	✓	✓	✗	✗	✗	✓	✓
Massively Parallel	✗	✗	✗	✗	✗	✓	✓
Safety Constraints	✗	✗	✗	✗	✗	✗	✓

simulation-based verification.

Scientific benchmarks impose stricter physical constraints. AutoBio (Lan et al., 2025) and Chemistry3D (Li et al., 2024) emphasize perception challenges involving transparent fluids and glassware, while LabUtopia (Li et al., 2025a) introduces fine-grained chemo-physical dynamics. However, these environments rely on fixed task sets and do not support interactive learning. In contrast, SafeLab integrates verified generative task synthesis with high-fidelity fluid simulation, enabling closed-loop reinforcement learning for irreversible safety failures.

Scalable Learning for Safety and Recovery. Generalist VLA models such as RT-2 (Zitkovich et al., 2023) and Octo (Octo Model Team et al., 2024) exhibit strong semantic generalization but are predominantly trained via Behavior Cloning (BC), which suffers from compounding errors under distribution shift (Ross et al., 2011). Interactive approaches such as DAgger (Ross et al., 2011) alleviate this issue but require expensive human supervision. SafeVLA (Zhang et al., 2025a) introduces constrained learning for safety alignment, demonstrating improved robustness in household navigation.

Nevertheless, existing benchmarks primarily emphasize rigid-body collision avoidance and lack the physical fidelity required to model irreversible fluid dynamics. SafeLab fills this gap by providing a safety-aware interaction loop with dense feedback on liquid manipulation, enabling reinforcement learning for active error recovery in zero-tolerance scientific environments.

3. Method

We introduce SafeLab, a generative simulation framework engineered to benchmark embodied safety in scientific environments. Unlike generalist platforms that prioritize visual diversity, SafeLab prioritizes the simulation of irreversible consequences, particularly the intricate dynamics of hazardous fluids and fragile interactions. Formally, we define the framework as a tuple $\mathcal{F} = \langle \mathcal{W}, \mathcal{M}, \mathcal{E}, \mathcal{L} \rangle$, comprising the high-fidelity scientific world $\mathcal{W}$, the verified generative engine $\mathcal{M}$, the automated expert $\mathcal{E}$, and the safety-aware learning interface $\mathcal{L}$. Built upon Isaac Lab (Mittal et al., 2025), the system leverages GPU-accelerated parallel simulation to scale the training of robust agents capable of active error recovery.

3.1. High-Fidelity Scientific World

To capture the zero-tolerance nature of laboratory manipulation, $\mathcal{W}$ minimizes the domain gap between simulation and reality through rigorous visual-physical fidelity.

Physically Calibrated Instrumentation. We construct a library of high-fidelity laboratory assets to address the limitations of generic household objects found in prior datasets. These assets are defined by precise physical properties, including mass distribution, friction coefficients, and restitution. Such calibration ensures the simulation replicates the actual handling dynamics of real-world glassware. On the visual front, we employ ray-tracing to model transparency, refraction, and specular reflections. Accurate simulation of these optical properties is essential, as they constitute the primary perception challenges for vision-based agents operating on transparent containers.

Fluid Dynamics and Irreversible Failure. Complementing these rigid assets, our framework simulates irreversible state changes through fluid dynamics. We utilize Position-Based Dynamics (PBD) to model liquids as continuous particle systems, rendered via real-time Metaball isosurface extraction. To replicate specific chemical reagents, we calibrate key rheological parameters such as viscosity, surface tension, and adhesion. Crucially, the simulation supports two-way

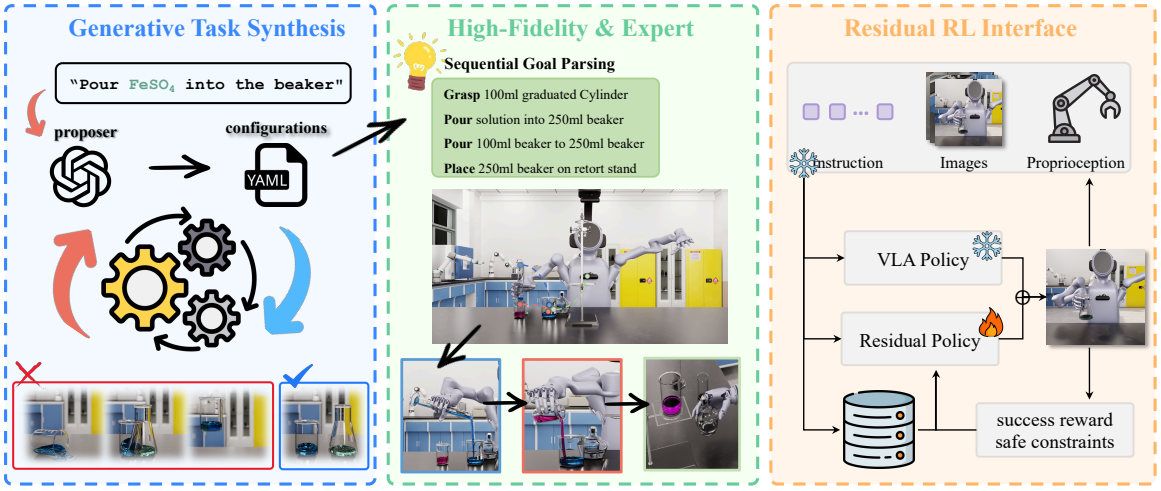


Figure 2. **Overview of the SafeLab Pipeline. Left:** An LLM-driven generative engine proposes task configurations that undergo hierarchical verification to strictly enforce physical validity. **Middle:** An automated expert collects large-scale demonstrations in a high-fidelity fluid simulation. **Right:** A safety-aware residual RL interface learns corrective actions atop a frozen VLA model, ensuring zero-tolerance manipulation.

coupling between fluids and rigid bodies. This enables the modeling of complex interactions where fluid momentum can destabilize a lightweight container. Capturing these dynamics exposes agents to "spillage" failure modes that are physically irreversible, thereby necessitating strict safety adherence during policy learning.

3.2. Generative Task Synthesis Engine

To scale task diversity without compromising physical validity, the generative engine $\mathcal{M}$ employs a "Propose-then-Verify" paradigm. In this workflow, an LLM expert acts as the proposer, translating abstract scientific instructions into structured YAML configurations, while a hierarchical physical system serves as the verifier to ensure execution feasibility.

We formally define a synthesized task instance encoded in these configurations as a tuple $\mathcal{T} = \langle \mathcal{S}, \mathcal{G}, \mathcal{C} \rangle$. Here, $\mathcal{S}$ denotes the scene configuration, including asset selection and initial states. $\mathcal{G}$ represents the sequence of semantic goals required for task completion. Crucially, $\mathcal{C}$ defines the set of safety constraints that must be maintained throughout the execution horizon. Unlike standard logic goals that only verify the final state, $\mathcal{C}$ imposes continuous boundary conditions on system dynamics, such as limits on contact forces and orientation deviations to prevent hazardous spillage or instrument damage.

Hierarchical Verification and Correction. Direct generation from LLMs often yields hallucinations, referring to configurations that are visually plausible but physically unrealizable. To enforce robustness, we implement a closed-loop verification cascade on the proposed YAML files. The process initiates with *syntactic parsing* to ensure the output strictly conforms to the simulator's schema. Subsequently, the system performs *geometric grounding* by decoupling high-level logic from low-level placement; a rejection-sampling solver generates collision-free poses that respect the robot's reachable workspace. To ensure logical soundness, a *causal logic check* employs a symbolic state machine to detect paradoxes, such as attempting to pour from a sealed container, before instantiation. Finally, *dynamic feasibility* conducts a "Sim-in-the-Loop" rollout, stepping through the physics engine to identify runtime instabilities like fluid explosion. If verification fails at any stage, the system serializes the error log into a natural language prompt, enabling the LLM to iteratively refine the configuration until all physical and safety constraints are satisfied.

3.3. Scalable Expert Data Collector

To provide safe behavioral priors, the automated expert $\mathcal{E}$ synthesizes high-quality demonstrations, bypassing the bottleneck of teleoperation. This pipeline translates abstract task goals into smooth, safety-compliant trajectories through a structured two-stage process.

The generation initiates by parsing the structured YAML configurations produced by the generative engine. By cross-referencing the defined semantic sub-goals with intrinsic grasp affordances from the asset registry, the system translates these high-level objectives into precise 6-DoF end-effector waypoints. Subsequently, we utilize cuRobo (Sundaralingam et al., 2023) to connect these waypoints via dynamically feasible paths. Operating directly within the simulation state, this CUDA-accelerated planner performs parallelized trajectory optimization. Unlike sampling-based methods that often yield jerky motion, this optimization ex-

plicitly minimizes jerk, a critical factor for preventing fluid instability during transport. Consequently, the pipeline produces a curated dataset of over 6,000 expert demonstrations, serving as a robust safety reference for policy learning.

3.4. Safety-Aware RL Interface

To bridge the gap between static demonstrations and robust deployment, the learning interface $\mathcal{L}$ is designed to facilitate a decoupled residual learning paradigm. By treating safety refinement as an additive correction to a frozen base policy, this formulation explicitly separates generalist manipulation skills from domain-specific safety constraints, preventing catastrophic forgetting while enabling sample-efficient alignment. Formally, we model the task as a Partially Observable Markov Decision Process (POMDP), defined by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$.

Action and Observation Spaces. We design a flexible interface that strictly aligns perception with action. The action space $\mathcal{A}$ supports four distinct control modes enabling diverse manipulation strategies: *joint position*, *joint delta*, *end-effector pose*, and *end-effector delta*. This flexibility allows researchers to evaluate policies across different levels of abstraction, from low-level motor control to high-level Cartesian planning. Correspondingly, to support generalist VLA policies, the observation space $\mathcal{O}$ is explicitly designed to be multi-modal. It includes: (1) *Visual Stream*: RGB-D images from egocentric and third-person views; (2) *Proprioception*: Joint states and end-effector poses; and (3) *Language Instruction*: The natural language task description l_{text} generated by $\mathcal{M}$. Crucially, to ensure state-action consistency, the proprioceptive stream is *isomorphic* to the selected control mode, automatically exposing the corresponding joint configurations or Cartesian coordinates. Finally, to facilitate rapid algorithmic verification, the interface optionally exposes privileged information, such as object velocities and contact forces. By significantly lowering the exploration barrier, this feature enables researchers to isolate algorithmic implementation errors from the intrinsic difficulty of long-horizon exploration.

Stage-wise Progressive Reward. To enable robust policy learning across long-horizon scientific procedures, we employ a multi-stage sparse reward structure reinforced by dense hybrid safety constraints. The reward function r_t at time step t is formulated as:

$$r_t = \underbrace{\mathbb{I}_{k,t} \cdot 2^{k-1} \cdot R_{\text{base}}}_{\text{Stage Bonus}} - \underbrace{(\lambda_g \mathcal{C}_{\text{gen}} + \lambda_s \mathcal{C}_{\text{task}})}_{\text{Safety Penalty}}, \qquad (1)$$

where $\mathbb{I}_{k,t}$ indicates the completion of the k-th logical subgoal (triggered by satisfying relaxed task-completion conditions). The term 2^{k-1} creates an exponential curriculum to incentivize progression. Crucially, while the task reward is sparse, the penalty term provides immediate, dense feedback at each timestep to enforce continuous safety adherence. $\mathcal{C}_{\text{gen}}$ applies generic regularization (e.g., minimizing motion jerk), while $\mathcal{C}_{\text{task}}$ strictly enforces domain-specific constraints. Specifically, we penalize the spatial error (ΔP) to encourage precise alignment, along with the orientation deviation (θ_{dev}) to prevent liquid sloshing and excessive contact force (F_{peak}) to avoid instrument damage.

Progress-Aware Termination. To balance exploration with execution efficiency in risk-sensitive environments, we implement an *adaptive time budgeting* mechanism. Rather than enforcing static horizons, the environment dynamically adjusts the maximum episode length T_{curr} conditioned on the empirical sub-goal completion rate. Each task phase operates within a duration window $[T_{min}, T_{max}]$. Initial training epochs maintain T_{max} to ensure sufficient exploration of sparse reward states. As policy competence improves, the horizon progressively contracts toward T_{min}. This adaptive truncation compels the agent to eliminate redundant actions, thereby minimizing the temporal window for potential cumulative errors and external disturbances.

4. Experiments

Our evaluation is designed to probe the safety boundaries of existing robotic agents and validate the efficacy of our generative safety-aware pipeline. We structure the experiments around four primary research questions: (Q1) Safety Gap: How do state-of-the-art generalist models perform in precision-critical scientific domains, particularly regarding irreversible fluid dynamics? (Q2) RL Efficacy: To what extent does our safety-aware RL fine-tuning improve policy robustness in out-of-distribution states where imitation learning typically fails? (Q3) Generative Generalization: How does training on diverse, logically synthesized variations contribute to zero-shot generalization across visual, spatial, and dynamic shifts? (Q4) Long-Horizon Consistency: Can current models sustain logical consistency in multi-stage sequential protocols, and how does adaptive time budgeting facilitate this efficiency?

4.1. Experimental Setup

Task and Dataset. We construct SafeLab, a comprehensive benchmark suite comprising 63 high-fidelity laboratory assets that replicate the visual and physical properties of real-world equipment ((see Figure 3)). Using our verified generative engine $\mathcal{M}$, we synthesize a hierarchy of tasks systematically categorized into three operational domains: (1) *Liquid Handling*: Focuses on fluid precision, including pouring and bimanual transport. The primary challenge is mitigating sloshing (θ_{dev}) to prevent hazardous spillage. (2) *Instrument Actuation*: Involves constrained mechanisms such as centrifuges and cabinet operation. This domain tests contact stability (F_{peak}) to ensure hardware integrity. (3) *Glassware Rearrangement*: Covers precise pick-and-

Figure 3. **Overview of the SafeLab Task Suite and Generated Dataset.** The benchmark utilizes 63 high-fidelity assets to construct 64 atomic tasks across 9 basic laboratory manipulation. To systematically test policy robustness, the generative engine $\mathcal{G}$ automatically synthesizes visual (e.g., lighting, colors) and physical (e.g., fluid viscosity, friction) variations for identical tasks. The resulting dataset comprises 100 successful expert trajectories per task, providing a comprehensive and diverse testbed for sim-to-real evaluation.

Table 2. **Main Results on Laboratory Manipulation Tasks.** We benchmark five state-of-the-art methods on the RoBoChem task suite. Performance is evaluated using standard and safe success rates (SR/SSR %), alongside domain-specific Safety Metrics (SM): orientation deviation (θ_{dev}), peak contact force (F_{peak}), and spatial residual (ΔP). **SSR** strictly counts trials that achieve goal completion without violating safety constraints.

Domain	**Task Behavior**	**Metric***	**DP** (Chi et al., 2023)		**DP3** (Ze et al., 2024)		**ACT** (Zhao et al., 2023)		**OpenVLA** (Kim et al., 2024)		$\pi_{0.5}$ (Black et al., 2025)	
			SR / SSR ↑	SM ↓	SR / SSR ↑	SM ↓	SR / SSR ↑	SM ↓	SR / SSR ↑	SM ↓	SR / SSR ↑	SM ↓
Liquid	Pour Liquid	θ_{dev} (rad)	72.4 / 35.8	0.48	78.5 / 42.1	0.42	61.2 / 28.4	0.55	58.9 / 22.5	0.58	91.2 / 52.4	0.32
	Lift Vessel		68.1 / 31.5	0.45	82.3 / 46.8	0.38	55.4 / 21.3	0.52	53.2 / 18.7	0.55	88.7 / 54.1	0.30
Actuation	Press Switch	F_{peak} (N)	75.6 / 38.2	28.4†	84.1 / 45.3	22.1†	64.8 / 25.1	36.5†	62.5 / 20.4	39.8†	93.4 / 54.8	14.2
	Open Cabinet		70.2 / 33.4	31.5†	81.5 / 43.1	24.6†	60.1 / 22.8	35.2†	59.2 / 19.5	38.2†	92.1 / 53.2	12.8
	Close Cabinet		73.8 / 36.5	27.9†	83.9 / 44.2	21.8†	63.5 / 24.3	34.8†	61.4 / 21.2	37.5†	94.5 / 55.0	13.5
Spatial	Grasp Vessel	ΔP (mm)	78.2 / 40.5	680	86.4 / 48.9	610	66.3 / 26.7	820	64.1 / 23.5	850	95.0 / 54.5	520
	Pick & Place		65.4 / 28.1	740	79.1 / 41.5	650	58.4 / 19.8	880	55.7 / 15.2	890	89.2 / 51.3	550
	Handover		58.7 / 20.2	810	72.5 / 35.6	720	52.1 / 15.5	895	50.4 / 12.1	910†	85.3 / 48.6	580
	Stack Vessel		61.3 / 22.5	790	75.2 / 38.4	690	54.8 / 18.2	870	51.2 / 14.8	895†	86.8 / 49.5	565

† denotes safety threshold violation, causing virtual hardware damage.

place and vertical stacking, requiring sub-centimeter spatial alignment (ΔP). In total, the suite spans 9 manipulation categories and 64 distinct atomic tasks. For each task, the automated expert $\mathcal{E}$ generates 100 successful demonstrations with randomized scene parameters, yielding a curated dataset of 6,400 trajectories for large-scale training.

Baselines. We benchmark five representative methods in robot learning. For foundation models, we evaluate *OpenVLA* (Kim et al., 2024) and $\pi_{0.5}$ *(Black et al., 2025)*, representing the state-of-the-art in VLA architectures. For continuous control policies, we evaluate the transformer-based *ACT* (Zhao et al., 2023), alongside diffusion-based policies including *Diffusion Policy (DP)* (Chi et al., 2023) and *DP3* (Ze et al., 2024). To evaluate RL post-training efficacy, all baselines are first trained via behavioral cloning. Subsequently, we adopt a *residual learning* paradigm for fine-tuning, where the pre-trained base policy remains frozen while RL optimizes a lightweight correction module. This formulation efficiently isolates safety refinement from basic manipulation skills, preventing catastrophic forgetting during exploration. We provide the detailed formulation and training hyperparameters in Appendix B.1.

Evaluation Metrics. Standard binary *Success Rate (SR)* is insufficient for hazardous chemical environments as it ignores intermediate dangers. We therefore introduce the *Safe Success Rate (SSR)*, where a trial is considered successful if and only if the agent reaches the goal without violating any safety constraints throughout the episode. We quantify these violations using three domain-specific kinematic metrics: (1) *Peak Contact Force (F_{peak} [N])* measures impact magnitude, employing object-specific thresholds to protect the diverse fragile vessels; (2) *Orientation Deviation (θ_{dev} [rad])* tracks the maximum angular tilt from the vertical axis during liquid transport, enforcing a strict limit of 0.25 rad

to prevent spillage; and (3) *Spatial Error (ΔP [mm])* calculates the Euclidean distance to the target pose, requiring a precision within 20 mm for successful arrangement.

4.2. Quantifying the Safety Gap in Embodied Agents

To address Q1, we benchmark the performance of five state-of-the-art IL policies on our generative expert data. As illustrated in Table 2, a profound safety gap exists between the standard Success Rate (SR) and the Safe Success Rate (SSR) across all tested methods. While $\pi_{0.5}$ achieves the highest in-domain SR, its SSR remains insufficient for autonomous laboratory standards. This discrepancy indicates that while current models can often achieve task goals, they frequently do so via unsafe behaviors that violate the rigorous safety constraints required for handling volatile chemical agents.

To diagnose the source of this gap, we categorize failure modes into three precision-critical domains. In terms of actuation and force compliance, maintaining precise interaction is vital for manipulating delicate instruments. We observe that ACT demonstrates sub-optimal performance primarily due to its action-chunking strategy, which often produces aggressive maneuvers with contact forces exceeding threshold. Similarly, OpenVLA struggles to capture the fine-grained visual cues necessary due to its single third-person perspective, limiting its ability to adjust manipulation behaviors. Regarding irreversible fluid dynamics, handling liquids imposes zero-tolerance constraints as spillage is an irreversible failure mode. Most tested methods exhibit excessive orientation jitter (typically $0.3 - 0.5$ rad), consistently failing the strict 0.25 rad safety threshold required to prevent splashes. Notably, only $\pi_{0.5}$ satisfies this condition in over half of its successful trials, whereas other models frequently generate jerky trajectories that would be catastrophic when transporting concentrated chemical solutions. For spatial rearrangement, high-precision tasks reveal the bottlenecks of bimanual coordination. The expanded workspace and complex temporal synchronization required for dual-arm tasks often impede policy convergence for diffusion-based methods. Furthermore, most models fail to meet the tight 20 mm error tolerance required for sensitive objects, where even minor spatial errors result in static equilibrium failure.

Based on these empirical findings, we identify three critical limitations in current models. First, foundation models provide robust semantic priors but lack the specialized safety awareness required for high-risk scenarios. Second, while point-cloud-based models exhibit better spatial grounding than pure image-to-action models, they still struggle with long-horizon precision in cluttered environments. Finally, the discrepancy between standard and safe success rates underscores the urgent need to transition from simple goal-reaching metrics to holistic, safety-constrained evaluations.

4.3. RL Post-Training for Robustness

To answer Q2, we analyze how safety-aware RL fine-tuning enhances robustness. We find that vanilla IL policies behave brittlely in out-of-distribution states. Lacking recovery mechanisms, minor execution drifts rapidly escalate into irreversible failures, such as crushing glassware (violating F_{peak}) or spilling fluids due to tilt.

As shown in Figure 5, our safety-aware residual RL framework effectively addresses these vulnerabilities. Quantitative results confirm a significant boost in safety: the SSR increases by 45.2% for DP and 33.1% for $\pi_{0.5}$. Crucially, this improvement stems from *active error correction*. Unlike static baselines that adhere to fixed trajectories, RL-optimized agents learn to dynamically adjust their pose to rectify deviations and minimize contact forces. This capability enables agents to recover from hazards, meeting the strict reliability standards of autonomous science.

4.4. Generalization across Diverse Domains

To address Q3, we systematically analyze how distinct distributional shifts impact the generalization of vanilla IL policies compared to our safety-aware RL agents. We isolate three axes of variation: visual interference (lighting), physical perturbations, and spatial misalignment.

As depicted in Figure 4, our empirical analysis identifies a clear hierarchy of difficulty for baseline models. Lighting variations present the most significant challenge, inducing severe performance degradation in purely visual policies (DP and $\pi_{0.5}$) due to feature drift. This is followed by physical parameter shifts, where discrepancies between training and test dynamics precipitate execution failures. However, our safety-aware RL framework effectively mitigates these vulnerabilities. As detailed in Table 9 and Table 10, our method yields consistent performance gains across all categories. Most notably, in the most challenging lighting scenarios, the residual module compensates for the base policy's perceptual errors by adjusting actions based on dynamic feedback. This leads to a substantial recovery of the SSR, validating the core premise of SafeLab: true embodied safety requires the resilience to maintain zero-tolerance constraints under severe environmental shifts, a critical dimension that static benchmarks fail to evaluate. Qualitative visualizations are provided in Figure 4.

4.5. Sim-to-Real Transfer on Physical Robot

We establish a physical twin using a PsiBot Robot and laboratory glassware filled with liquid. To evaluate the fidelity of our physics engine, we employ a *stratified sampling* strategy: we select 10 trajectories where the simulation predicts a safety violation (e.g., $\theta_{dev} > 0.25$ rad) and 10 trajectories predicted as safe. We then execute these joint-space trajectories open-loop on the hardware.

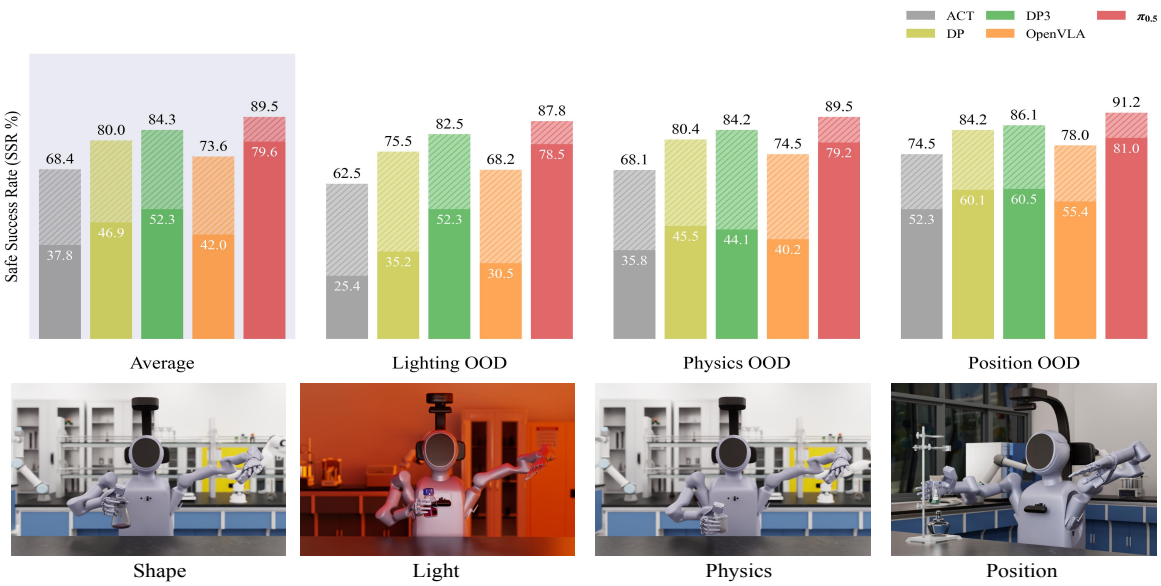


Figure 4. **Robustness against Distributional Shifts. (Top)** We compare the Safe Success Rate (SSR) of vanilla IL baselines against safety-aware RL agents across four categories: Average performance, Visual interference, Physical perturbations, and Spatial misalignment. The residual RL refinement achieves consistent gains over static policies (DP and $\pi_{0.5}$), particularly in high-uncertainty scenarios. **(Bottom)** Visualizations of the RL-tuned agent operating under perturbations. Despite severe lighting shifts and dynamic mismatches, the agent maintains zero-tolerance safety constraints, validating the efficacy of the proposed RL interface for ensuring intrinsic resilience.

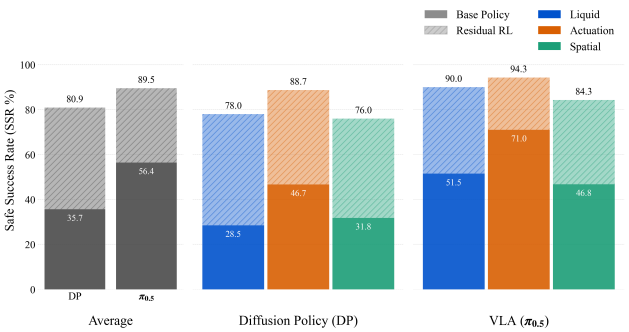


Figure 5. **Quantitative Improvements in Safe Success Rate.** We compare the Safe Success Rate (SSR) of baselines against their RL-finetuned counterparts across three task categories. **Average** denotes the mean performance across all domains. Results demonstrate our *safety-aware RL* refinement yields substantial and consistent improvements over base policies, validating the necessity of active error correction in reliable laboratory automation.

We observe a high degree of consistency between simulated safety predictions and physical outcomes. As illustrated in Figure 6 (Left), trajectories identified as safety violations in simulation, such as the instability of a graduated cylinder, correspond to hazardous states on the physical hardware. In contrast, trajectories classified as safe in simulation maintain stability during real-world execution. This correlation demonstrates that the kinematic constraints modeled in SafeLab serve as an effective proxy for real-world physical risks, validating the utility of our benchmark for evaluating embodied safety.

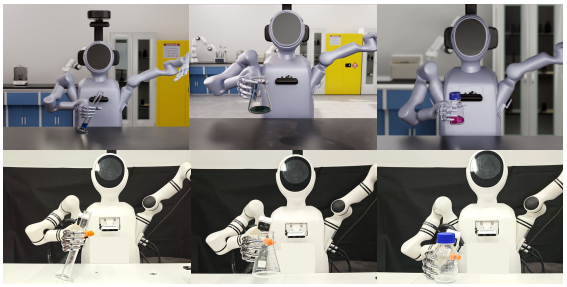


Figure 6. **Sim-to-Real Physical Validation.**

5. Conclusion and Limitations

In this work, we present SafeLab, a generative simulation benchmark for evaluating embodied safety in scientific robotics. By synthesizing a physics-grounded dataset across 64 diverse tasks, we reveal that current VLA models struggle to maintain kinematic safety under strict constraints. To address this limitation, we demonstrate a decoupled residual RL interface that enables active error recovery, substantially improving safe success rates across all evaluated domains. While SafeLab provides high-fidelity modeling of fluid–rigid interactions, our current PBD solver focuses on macroscopic liquid behavior and may exhibit volume drift over long horizons compared to Navier–Stokes methods. Future work will explore hybrid solvers to address this limitation, alongside extensions to mobile manipulation and scientific LLMs. Ultimately, SafeLab establishes a scalable foundation for developing reliable "AI Chemists" in automated scientific discovery.

Impact Statement

The automation of scientific discovery necessitates agents that are not only capable but fundamentally trustworthy. Our work exposes a critical vulnerability in contemporary generalist policies, specifically the inability of static imitation to accommodate the zero-tolerance dynamics of physical laboratories where minor execution drift precipitates irreversible hazards. By transitioning the paradigm from passive behavioral cloning to interactive safety learning, we provide the essential infrastructure to train agents capable of active perception and deviation recovery. This establishes a rigorous testbed for embodied safety, thereby facilitating future research into the robustness of embodied agents in risk-intolerant environments.

References

Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M. R., 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. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In Lim, J., Song, S., and Park, H.-W. (eds.), *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 17–40. PMLR, 27–30 Sep 2025.

Burger, B., Maffettone, P. M., Gusev, V. V., Aitchison, C. M., Bai, Y., yan Wang, X., Li, X., Alston, B. M., Li, B., Clowes, R., Rankin, N., Harris, B., Sprick, R. S., and Cooper, A. I. A mobile robotic chemist. *Nature*, 583: 237–241, 2020.

Chemspeed Technologies AG. Chemspeed high output solutions for lab automation. https://www.chemspeed.com, 2024. Accessed: 2026-01-27.

Chen, T., Chen, Z., Chen, B., Cai, Z., Liu, Y., Li, Z., Liang, Q., Lin, X., Ge, Y., Gu, Z., et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

Chi, C., Feng, S., Du, Y., Xu, Z., Cousineau, E., Burchfiel, B., and Song, S. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems (RSS)*, 2023.

Cooper, A. I., Courtney, P., Darvish, K., Eckhoff, M., Fakhruldeen, H., Gabrielli, A., Garg, A., Haddadin, S., Harada, K., Hein, J., Hübner, M., Knobbe, D., Pizzuto, G., Shkurti, F., Shrestha, R., Thurow, K., Vescovi, R., Vogel-Heuser, B., Ádám Wolf, Yoshikawa, N., Zeng, Y., Zhou, Z., and Zwirnmann, H. Accelerating discovery in natural science laboratories with ai and robotics: Perspectives and challenges. *Science Robotics*, 10(106):eadv7932, 2025. doi: 10.1126/scirobotics.adv7932. URL https://www.science.org/doi/abs/10.1126/scirobotics.adv7932.

Dai, T., Vijayakrishnan, S., Szczypiński, F. T., Ayme, J.-F., Simaei, E., Fellowes, T., Clowes, R., Kotopanov, L., Shields, C. E., Zhou, Z., Ward, J. W., and Cooper, A. I. Autonomous mobile robots for exploratory synthetic chemistry. *Nature*, 635:890–897, 2024.

Darvish, K., Skreta, M., Zhao, Y., Yoshikawa, N., Som, S., Bogdanovic, M., Cao, Y., Hao, H., Xu, H., Aspuru-Guzik, A., et al. Organa: A robotic assistant for automated chemistry experimentation and characterization. *Matter*, 8(2), 2025.

de Haan, P., Jayaraman, D., and Levine, S. Causal confusion in imitation learning. In *International Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

James, S., Ma, Z., Arrojo, D. R., and Davison, A. J. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.

Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E. P., Sanketi, P. R., Vuong, Q., Kollar, T., Burchfiel, B., Tedrake, R., Sadigh, D., Levine, S., Liang, P., and Finn, C. OpenVLA: An open-source vision-language-action model. In *Conference on Robot Learning (CoRL)*, 2024.

Lan, Z., Jiang, Y., Wang, R., Xie, X., Zhang, R., Zhu, Y., Li, P., Yang, T., Chen, T., Gao, H., et al. Autobio: A simulation and benchmark for robotic automation in digital biology laboratory. *arXiv preprint arXiv:2505.14030*, 2025.

Li, R., Hu, Z., Qu, W., Zhang, J., Yin, Z., Zhang, S., Huang, X., Wang, H., Wang, T., Pang, J., et al. Labutopia: High-fidelity simulation and hierarchical benchmark for scientific embodied agents. *arXiv preprint arXiv:2505.22634*, 2025a.

Li, S., Huang, Y., Guo, C., Wu, T., Zhang, J., Zhang, L., and Ding, W. Chemistry3d: Robotic interaction benchmark for chemistry experiments. *arXiv preprint arXiv:2406.08160*, 2024.

Li, Z., Yang, J., Xu, J., Xie, S., Wang, Y., Shen, Z., Chen, T., Shen, Y., Li, W., Zheng, Y., Zhang, C., Chen, M., Xie, C., and Wu, R. Lehome: A simulation environment for deformable object manipulation in household scenarios.

In *IROS 2025 - 5th Workshop on RObotic MAnipulation of Deformable Objects: holistic approaches and challenges forward*, 2025b.

Liu, B., Zhu, Y., Gao, C., Feng, Y., qiang liu, Zhu, Y., and Stone, P. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Mittal, M., Roth, P., Tigue, J., Richard, A., Zhang, O., Du, P., Serrano-Muñoz, A., Yao, X., Zurbrügg, R., Rudin, N., et al. Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning. *arXiv preprint arXiv:2511.04831*, 2025.

Nasiriany, S., Maddukuri, A., Zhang, L., Parikh, A., Lo, A., Joshi, A., Mandlekar, A., and Zhu, Y. Robocasa: Large-scale simulation of everyday tasks for generalist robots. In *RSS 2024 Workshop: Data Generation for Robotics*, 2024.

Octo Model Team, Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Xu, C., Luo, J., Kreiman, T., Tan, Y., Chen, L. Y., Sanketi, P., Vuong, Q., Xiao, T., Sadigh, D., Finn, C., and Levine, S. Octo: An open-source generalist robot policy. In *Robotics: Science and Systems (RSS)*, Delft, Netherlands, 2024.

Opentrons Labworks Inc. Opentrons ot-2: High-precision liquid handling robot. https://opentrons.com, 2024. Accessed: 2026-01-27.

Ross, S. and Bagnell, D. Efficient reductions for imitation learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 9, pp. 661–668, 2010.

Ross, S., Gordon, G., and Bagnell, D. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics (AISTAT)*, volume 15, pp. 627–635, 2011.

Sundaralingam, B., Hari, S. K. S., Fishman, A., Garrett, C. R., Wyk, K. V., Blukis, V., Millane, A., Oleynikova, H., Handa, A., Ramos, F., Ratliff, N. D., and Fox, D. Curobo: Parallelized collision-free minimum-jerk robot motion generation. *CoRR*, 2023.

Szymanski, N. J., Rendy, B., Fei, Y., Kumar, R. E., He, T., Milsted, D., McDermott, M. J., Gallant, M. C., Cubuk, E. D., Merchant, A., Kim, H., Jain, A., Bartel, C. J., Persson, K. A., Zeng, Y., and Ceder, G. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*, 624:86–91, 2023.

Tao, S., Xiang, F., Shukla, A., Qin, Y., Hinrichsen, X., Yuan, X., Bao, C., Lin, X., Liu, Y., Chan, T.-K., Gao, Y., Li, X., Mu, T., Xiao, N., Gurha, A., N, V., Choi, Y. W., Chen, Y.-R., Huang, Z., Calandra, R., Chen, R., Luo, S., and Su, H. Maniskill3: GPU parallelized robot simulation and rendering for generalizable embodied AI. In *CoRL Workshop: Towards Robots with Human-Level Abilities*, 2025.

Wang, L., Ling, Y., Yuan, Z., Shridhar, M., Bao, C., Qin, Y., Wang, B., Xu, H., and Wang, X. Gensim: Generating robotic simulation tasks via large language models. In *International Conference on Learning Representations (ICLR)*, 2024a.

Wang, Y., Xian, Z., Chen, F., Wang, T.-H., Wang, Y., Fragkiadaki, K., Erickson, Z., Held, D., and Gan, C. RoboGen: Towards unleashing infinite data for automated robot learning via generative simulation. In *International Conference on Machine Learning (ICML)*, pp. 51936–51983, 2024b.

Yarats, D., Fergus, R., Lazaric, A., and Pinto, L. Mastering visual continuous control: Improved data-augmented reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Yu, T., Quillen, D., He, Z., Julian, R., Hausman, K., Finn, C., and Levine, S. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning (CoRL)*, pp. 1094–1100, 2020.

Ze, Y., Zhang, G., Zhang, K., Hu, C., Wang, M., and Xu, H. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In *Robotics: Science and Systems (RSS)*, 2024.

Zhang, B., Zhang, Y., Ji, J., Lei, Y., Dai, J., Chen, Y., and Yang, Y. Safevla: Towards safety alignment of vision-language-action model via constrained learning. In *International Conference on Neural Information Processing Systems (NeurIPS)*, 2025a.

Zhang, Z., Ren, Z., Hsu, C.-W., Chen, W., Hong, Z.-W., Lee, C.-F., Penn, A., Xu, H., Zheng, D. J., Miao, S., Huang, Y., Gao, Y., Chen, W., Smith, H., Niu, Y., Tian, Y., Lu, Y.-R., Shao, Y.-C., Li, S., Wang, H.-T., Abate, I. I., Agrawal, P., Shao-Horn, Y., and Li, J. A multimodal robotic platform for multi-element electrocatalyst discovery. *Nature*, 647 (8089):390–396, November 2025b. ISSN 1476-4687. doi: 10.1038/s41586-025-09640-5. URL https://doi.org/10.1038/s41586-025-09640-5.

Zhao, T. Z., Kumar, V., Levine, S., and Finn, C. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems*, 2023.

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

Supplementary Material

A. Simulation and Assets

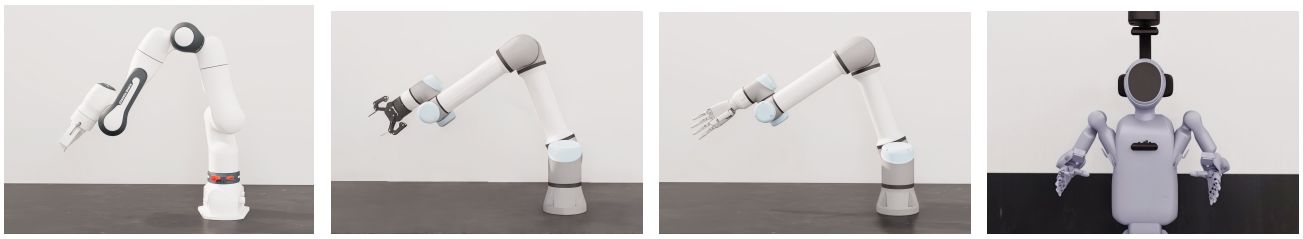
Figure 7. **Overview of the High-Fidelity Digital Asset Library.** The inventory comprises 63 calibrated objects spanning three categories: glassware, instruments and laboratory tools.

Table 3. **Catalog of the Simulation Asset Library.** The inventory consists of 63 high-fidelity assets calibrated with realistic physical and optical properties. Items are categorized by their primary functional role in the laboratory.

Category	Sub-category	Asset Instances
Glassware	Flasks	Standard Erlenmeyer flasks, Stoppered Erlenmeyer flasks, Volumetric flasks, Three-neck round-bottom flasks
	Beakers	Low-form Griffin beakers (50mL - 1000mL)
	Containers	Test tubes, Clear reagent bottles, Amber reagent bottles, Sample vials
	Measurement	Graduated cylinders, Volumetric pipettes, Pasteur pipettes
	Separation	Liebig condensers, Allihn condensers, Graham condensers, Separatory funnels
	Specialized	Liquid-in-glass thermometers, Thermometer adapters, Mortar and pestle sets
Instruments	Analytical	Analytical balances, pH meters, Spectrophotometers, Automatic polarimeters, Potentiometric titrators
	Thermal	Drying ovens, Muffle furnaces, Heating mantles, Oil baths, Electric heating mantle
	Mixing	Magnetic stirrers, Hot plates, Digital overhead stirrers
	Separation	High-speed centrifuges, Desktop centrifuges
	Environmental	Digital thermo-hygrometers
Ancillary	Handling	Micropipettes, Pipette tips, Bottle-top dispensers
	Support	Retort stands, Funnel racks, Test tube racks, Crucible tongs, 3-prong clamps
	Consumables	Weighing paper, Centrifuge tubes, Crucibles, Cuvettes, Filter paper
	Storage	Reagent cabinets, Safety waste containers

A.1. High-Fidelity Assets

To capture the complexity of real-world scientific environments, we construct a high-fidelity asset library comprising 63 distinct digital objects. Unlike generic household items used in prior robotic benchmarks, these assets are modeled with strict adherence to the geometric and functional standards of modern laboratories. We rigorously calibrate physical



(a) Franka Research 3 *(b)* UR5 with 2F-85 Gripper *(c)* UR5 with Aoyi Hand *(d)* Psibot with Aoyi Hand

Figure 8. **Overview of Supported Hardware Embodiments.** The benchmark provides native support for four distinct kinematic configurations, ranging from standard parallel grippers to high-DoF dexterous hands, enabling cross-embodiment validation.

properties by aligning inertial parameters with the density of borosilicate glass and fine-tuning friction coefficients to ensure realistic grasp dynamics. To facilitate sim-to-real transfer for vision-based agents, we employ physically-based rendering materials with calibrated Index of Refraction to accurately reproduce the specular reflections and distortions inherent to transparent glassware. Furthermore, complex concave geometries are processed with precise collision mesh decomposition to enable stable fluid-structure interactions. The inventory is organized into three functional categories: glassware, analytical instruments, and ancillary tools. Figure 7 shows the diversity of these assets, while Table 3 provides the complete catalog.

A.2. Hardware and Embodiments

To promote cross-platform generalization and accessibility, our framework natively supports four distinct robotic embodiments. We integrate three widely used manipulators, the *Franka Research 3 (FR3)*, *Universal Robots UR5*, and *Psibot*—with specialized end-effectors to encompass a broad spectrum of manipulation modalities. As illustrated in Figure 8, these configurations are strategically selected to span diverse kinematic structures. For tasks requiring high-precision rigid manipulation, we employ the *UR5* paired with a *Robotiq 2F-85* parallel gripper. To facilitate human-like dexterous interaction, we integrate the *Aoyi Hand* multi-fingered hand with both the *UR5* and *Realman* platforms. Additionally, the standalone *FR3* configuration is included to facilitate research into torque-controlled compliant manipulation. This modular architecture ensures the benchmark remains applicable across varying degrees of freedom and hardware constraints, establishing a unified standard for both parallel and dexterous scientific robotics.

A.3. Domain Randomization Setting

Domain randomization serves as a cornerstone for bridging the simulation-to-reality gap, particularly in scientific robotics where strict safety constraints mandate zero-tolerance for operational hazards. To rigorously evaluate the capacity of RL to enhance out-of-distribution robustness, we implement a staged randomization protocol tailored to these safety-critical scenarios. During the policy initialization phase via IL, we limit environmental variations primarily to spatial perturbations within the nominal workspace, ensuring the agent acquires stable fundamental primitives. In the subsequent RL refinement phase, we significantly expand the randomization scope to benchmark high-level generalization and active hazard avoidance. We broaden spatial distributions to include edge-case configurations that challenge kinematic limits and introduce extensive visual variability through diverse lighting intensities and material textures. Crucially, we systematically perturb fundamental physical dynamics, including friction coefficients, restitution, and object mass, to model the unpredictable interactions between robotic hands and varied laboratory apparatus. By exposing the scientific embodied agent to these rigorous uncertainties, the framework ensures that learned policies can effectively transfer to the complex dynamics of real-world chemical automation.

B. Reinforcement Learning Details

B.1. Residual RL Formulation

Formulation. To efficiently adapt pre-trained policies to safety-critical constraints, we employ a residual policy architecture. Let $\pi_{\text{base}}(s)$ denote the policy trained via behavioral cloning on expert demonstrations. During the RL phase, we freeze the weights of π_{base} and introduce a learnable residual policy $\pi_{\text{res}}(s)$ defined over a normalized action space $\mathcal{A}_{\text{res}} = [-1, 1]^{d_{\text{action}}}$. The final action $\mathbf{a}_t$ is computed as:

$$\mathbf{a}_t = \pi_{\text{base}}(\mathbf{s}_t) + \lambda \cdot \mathbf{a}_t^{\text{res}}, \quad \text{where } \mathbf{a}_t^{\text{res}} \sim \pi_{\text{res}}(\mathbf{s}_t). \tag{2}$$

Here, α represents an explicit scaling factor fixed at 0.1 that maps the normalized residual output to the physical joint limits, thereby regulating the maximum magnitude of the corrective action.

Strategic Rationale. We explicitly constrain the residual policy to operate within a normalized numerical range $[-1, 1]$ rather than directly outputting physical joint angles. This design choice serves two critical purposes. First, it decouples the learning process from the varying physical scales of different joints to ensure consistent gradient dynamics across all degrees of freedom. Second, it imposes a hard safety constraint on policy deviation. By strictly bounding the residual term $\mathbf{a}_t^{\text{res}}$, the scalar α defines a permissible trust region around the expert policy. This guarantees that the RL agent cannot theoretically override the fundamental manipulation priors learned by the base model, such as grasping poses, but instead limits its authority to the subtle kinematic micro-adjustments required for safety compliance.

B.2. Observation and Action Space

To enable precise manipulation under strict safety constraints, we formulate the observation and action spaces to explicitly couple visual perception with proprioceptive feedback. The state input is represented as a multi-modal tuple $\mathbf{o}_t = \langle \mathbf{I}_{\text{wrist}}, \mathbf{I}_{\text{third}}, \mathbf{q}_{\text{prop}} \rangle$. The visual component comprises dual-stream RGB images captured from an egocentric wrist-mounted camera and an exocentric third-person camera. To balance perceptual granularity with computational efficiency during the high-frequency RL loop, both streams are resized to a resolution of 224×224 pixels. Complementing the visual data, the proprioceptive vector $\mathbf{q}_{\text{prop}} \in \mathbb{R}^{d_{\text{prop}}}$ encodes the precise kinematic state, encompassing the joint positions and velocities of the robotic arm concatenated with the full joint states of the dexterous hand.

Regarding the action space, we operate directly in the joint space to ensure kinematic smoothness. The final action $\mathbf{a}_t \in \mathbb{R}^{d_{\text{action}}}$ represents the target joint position deltas. It is important to note that while the final execution command is in physical units of radians, the residual policy itself operates within the normalized action space $[-1, 1]^{d_{\text{action}}}$. This normalization decouples the learning process from the physical magnitude of the joints, allowing the RL agent to optimize a consistent displacement distribution before it is scaled by the safety factor α.

B.3. Reward and Safety Constraints

To guide the agent through long-horizon chemical manipulation tasks while enforcing safety learning, we employ the stage-wise progressive reward function formulated in Eq. (1). This structure balances exploration efficiency with precise safety compliance through two distinct components: a sparse stage-completion bonus and a dense hybrid penalty term.

Stage-wise Curriculum Bonus. The positive reward component is governed by the term $\mathbb{I}_{k,t} \cdot 2^{k-1} \cdot R_{\text{base}}$. Here, $\mathbb{I}_{k,t}$ serves as a binary indicator triggered when the agent satisfies the logical conditions of the k-th sub-goal. To facilitate exploration during early training phases, we utilize relaxed trigger conditions for these milestones, such as placing a beaker within a lenient radius of the target. The factor 2^{k-1} introduces an exponential curriculum that assigns significantly higher value to later stages of the workflow. This design effectively addresses the credit assignment problem inherent in long-horizon tasks, incentivizing the agent to overcome intermediate bottlenecks and progress toward the final experimental outcome.

Hybrid Safety Penalties. To refine the coarse policy learned from sparse bonuses into a precise and safe controller, we apply a dense penalty structure composed of generic regularization ($\mathcal{C}_{\text{gen}}$) and task-specific constraints ($\mathcal{C}_{\text{task}}$). The generic term $\mathcal{C}_{\text{gen}}$ promotes kinematic smoothness and operational efficiency by aggregating three physical costs: joint jerk $\mathbf{j}_t = \dddot{\mathbf{q}}_t$ to prevent sudden accelerations that could induce fluid instability; end-effector velocity $\mathbf{v}_{\text{ee}}$ to enforce safe kinetic energy limits; and mechanical energy consumption $E_t \approx \sum |\boldsymbol{\tau}_t \cdot \dot{\mathbf{q}}_t|$ to discourage high-frequency oscillations. Empirically, we calibrate the weighting coefficient λ_g within the range of $[10^{-6}, 10^{-5}]$ to provide necessary regularization for these auxiliary objectives without overshadowing the primary task rewards.

Complementing these kinematic priors, the task-specific term $\mathcal{C}_{\text{task}}$ enforces the domain-specific constraints required for chemical safety through a weighted sum of three physically grounded penalties. First, we penalize the pose residual (ΔP), defined as the Euclidean distance and rotational discrepancy between the end-effector and the strict target pose. This dense signal guides the agent from the relaxed milestone regions toward the precise interaction points required for chemical transfers. Second, to prevent liquid sloshing during transport, we penalize the orientation deviation (θ_{dev}) between the container's vertical axis and the global gravity vector, effectively constraining the workspace to upright manipulation. Finally, to protect fragile glassware, we penalize excessive contact forces (F_{peak}) that exceed a calibrated safety threshold. Unlike binary termination signals, this continuous penalty provides the gradient information necessary for the agent to learn force modulation and gentle interaction strategies.

B.4. Network Architecture

We implement the residual policy using an Actor-Critic architecture specifically tailored for pixel-based control. Unlike prior approaches that freeze visual encoders to save compute, we train the residual encoders from scratch. This design choice is critical as it allows the residual policy to focus on safety-critical visual cues, such as subtle liquid surface oscillations, that may have been treated as noise by the base VLA model. The visual inputs are processed by a standard NatureCNN architecture consisting of three convolutional layers followed by a linear projection. These visual features are then concatenated with the proprioceptive state to form the latent embedding.

The policy and value networks are modeled as Multi-Layer Perceptrons (MLPs). The Actor network parameterizes a diagonal Gaussian distribution using a 3-layer MLP with hidden units $[512, 256, 128]$. It outputs the mean and log standard deviation of the residual action. The Critic network employs a Double Q-network structure to mitigate overestimation bias, with each Q-network utilizing a larger 3-layer MLP of size $[1024, 512, 256]$. To stabilize training dynamics across diverse tasks, we apply Layer Normalization prior to the ReLU activation in each hidden layer of the Critic.

B.5. Training Implementation Details

We utilize DrQ-v2 (Yarats et al., 2022), a state-of-the-art off-policy algorithm designed for sample-efficient visual reinforcement learning. To improve generalization and data efficiency, we employ random shift augmentation on the input images. The training is performed across 128 parallel environments to accelerate data collection. We use the Adam optimizer for both actor and critic updates, with the entropy temperature α automatically tuned to balance exploration and exploitation. The detailed hyperparameters are listed in Table 4.

Table 4. **Hyperparameters for Residual RL Training.**

Category	Hyperparameter	Value
Training	Total Training Steps	1.0×10^7
	Parallel Environments	128
	Replay Buffer Capacity	250,000
	Batch Size	1024
	Discount Factor (γ)	0.99
Optimization	Optimizer	Adam
	Actor Learning Rate	3×10^{-4}
	Critic Learning Rate	3×10^{-4}
	Alpha Learning Rate	5×10^{-2}
	Critic Target Update (τ)	0.05
	Weight Decay	1×10^{-2}
Architecture	Image Augmentation	Random Shifts
	Actor Hidden Layers	[512, 256, 128]
	Critic Hidden Layers	[1024, 512, 256]

C. Baseline Implementation Details

C.1. Visual and Data Modalities

We employ a robust multi-view perception system comprising three distinct perspectives: a chest-mounted camera, a head-mounted camera, and a third-person panoramic camera. Each sensor captures RGB images at a resolution of 480×640. Regarding model-specific inputs, *Diffusion Policy* (DP) (Chi et al., 2023), *ACT* (Zhao et al., 2023), and $\pi_{0.5}$ (Black et al., 2025) leverage the full three-camera configuration to maximize spatial awareness. In contrast, *OpenVLA* (Kim et al., 2024) relies exclusively on the third-person panoramic stream to align with its pre-training paradigm. Beyond standard RGB data, our dataset provides high-fidelity depth maps, surface normals, semantic segmentation masks, and raw point clouds to support a diverse range of perception-centric research.

C.2. Proprioceptive State Representations

Proprioceptive observation spaces are tailored to the architectural requirements of each baseline. For the diffusion-based models, specifically *DP* and *DP3* (Ze et al., 2024), the state vector includes hand and arm joint positions, end-effector

Cartesian coordinates, quaternions, and the ground-truth coordinates of the target object. *ACT* adopts a higher-order representation by incorporating both joint positions and velocities for the arm and grippers. The $\pi_{0.5}$ model utilizes joint positions alongside end-effector rotation encoded as Euler angles. Notably, *OpenVLA* operates as a image-only policy without explicit proprioceptive input. For all models, the observation includes the absolute values of the robot's current joint positions, and the output corresponds to the predicted joint positions for subsequent time steps.

C.3. Training Configurations and Baselines

To ensure a rigorous comparison, we standardize training configurations across all baselines utilizing their official open-source repositories to guarantee reproducibility.

VLA Models. We evaluate two high-capacity models tailored for precision manipulation. The $\pi_{0.5}$[1] model is initialized with `pi0.5-base` weights and undergoes full-parameter fine-tuning for 10,000 steps. Training is conducted on a distributed cluster of 8 NVIDIA A800 (80GB) GPUs with a global batch size of 256 and an action chunking size of $T_a = 8$. Similarly, *OpenVLA*[2] utilizes the OpenVLA-7B base model and performs full-parameter fine-tuning for 10,000 steps on the same hardware infrastructure. However, due to its substantial memory footprint, we adjust the batch size to 16 while maintaining an action chunking size of $T_a = 8$.

Continuous Control Baselines. We implement domain-specific architectural and hyperparameter adjustments to standard control policies. *ACT*[3] employs a ResNet-18 backbone and a transformer architecture configured with 4 encoder layers, 7 decoder layers, and 8 attention heads. It is trained for 5,000 epochs on a single GPU with a batch size of 64, an action chunking size of $T_a = 4$, and a learning rate of 1×10^{-5}. We explicitly disable temporal aggregation during inference to evaluate raw decision-making capability. *DP*[4] utilizes a ResNet-hybrid Vision Transformer encoder featuring a ResNet-26 convolutional stem and is optimized for 3,000 epochs with a batch size of 128 and a learning rate of 3×10^{-4}. The temporal configuration includes an observation horizon of $T_{obs} = 1$, an action chunking size of $T_a = 8$, and 16 DDIM inference steps. Finally, *DP3*[5] processes point cloud inputs at a resolution of 1,024 points, where each point is represented by a 6-dimensional vector containing Cartesian coordinates and zero-padded features. Trained for 5,000 epochs with a batch size of 256, the model operates with an observation horizon of $T_{obs} = 1$, a prediction horizon of $T_p = 16$, and an action chunking size of $T_a = 8$, utilizing 16 denoising steps.

C.4. Dataset Details

Trajectories are recorded at a sampling frequency of 30 Hz, capturing comprehensive dual-arm joint positions, velocities, and end-effector poses. To support advanced multimodal research, we provide synchronized auxiliary data across all camera views, including semantic segmentation masks, depth maps, surface normals, and raw point clouds. During policy deployment, we implement a temporal decimation factor of $k = 4$. Under this scheme, for each single inference step of the high-level policy, the low-level controller sequentially executes four consecutive sub-steps of the predicted trajectory. This strategy effectively optimizes the trade-off between the high-frequency requirements of motor control (30 Hz) and the computational throughput of the decision-making policy.

[1] https://github.com/physical-intelligence/openpi
[2] https://github.com/openvla/openvla
[3] https://github.com/tonyzhaozh/act
[4] https://github.com/real-stanford/diffusion_policy
[5] https://github.com/YanjieZe/3D-Diffusion-Policy

D. Additional Experiments

D.1. Detailed Atomic Task Performance Breakdown

All performance evaluations in this section follow a consistent reporting format. **SR** and **SSR** denote the Success Rate (%) and Safe Success Rate (%), respectively. The latter represents the percentage of trials successfully completed without violating safety constraints, such as liquid spillage, unintended collisions, or application of excessive force.

Table 5. **Performance breakdown for liquid handling and instrument actuation tasks (Aligned to Main Table).**

Domain	Behavior	Task Instance	DP		DP3		ACT		OpenVLA		$\pi_{0.5}$	
			SR	SSR	SR	SSR	SR	SSR	SR	SSR	SR	SSR
Liquid	Pour	100ml Glass → 250ml Beaker	74	38	80	44	62	30	60	24	92	54
		100ml Cylinder → 250ml Beaker	71	34	77	40	60	27	58	21	90	51
	Lift	Bimanual 500ml Volumetric Flask	70	33	84	48	56	22	55	20	90	56
		Bimanual 1000ml Beaker	66	30	81	45	54	20	51	17	87	52
Act.	Switch	High-speed Centrifuge Panel	76	38	84	45	65	25	62	20	93	55
	Open	Drying Oven Door	71	35	82	44	61	24	60	21	93	54
		Reagent Cabinet Door	69	32	81	42	59	21	58	18	91	52
	Close	Drying Oven Door	75	38	85	46	65	26	63	23	95	56
		Reagent Cabinet Door	73	35	83	42	62	22	60	19	94	54

Table 6. **Performance evaluation of 21 vessel grasping tasks (Aligned to Main Table).**

Domain	Behavior	Task Instance	DP		DP3		ACT		OpenVLA		$\pi_{0.5}$	
			SR	SSR	SR	SSR	SR	SSR	SR	SSR	SR	SSR
Spatial	Grasp	100ml Glass Beaker	78	40	86	48	66	26	64	23	95	54
		250ml Brown Volumetric Flask	77	39	85	47	65	25	63	22	94	53
		250ml Glass Beaker	80	42	88	51	68	28	66	25	97	57
		500ml Glass Beaker	82	45	90	54	70	30	68	27	98	59
		50ml Glass Beaker	75	37	83	45	63	23	61	20	92	51
		Crucible	72	34	80	42	60	20	58	17	89	48
		Large Brown Reagent Bottle	85	48	93	57	73	33	71	30	99	61
		Small Brown Reagent Bottle	80	42	88	51	68	28	66	25	97	57
		Large Clear Reagent Bottle	84	47	92	56	72	32	70	29	99	60
		Small Clear Reagent Bottle	79	41	87	50	67	27	65	24	96	56
		100ml Plastic Cylinder	78	40	86	48	66	26	64	23	95	54
		100ml Glass Cylinder	76	38	84	46	64	24	62	21	93	52
		500ml Plastic Cylinder	80	42	88	51	68	28	66	25	97	57
		500ml Glass Cylinder	78	40	86	48	66	26	64	23	95	54
		250ml Clear Volumetric Flask	76	38	84	46	64	24	62	21	93	52
		500ml Clear Volumetric Flask	78	40	86	48	66	26	64	23	95	54
		1000ml Clear Volumetric Flask	80	42	88	51	68	28	66	25	97	57
		Erlenmeyer Flask	78	40	86	48	66	26	64	23	95	54
		Stoppered Erlenmeyer Flask	75	37	83	45	63	23	61	20	92	51
		Funnel	68	30	76	38	58	18	56	15	87	46
		Spirit Lamp	71	33	79	41	61	21	59	18	90	49

Table 7. **Performance on pick-and-place tasks (Aligned to Main Table).**

Domain	Behavior	Task Instance	DP		DP3		ACT		OpenVLA		$\pi_{0.5}$	
			SR	SSR	SR	SSR	SR	SSR	SR	SSR	SR	SSR
Spatial	Pick/Place	50ml Glass Beaker	65	28	79	41	58	20	55	15	89	51
		100ml Glass Beaker	68	31	82	44	61	23	58	18	92	54
		250ml Glass Beaker	70	33	84	46	63	25	60	20	94	56
		100ml Glass Cylinder	65	28	79	41	58	20	55	15	89	51
		100ml Plastic Cylinder	66	29	80	42	59	21	56	16	90	52
		250ml Brown Volumetric Flask	63	26	77	39	56	18	53	13	87	49
		250ml Clear Volumetric Flask	64	27	78	40	57	19	54	14	88	50
		Large Brown Reagent Bottle	72	35	86	48	65	27	62	22	96	58
		Large Clear Reagent Bottle	73	36	87	49	66	28	63	23	97	59
		Erlenmeyer Flask	65	28	79	41	58	20	55	15	89	51
		Stoppered Erlenmeyer Flask	63	26	77	39	56	18	53	13	87	49
		Crucible	58	21	72	34	51	13	48	08	82	44
		Spirit Lamp	58	21	72	34	51	13	48	08	82	44

Table 8. **Performance breakdown for bimanual handover tasks.**

Domain	Behavior	Task Instance	DP		DP3		ACT		OpenVLA		$\pi_{0.5}$	
			SR	SSR	SR	SSR	SR	SSR	SR	SSR	SR	SSR
Spatial	Handover	100ml Glass Cylinder	59	20	73	36	52	15	51	12	86	49
		100ml Plastic Cylinder	61	22	75	38	54	17	53	14	88	51
		500ml Glass Cylinder	57	18	71	34	50	13	49	10	84	47
		500ml Plastic Cylinder	58	19	72	35	51	14	50	11	85	48
		250ml Glass Beaker	62	23	76	39	56	19	54	15	89	52
		500ml Glass Beaker	60	21	74	37	53	16	52	13	87	50
		250ml Clear Volumetric Flask	56	18	70	33	49	13	47	10	82	46
		250ml Brown Volumetric Flask	57	19	71	34	50	14	48	11	83	47
		500ml Clear Volumetric Flask	58	21	71	35	53	17	50	12	84	49
		1000ml Clear Volumetric Flask	59	21	72	35	53	17	50	13	85	48

Table 9. **Ablation Study: Quantifying the Impact of Residual RL Refinement on OOD Robustness.** This table presents a comparative analysis between the **Pure Diffusion Policy (DP)** and our **DP + RL** refined version. To evaluate generalization limits, we report the Success Rate (**SR** % ↑) and the Safe Success Rate (**SSR** % ↑), where SSR represents a stringent criterion requiring task completion without any safety threshold violations (e.g., fluid spillage or excessive contact force). We benchmark performance across four Out-of-Distribution (OOD) scenarios: **Full** denotes the concurrent presence of all environmental perturbations, while the remaining columns isolate individual shifts in **Lighting**, **Physics**, and **Position**. Results are formatted as **Pure DP / DP + RL**. The significant gap between SR and SSR in baseline policies, particularly under Physics OOD, underscores the necessity of closed-loop RL refinement for safety-critical laboratory automation.

Domain	Task	Full		Lighting OOD		Physics OOD		Position OOD	
		SR	SSR	SR	SSR	SR	SSR	SR	SSR
Liquid	Pour Liquid	62 / **86**	35 / **81**	65 / **91**	38 / **88**	58 / **84**	28 / **78**	70 / **88**	45 / **82**
	Lift Vessel	45 / **79**	22 / **75**	48 / **84**	25 / **80**	40 / **78**	18 / **72**	52 / **81**	30 / **76**
Actu.	Press Switch	80 / **94**	48 / **90**	82 / **96**	52 / **93**	78 / **92**	42 / **88**	85 / **95**	55 / **92**
	Open Cabinet	65 / **89**	40 / **85**	68 / **92**	42 / **89**	62 / **87**	35 / **82**	72 / **91**	48 / **87**
	Close Cabinet	70 / **93**	52 / **91**	74 / **96**	55 / **94**	68 / **91**	45 / **88**	78 / **95**	58 / **92**
Spat.	Grasp Vessel	72 / **91**	50 / **88**	75 / **94**	54 / **91**	70 / **89**	45 / **85**	78 / **92**	58 / **89**
	Pick & Place	55 / **82**	30 / **78**	58 / **86**	32 / **82**	52 / **80**	25 / **75**	62 / **84**	38 / **80**
	Handover	35 / **68**	15 / **62**	38 / **72**	18 / **68**	32 / **65**	12 / **60**	45 / **71**	25 / **65**
	Stack Vessel	52 / **80**	32 / **76**	55 / **84**	35 / **80**	48 / **78**	28 / **74**	60 / **82**	42 / **78**

Table 10. **Ablation Study: Evaluating OOD Robustness and Safety Reliability for $\pi_{0.5}$.** This table benchmarks the performance of the $\pi_{0.5}$ Vision-Language-Action (VLA) model across isolated and concurrent environmental perturbations. **Full** denotes the simultaneous application of Lighting, Physics, and Position OOD shifts, whereas other columns examine individual axes of generalization. Performance is quantified via Success Rate (**SR** % ↑) and Safe Success Rate (**SSR** % ↑), the latter requiring strict adherence to zero-tolerance safety constraints (e.g., vessel integrity and spillage prevention). Results are presented as **Pure** $\pi_{0.5}$ **/** $\pi_{0.5}+$ **RL**. The persistent performance gap in **Pure** $\pi_{0.5}$, despite its large-scale pre-training, highlights that semantic mastery does not inherently grant physical safety—a critical gap successfully bridged by our **Residual RL** refinement.

Domain	Task	Full		Lighting OOD		Physics OOD		Position OOD	
		SR	SSR	SR	SSR	SR	SSR	SR	SSR
Liquid	Pour Liquid	76 / **92**	48 / **88**	79 / **95**	52 / **92**	75 / **91**	44 / **86**	82 / **94**	60 / **90**
	Lift Vessel	82 / **95**	55 / **92**	84 / **97**	58 / **94**	80 / **93**	52 / **90**	86 / **96**	64 / **93**
Actu.	Press Switch	89 / **97**	65 / **94**	92 / **99**	70 / **97**	88 / **96**	62 / **92**	94 / **98**	75 / **95**
	Open Cabinet	91 / **96**	70 / **93**	92 / **98**	74 / **96**	89 / **95**	68 / **91**	93 / **97**	76 / **94**
	Close Cabinet	95 / **99**	78 / **96**	96 / **99**	80 / **98**	93 / **98**	74 / **94**	97 / **99**	82 / **97**
Spat.	Grasp Vessel	84 / **95**	62 / **91**	87 / **98**	65 / **95**	82 / **95**	58 / **89**	88 / **96**	70 / **92**
	Pick & Place	77 / **92**	55 / **88**	80 / **94**	58 / **91**	75 / **90**	50 / **85**	81 / **92**	62 / **89**
	Handover	67 / **88**	38 / **82**	70 / **91**	42 / **86**	64 / **87**	35 / **78**	74 / **90**	48 / **83**
	Stack Vessel	59 / **84**	32 / **76**	65 / **88**	38 / **82**	58 / **83**	28 / **72**	68 / **85**	42 / **79**