BEYOND MOTION EXECUTION: BENCHMARKING EMBODIED AGENTS ON WET-LAB OPERATIONS AND PROCEDURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Completing a manipulation sequence does not guarantee wet-lab task success: an agent may still spill a reagent, miss a target mass, or overshoot a titration endpoint. We therefore introduce a 20-task dual-arm simulation benchmark, called WET-OP Bench, which *couples manipulation with evolving experimental states*. Tasks are organized around laboratory operations composed of shared atomic manipulation skills, spanning basic operations and experimental procedures of varying horizons. Task-specific demonstrations are provided for fourteen tasks, while six are held out to assess *zero-shot generalization* to new operation compositions and experimental conditions. For demonstrated tasks, controlled variations in task configuration, background, and instruction enable systematic evaluation of within-task generalization. The simulator couples manipulation with dynamic liquid transfer, quantitative instrument feedback, and procedure-conditioned appearance changes; the same underlying states drive observable feedback and success checks for containment, quantitative targets, and reaction outcomes. Reusable interaction graphs implement these state–action couplings across tasks, while Lab2Sim, an agentic asset-compilation pipeline, constructs specification-grounded, validated manipulation-ready assets. Finally, we conduct a capability-oriented evaluation of representative vision-language-action models and harness-based agents under controlled within-task and zero-shot settings, combining task success rates with stage scores for partial-progress tracking and failure diagnosis. The project website is available at `https://beyond-motion-execution.pages.dev/`.

1 INTRODUCTION

Recent advances in embodied intelligence have enabled general-purpose agents to operate in laboratory environments. Yet wet-lab execution differs from conventional manipulation: success depends not only on object motion, but also on evolving experimental states such as material distribution, instrument readouts, and reaction-dependent appearance. These states determine both subsequent actions and whether a procedure has been successfully completed. This raises a central question: *Can embodied policies and agents execute laboratory operations and procedures under evolving experimental states, and generalize to new compositions and experimental conditions?*

Existing laboratory benchmarks provide important pieces of this evaluation problem, but typically emphasize different subsets of the required capabilities. AutoBio (Lan et al., 2026), for example, supports precise laboratory-instrument interaction, while its quasi-static liquid model does not model pouring dynamics. LabUtopia (Li et al., 2026) substantially broadens scientific manipulation with chemical-state changes and hierarchical laboratory tasks; however, its reported tabletop benchmark primarily uses a single arm and does not incorporate liquid dynamics into evaluation. More recently, Labimus (Wu et al., 2026) introduces quantitative precision and stage-level evaluation for chemistry manipulation, but its benchmark centers on solid weighing and a corresponding procedural workflow. Together, these efforts establish important foundations for evaluating embodied agents in scientific laboratories, yet leave open a broader setting in which *dual-arm coordination, liquid transfer, instrument interaction, quantitative targets, and reaction-dependent feedback jointly determine successful execution*. Our benchmark targets this setting and further assesses whether

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

demonstrated laboratory operations generalize to new combinations and experimental conditions without task-specific demonstrations.

To study this problem, we introduce a 20-task dual-arm simulation benchmark, called WET-OP Bench, in which manipulation is coupled to evolving experimental states. Rather than treating isolated actions or individual instruments as the unit of evaluation, we organize tasks around laboratory operations composed of shared atomic manipulation skills. The demonstrated tasks span four operation families: *Handling and Placement (HP)*, *Transfer and Mixing (TM)*, *Instrument Operation (IO)*, and *Measurement and Dosing (MD)*, with tasks ranging from basic operations to experimental procedures of varying horizons. Fourteen tasks provide task-specific demonstrations, while six are held out without such demonstrations to evaluate zero-shot generalization to new operation compositions and experimental conditions. For demonstrated tasks, we further introduce controlled variations in task configuration, background, and instruction, enabling within-task generalization to be evaluated separately from transfer to held-out tasks.

To capture the consequences of laboratory actions, the simulator maintains experimental states that evolve with manipulation, including dynamic liquid transfer, quantitative instrument feedback, and procedure-conditioned appearance changes. These states provide observable feedback during execution, while the same underlying representations support checks for containment, quantitative targets, instrument states, and observable reaction outcomes. Policies must therefore respond to evolving experimental feedback and determine when task conditions have been satisfied, rather than merely replaying a prescribed manipulation sequence. We implement this coupling in two layers. *Lab2Sim*, an agentic asset-compilation pipeline, converts laboratory-object requests into specification-grounded, validated manipulation-ready assets. After asset admission, reusable interaction graphs couple robot–object interactions to evolving experimental states and their observable outcomes. This separation allows physical assets and experimental interactions to be reused across tasks while keeping the states observed by the agent consistent with those used for evaluation.

We further design the benchmark to characterize how an agent succeeds or fails rather than reducing every trial to a single endpoint metric. Tasks are annotated along four execution dimensions: task horizon, manipulation precision, quantitative-target requirement, and reaction-observation requirement. Horizon reflects the number of atomic-skill steps required for execution, while precision reflects the spatial accuracy demanded by the task; quantitative-target and reaction-observation requirements capture state-dependent experimental demands beyond manipulation alone. For each task, binary task success is complemented by stage scores that expose partial progress and localize failures within the procedure. We use this framework to evaluate representative vision-language-action models and harness-based agents under controlled within-task variations and zero-shot held-out tasks. Together, task-level success and stage-wise diagnostics reveal where execution breaks down, which experimental requirements remain unmet, and whether capabilities acquired on demonstrated tasks transfer to new procedural and experimental contexts.

In summary, our contributions are threefold:

- **Operation-centric wet-lab benchmarking.** We introduce a 20-task dual-arm benchmark organized around laboratory operations, with atomic manipulation skills composed into basic laboratory operations and experimental procedures of varying horizons. The benchmark evaluates within-task generalization under controlled variations and cross-task zero-shot generalization to new operation compositions and experimental conditions.
- **Interactive simulation of evolving experimental states.** We couple dual-arm manipulation with dynamic liquid transfer, quantitative instrument feedback, and procedure-conditioned appearance changes. Lab2Sim compiles specification-grounded, validated manipulation-ready assets, while reusable interaction graphs couple these assets to task-relevant state transitions. The same evolving states drive both policy-visible feedback and state-dependent evaluation of containment, quantitative targets, instrument states, and observable reaction outcomes.
- **Capability profiling and failure diagnosis.** We characterize wet-lab execution along task horizon, manipulation precision, quantitative-target requirements, and reaction-observation requirements. Using task success rates and stage scores, we evaluate representative VLA and harness-based agents under controlled within-task variations and zero-shot held-out tasks, enabling both capability-level comparison and localized failure analysis.



Figure 1: Overview of **WET-OP Bench**. Fourteen demonstrated tasks and six zero-shot protocols span four laboratory-operation families and are annotated by horizon, spatial precision, quantitative-target, and reaction-observation requirements. The benchmark combines controlled within-task variations with cross-task zero-shot evaluation, while stage-wise scoring distinguishes partial procedural progress from experimentally valid completion. Representative capability profiles summarize performance across these dimensions.

2 Related Work

Embodied laboratory benchmarks. Recent embodied laboratory benchmarks have broadened scientific manipulation, yet key wet-lab capabilities are still evaluated only in part. AutoBio (Lan et al., 2026) supports dual-arm instrument manipulation but excludes pouring dynamics, while LabUtopia (Li et al., 2026) spans hierarchical laboratory tasks and supports fluids and chemical transformations, yet its reported tabletop benchmark remains single-arm and does not incorporate liquid dynamics into evaluation. Our benchmark brings bimanual coordination, liquid transfer, instrument interaction, and quantitative and containment requirements into a unified evaluation setting.

Generalization and progress evaluation remain similarly fragmented. LabUtopia emphasizes object- and scene-level variations, while AtomBridge (Pang et al., 2026) addresses state mismatches across skill transitions for long-horizon execution. We further reserve tasks without task-specific demonstrations to evaluate **zero-shot generalization to new combinations of operations and experimental conditions**. Chemistry3D (Li et al., 2024) and MATTERIX (Darvish et al., 2026) model reaction processes, while Labimus (Wu et al., 2026) incorporates quantitative precision and stage-level evaluation for solid weighing. Our benchmark further combines task success with stage scores to analyze performance by task horizon, spatial precision, and reaction-induced appearance changes. Further comparisons are provided in the Supp. A.

3 Benchmark Design

We formulate wet-lab manipulation evaluation around laboratory operations and their composition into experimental procedures. The benchmark contains 20 dual-arm tasks, each paired with task-specific success criteria and stage annotations. The task suite is designed to evaluate not only whether an agent completes a manipulation sequence, but whether the resulting experimental state satisfies the intended laboratory objective.

3.1 Task Organization: From Operations to Procedures

We structure the benchmark around laboratory operations rather than isolated motion primitives or individual pieces of equipment. Tasks are built from shared atomic manipulation skills, which are composed into semantically meaningful laboratory operations and, when required, into broader

Table 1: Overview of the WET-OP Bench. **Scope**: OP., basic operation; PROC., procedure; **Hor.**: horizon (S/M/L: short/medium/long); **Prec.**: precision (L/M/H: low/medium/high); **Q/R**: quantitative target/reaction observation. **Distinctive skills** lists representative atomic skills.

	Family	Task	Scope	Hor.	Prec.	Q	R	Distinctive skills
Demonstrated Tasks	Handling & Placement (HP)	Magnetic Stirring Setup and Operation	OP.	M	M	–	–	align, place
		Bimanual Stopper Removal	OP.	S	L	–	–	grasp, pick, place
		Fehling Water-Bath Test	PROC.	M	H	–	✔	place, insert
	Transfer & Mixing (TM)	Controlled Flask-to-Cylinder Transfer	OP.	M	M	–	–	pour, align, place
		Liquid Transfer and Stirring	OP.	M	L	–	–	pour, stir, place
		Funnel-Assisted Liquid Transfer	OP.	M	M	–	–	insert, pour, align
		Blue-Bottle Redox Cycle	PROC.	L	M	–	✔	pour, insert, shake
	Instrument Operation (IO)	Centrifuge Tube Unloading	OP.	M	H	–	–	press, pick, insert
		Centrifuge Tube Loading	OP.	M	H	–	–	press, insert
		Oven Loading and Temperature Setting	OP.	L	H	–	–	handover, push, turn
		Oven Unloading	OP.	M	M	–	–	pull, handover, push
	Measurement & Dosing (MD)	Whole-Sample Solid Weighing	OP.	L	M	–	–	press, pour, align
		Color-Endpoint Titration	PROC.	S	M	–	✔	turn
		Target-Mass Powder Weighing	OP.	L	M	✔	–	scoop, pour, press
Zero-Shot Tasks	IO+TM	Oven Retrieval, Liquid Transfer, and Shaking	PROC.	L	M	–	–	pull, pour, shake
	HP + TM	MOF Funnel Transfer and Stirrer Setup	PROC.	L	M	–	–	insert, pour, align
		Evaporative-Crystallization Setup	PROC.	L	M	–	–	place, pour, align
	TM	Copper-Sulfate Dissolution	PROC.	L	M	–	✔	pour, stir
	HP	Silver-Mirror Water-Bath Test	PROC.	M	H	–	✔	place, insert
	MD	Target-Mass Weighing at a Held-Out Setpoint	OP.	L	M	✔	–	scoop, pour, press

experimental procedures. The benchmark therefore uses operations, rather than individual skills, as its primary unit of task organization.

At the operation level, we distinguish four families according to the dominant laboratory operation required for task completion: *Handling and Placement (HP)*, *Transfer and Mixing (TM)*, *Instrument Operation (IO)*, and *Measurement and Dosing (MD)*. HP covers manipulation whose primary objective is to position, place, remove, or otherwise arrange laboratory objects in task-relevant configurations. TM covers controlled material transfer and mixing, including pouring, funnel-assisted transfer, stirring, and shaking. IO captures direct manipulation of the mechanical or control state of laboratory equipment, such as centrifuge and oven operation. MD covers manipulation coupled to measurement or controlled dosing, including weighing, target-mass dosing, and endpoint-controlled titration.

The families are not mutually exclusive at the task level. A task may compose operations from multiple families when its laboratory objective requires them in sequence. For example, retrieving a vessel from an oven and subsequently transferring and mixing liquids combines IO and TM, whereas funnel-based transfer followed by apparatus placement combines HP and TM. Such multi-family labels describe operation composition rather than a separate task category.

Orthogonal to operation family, we annotate each task by its *scope*. A *basic operation* (OP.) is organized around a single laboratory-operation objective, even when completing it requires multiple atomic skills or a relatively long action sequence. An *experimental procedure* (PROC.) requires a broader sequence of operation-level steps or execution conditioned on an evolving experimental state. This distinction is semantic rather than trajectory-based: oven loading and temperature setting is a long-horizon basic operation, whereas color-endpoint titration is a short-horizon procedure because successful termination depends on the evolving state of the solution.

Together, operation family and task scope provide two complementary views of task structure: family specifies *what kind of laboratory operation* is performed, while scope specifies *how that oper-*

Table 2: Data splits and generalization protocol. **C/B/I** denote task-configuration, background, and instruction variations.

Task split	Data	Construction	Test setting	Eps.	Evaluation target
Demo	**Train**	**500 episodes/task** C/B/I variations are jointly composed per episode.	–	500	–
	Test	Same task identities; task objectives and operation structures remain unchanged.	**C-only / B-only / I-only / Mixed** 10 episodes/setting	40	**Within-task generalization** Controlled task-configuration, background, and language shifts
Zero-shot	**Test**	**No task-specific demonstrations**	**Mixed (C+B+I)**	20	**Cross-task generalization** New operation compositions and/or experimental conditions

ation participates in the experimental objective. Table 1 summarizes the 20 tasks using these annotations, together with their capability attributes and distinctive atomic skills. The listed skills are representative rather than exhaustive, highlighting the manipulation primitives that most distinguish each task.

3.2 Demonstration and Generalization Protocol

We organize evaluation into two task-level regimes based on the availability of task-specific demonstrations. The 14 *demonstrated tasks* provide training demonstrations and held-out test episodes of the same tasks, while the six *zero-shot tasks* provide no task-specific demonstrations and are used exclusively for cross-task evaluation. Table 2 summarizes the resulting data splits and evaluation settings. Our evaluation protocol builds on the evaluation framework of GenManip Gao et al. (2025), while adapting it to the requirements of wet-lab manipulation.

For each demonstrated task, we construct 500 training episodes with controlled variation along three dimensions: *task configuration (C)*, *background (B)*, and *instruction (I)*. Task-configuration variations modify task-relevant physical or experimental parameters, including object positions, liquid amount or color when applicable, and duration for time-dependent tasks. Background variations modify task-irrelevant visual context, including tabletop appearance, room background, and illumination. Instruction variations change the level of descriptive detail while preserving task semantics.

To evaluate *within-task generalization*, demonstrated tasks are tested under three isolated shifts—C-only, B-only, and I-only—and a *Mixed* setting in which all three dimensions vary jointly. Each setting contains 10 episodes per task, for a total of 40 test episodes. The task identity, objective, and underlying operation structure remain unchanged, allowing these settings to isolate robustness to controlled physical, visual, and linguistic variation.

The six zero-shot tasks are evaluated without task-specific demonstrations. Each task is tested on 20 episodes under the Mixed setting. Unlike within-task evaluation, these tasks introduce previously unseen operation compositions and/or experimental conditions while reusing operation families and atomic skills represented in the demonstrated set. This protocol therefore evaluates whether learned laboratory capabilities transfer beyond the tasks for which demonstrations are provided.

3.3 Capability Dimensions

Beyond operation family and procedural scope, we annotate each task along four execution dimensions: *horizon*, *precision*, *quantitative-target requirement*, and *reaction-observation requirement*. Horizon and precision are discretized into low, medium, and high levels, while the latter two are binary attributes, as presented in Table 1.

Horizon and precision. *Horizon* is determined by the number of atomic-skill steps required to complete a task, with tasks grouped into short, medium, and long execution horizons. *Precision* reflects the spatial accuracy required for successful manipulation, with tasks grouped into low, medium, and high precision levels according to their tolerance to positioning and alignment er-

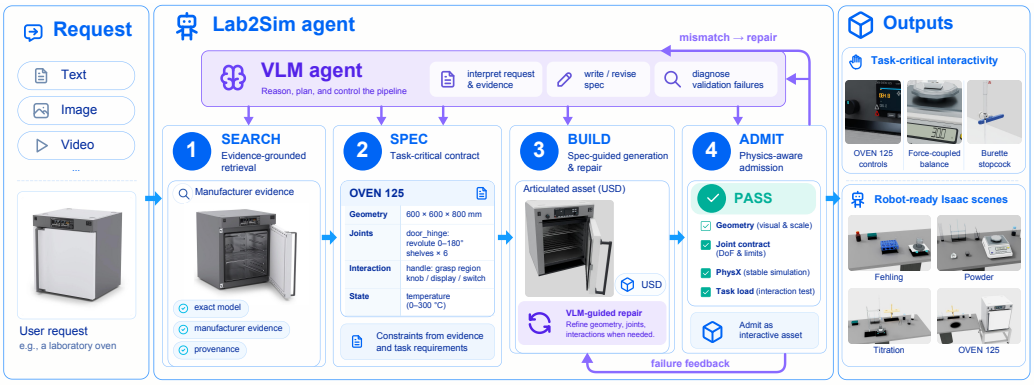


Figure 2: **Lab2Sim request-to-admission pipeline.** A multimodal asset request is grounded in retrieved evidence, converted into a task-critical specification, instantiated as an articulated USD candidate, and admitted only after geometry, articulation, physics, and interaction validation. Validation mismatches are fed back to generation for specification- preserving repair. The admitted assets provide robot-ready scene components; evolving experimental behaviors are attached subsequently through reusable interaction graphs.

rors. These two dimensions characterize complementary aspects of execution difficulty: sequence length and spatial accuracy.

State-dependent requirements. A task is marked Q when successful completion requires reaching a specified quantitative target, rather than merely obtaining a numerical readout. A task is marked R when successful completion requires observing a reaction-induced visual state, such as a titration endpoint, precipitate formation, or color change. These annotations capture experimental requirements that depend on evolving task states rather than motion execution alone.

3.4 Success Criteria and Stage-Wise Evaluation

We evaluate each episode at two complementary levels: *task success* and *stage-wise progress*. Task success is a binary outcome defined by task-specific terminal conditions rather than by completion of a prescribed motion sequence. These conditions are grounded in the resulting experimental state and may include successful material transfer and containment, instrument configuration, satisfaction of a quantitative target, or attainment of an observable reaction outcome. Consequently, executing the expected manipulation sequence is insufficient when the resulting experimental state does not satisfy the task objective.

To expose partial progress, each task is further decomposed into stages associated with task-specific intermediate criteria and scores. These stages correspond to meaningful execution milestones, such as placing or aligning an object, completing a transfer, reaching an instrument state, or satisfying an intermediate experimental condition. The resulting stage score provides a graded measure of task progress even when the final success condition is not reached. Figure 3 illustrates these milestones and their rubric weights for three tasks; the full rubrics of all 20 tasks are given in Appendix D.

Task success and stage-wise scores therefore serve complementary roles. The former measures whether the complete laboratory objective is achieved, whereas the latter localizes where execution breaks down and which requirements remain unsatisfied. Because both intermediate and terminal checks are derived from the same simulated task states that generate observable feedback, the evaluation captures the consequences of manipulation rather than motion execution alone.

4 Interactive Wet-Lab Simulation

The benchmark requires both robot-operable laboratory assets and experimental states that evolve with manipulation. We implement these requirements in two layers: Lab2Sim constructs and val-

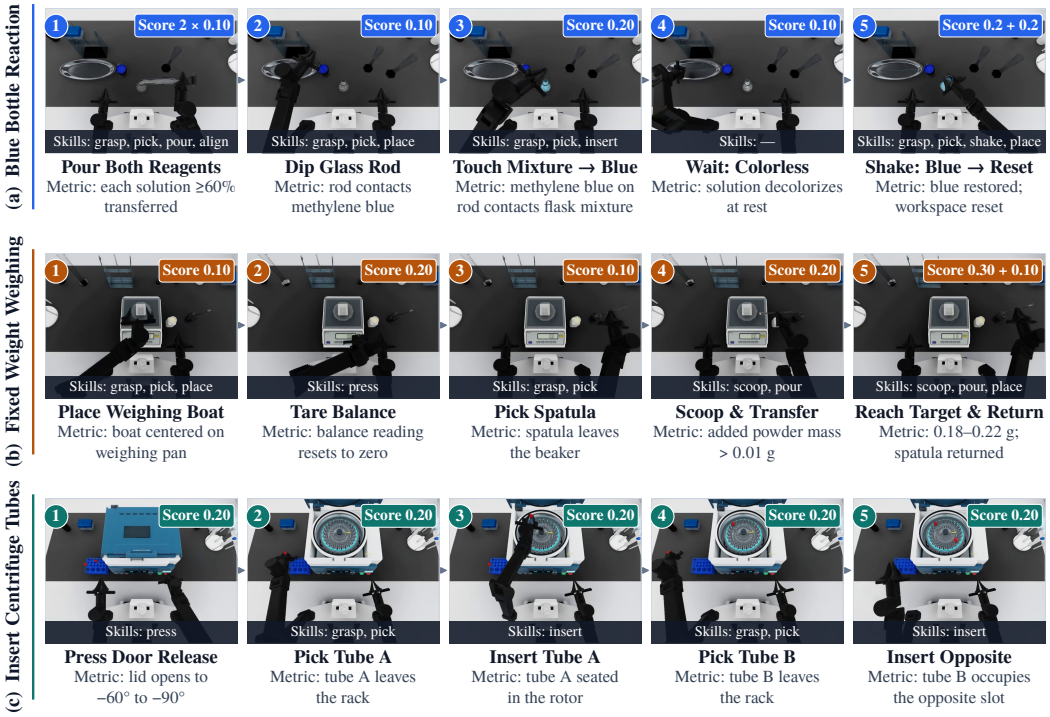


Figure 3: Representative workflows and rubric-aligned milestones for three simulated scientific manipulation tasks. Each row shows temporally ordered key frames from episode 130 in the overlook view. Score badges indicate rubric weights, and the bottom bars list the associated atomic skills.

idates robot-ready assets, while reusable interaction graphs augment the admitted assets with task-relevant experimental dynamics.

4.1 Lab2Sim: From Requests to Robot-Ready Assets

Lab2Sim is an agentic asset-compilation pipeline that converts laboratory object requests into validated, manipulation-ready USD assets. As illustrated in Figure 2, a request progresses through evidence retrieval, task-critical specification, asset generation, and physics-aware admission. The retrieved evidence is first distilled into a specification that defines the geometry, articulation, and interaction interfaces required for robotic execution. This specification serves as a persistent contract for subsequent generation and repair: validation failures may trigger revision of the asset candidate, but not silent relaxation of the task-critical constraints.

A generated candidate is admitted only after runtime and interaction checks, with a designed ConvertAsset used for the underlying asset conversion and physics qualification. Failed candidates are returned to the generation stage with the detected mismatch, forming a specification-preserving repair loop. The output of Lab2Sim is therefore a robot-ready asset with validated geometry and articulation, rather than merely a visually plausible USD. Lab2Sim stops at asset admission; time-varying processes such as liquid transfer, reaction-dependent appearance, and quantitative readouts are introduced by the interaction layer below.

4.2 Reusable Interactions for Evolving Experimental States

Robot-ready assets alone do not capture the evolving experimental states required by wet-lab execution. We therefore bind reusable OmniGraph interaction graphs to admitted assets, coupling task-relevant robot–object interactions to state transitions in the scene. The graphs are reused across tasks rather than implemented as task-specific callbacks, and support three classes of benchmark interac-

Table 3: Comparison on the test set of demonstrated tasks: SR (%) and stage score (0–100). **(a)** Results under the Mixed generalization setting. **(b)** Results grouped by operation family and scope, and capacity profile.

(a) Per-task results

#	Task	GPT-6-Astra +$\pi_{0.5}$ SR	GPT-6-Astra +$\pi_{0.5}$ Score	$\pi_{0.5}$ SR	$\pi_{0.5}$ Score
01	Magnetic Stirring Setup	**60**	**86.0**	0	72.0
02	Bimanual Stopper Removal	70	85.0	**80**	**90.0**
03	Fehling Water-Bath Test	**30**	**86.0**	10	67.0
04	Flask-to-Cylinder Transfer	10	**50.0**	10	27.0
05	Liquid Transfer & Stirring	0	**18.0**	0	6.0
06	Funnel-Assisted Transfer	0	21.0	**20**	**67.0**
07	Blue-Bottle Redox Cycle	0	14.0	**20**	**22.0**
08	Centrifuge Tube Unloading	0	32.0	0	**71.0**
09	Centrifuge Tube Loading	50	84.0	50	**88.0**
10	Oven Loading & Temp. Setting	10	46.0	**50**	**84.0**
11	Oven Unloading	70	**94.0**	70	81.0
12	Whole-Sample Solid Weighing	**40**	**64.0**	30	46.0
13	Color-Endpoint Titration	**30**	**52.0**	10	34.0
14	Target-Mass Powder Weighing	0	**50.0**	0	49.0
	Average	**26.4**	55.9	25.0	**57.4**

(b) Grouped results

Group	GPT-6-Astra +$\pi_{0.5}$ SR	GPT-6-Astra +$\pi_{0.5}$ Score	$\pi_{0.5}$ SR	$\pi_{0.5}$ Score
HP (3)	**53.3**	**85.7**	30.0	76.3
TM (4)	2.5	25.8	**12.5**	**30.5**
IO (4)	32.5	64.0	**42.5**	**81.0**
MD (3)	**23.3**	**55.3**	13.3	43.0
OP. (11)	28.2	57.3	28.2	**61.9**
PROC. (3)	**20.0**	**50.7**	13.3	41.0
Q: no (13)	**28.5**	56.3	26.9	**58.1**
Q: yes (1)	0.0	**50.0**	0.0	49.0
R: no (11)	28.2	57.3	28.2	**61.9**
R: yes (3)	**20.0**	**50.7**	13.3	41.0

tions: liquid transfer and containment, procedure-conditioned appearance changes, and quantitative instrument feedback.

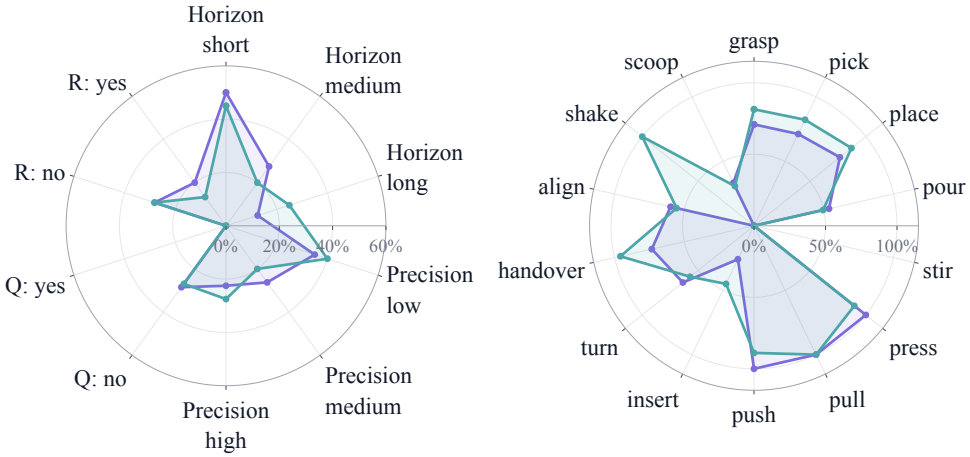
The same graph-maintained states drive both policy-visible feedback and task evaluation. Liquid states determine visible transfer and spill outcomes, reaction states determine appearance changes such as color, precipitate, or surface formation, and instrument states determine numerical feedback such as balance readings. This shared state representation keeps observation and evaluation consistent, so task success reflects the experimental consequences of manipulation rather than motion completion alone. Detailed interaction bindings and implementation are provided in Supp. B.

5 EXPERIMENTS

Setup. We evaluate a fine-tuned VLA policy and a harness-based agent built on top of it. $\pi_{0.5}$ (Physical Intelligence et al., 2025) is fine-tuned as a single multi-task policy on demonstrations from all 14 demonstrated tasks for 100k steps with a global batch size of 256. GPT-6-Astra+$\pi_{0.5}$ wraps the frozen fine-tuned $\pi_{0.5}$ within the Harness VLA framework (Zhang et al., 2026), where GPT-6-Astra performs high-level planning and invokes $\pi_{0.5}$ as a retryable low-level policy. Task-specific and global memory modules are disabled; following Yao et al. (2026), GPT-6-Astra+$\pi_{0.5}$ is instead conditioned in context on one annotated demonstration per task from the training split, without parameter updates. The main comparison uses the most difficulty Mixed (C+B+I) generalization setting of the test tasks, with the same 10 test scenes per task for both agents (140 episodes each); $\pi_{0.5}$ is further evaluated under the full protocol of Section 3.2 (680 episodes). We report success rate (SR, %) and stage score (0–100; Section 3.4), averaged per task and then across tasks; as one episode shifts a per-task SR by 10 points, we focus on trends over task groups.

Main results. Overall task success remains low for both agents despite substantial intermediate progress. Under the Mixed setting, GPT-6-Astra+$\pi_{0.5}$ and $\pi_{0.5}$ achieve similar average SRs (26.4% vs. 25.0%), while their stage scores remain above 55 (Table 3(a)). This gap is pronounced on tasks whose success depends on a precise experimental outcome: both agents reach roughly half of the stages in Target-Mass Powder Weighing but never satisfy the target-mass criterion, while $\pi_{0.5}$ attains a stage score of 71.0 on Centrifuge Tube Unloading with zero successful trials. These results reinforce the distinction between procedural progress and experimentally valid completion.

Adding the harness does not yield a uniform improvement. Because GPT-6-Astra+$\pi_{0.5}$ uses the same frozen $\pi_{0.5}$ as its low-level policy, the comparison measures the effect of augmenting that policy with high-level planning and retry. The overall SR changes little, but performance shifts substantially across task groups (Table 3(b)): GPT-6-Astra+$\pi_{0.5}$ improves HP and MD tasks and



(a) Capability profile (success rate)

(b) Atomic-skill (success rate)

Figure 4: **Capability analysis** on the mixed generalization setting of the test set in demonstrated tasks, where GPT-6-Astra+$\pi_{0.5}$ is purple and $\pi_{0.5}$ is green.

raises procedure-level SR from 13.3% to 20.0%, whereas $\pi_{0.5}$ remains stronger on TM and IO. For example, the harness improves Magnetic Stirring Setup (60% vs. 0%) and Color-Endpoint Titration (30% vs. 10%), but degrades Oven Loading and Temperature Setting (10% vs. 50%). Thus, high-level orchestration changes where the shared low-level policy succeeds, rather than consistently increasing task completion.

Capability analysis. As shown in Figure 4(a), SR drops from short to medium horizons for both agents (GPT-6-Astra+$\pi_{0.5}$: 50.0% to 27.5%; $\pi_{0.5}$: 45.0% to 20.0%), while on long horizons the harness halves SR (12.5% vs. 25.0%), suggesting that planning errors compound over long sequences. High-precision tasks show the largest score–success gap (scores 62.0/77.5, SR 22.5/27.5%): coarse stages succeed while the final insertion or alignment fails. Reaction-observation tasks are harder than the rest (SR 20.0 vs. 28.2% and 13.3 vs. 28.2%), so acting on reaction-induced visual states remains a bottleneck, which the harness only partially mitigates; the quantitative-target task, currently a single task, is never solved. The atomic-skill profile (Figure 4(b)) localizes these failures. Both agents reliably execute instrument interactions (press, pull, push: 89–100%), but struggle with skills demanding fine spatial or material control, such as scoop (33.3/30.8%) and pour (53.8/49.5%). The harness further lowers the success of contact-rich skills executed by $\pi_{0.5}$, e.g., grasp (70.9 vs. 81.5%), insert (26.0 vs. 45.1%), and handover (73.3 vs. 95.8%), suggesting that planner-driven restaging can move the frozen policy away from its demonstrated state distribution.

Full results of $\pi_{0.5}$ under all within-task settings and on the six zero-shot tasks are reported in Appendix C. Task-configuration shifts are the most damaging within-task variation (19.3% SR under C-only vs. 34.3% under B-only), and $\pi_{0.5}$ succeeds in only one of 120 zero-shot episodes, reusing demonstrated motion segments rather than recomposing operations into new procedures.

6 Conclusion and Future Work

We introduced a 20-task dual-arm benchmark for evaluating embodied agents on wet-lab operations and experimental procedures. The benchmark organizes tasks around reusable laboratory operations, separates controlled within-task generalization from zero-shot transfer to held-out tasks, and characterizes execution requirements through horizon, precision, quantitative-target, and reaction-observation dimensions. Crucially, manipulation is coupled to evolving experimental states, including liquid transfer, quantitative instrument feedback, and procedure-conditioned appearance changes. The same simulated states drive both policy-visible feedback and task evaluation, allowing success to reflect experimental outcomes rather than motion completion alone. Together with task-level success and stage scores, this provides a structured basis for identifying where current vision-language-action models and harness-based agents succeed or fail in wet-lab execution.

AI USE STATEMENT

In this work, we used generative AI tools to support the generation of simulation assets and scenes through Lab2Sim, which form part of the synthetic benchmark data construction process. We also used generative AI tools to polish writing, including improving grammar, clarity, and presentation, without relying on them to determine the technical content or scientific claims. We have not used generative AI tools for retrieval and discovery, research ideation or execution, drafting sections of the paper, or proving mathematical claims; other required disclosure tasks are not applicable to this work. All AI-assisted outputs were reviewed by the authors. Assets generated through Lab2Sim were subject to the specification and validation procedures described in the paper before being admitted into the benchmark scenes. All AI-assisted language edits were reviewed to ensure technical accuracy and fidelity to the authors' intended meaning. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the information required to reproduce the benchmark construction, evaluation protocol, and reported experiments across the main paper and supplementary material. Section 3 defines the 20-task suite, operation families, task scope, demonstrated and zero-shot splits, controlled variations, capability annotations, and task-level and stage-wise evaluation metrics. The simulation section specifies the evolving experimental states, instrument interactions, and reusable interaction graphs used to produce observable feedback and task-specific success checks, while the supplementary material documents the Lab2Sim asset-construction pipeline and additional implementation details. For all evaluated systems, we report the observation–action interface, training or adaptation protocol, evaluation settings, and metrics used for comparison. The supplementary material further provides complete task specifications, success predicates, data-generation and split details, and additional experimental results. We will release the benchmark assets, task definitions, evaluation code, and configuration files needed to reproduce the reported results once accepted.

REFERENCES

Jianbao Cao, Zhangrui Zhao, Bohan Feng, Zixuan Hu, Rui Li, Haiyuan Wan, Chenxi Li, Jingyuan Li, Wenzhe Cai, Lei Bai, et al. Labbuilder: Protocol-grounded 3d layout generation for interactable and safe laboratory. *arXiv preprint arXiv:2605.02288*, 2026.

Kourosh Darvish, Arjun Sohal, Abhijoy Mandal, Hatem Fakhruldeen, Nikola Radulov, Zhengxue Zhou, Satheeshkumar Veeramani, Joshua Choi, Sijie Han, Brayden Zhang, et al. Matterix: toward a digital twin for robotics-assisted chemistry laboratory automation. *Nature computational science*, 6(1):67–82, 2026.

Ning Gao, Yilun Chen, Shuai Yang, Xinyi Chen, Yang Tian, Hao Li, Haifeng Huang, Hanqing Wang, Tai Wang, and Jiangmiao Pang. Genmanip: Llm-driven simulation for generalizable instruction-following manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12187–12198, 2025.

Zhiqian Lan, Yuxuan Jiang, Ruiqi Wang, Xuanbing Xie, Rongkui Zhang, Yicheng Zhu, LI PEIHANG, Tianshuo Yang, Tianxing Chen, Haoyu Gao, et al. Autobio: A simulation and benchmark for robotic automation in digital biology laboratory. In *International Conference on Learning Representations*, volume 2026, pp. 146176–146196, 2026.

Rui Li, Zixuan Hu, Wenxi Qu, Jinouwen Zhang, Zhenfei Yin, Sha Zhang, Xuantuo Huang, Hanqing Wang, Tai Wang, Jiangmiao Pang, et al. Labutopia: High-fidelity simulation and hierarchical benchmark for scientific embodied agents. *Advances in Neural Information Processing Systems*, 38, 2026.

Shoujie Li, Yan Huang, Changqing Guo, Tong Wu, Jiawei Zhang, Linrui Zhang, and Wenbo Ding. Chemistry3d: Robotic interaction benchmark for chemistry experiments. *arXiv preprint arXiv:2406.08160*, 2024.

Yiwen Pang, Bo Zhou, Changjin Li, Xuanhao Wang, Shengxiang Xu, Deng-Bao Wang, Min-Ling Zhang, and Shimin Di. AtomBridge: Agentic VLA inference plugin for long-horizon tasks in scientific experiments. *arXiv preprint arXiv:2602.09430*, 2026.

Physical Intelligence et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Baochang Ren, Xinjie Liu, Xi Chen, Yanshuo Liu, Chenxi Li, Daqi Gao, Zeqin Su, Jintao Xing, Zirui Xue, Rui Li, et al. Labvla: Grounding vision-language-action models in scientific laboratories. *arXiv preprint arXiv:2606.13578*, 2026.

Yuhan Wu, Zhao Jin, Tao Li, Yuheng Zhang, Zhichao Wang, Shuo Wang, Jun Jiang, Xiaobo Li, Yanyong Zhang, Jian Tang, et al. Labimus: A simulation and benchmark for humanoid dexterous manipulation in chemical laboratory. *arXiv preprint arXiv:2606.31037*, 2026.

Yifei Yao, Yude Zou, Guanbo Jiao, Ning Gao, Yukai Wang, Jiantong Chen, Zihou Zhu, Xinyu Li, Ningyu Yan, Shuai Wang, Jianming Ma, Haoyu Guo, Xing Gao, Weinan Zhang, and Chunhua Shen. How far does a frontier agent get as an embodied policy? A comprehensive evaluation of GPT-6-Astra on EBench. Technical report, 2026. URL `https://internrobotics.github.io/EBench-Astra-Tech-Report/`.

Yixian Zhang, Huanming Zhang, Feng Gao, Xiao Li, Zhihao Liu, Yi Nie, Chunyang Zhu, Jiaxing Qiu, Yuchen Yan, Jiyuan Liu, Wenhao Tang, Jiaji Rao, Zhengru Fang, Changxu Wei, Yu Wang, Wenbo Ding, and Chao Yu. Harness VLA: Steering frozen VLAs into reliable manipulation primitives via memory-guided agents. *arXiv preprint arXiv:2607.08448*, 2026.

A More Discussion of Related Work

Embodied laboratory benchmarks. Recent embodied laboratory benchmarks have substantially expanded the scope of scientific manipulation, but they emphasize different subsets of the capabilities required for wet-lab execution. LabUtopia (Li et al., 2026) spans atomic actions and long-horizon workflows and supports fluids and chemical transformations, while its reported tabletop evaluation primarily uses a single arm and does not incorporate liquid dynamics into evaluation. AutoBio (Lan et al., 2026) supports dual-arm instrument manipulation, but its quasi-static liquid model excludes pouring dynamics. These settings leave open the joint evaluation of bimanual coordination and material-state change, which is central to operations such as holding, pouring, and containment-aware transfer. Our benchmark explicitly couples dual-arm manipulation with simulated liquid transfer and task-specific containment criteria.

A second challenge concerns generalization beyond demonstrated tasks. LabUtopia evaluates generalization under object and scene variations (Li et al., 2026), while AtomBridge (Pang et al., 2026) targets state mismatches at skill transitions to support longer-horizon execution. In contrast, we organize demonstrations around laboratory operations composed of atomic manipulation skills, and reserve tasks without task-specific demonstrations to assess **zero-shot generalization to new combinations of operations and experimental conditions**. This setting tests whether previously demonstrated operations remain effective when recomposed under new procedural or quantitative requirements, rather than only under variations of an existing task.

Finally, wet-lab success often depends on experimental states that evolve during execution. Chemistry3D (Li et al., 2024) and MATTERIX (Darvish et al., 2026) model reaction processes, while Labimus (Wu et al., 2026) introduces quantitative precision and stage-level evaluation in solid weighing. Our benchmark extends this style of evaluation across liquid handling, instrument operation, analytical weighing, and reactive procedures, using spill limits, numerical targets, and observable reaction outcomes as task-specific success criteria. We further analyze performance by task horizon, spatial precision, and the need to observe reaction-induced appearance changes, combining task success rates with skill-aligned stage scores to expose partial progress and unmet experimental requirements.

Laboratory scene generation. Laboratory scene generation has evolved from procedural layout construction toward increasingly task-conditioned simulation. LabScene, part of LabUtopia (Li et al., 2026), procedurally assembles curated laboratory assets under layout constraints. LabBuilder (Cao et al., 2026) conditions layout generation on experimental protocols and checks geometric validity, chemical safety, and navigation reachability, but does not validate physical manipulation feasibility. RoboGenesis, the data engine in LabVLA (Ren et al., 2026), further connects asset and scene generation with workflow composition and success-filtered demonstration synthesis.

Our pipeline focuses on the complementary problem of constructing manipulation-ready assets and binding them to executable experimental interactions. Lab2Sim derives specification-grounded asset descriptions from manufacturer evidence and validates generated geometry and articulation against those specifications. Reusable interaction graphs are then bound to the validated assets to implement dynamic liquid transfer, procedure-conditioned appearance changes, and numerical instrument readouts. Because these interactions and the corresponding success checks share the same simulated states, the resulting scenes support executable wet-lab evaluation rather than serving only as static layouts or visual assets.

B Interactive Simulation Implementation Details

This appendix provides implementation details for the two simulation components introduced in Section 4: (i) specification-grounded construction and admission of robot-ready assets through Lab2Sim, and (ii) reusable interaction graphs that couple robot actions to evolving experimental states. We focus on details needed to reproduce the benchmark behavior and evaluation, while omitting asset-generation internals that do not affect task execution.

B.1 Lab2Sim Asset Compilation and Admission

Lab2Sim converts a laboratory-object request into a validated USD asset through four stages: evidence retrieval, task-critical specification, asset generation, and physics-aware admission. A request may be provided as text, an image, or a video. Retrieved evidence is used to define an asset specification containing the properties required by downstream robot interaction, including object scale, link decomposition, articulation type and limits, grasp or contact regions, placement surfaces, and task-relevant control interfaces.

The specification is treated as a persistent contract during asset generation. Generated candidates are checked against it before being admitted into benchmark scenes. Admission verifies four properties: geometry and scale, articulation, stable physics execution, and task-relevant interaction. Candidates that fail a check are returned to the generation stage together with the detected mismatch; the specification itself is not relaxed during this repair process.

The output of Lab2Sim is therefore a manipulation-ready asset with validated geometry, articulation, collision, and interaction interfaces. Lab2Sim does not implement experimental dynamics such as liquid transfer, reaction-dependent appearance, or quantitative instrument readouts. These behaviors are introduced after asset admission through the reusable interaction graphs described below.

B.2 Reusable Interaction Graphs

Experimental dynamics are implemented as reusable OmniGraph action graphs bound to the task-relevant prims of admitted assets. The graphs are stepped together with the physics simulation and maintain the experimental states used both for policy-visible feedback and for evaluation. A graph is instantiated for a new scene by binding its required prim paths and parameters rather than by implementing a task-specific callback.

Table 4 summarizes the three interaction classes used in the benchmark.

Liquid transfer and containment. Liquids are represented using PhysX PBD particle sets initialized inside vessel volumes. Transfer is therefore induced by physical manipulation of the vessels rather than by a scripted visual transition. The interaction graph tracks the amount of liquid associated with each vessel and the residual volume outside valid containers, which is treated as spill.

Table 4: Reusable interaction bindings for evolving experimental states. The same simulated states determine policy-visible feedback and evaluation predicates.

Interaction	Maintained state	Observable / evaluated outcome
Liquid transfer and containment	PBD liquid particles, vessel occupancy, transferred and spilled volume, and articulated flow controls such as the burette stopcock	Visible liquid transfer and surface state; transferred-volume, containment, and spill predicates
Reaction-conditioned appearance	Procedure-dependent reaction progress conditioned on events such as mixing, immersion, waiting, or shaking	Visible color, precipitate, or surface change; reaction-state and endpoint predicates
Quantitative instrument feedback	Pan occupancy, accumulated mass, and tare offset	Rendered balance readout; target-mass and weighing predicates

For *Controlled Flask-to-Cylinder Transfer*, a transfer stage is reached when liquid enters the target cylinder, while containment is evaluated from the fraction remaining in the source and target vessels. In *Liquid Transfer and Stirring*, intermediate scores track transfer progress before rod insertion and placement. In *Funnel-Assisted Liquid Transfer*, funnel placement and transferred volume are evaluated separately. For *Color-Endpoint Titration*, liquid delivery is controlled by the simulated burette stopcock, so dispensing begins and ends with the agent's manipulation of the articulated control rather than with a predefined animation.

Reaction-conditioned appearance. Reaction-visible tasks update material appearance from procedure-specific state conditions. These transitions are written directly to the simulated liquid or surface material and are therefore observable through the same RGB observations provided to the agent.

For *Color-Endpoint Titration*, the relevant terminal observation is a stable pale-pink endpoint after the stopcock has been closed. In the *Fehling Water-Bath Test*, reaction progress begins only after the test tube satisfies the required immersion condition, with the reaction latency sampled per episode. In the *Blue-Bottle Redox Cycle*, the graph models the transition from blue to colorless while standing and the return to blue after shaking. The held-out *Silver-Mirror Water-Bath Test* reuses the same event-conditioned appearance mechanism for surface formation.

Quantitative instrument feedback. The analytical balance maintains the mass currently supported by the pan and a tare offset m_0. Pressing tare stores the current mass as m_0, and the display reports the resulting relative mass. The displayed value is rendered in the scene and must therefore be observed visually by the policy.

In *Whole-Sample Solid Weighing*, stage predicates track placement on the pan, tare operation, material transfer, and containment. In *Target-Mass Powder Weighing*, success additionally requires the displayed mass to fall within the prescribed target interval. The zero-shot *Target-Mass Weighing at a Held-Out Setpoint* reuses the same balance interaction while changing the quantitative target, allowing cross-task generalization to be evaluated without introducing a new instrument model.

B.3 Shared State for Observation and Evaluation

The interaction graphs expose observable consequences and evaluation predicates from the same underlying simulated states. Liquid occupancy determines both visible transfer and spill measurements; reaction progress determines both appearance changes and reaction-state checks; and the balance state determines both the rendered numerical readout and quantitative-target evaluation. Consequently, benchmark success cannot be obtained by completing a nominal motion sequence while leaving the corresponding experimental objective unsatisfied.

Adding a new task therefore requires three components: admitted robot-ready assets, bindings from those assets to the required reusable interaction graphs, and task-specific terminal and stage predicates over the resulting states. The underlying liquid, reaction, and instrument mechanisms are reused across tasks rather than reimplemented for each procedure.

C Full Evaluation of $\pi_{0.5}$

Table 5 reports the complete evaluation of $\pi_{0.5}$ at 100k training steps, covering all four within-task test settings of the demonstrated tasks and the six zero-shot tasks (680 episodes in total).

Generalization. Within demonstrated tasks (Table 5(b)), $\pi_{0.5}$ is most sensitive to task-configuration shifts (19.3% SR under C-only) and least to background shifts (34.3% under B-only): changes that alter physical execution matter more than visual context or phrasing. On the six zero-shot tasks (Table 5(c)), it succeeds in only one of 120 episodes (0.8% SR, 19.0 score). Partial progress appears only where a held-out task mirrors a demonstrated one, e.g., the silver-mirror test (score 44.2; cf. the Fehling test) and target-mass weighing at a held-out setpoint (40.5), whereas evaporative-crystallization setup and copper-sulfate dissolution score zero. The fine-tuned VLA thus reuses demonstrated motion segments but does not recompose laboratory operations into new procedures—the central gap this benchmark is designed to expose.

Table 5: Full evaluation of $\pi_{0.5}$ (100k training steps): SR (%) and stage score (0–100). **(a)** Demonstrated tasks, each evaluated on 40 episodes (10 per C-only, B-only, I-only, and Mixed setting). **(b)** Averages over the 14 demonstrated tasks per test setting. **(c)** Zero-shot tasks under the Mixed setting (20 episodes per task).

(a) Demonstrated tasks (40 episodes each)

#	Task	SR	Score
01	Magnetic Stirring Setup	17.5	76.5
02	Bimanual Stopper Removal	70.0	83.8
03	Fehling Water-Bath Test	5.0	58.8
04	Flask-to-Cylinder Transfer	20.0	42.5
05	Liquid Transfer & Stirring	7.5	23.7
06	Funnel-Assisted Transfer	22.5	66.8
07	Blue-Bottle Redox Cycle	12.5	17.2
08	Centrifuge Tube Unloading	10.0	73.0
09	Centrifuge Tube Loading	47.5	88.5
10	Oven Loading & Temp. Setting	60.0	89.0
11	Oven Unloading	45.0	67.5
12	Whole-Sample Solid Weighing	27.5	52.0
13	Color-Endpoint Titration	17.5	42.0
14	Target-Mass Powder Weighing	0.0	43.5
	Average	25.9	58.9

(b) Demonstrated tasks by test setting

Setting	Eps.	SR	Score
C-only	140	19.3	54.8
B-only	140	34.3	66.6
I-only	140	25.0	56.8
Mixed (C+B+I)	140	25.0	57.4
All settings	560	25.9	58.9

(c) Zero-shot tasks (20 episodes each)

#	Task	SR	Score
15	Oven Retrieval, Transfer & Shaking	0.0	22.0
16	MOF Funnel Transfer & Stirrer	0.0	7.0
17	Evaporative-Crystallization Setup	0.0	0.0
18	Copper-Sulfate Dissolution	0.0	0.0
19	Silver-Mirror Water-Bath Test	5.0	44.2
20	Target Mass, Held-Out Setpoint	0.0	40.5
	Average	0.8	19.0

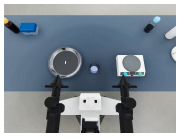
D Task Details and Evaluation Rubrics

Figures 5–7 document all 20 tasks in the order of Table 1, each with a representative scene, its instruction, capability annotations, and stage-wise rubric (metrics S1, S2, … with score weights summing to 1 and associated atomic skills; "$\times N$" marks N independently scored instances).

Figure 8 illustrates the controlled within-task variations of Section 3.2 on color-endpoint titration, including the full instruction set used for instruction variation.

Demonstrated Tasks: Handling & Placement / Transfer & Mixing

Tasks are numbered 01–20 in Table 1 order. Scope: basic operation / experimental procedure.



Family: HP **Scope:** Operation
Horizon: Medium **Precision:** Medium
Quantitative-target requirement: No
Reaction-observation requirement: No

01 Magnetic Stirring Setup and Operation

Instruction: Place the magnetic stir bar into the liquid-filled beaker, then place the beaker at the center of the magnetic stirrer platform.

Metric (stage)	Score	Skills
S1 Lift the stir bar above the tray.	**0.20**	grasp, pick
S2 Place the stir bar inside the beaker.	**0.20**	align, place
S2 Place the beaker on the magnetic stirrer.	**0.20**	grasp, pick, place
S2 Center the beaker on the stirrer plate.	**0.20**	align
S2 Retain ≥95% of the original liquid in the beaker.	**0.10**	—
S2 Keep the beaker upright.	**0.10**	place



Family: HP **Scope:** Operation
Horizon: Short **Precision:** Low
Quantitative-target requirement: No
Reaction-observation requirement: No

02 Bimanual Stopper Removal

Instruction: Use the left robotic arm to hold the transparent reagent bottle on the left steady. Use the right robotic arm to remove its stopper and place the stopper on the clean tray. Return both arms to their home positions without touching any other objects.

Metric (stage)	Score	Skills
S1 Separate the stopper from the bottle.	**0.50**	grasp, pick
S2 Place the stopper on the tray.	**0.50**	place

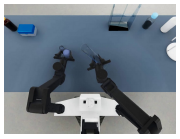


Family: HP **Scope:** Procedure
Horizon: Medium **Precision:** High
Quantitative-target requirement: No
Reaction-observation requirement: Yes

03 Fehling Water-Bath Test

Instruction: Move the test tube containing Fehling's reagent and the standard sugar solution from the test tube rack into the heated water bath, wait until a brick-red cuprous oxide precipitate forms, and return the test tube to the test tube rack.

Metric (stage)	Score	Skills
S1 Lift the test tube clear of its rack.	**0.15**	grasp, pick
S2 Reinsert the test tube into the rack.	**0.15**	insert
S2 Reach >50% reaction progress.	**0.35**	align, place
S2 Reach >95% reaction progress.	**0.35**	align, place



Family: TM **Scope:** Operation
Horizon: Medium **Precision:** Medium
Quantitative-target requirement: No
Reaction-observation requirement: No

04 Controlled Flask-to-Cylinder Transfer

Instruction: Transfer a portion of the liquid from the Erlenmeyer flask into the graduated cylinder without spilling any liquid.

Metric (stage)	Score	Skills
S1 Tilt the graduated cylinder by 5–45°.	**0.20**	grasp, pick, pour
S2 Transfer ≥1% of the flask liquid into the cylinder.	**0.40**	grasp, pick, pour, align
S3 Retain ≥95% of the original liquid in the two vessels.	**0.20**	grasp, pick, pour, align
S3 Return the cylinder and flask upright.	**0.10 ×2**	place



Family: TM **Scope:** Operation
Horizon: Medium **Precision:** Low
Quantitative-target requirement: No
Reaction-observation requirement: No

05 Liquid Transfer and Stirring

Instruction: Pour all liquid from the graduated cylinder into the beaker, stir the liquid with the glass stirring rod, and place the glass stirring rod on the clean tray.

Metric (stage)	Score	Skills
S1 Transfer ≥45% of the cylinder liquid into the beaker.	**0.20**	grasp, pick, pour, align
S2 Transfer ≥90% of the cylinder liquid into the beaker.	**0.20**	grasp, pick, pour, align
S3 Insert ≥10% of the glass rod into the beaker.	**0.40**	grasp, pick, place, stir
S4 Retain ≥90% of the original beaker liquid.	**0.04**	—
S4 Keep ≥90% of the cylinder liquid in the beaker.	**0.04**	grasp, pick, pour, align
S4 Keep the beaker upright.	**0.04**	—
S4 Return the cylinder upright; place the rod on the tray.	**0.04 ×2**	place



Family: TM **Scope:** Operation
Horizon: Medium **Precision:** Medium
Quantitative-target requirement: No
Reaction-observation requirement: No

06 Funnel-Assisted Liquid Transfer

Instruction: Use the right robotic arm to pick up the funnel from the funnel stand and place it on the reagent bottle.

Metric (stage)	Score	Skills
S1 Lift the funnel clear of its rack.	**0.30**	grasp, pick
S2 Insert the funnel upright into the bottle neck, with contact and aligned centers.	**0.30**	align, insert
S3 Transfer ≥80% of the cylinder liquid into the bottle.	**0.30**	grasp, pick, pour, align
S3 Return the cylinder upright.	**0.10**	place



Family: TM **Scope:** Procedure
Horizon: Long **Precision:** Medium
Quantitative-target requirement: No
Reaction-observation requirement: Yes

07 Blue-Bottle Redox Cycle

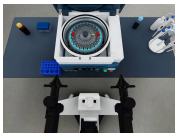
Instruction: Pour the aqueous glucose solution and then the aqueous sodium hydroxide solution into the empty Erlenmeyer flask. Add a small amount of methylene blue solution using a glass rod. Wait until the mixture becomes colorless, then shake the flask until the mixture turns blue.

Metric (stage)	Score	Skills
S1 Transfer ≥60% of each cylinder's liquid into the flask.	**0.10 ×2**	grasp, pick, pour, align
S2 Dip the glass rod into methylene blue.	**0.10**	grasp, pick, place
S3 Touch the flask mixture with the wetted rod.	**0.20**	grasp, pick, insert
S4 Allow the stationary solution to become colorless.	**0.10**	—
S5 Lift and shake the flask until the solution turns blue.	**0.20**	grasp, pick, shake
S6 Return both cylinders and the flask upright; place the rod on the tray.	**0.05 ×4**	place

Figure 5: Demonstrated tasks 01–07 (Handling & Placement and Transfer & Mixing): scenes, instructions, capability annotations, and stage-wise evaluation rubrics.

Demonstrated Tasks: Instrument Operation / Measurement & Dosing

Tasks are numbered 01–20 in Table 1 order. Scope: basic operation / experimental procedure.

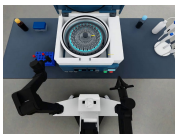


08 Centrifuge Tube Unloading

Instruction: Press the centrifuge door-opening button, remove the two centrifuge tubes, and place both centrifuge tubes in the centrifuge tube rack on the left.

Metric (stage)	Score	Skills
S1 Open the centrifuge lid to an angle of −90° to −60°.	**0.20**	press
S2 Remove each of the two tubes from the rotor.	**0.20 ×2**	grasp, pick
S2 Insert ≥40% of each tube into the rack.	**0.10 ×2**	insert
S2 Keep each tube upright.	**0.10 ×2**	insert

Family: IO **Scope:** Operation
Horizon: Medium **Precision:** High
Quantitative-target requirement: No
Reaction-observation requirement: No



09 Centrifuge Tube Loading

Instruction: Open the centrifuge and place two centrifuge tubes from the left-side centrifuge tube rack into two opposite rotor slots.

Metric (stage)	Score	Skills
S1 Open the centrifuge lid to an angle of −90° to −60°.	**0.20**	press
S1 Insert each of the two tubes into a rotor slot.	**0.20 ×2**	insert
S1 Lift each tube clear of the tube rack.	**0.20 ×2**	grasp, pick

Family: IO **Scope:** Operation
Horizon: Medium **Precision:** High
Quantitative-target requirement: No
Reaction-observation requirement: No



10 Oven Loading and Temperature Setting

Instruction: Open the drying oven door, place the empty beaker and empty Erlenmeyer flask on the internal metal rack, close the oven door, and rotate the temperature-control knob clockwise by approximately 90°.

Metric (stage)	Score	Skills
S1 Open the oven door to 10–60°.	**0.20**	pull
S2 Place the beaker on the oven rack.	**0.15**	grasp, pick, place
S2 Place the flask on the oven rack.	**0.15**	grasp, pick, handover, place
S3 Close the oven door to 0–5°.	**0.20**	push
S4 Rotate the temperature knob clockwise by 60–120°.	**0.20**	grasp, turn
S4 Keep the beaker and flask upright.	**0.05 ×2**	place

Family: IO **Scope:** Operation
Horizon: Long **Precision:** High
Quantitative-target requirement: No
Reaction-observation requirement: No



11 Oven Unloading

Instruction: Open the oven door, remove the beaker and the Erlenmeyer flask from the metal rack, place both vessels on the clean tray, and close the oven door.

Metric (stage)	Score	Skills
S1 Open the oven door to 10–60°.	**0.20**	pull
S2 Remove the beaker and flask from the oven.	**0.10 ×2**	grasp, pick
S3 Close the oven door to 0–5°.	**0.20**	push
S4 Place the beaker and flask on the tray.	**0.20 ×2**	grasp, pick, handover, place

Family: IO **Scope:** Operation
Horizon: Medium **Precision:** Medium
Quantitative-target requirement: No
Reaction-observation requirement: No

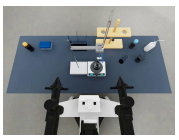


12 Whole-Sample Solid Weighing

Instruction: Weigh all solid material from the weighing boat using the balance.

Metric (stage)	Score	Skills
S1 Place the empty weighing bottle on the balance.	**0.10**	grasp, pick, place
S2 Press the tare button.	**0.10**	press
S3 Remove the bottle from the balance.	**0.10**	grasp, pick, place
S4 Transfer each of five solid particles into the bottle.	**0.10 ×5**	grasp, pick, pour, align
S5 Return the filled bottle to the balance.	**0.10**	grasp, pick, place
S5 Keep all five particles inside the bottle.	**0.10**	grasp, pick, pour, align

Family: MD **Scope:** Operation
Horizon: Long **Precision:** Medium
Quantitative-target requirement: No
Reaction-observation requirement: No



13 Color-Endpoint Titration

Instruction: Perform the acid–base titration by opening the burette stopcock and adding titrant until the solution in the Erlenmeyer flask changes from colorless to a stable pale pink, then stop the titration immediately.

Metric (stage)	Score	Skills
S1 Open the burette stopcock.	**0.20**	turn
S2 Close the burette stopcock.	**0.20**	turn
S2 Reach a stable pale-pink solution in the flask.	**0.60**	turn

Family: MD **Scope:** Procedure
Horizon: Short **Precision:** Medium
Quantitative-target requirement: No
Reaction-observation requirement: Yes



14 Target-Mass Powder Weighing

Instruction: Weigh 0.18–0.22 g of powder in the weighing boat. Place the weighing boat on the balance and tare the balance before adding powder with the laboratory spoon. Pour any powder remaining in the laboratory spoon into the reagent bottle, then place the laboratory spoon in the beaker.

Metric (stage)	Score	Skills
S1 Place the weighing boat on the balance.	**0.10**	grasp, pick, place
S2 Press tare to reset the balance reading.	**0.20**	press
S3 Lift the spoon out of the beaker.	**0.10**	grasp, pick
S4 Add powder until the balance reads >0.01 g.	**0.20**	scoop, pour
S5 Reach 0.20 ± 0.02 g.	**0.30**	scoop, pour
S5 Return the spoon to the beaker.	**0.10**	place

Family: MD **Scope:** Operation
Horizon: Long **Precision:** Medium
Quantitative-target requirement: Yes
Reaction-observation requirement: No

Figure 6: Demonstrated tasks 08–14 (Instrument Operation and Measurement & Dosing): scenes, instructions, capability annotations, and stage-wise evaluation rubrics.

Zero-Shot Tasks

Tasks are numbered 01–20 in Table 1 order. Scope: basic operation / experimental procedure.



15 Oven Retrieval, Liquid Transfer, and Shaking

Instruction: Open the oven door, remove only the clean beaker, and close the oven door. Pour the liquids from the two graduated cylinders into the beaker sequentially. Lift the beaker, shake the beaker from side to side to mix the liquids uniformly, and place the beaker upright on the table.

Metric (stage)	Score	Skills
S1 Open the oven door to 10–60°.	**0.20**	pull
S2 Remove the beaker from the oven.	**0.10**	grasp, pick
S3 Close the oven door to 0–5°.	**0.20**	push
S4 Transfer ≥60% of each cylinder's liquid into the beaker.	**0.10 ×2**	grasp, pick, pour, align
S5 Lift and shake the beaker.	**0.30**	grasp, pick, shake

Family: IO+TM **Scope:** Procedure
Horizon: Long **Precision:** Medium
Quantitative-target requirement: No
Reaction-observation requirement: No



16 MOF Funnel Transfer and Stirrer Setup

Instruction: Place the funnel in the neck of the Erlenmeyer flask, then pour the liquid from the graduated cylinder through the funnel. Set the graduated cylinder upright on the table. Remove the funnel and return the funnel to the funnel rack. Move the flask to the center of the magnetic hotplate stirrer plate.

Metric (stage)	Score	Skills
S1 Lift the funnel clear of its rack.	**0.10**	grasp, pick
S2 Insert the funnel upright into the flask neck, with contact and aligned centers.	**0.30**	align, insert
S3 Transfer ≥80% of the cylinder liquid into the flask.	**0.20**	grasp, pick, pour, align
S4 Return the funnel to its rack.	**0.10**	grasp, pick, place
S5 Place the flask on the stirrer.	**0.10**	grasp, pick, place
S5 Center the flask on the stirrer plate.	**0.20**	align

Family: HP+TM **Scope:** Procedure
Horizon: Long **Precision:** Medium
Quantitative-target requirement: No
Reaction-observation requirement: No



17 Evaporative-Crystallization Setup

Instruction: Place the alcohol lamp on the ring stand base directly below the iron ring, place the evaporating dish level on the iron ring, pour all liquid from the beaker into the evaporating dish, and return the beaker upright to the tabletop.

Metric (stage)	Score	Skills
S1 Place the lamp below the ring with center offset <3 cm.	**0.40**	grasp, pick, place
S1 Place the evaporating dish nearly level on the ring.	**0.40**	grasp, pick, align, place
S1 Transfer ≥80% of the beaker liquid into the dish.	**0.20**	grasp, pick, pour, align, place

Family: HP+TM **Scope:** Procedure
Horizon: Long **Precision:** Medium
Quantitative-target requirement: No
Reaction-observation requirement: No



18 Copper-Sulfate Dissolution

Instruction: Transfer the weighed copper sulfate crystals from the balance boat into the beaker, add the water from the graduated cylinder, and stir the mixture with a glass rod.

Metric (stage)	Score	Skills
S1 Transfer each of five solid particles into the beaker.	**0.05 ×5**	grasp, pick, pour, align
S2 Transfer ≥80% of the cylinder liquid into the beaker.	**0.20**	grasp, pick, pour, align
S3 Insert ≥10% of the glass rod into the beaker.	**0.30**	grasp, pick, place, stir
S4 Keep the beaker upright.	**0.05**	—
S4 Return the cylinder upright.	**0.10**	place
S4 Place the rod on the tray.	**0.10**	place

Family: TM **Scope:** Procedure
Horizon: Long **Precision:** Medium
Quantitative-target requirement: No
Reaction-observation requirement: Yes

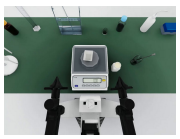


19 Silver-Mirror Water-Bath Test

Instruction: Heat the test tube containing freshly prepared Tollens' reagent and glucose standard solution in a 50–60 °C water bath without shaking. Return the test tube to the test tube rack when a clearly visible, continuous, bright silver mirror forms on the inner wall.

Metric (stage)	Score	Skills
S1 Lift the test tube clear of its rack.	**0.15**	grasp, pick
S2 Reinsert the test tube into the rack.	**0.15**	insert
S2 Reach >50% reaction progress.	**0.35**	align, place
S2 Reach >95% reaction progress.	**0.35**	align, place

Family: HP **Scope:** Procedure
Horizon: Medium **Precision:** High
Quantitative-target requirement: No
Reaction-observation requirement: Yes



20 Target-Mass Weighing at a Held-Out Setpoint

Instruction: Set the empty weighing boat on the balance and establish a zero reading using the tare button. Add powder with the laboratory spoon to obtain 0.12 g within ±0.02 g. Empty residual powder from the laboratory spoon into the reagent bottle, then place the laboratory spoon inside the beaker.

Metric (stage)	Score	Skills
S1 Place the weighing boat on the balance.	**0.10**	grasp, pick, place
S2 Press tare to reset the balance reading.	**0.20**	press
S3 Lift the spoon out of the beaker.	**0.10**	grasp, pick
S4 Add powder until the balance reads >0.01 g.	**0.20**	scoop, pour
S5 Reach the specified target mass within ±0.02 g.	**0.30**	scoop, pour
S5 Return the spoon to the beaker.	**0.10**	place

Family: MD **Scope:** Operation
Horizon: Long **Precision:** Medium
Quantitative-target requirement: Yes
Reaction-observation requirement: No

Figure 7: Zero-shot tasks 15–20, which have no task-specific demonstrations: scenes, instructions, capability annotations, and stage-wise evaluation rubrics.

Generalization in acid–base titration



1 Standard Scene
Seen laboratory arrangement



2 Background Variation
Green laboratory tabletop



3 Task Configuration
Burette and stand reconfigured



4 Mixed Variation
Green tabletop + shifted apparatus

Instruction generalization

20 seen dataset instructions + 10 held-out test instructions

Held-out test instruction example	Use the right-hand gripper to engage the burette stopcock and turn it counterclockwise to start the titrant flow. Make successive clockwise adjustments while observing the flask, and fully shut the stopcock when the solution remains a faint pink.

Complete instruction set from the wet-lab dataset

01 Perform the acid–base titration by opening the burette stopcock and adding titrant until the solution in the Erlenmeyer flask changes from colorless to a stable pale pink, then stop the titration immediately.

02 Use the right robotic arm to operate the burette stopcock, gradually deliver the titrant into the phenolphthalein-containing acid solution, and close the stopcock as soon as a stable pale-pink endpoint appears.

03 Position the right gripper toward the burette stopcock and rotate it counterclockwise to begin dispensing. Then rotate the stopcock clockwise in controlled stages until the Erlenmeyer-flask solution becomes uniformly and persistently pale pink, and immediately close the stopcock completely.

04 Start the titration by turning the burette stopcock counterclockwise with the right gripper. Add titrant in small, controlled increments, and fully shut the stopcock when the initially colorless solution just becomes stable pale pink.

05 With the right robotic arm facing the burette stopcock, rotate the stopcock counterclockwise to release titrant. Continue the titration using intermittent clockwise adjustments, stopping and closing the stopcock completely when a stable pale-pink color is observed.

06 Operate the burette stopcock with the right gripper: turn it counterclockwise to initiate flow, then regulate the flow through successive clockwise turns. Stop immediately when the acid solution containing phenolphthalein reaches a uniform, stable pale-pink color.

07 Rotate the burette stopcock counterclockwise to start adding titrant to the Erlenmeyer flask. Turn it clockwise step by step while monitoring the solution, and completely close it at the first persistent pale-pink endpoint.

08 Use the right arm to open the burette stopcock by counterclockwise rotation. Deliver the titrant gradually, and when the colorless solution changes to a stable light pink corresponding to the titration endpoint, close the stopcock immediately.

09 Face the right gripper toward the burette stopcock and rotate counterclockwise to allow titrant to enter the phenolphthalein-containing acid solution. Regulate the addition with clockwise rotations and terminate the operation when a persistent pale-pink color develops.

10 Begin the acid–base titration by turning the stopcock counterclockwise with the right robotic gripper. Adjust it clockwise in stages to control titrant delivery, then turn it to the fully closed position as soon as the solution remains pale pink.

11 Open the burette stopcock using a counterclockwise rotation of the right gripper. Add titrant under controlled flow conditions until the solution changes from colorless to just-persistent pale pink, then immediately stop the flow by fully closing the stopcock.

12 Use the right robotic arm to rotate the burette stopcock counterclockwise and initiate titrant flow. Apply successive clockwise rotations to reduce and control the flow, and close the stopcock completely once the solution has a stable pale-pink appearance.

13 Start dispensing titrant by rotating the stopcock counterclockwise. While observing the Erlenmeyer flask, make gradual clockwise adjustments and stop the titration immediately when the phenolphthalein endpoint is reached, indicated by a uniform stable pale pink.

14 Direct the right gripper at the burette stopcock, turn it counterclockwise to begin titration, and then control the addition with phased clockwise turns. Fully close the stopcock at the moment the solution becomes persistently pale pink.

15 Carry out the titration with the right robotic arm by first rotating the burette stopcock counterclockwise. Add titrant progressively using clockwise adjustments, and terminate delivery immediately when the solution changes from colorless to stable pale pink.

16 Rotate the burette stopcock counterclockwise to release titrant into the acid sample containing phenolphthalein. Continue with controlled clockwise rotation until a stable light-pink color appears throughout the solution, then close the stopcock fully without delay.

17 Use the right gripper to face and turn the burette stopcock counterclockwise, initiating the titrant flow. Regulate the flow through discrete clockwise turns and immediately return the stopcock to the fully closed position at the stable pale-pink endpoint.

18 Begin adding titrant by counterclockwise operation of the burette stopcock with the right arm. Monitor the Erlenmeyer flask continuously, make clockwise adjustments in stages, and stop as soon as the colorless solution is stably pale pink.

19 Perform the endpoint titration by opening the stopcock counterclockwise, delivering titrant in controlled increments, and closing the stopcock completely at the first stable pale-pink coloration of the phenolphthalein-containing solution.

20 Position the right robotic gripper toward the burette stopcock and rotate counterclockwise to start titrant delivery. Rotate clockwise stepwise while monitoring the flask, and immediately stop and fully close the stopcock when the solution has changed to a persistent pale pink, indicating the endpoint.

Figure 8: Within-task generalization settings for color-endpoint titration. Top: the standard scene and its background (B), task-configuration (C), and mixed variations. Bottom: a held-out test instruction and the 20 training instructions used for instruction (I) variation.