LABENGINE-CHEM: FLUX-BASED GROUNDING OF SYMBOLIC CHEMISTRY IN EMBODIED WORLDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Autonomous laboratories require simulated worlds in which physical execution and chemical evolution form a single causal process. Existing chemistry environments typically apply nominal laboratory actions directly to symbolic chemical state, while embodied simulators model how those actions are physically realized. We introduce **ChemWorld**, a conservative synchronization framework that bridges these two views through realized material, heat, and phase fluxes. Rather than allowing nominal actions to directly modify chemistry, ChemWorld measures what physically happened, commits these consequences through a single conservative world integrator, and then advances the chemical state and its observable properties. This abstraction is independent of a particular chemistry backend and extends across liquid, solid, and gas-phase processes. Experiments show that execution errors in pouring and heating propagate into distinct chemical and observable outcomes, while continuous flux coupling preserves transient dynamics that are lost by nominal-action or terminal-event interfaces. More importantly, the choice of world abstraction substantially changes agent evaluation: across 200 paired physical cases, an open-loop protocol agent achieves 100.0% success under symbolic execution but only 54.5% under embodied execution, while grounded observations and corrective actions recover performance to 66.5%. These results position ChemWorld as a bridge from protocol-level chemical reasoning to physically grounded autonomous experimentation.

1 INTRODUCTION

Autonomous laboratories require a simulated experimental world in which physical execution and chemical evolution are causally connected (Häse et al., 2019; Abolhasani & Kumacheva, 2023; Tom et al., 2024; Seifrid et al., 2022; Zhao & Zeng, 2026; MacLeod et al., 2020; Burger et al., 2020;

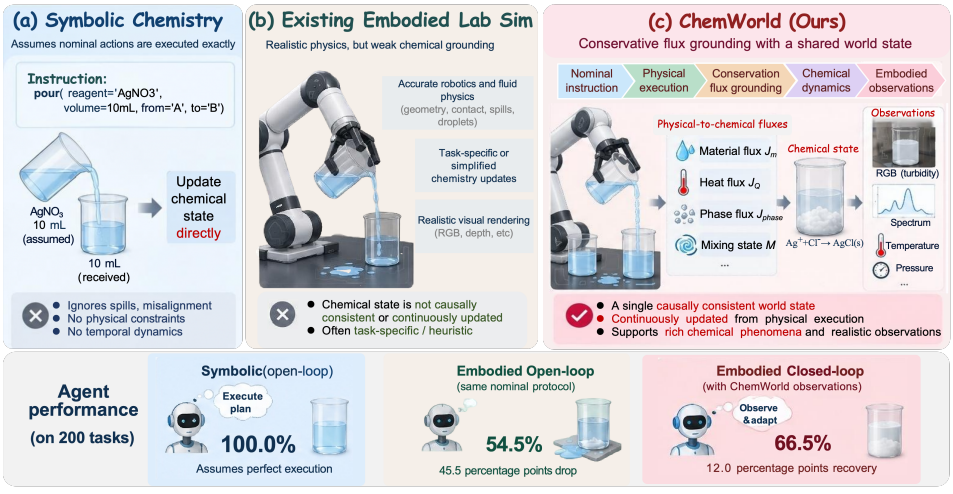


Figure 1: Advantages of our ChemWorld compared with symbolic chemistry and existing embodied laboratory simulations.

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.

Granda et al., 2018; Steiner et al., 2019; Coley et al., 2019; Shields et al., 2021; Caramelli et al., 2021; Gromski et al., 2020). Embodied simulators capture how robots move, make contact, and interact with laboratory devices, whereas symbolic chemistry environments describe how compositions, reactions, phases, and chemical properties evolve (Makoviychuk et al., 2021; Todorov et al., 2012; Xian et al., 2023; Beeler et al., 2024). While both are individually useful, they leave a critical interface unresolved: chemical evolution should be driven by the consequences of what physically happened, rather than by what an agent merely intended to do. The key challenge is therefore to *construct a principled interface that connects realized physical execution to chemical evolution while preserving causal consistency.*

Existing interfaces between physical and chemical environments (Li et al., 2025b;a; Darvish et al., 2026) are insufficient in two important ways, as shown in Fig. 1. *One the one hand, nominal actions do not uniquely determine chemical state transitions.* Consider the instruction "pour 10 mL of reagent A into vessel B". A symbolic environment can apply this operation directly as a 10 mL update, whereas an embodied system must physically realize the transfer (Beeler et al., 2024). However, a spill, trajectory error, or imperfect interaction with the apparatus may cause a different amount to reach the target. *On the other hand, even the realized terminal outcome is not always sufficient.* One could report after execution that, for example, 7 mL was actually transferred, thereby correcting the final amount. However, this still compresses an extended physical process into a single event. When material transfer, heating, reaction, or phase change occur concurrently, the temporal structure of execution also affects chemical evolution: two operations may produce the same terminal transfer while inducing different chemical trajectories. Thus, neither nominal actions nor terminal events provide a complete interface between embodied physics and chemical dynamics. *A continuous, execution-grounded representation of the physical consequences that drive chemical evolution* is expected.

We introduce ChemWorld, a conservative synchronization framework that connects continuous embodied physics with symbolic chemical dynamics through realized material, heat, and phase fluxes. Physical execution determines what material or energy is actually transferred; these fluxes are committed through a single conservative world integrator; and the chemistry backend evolves from the grounded state. Nominal actions and semantic laboratory events never directly modify chemistry. The resulting chemical state is then projected back into observable quantities such as spectra, RGB appearance, pH, turbidity, pressure, and phases. In a nutshell, symbolic chemistry assumes that commanded actions happen exactly; ChemWorld makes chemistry evolve from the conserved physical consequences that actually occurred.

Our experiments show that the choice of interface materially affects both chemical dynamics and agent evaluation. Terminal event coupling, despite recovering the final transferred amount, incurs a 0.00797 mol product-trajectory error for a 10 s pour relative to continuous flux coupling. More importantly, across 200 paired physical cases, the same open-loop protocol agent achieves 100.0% success under symbolic execution but only 54.5% under embodied execution, revealing a 45.5 percentage-point embodiment gap. Allowing grounded observations and corrective actions raises success to 66.5%. Under held-out combinations of physical perturbations, closed-loop execution improves success from 11.0% to 31.5%. These results show that symbolic evaluation can substantially overestimate autonomous-laboratory capability by abstracting away execution failures that only emerge in the physical world (Jakobi et al., 1995).

ChemWorld is designed as an interface rather than as a new chemistry solver or robot simulator. The same physical traces can drive different chemistry backends without changing the physical or integration layers, making the coupling abstraction largely independent of any particular reaction system.

The main contributions of our work are summarized as follows:

- We introduce a flux-based interface that grounds symbolic chemistry in realized material, heat, and phase transfer rather than nominal actions or terminal events.
- ChemWorld maintains a causally consistent shared state across physics, chemistry, and observation, while remaining compatible with different chemistry backends and multiphase processes.

- We show that symbolic environments can substantially overestimate autonomous laboratory performance and that physically grounded observations enable measurable closed-loop recovery.

2 Related Works

2.1 Symbolic Chemistry Environments

Symbolic chemistry environments provide structured abstractions for modeling chemical composition, reaction dynamics, separations, and laboratory observations, making them useful for protocol optimization and reinforcement learning (Zhou et al., 2017; Roch et al., 2020). ChemGymRL (Beeler et al., 2024), for example, offers interconnected virtual chemical benches for reaction, extraction, distillation, and characterization tasks, allowing agents to interact with chemical processes in a compact symbolic state space. These environments are efficient and expressive at the chemical level, but they largely abstract away the physical realization of laboratory operations. ChemWorld is complementary to this line of work: rather than replacing symbolic chemistry engines, it treats them as chemical backends and focuses on how their dynamics should be driven by realized physical execution rather than nominal laboratory actions.

2.2 Embodied Laboratory Simulation

Recent work has developed increasingly realistic simulators for robotic chemistry and scientific embodied agents. Chemistry3D (Li et al., 2025b) integrates robotic manipulation with chemical experiment simulation in NVIDIA Omniverse, supporting liquid handling and real-time visualization of temperature, color, and pH changes during reactions. LabUtopia (Li et al., 2025a) provides a broader simulation and benchmarking suite for scientific embodied agents, combining multi-physics simulation, chemically meaningful interactions, procedural laboratory scenes, and a hierarchical benchmark spanning 30 tasks and more than 200 assets. MATTERIX (Darvish et al., 2026) further develops a GPU-accelerated digital-twin framework for robotics-assisted chemistry, incorporating robotic manipulation, powder and liquid dynamics, device functionality, heat transfer, and basic reaction kinetics through a modular semantics engine. These systems substantially advance the realism and breadth of embodied laboratory simulation. ChemWorld differs in the abstraction it targets: rather than primarily expanding simulator fidelity or task coverage, it explicitly focuses on the interface that synchronizes physical execution with chemical evolution.

2.3 Physical-Chemical Coupling for Autonomous Laboratories

Existing laboratory simulators (Li et al., 2025b;a; Darvish et al., 2026) typically connect physical and chemical state through simulator-specific semantics, high-level actions, or discrete operation events. Such abstractions are often sufficient when only the terminal outcome matters, but they can lose information when chemistry evolves concurrently with transfer, heating, reaction, or phase change. ChemWorld instead treats physical–chemical synchronization itself as the modeling problem. Realized material, heat, and phase fluxes form the interface between continuous embodied physics and symbolic chemical dynamics, while a conservative integrator maintains a shared causal state. This design allows the same physical execution traces to drive different chemistry backends without modifying the physical or integration layers, and extends the same coupling abstraction across liquid, solid, and gas processes. In this sense, ChemWorld is not intended as a broader chemistry simulator, but as an execution-grounded synchronization layer that makes the causal connection between physical actions and chemical evolution explicit.

3 Method

Figure 2 summarizes the execution-to-chemistry interface: physical monitors produce realized fluxes, a conservative integrator maintains the shared state, and a pluggable chemistry backend supplies species dynamics and projected observations.

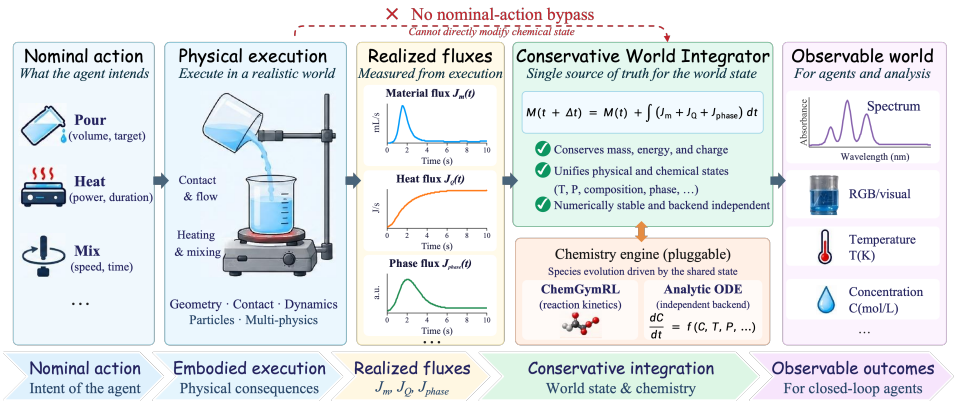


Figure 2: **ChemWorld architecture.** Nominal actions execute in the physical world; realized fluxes connect physical execution to a conservative world integrator and a pluggable chemistry backend. The shared chemical state is projected into observable quantities. Transfer histories are schematic.

3.1 Problem Formulation

We consider an autonomous laboratory in which embodied physics and chemical dynamics describe two views of the same evolving experiment. Let x_t^{phys} denote the physical state at time t, including robot and vessel poses, device states, contact geometry, and spatial material distribution, and let x_t^{chem} denote the chemical state, including species inventories, reaction progress, temperature, and phase composition. An agent issues a nominal laboratory action a_t, such as pouring, heating, mixing, or draining.

A symbolic chemistry environment (Beeler et al., 2024) may apply a_t directly as a chemical state transition. In an embodied laboratory (Makoviychuk et al., 2021), however, the action must first be realized through physical dynamics:

$$x_{t+1}^{\text{phys}} = f_{\text{phys}}\left(x_t^{\text{phys}}, a_t\right). \tag{1}$$

The realized execution can differ from the intended operation. We therefore seek an interface g that grounds chemical evolution in the physical consequences of execution:

$$x_{t+1}^{\text{chem}} = f_{\text{chem}}\left(x_t^{\text{chem}}, g(x_{t:t+1}^{\text{phys}})\right). \tag{2}$$

The goal of ChemWorld is to construct g such that the two states remain causally consistent while preserving temporal dynamics and conservation.

3.2 Conservative Flux Grounding

ChemWorld represents the interface g through physically realized fluxes. Rather than translating a laboratory command into a chemistry-specific state update, the physical layer measures what is actually transported or exchanged during each synchronization interval.

For material transport from source u to destination v, we define

$$J_{m,i}^{u\to v}(t) = \frac{\Delta n_i^{u\to v}}{\Delta t}, \tag{3}$$

where $\Delta n_i^{u\to v}$ is the amount of species i transferred during $[t, t+\Delta t]$. Physics determines *how much material moved*, while the chemistry backend determines the composition of the material at the time of transfer. Material that leaves a source without entering the intended destination is explicitly assigned to spill or the environment, rather than silently credited to the target vessel.

Thermal interaction is represented by a realized heat flux

$$J_Q(t) = \frac{\Delta Q(t)}{\Delta t}, \tag{4}$$

whose magnitude is determined by the physical heater–vessel interaction. Cross-phase transport is represented by $J^{p\to q}_{\mathrm{phase},i}(t)$, which transfers species i from phase p to phase q. Mixing is represented by a continuous state $M(t)$ that modulates reaction or transport dynamics without creating or removing material.

The physical–chemical interface is therefore

$$\mathcal{J}_t = \{J_m(t), J_Q(t), J_{\mathrm{phase}}(t), M(t)\}\,. \tag{5}$$

Importantly, $\mathcal{J}_t$ represents the *realized consequence* of an action, not its semantic description. A command to pour 10 mL therefore does not imply a 10 mL chemical update; chemistry receives only the material that was actually delivered by physical execution.

3.3 Conservative World Integration

All physical–chemical state changes pass through a single *Conservative World Integrator* (Blochwitz et al., 2012; Gomes et al., 2018). Monitors, flux estimators, semantic events, and renderers are read-only with respect to the authoritative chemical state. This gives ChemWorld a unique causal update path:

$$\text{physical execution} \to \mathcal{J}_t \to \text{world integration} \to \text{chemical evolution}. \tag{6}$$

Within each chemistry interval, ChemWorld aggregates physical fluxes over the corresponding simulation steps, applies realized material transport using the current source composition, updates thermal and mixing conditions, and advances the chemistry backend (Hairer et al., 1993; Hindmarsh et al., 2005). Phase transfer is then applied through the same shared state before conservation checks and commit.

For species i in phase p, the generic balance is

$$\frac{dn^p_i}{dt} = J^{p,\mathrm{in}}_{m,i} - J^{p,\mathrm{out}}_{m,i} + \sum_r \nu^p_{ir} r_r + \sum_{q\neq p}\left(J^{q\to p}_{\mathrm{phase},i} - J^{p\to q}_{\mathrm{phase},i}\right), \tag{7}$$

where r_r is the rate of reaction r and ν^p_{ir} is its stoichiometric coefficient. Summing over all phases cancels internal phase transfer terms, so cross-phase processes redistribute material without creating or destroying it.

Temperature is likewise maintained as part of the shared world state rather than independently integrated by the physics and chemistry modules. The chemistry backend is responsible for reaction kinetics given the grounded material, temperature, and mixing conditions; ChemWorld is responsible for supplying these inputs consistently.

This separation makes the chemistry engine replaceable. ChemWorld exposes a common backend interface for vessel state, composition queries, material transfer, temperature and mixing updates, reaction advancement, and chemical observations. In our experiments, the same physical flux traces drive both ChemGymRL (Beeler et al., 2024) and an independent analytic ODE backend (Virtanen et al., 2020) without modifying the physics, monitoring, or integration layers.

At every commit, the integrator checks conservation, non-negativity, physical–chemical transfer consistency, and thermal synchronization. Accordingly, near-machine-precision conservation errors reported in our experiments measure consistency of the synchronization and bookkeeping mechanism, rather than the fidelity of the underlying physical model.

3.4 Observable and Multiphase Chemical World

ChemWorld projects the authoritative chemical state back into observations that can be perceived by an autonomous agent. Internal species inventories are not directly exposed unless they correspond to an explicit measurement. For optical observations, the chemistry backend produces an absorbance spectrum $A(\lambda)$ (Mayerhöfer et al., 2020), from which transmission is computed as

$$T(\lambda) = 10^{-A(\lambda)}. \tag{8}$$

The resulting spectrum is mapped to visible appearance and written back to the simulated material. Other observables are generated from the same shared state: temperature from the synchronized

thermal state, pH from the corresponding chemistry model, suspended solids from turbidity, gas inventory from headspace pressure, and gas evolution from bubble-rate projections. Visual precipitates and bubbles are therefore consequences of the chemical state, rather than independent state variables that drive chemistry.

The same abstraction extends to multiphase chemistry. For species i, we represent its inventory across liquid, solid, and gas phases as

$$\mathbf{n}_i = \left[n_i^L, n_i^S, n_i^G\right], \qquad n_i^{\mathrm{tot}} = n_i^L + n_i^S + n_i^G. \tag{9}$$

Processes such as precipitation, dissolution, evaporation, condensation, and gas-liquid transfer (Sander, 2023) are expressed as conservative transfers between these phase inventories through J_{phase}, without introducing new state-mutation paths.

Our current implementation instantiates this abstraction with precipitation, sedimentation, gas evolution, gas-liquid transfer, venting, filtration, and crystallization. These are conservative lumped models rather than CFD, DEM, or molecular simulations; their purpose is to test the generality of the physical-chemical synchronization interface. We distinguish this numerical multiphase validation from our Live Isaac demonstrations, where precipitation and gas evolution are separately bound to observable solid and bubble proxies.

4 Experiments

We evaluate ChemWorld along five questions: **(1)** Does physical execution change chemical outcomes? **(2)** Is continuous flux-based coupling necessary beyond nominal actions or terminal events? **(3)** Is the resulting synchronization numerically stable, robust, and chemistry-backend independent? **(4)** Does execution grounding change the evaluation of autonomous laboratory agents? and **(5)** Does the same abstraction extend beyond liquid-phase reactions? Unless otherwise stated, all chemical state updates follow the conservative integration path described in Sec. 3; no experiment directly modifies the authoritative chemical state outside the world integrator.

4.1 Physical Execution Changes Chemical Outcomes

We first test whether the same nominal laboratory operation can produce different chemical outcomes under different physical executions. For pouring, we fix the nominal command to transfer 10 mL and vary only the lateral pouring offset from 0 to 30 mm across 35 Live Isaac (Makoviychuk et al., 2021) episodes. At 0 mm offset, the full 10.00 mL reaches the target vessel with no spill. At 30 mm, only 1.63 mL reaches the target while 8.37 mL is spilled. This execution difference propagates into chemistry, reducing the final product from 0.01479 mol to 0.00527 mol. Thus, identical symbolic actions can induce substantially different chemical outcomes once physically realized. Figure 3 presents the material and thermal trends; Table S1 provides the precise endpoint values and consistency errors.

The same effect holds for thermal execution. Using Live Isaac geometry, we vary heater–vessel contact while keeping the nominal heating operation fixed. The measured contact ratio changes cumulative heat input from 1201.8 J to 0 J and final temperature from 333.7 K to 298.2 K. When pouring and heating occur concurrently, material and heat fluxes are committed in the same synchronization interval, resulting in distinct coupled temperature and reaction trajectories. These results show that chemical evolution is determined by realized material and energy transfer rather than by nominal commands alone.

4.2 Why Continuous Flux-Based Coupling?

We next isolate the effect of the physical–chemical interface itself. We compare three coupling strategies: **Nominal**, which applies the commanded action directly to chemistry; **Event**, which applies the physically realized terminal outcome only after the operation completes; and **Flux**, which continuously applies the realized transfer at every chemistry step.

Nominal coupling fails whenever execution deviates from intent. For the 30 mm off-center pour, the nominal baseline still applies the full 10 mL command despite an 8.37 mL discrepancy between commanded and delivered volume. Event coupling corrects this terminal grounding error but loses

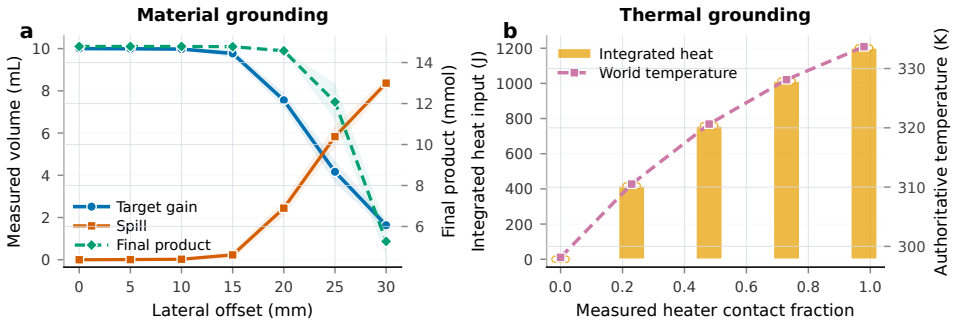


Figure 3: **Physical execution grounds material and energy transfer.** (a) In 35 real-Isaac scripted-pour episodes (five repeats per offset), increasing lateral offset reduces measured target gain and increases spill; the resulting actual target gain, rather than the nominal 10 mL command, determines final product formation in ChemGymRL. Error bars show one standard deviation. From 0 to 30 mm, mean target delivery decreases by 8.37 mL and final product by 9.51 mmol. (b) In 15 real-geometric-contact heating trajectories (three runs per requested contact condition), measured vessel–heater contact determines integrated heat input, which drives the single authoritative world temperature. Temperature is synchronized to the chemistry and MatteriX backends with a maximum discrepancy of 5.03×10^{-8} K. Across the sweep, measured contact changes from 0 to 0.980, integrated heat from 0 to 1201.8 J, and final temperature from 298.15 to 333.68 K.

intra-operation timing. We therefore compare 10 mL pours executed over 1, 2, 5, and 10 s. Although the terminal transferred amount is identical, the continuous chemical trajectories differ. For the 10 s pour, Event coupling incurs trajectory distances of 1.208 in concentration, 0.00797 mol in product, and 0.448 in absorbance relative to per-step flux coupling. Flux coupling matches the per-step physical-flux reference. A separate comparison against an independent RK4 reference is reported in Appendix B; its error metric and reference differ from the distances reported here.

This difference becomes more important when multiple physical processes occur concurrently. In the pour–heat–reaction benchmark, J_m and J_Q evolve within the same interval; collapsing either interaction into a terminal event alters the temporal input seen by the chemistry backend. These results support the central design choice of ChemWorld: the physics-chemistry interface should preserve realized transfer over time rather than only nominal intent or terminal outcome.

4.3 Stability, Robustness, and Backend Portability

We evaluate whether the observed effects depend on a specific synchronization timestep or on a small set of curated executions. We sweep the chemistry timestep over $\{0.025, 0.05, 0.1, 0.2, 0.5, 1.0\}$s. All configurations run successfully, and the preregistered stable region of 0.025–0.2s maintains conservation errors below 2.53×10^{-15}. At the default 0.1s timestep, the final concentration error is 2.44×10^{-4}mol/L for the representative pouring task and 2.73×10^{-4}mol/L for the coupled pour-heat task. No reaction, thermal, or physical parameters are retuned across timesteps.

We further evaluate 500 randomized physical episodes: 100 use real-Isaac-derived traces and 400 use explicitly labeled fast-physics executions. All 500 satisfy the predefined conservation, non-negativity, grounding-consistency, and thermal-synchronization checks, with no NaNs, duplicate mutations, or silent fallbacks. The maximum conservation and grounding errors remain on the order of 10^{-15}. These values quantify the consistency of the synchronization mechanism, rather than physical-model accuracy.

Finally, we test whether the interface is specific to ChemGymRL. The same centered and off-center real-Isaac-derived physical flux traces are replayed through both ChemGymRL (Beeler et al., 2024) and an independent analytic ODE chemistry backend (Virtanen et al., 2020), without modifying the physics, monitor, flux estimator, world integrator, or observation layer. Both backends respond to the same physical perturbations while producing backend-specific chemical trajectories. This demonstrates that ChemWorld's synchronization mechanism is not tied to a single chemistry implementation. Additional grounding consistency and computational overhead are in Appendix A.

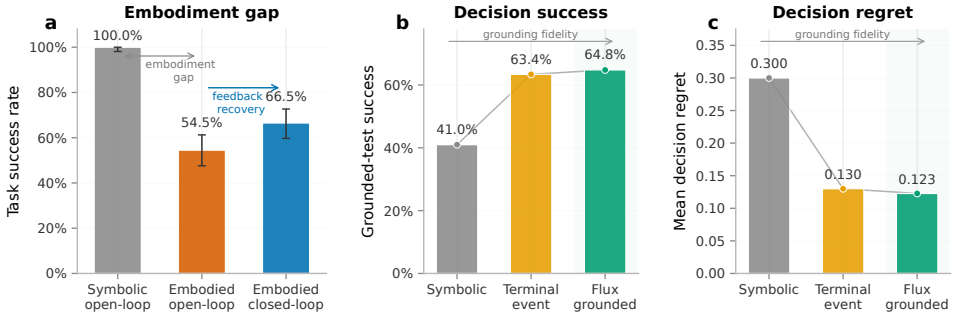


Figure 4: **Grounding changes evaluation and downstream decisions.** (a) Success over 200 paired benchmark contexts; error bars are Wilson 95% intervals. Symbolic, embodied open-loop, and embodied closed-loop success rates are 100.0%, 54.5%, and 66.5%. (b-c) Fixed ridge selectors trained on three label sources are evaluated on the same 100-context grounded test set over five seeds. Grounded labels improve mean success and reduce mean regret relative to symbolic labels.

4.4 Embodiment Changes Agent Evaluation

We next ask whether execution grounding changes the measured capability of autonomous laboratory agents. We evaluate 200 paired physical cases across four tasks under three settings: **Symbolic (S)**, where nominal actions are executed exactly; **Embodied Open-loop (E)**, where the same protocol is subject to physical execution errors; and **Embodied Closed-loop (EO)**, where the agent receives permitted observations and may issue corrective actions. Across symbolic, real-Isaac-derived replay, and validated fast-physics executions, the 200 paired cases produce 1400 evaluation rows. All thresholds are fixed before evaluation.

The effect is substantial. The open-loop protocol agent achieves 100.0% success under symbolic execution but only 54.5% under embodied execution, revealing a 45.5 percentage-point embodiment gap. Allowing grounded observations and corrective actions raises success to 66.5%, corresponding to a 12.0 percentage-point recovery. RGB alone achieves 57.5%, while RGB+temperature reaches 66.5%. Adding UV–Vis does not further improve success in this benchmark because RGB is deterministically projected from the same spectrum. Figure 4(a) summarizes the evaluation gap. Full numerical results and sensor-model robustness variants are provided in Tables S4 and S5.

We also examine whether grounding changes downstream protocol selection. Fixed ridge protocol selectors trained with symbolic, terminal-event, or flux-grounded labels are evaluated on the same 100-context grounded test set over five model seeds. Symbolic labels yield 41.0% success and 0.300 mean regret; terminal-event labels yield 63.4% and 0.130; flux-grounded labels yield 64.8% and 0.123 (Fig. 4b–c). The 500 evaluation rows correspond to five seeds on 100 contexts. These means support the value of grounded supervision; we do not claim a statistically significant difference between terminal-event and flux-grounded labels without an additional paired test.

We further test held-out combinations of execution errors without retuning the agent. Across 200 paired cases combining pouring offset, fast pouring, and poor heating, open-loop success falls to 11.0%, while closed-loop execution reaches 31.5%, yielding a 20.5 percentage-point recovery. A separate liquid-liquid extraction benchmark shows the same qualitative effect: symbolic execution achieves 100.0% success, whereas embodied execution reaches 53.5%. This indicates that the symbolic-embodied gap is not limited to one perturbation or one reaction task. Table S6 summarizes these external-validity evaluations.

To validate replay-based correction against online physical execution, we additionally run 30 Live Isaac closed-loop episodes: 15 pouring-error and 15 heating-error cases. Every corrective action is re-executed through Isaac physics. Live closed-loop success is 50.0%, exactly matching replay closed-loop success, with no replay-success/live-fail cases. The current frozen control budget does not recover live heating failures; therefore, the observed live recovery comes from pouring correction, and we do not claim successful live heat recovery.

4.5 Generality Beyond Liquid-Phase Reactions

Finally, we evaluate whether the same synchronization abstraction extends beyond a single reaction family or a single phase. The liquid-liquid extraction benchmark already provides a distinct process

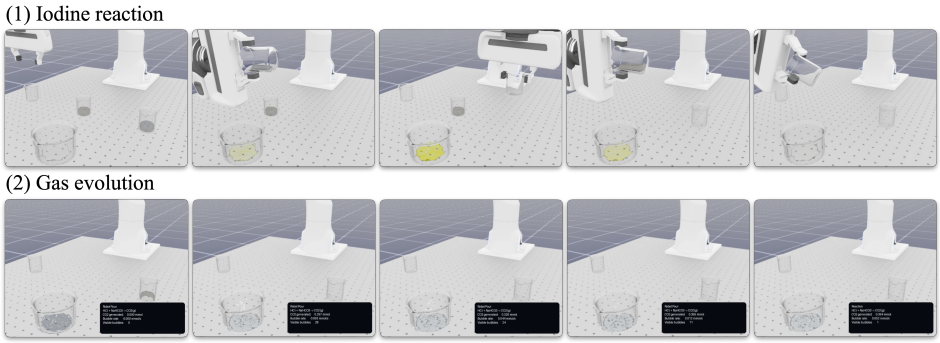


Figure 5: **Qualitative embodied chemistry demonstrations.** Frames from Isaac Sim robot-execution videos show (1) iodine reaction, (2) acid-bicarbonate gas evolution. Visible bubbles are renderer proxies driven by authoritative lumped phase state, not DEM- or CFD-resolved entities.

family in which mixing, settling, and bottom draining alter phase distribution and recovery. We further evaluate a multiphase extension with explicit liquid, solid, and gas inventories. Using the same conservative transfer abstraction, the system supports precipitation, sedimentation, gas evolution, gas-liquid transfer (Sander, 2023), venting, filtration, crystallization, and a coupled precipitation-settling-filtration process.

The numerical multiphase benchmark uses explicitly labeled lumped physical execution rather than Live Isaac multiphase physics. Across these processes, phase transfers reuse the same world-integrator path and preserve phase-wise conservation to numerical precision. Separately, Live Isaac demonstrations bind AgCl formation to progressively visible precipitate and CO_2 generation to bubble-rate visualization after robot-executed reagent transfer. We therefore distinguish between *numerical multiphase validation* of the synchronization abstraction and *Live Isaac visual grounding* of selected solid and gas outcomes. Table S7 reports process-level results, while Fig. 5 shows selected rendered demonstrations. The evidence sources are summarized in Appendix D.

5 Discussion

ChemWorld suggests a different way to construct chemical worlds for embodied agents: instead of treating laboratory operations as symbolic state transitions, it treats their realized physical consequences as the interface to chemistry. This separates physical execution from chemical modeling while preserving a shared causal state, allowing the same synchronization mechanism to be reused across different chemistry backends and extended from liquid reactions to solid- and gas-phase processes. This distinction also changes how autonomous laboratory agents should be evaluated. In our experiments, symbolic execution hides a substantial class of failures that only emerge during physical realization, while grounded observations enable agents to detect and correct part of these failures. ChemWorld therefore provides a bridge between protocol-level reasoning and embodied experimentation, enabling agents to be evaluated not only on whether they choose the right operation, but also on whether they can succeed when that operation must be physically realized.

6 Conclusion

We presented ChemWorld, a physically grounded chemical world in which chemical evolution is driven by the realized consequences of embodied execution rather than by nominal laboratory actions. Its central abstraction is a conservative flux-based interface that connects continuous physical processes with symbolic chemical dynamics through a single shared world state. Across material transfer, heating, concurrent operations, and multiphase processes, this interface preserves causal and conservative synchronization while remaining independent of a particular chemistry backend. Our agent experiments further show that this distinction matters for evaluation: symbolic execution can substantially overestimate performance, whereas grounded observations enable closed-loop recovery from physical execution errors. We hope ChemWorld provides a useful foundation for autonomous laboratories in which reasoning, physical execution, chemical evolution, and observation are evaluated within the same causal loop.

AI USE STATEMENT

In this work, we used generative AI tools for polishing our paper and helping us with coding. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Milad Abolhasani and Eugenia Kumacheva. The rise of self-driving labs in chemical and materials sciences. *Nature Synthesis*, 2:483–492, 2023. doi: 10.1038/s44160-022-00231-0.

Chris Beeler, Sriram Ganapathi Subramanian, Kyle Sprague, Mark Baula, Nouha Chatti, Amanuel Dawit, Xinkai Li, Nicholas Paquin, Mitchell Shahen, Zihan Yang, Colin Bellinger, Mark Crowley, and Isaac Tamblyn. Chemgymrl: a customizable interactive framework for reinforcement learning for digital chemistry. *Digital Discovery*, 3:742, 2024. doi: 10.1039/D3DD00183K.

Torsten Blochwitz, Martin Otter, Johan Åkesson, Martin Arnold, Christoph Clauss, Hilding Elmqvist, Markus Friedrich, Andreas Junghanns, Jakob Mauss, Dietmar Neumerkel, Hans Olsson, and Antoine Viel. Functional mockup interface 2.0: The standard for tool independent exchange of simulation models. In *Proceedings of the 9th International MODELICA Conference*, pp. 173–184, 2012. doi: 10.3384/ecp12076173.

Benjamin Burger, Phillip M. Maffettone, Vladimir V. Gusev, Catherine M. Aitchison, Yang Bai, Xiaoyan Wang, Xiaobo Li, Ben M. Alston, Buyi Li, Rob Clowes, Nicola Rankin, Brandon Harris, Reiner Sebastian Sprick, and Andrew I. Cooper. A mobile robotic chemist. *Nature*, 583(7815): 237–241, 2020. doi: 10.1038/s41586-020-2442-2.

Dario Caramelli, Jaroslaw M. Granda, S. Hessam M. Mehr, Dario Cambié, Alon B. Henson, and Leroy Cronin. Discovering new chemistry with an autonomous robotic platform driven by a reactivity-seeking neural network. *ACS Central Science*, 7(11):1821–1830, 2021. doi: 10.1021/acscentsci.1c00435.

Connor W. Coley, Dale A. Thomas, Justin A. M. Lummiss, Jonathan N. Jaworski, Christopher P. Breen, Victor Schultz, Travis Hart, Joshua S. Fishman, Luke Rogers, Hanyu Gao, Robert W. Hicklin, Pieter P. Plehiers, Joshua Byington, John S. Piotti, William H. Green, A. John Hart, Timothy F. Jamison, and Klavs F. Jensen. A robotic platform for flow synthesis of organic compounds informed by ai planning. *Science*, 365(6453):eaax1566, 2019. doi: 10.1126/science.aax1566.

Kourosh Darvish, Arjun Sohal, Abhijoy Mandal, Hatem Fakhruldeen, Nikola Radulov, Zhengxue Zhou, Satheeshkumar Veeramani, Joshua Choi, Sijie Han, Brayden Zhang, Jeeyeoun Chae, Alex Wright, Yijie Wang, Hossein Darvish, Yuchi Zhao, Gary Tom, Han Hao, Miroslav Bogdanovic, Gabriella Pizzuto, Andrew I. Cooper, Alán Aspuru Guzik, Florian Shkurti, and Animesh Garg. Matterix: toward a digital twin for robotics-assisted chemistry laboratory automation. *Nature Computational Science*, 6(1):67–82, 2026. doi: 10.1038/s43588-025-00924-4.

Cláudio Gomes, Casper Thule, David Broman, Peter Gorm Larsen, and Hans Vangheluwe. Co-simulation: a survey. *ACM Computing Surveys*, 51(3):49, 2018. doi: 10.1145/3179993.

Jarosław M. Granda, Liva Donina, Vincenza Dragone, De-Liang Long, and Leroy Cronin. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature*, 559 (7714):377–381, 2018. doi: 10.1038/s41586-018-0307-8.

Piotr S. Gromski, Jarosław M. Granda, and Leroy Cronin. Universal chemical synthesis and discovery with 'The Chemputer'. *Trends in Chemistry*, 2(1):4–12, 2020. doi: 10.1016/j.trechm.2019.07.004.

Ernst Hairer, Syvert P. Nørsett, and Gerhard Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer, 2nd edition, 1993. doi: 10.1007/978-3-540-78862-1.

Florian Häse, Loïc M. Roch, and Alán Aspuru-Guzik. Next-generation experimentation with self-driving laboratories. *Trends in Chemistry*, 1(3):282–291, 2019. doi: 10.1016/j.trechm.2019.02.007.

Alan C. Hindmarsh, Peter N. Brown, Keith E. Grant, Steven L. Lee, Radu Serban, Dan E. Shumaker, and Carol S. Woodward. Sundials: Suite of nonlinear and differential/algebraic equation solvers. *ACM Transactions on Mathematical Software*, 31(3):363–396, 2005. doi: 10.1145/1089014.1089020.

Nick Jakobi, Phil Husbands, and Inman Harvey. Noise and the reality gap: The use of simulation in evolutionary robotics. In *Advances in Artificial Life*, pp. 704–720, 1995. doi: 10.1007/3-540-59496-5_337.

Rui Li, Zixuan Hu, Wenxi Qu, Jinouwen Zhang, Zhenfei Yin, Sha Zhang, Xuantuo Huang, Hanqing Wang, Tai Wang, Jiangmiao Pang, Wanli Ouyang, Lei Bai, Wangmeng Zuo, Lingyu Duan, and Dongzhan Zhou. Labutopia: High-fidelity simulation and hierarchical benchmark for scientific embodied agents, 2025a. Accepted by NeurIPS 2025 Dataset and Benchmark Track.

Shoujie Li, Yan Huang, Changqing Guo, Tong Wu, Jiawei Zhang, Linrui Zhang, and Wenbo Ding. Chemistry3d: Robotic interaction toolkit for chemistry experiments. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025b. doi: 10.1109/ICRA55743.2025.11128846.

Benjamin P. MacLeod, Fraser G. L. Parlane, Thomas D. Morrissey, Florian Häse, Loïc M. Roch, Kevan E. Dettelbach, Ricardo Moreira, Lars P. E. Yunker, Michael B. Rooney, Jason R. Deeth, Vivian Lai, Grace J. Ng, Henry Situ, Rory H. Zhang, and Curtis P. Berlinguette. Self-driving laboratory for accelerated discovery of thin-film materials. *Science Advances*, 6(20):eaaz8867, 2020. doi: 10.1126/sciadv.aaz8867.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021. doi: 10.48550/arXiv.2108.10470.

Thomas G. Mayerhöfer, Susanne Pahlow, and Jürgen Popp. The bouguer-beer-lambert law: Shining light on the obscure. *ChemPhysChem*, 21(18):2029–2046, 2020. doi: 10.1002/cphc.202000464.

Loïc M. Roch, Florian Häse, Christoph Kreisbeck, Teresa Tamayo-Mendoza, Lars P. E. Yunker, Jason E. Hein, and Alán Aspuru-Guzik. Chemos: An orchestration software to democratize autonomous discovery. *PLOS ONE*, 15(4):e0229862, 2020. doi: 10.1371/journal.pone.0229862.

Rolf Sander. Compilation of henry's law constants (version 5.0.0) for water as solvent. *Atmospheric Chemistry and Physics*, 23(19):10901–12440, 2023. doi: 10.5194/acp-23-10901-2023.

Martin Seifrid, Robert Pollice, Andrés Aguilar-Granda, Zamyla Morgan Chan, Kazuhiro Hotta, Cher Tian Ser, Jenya Vestfrid, Tony C. Wu, and Alán Aspuru-Guzik. Autonomous chemical experiments: Challenges and perspectives on establishing a self-driving lab. *Accounts of Chemical Research*, 55(17):2454–2466, 2022. doi: 10.1021/acs.accounts.2c00220.

Benjamin J. Shields, Jason Stevens, Jun Li, Marvin Parasram, Farhan Damani, Jesus I. Martinez Alvarado, Jacob M. Janey, Ryan P. Adams, and Abigail G. Doyle. Bayesian reaction optimization as a tool for chemical synthesis. *Nature*, 590(7844):89–96, 2021. doi: 10.1038/s41586-021-03213-y.

Sebastian Steiner, Jakob Wolf, Stefan Glatzel, Anna Andreou, Jarosław M. Granda, Graham Keenan, Trevor Hinkley, Gerardo Aragon-Camarasa, Philip J. Kitson, Davide Angelone, and Leroy Cronin. Organic synthesis in a modular robotic system driven by a chemical programming language. *Science*, 363(6423):eaav2211, 2019. doi: 10.1126/science.aav2211.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109.

Gary Tom, Stefan P. Schmid, Sterling G. Baird, Yang Cao, Kourosh Darvish, Han Hao, Stanley Lo, Sergio Pablo-García, Ella M. Rajaonson, Marta Skreta, Naruki Yoshikawa, Samantha Corapi, Gun Deniz Akkoc, Felix Strieth-Kalthoff, Martin Seifrid, and Alán Aspuru-Guzik. Self-driving laboratories for chemistry and materials science. *Chemical Reviews*, 124(16):9633–9732, 2024. doi: 10.1021/acs.chemrev.4c00055.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Zhou Xian, Bo Zhu, Zhenjia Xu, Hsiao-Yu Fish Tung, Antonio Torralba, Katerina Fragkiadaki, and Chuang Gan. Fluidlab: A differentiable environment for benchmarking complex fluid manipulation. *arXiv preprint arXiv:2303.02346*, 2023. doi: 10.48550/arXiv.2303.02346.

Tong Zhao and Yan Zeng. Digital twins for self-driving chemistry laboratories. *Nature Computational Science*, 6:15–16, 2026. doi: 10.1038/s43588-025-00908-4.

Zhenpeng Zhou, Xiaocheng Li, and Richard N. Zare. Optimizing chemical reactions with deep reinforcement learning. *ACS Central Science*, 3(12):1337–1344, 2017. doi: 10.1021/acscentsci.7b00492.

A Grounding Consistency and Computational Overhead

Table S1 reports material and thermal grounding endpoints and bookkeeping errors. These residuals measure numerical consistency, not physical-model accuracy. Table S2 reports runtime measurements; peak memory is for the entire process, not the incremental cost of ChemWorld.

Table S1: Physical execution grounds material and energy updates.

Channel	Physical perturbation	Measured consequence	World consequence	N	Max error
Material	Offset: 0 to 30 mm	Target 10.000 to 1.635 mL; spill 0.000 to 8.365 mL	Product 14.787 to 5.274 mmol	35 episodes	mass 1.80e-15; grounding 1.58e-15
Thermal	Measured contact: 0 to 0.980	Integrated heat 0.0 to 1201.8 J	Temperature 298.15 to 333.68 K	15 trajectories	temperature sync 5.03e-08 K; mass 1.98e-15

Material values are means at the endpoint offsets. Thermal values use the measured geometric contact, not the requested contact. Errors are absolute conservation, physics-to-chemistry grounding, or backend synchronization errors.

Table S2: Computational overhead per 1,000 physics steps.

Vessels	Isaac-only (s)	Monitor/Flux (s)	Full (s)	Full overhead	Full FPS	Chem tick (ms)	Peak MB
1	4.01	4.38	5.17	28.8%	193.6	4.68	3157.9
5	12.13	13.37	15.23	25.6%	65.7	2.23	3161.9
10	22.12	24.14	27.45	24.1%	36.4	1.98	3173.8

Wall-clock values come from the frozen R4 runtime study. Peak memory is process-level and is therefore not an isolated incremental-memory measurement for ChemWorld.

B Independent Temporal Reference Comparison

This additional experiment compares coupling schedules against an independent scalar RK4 reference over 24 material/thermal traces. It is distinct from the comparison to per-step flux in Section 4.2. Figure S1 illustrates a 10 s pour and Table S3 reports product-trajectory mean absolute error at two chemistry step sizes. Fine-window rankings depend on discretization: the 0.2 s window has the lowest error at a 0.10 s chemistry step, while per-tick flux has the lowest error at 0.05 s. The reference is a numerical model, not wet-lab ground truth.

Table S3: Product-trajectory error relative to the independent RK4 reference.

Schedule	Material-update semantics	MAE (mmol) $\Delta t = 0.05$ s	MAE (mmol) $\Delta t = 0.10$ s
Nominal at start	Commanded amount at t=0	2.6765	2.6713
Actual terminal	Actual amount after pour	0.9123	0.8929
1.0 s window	Actual flux batched every 1.0 s	0.1891	0.1684
0.5 s window	Actual flux batched every 0.5 s	0.0850	0.0643
0.2 s window	Actual flux batched every 0.2 s	0.0213	0.0011
Per-tick flux	Actual increment each chemistry tick	0.0104	0.0205

Mean absolute error over the same 24 physical material/thermal traces; lower is better. The unusually low 0.2 s-window error at dt=0.10 s is a grid-alignment error cancellation, so fine-schedule rankings should not be interpreted independently of numerical step size.

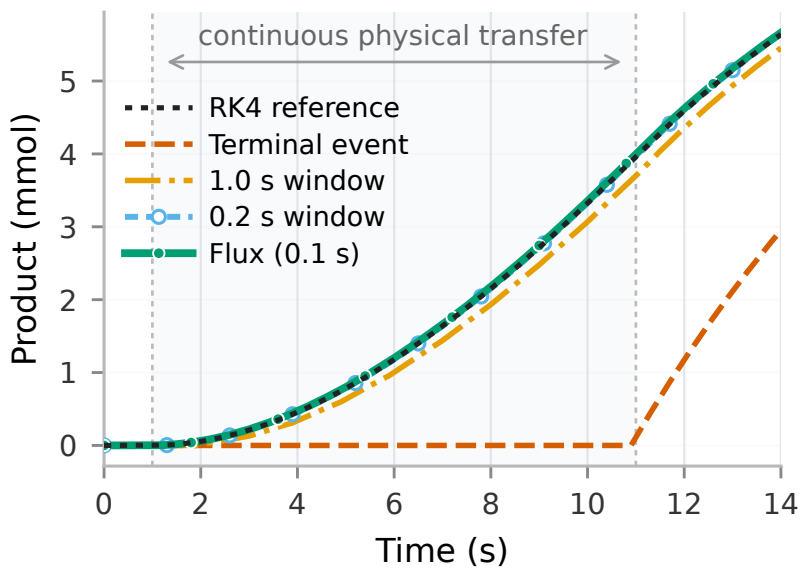


Figure S1: **Temporal coupling against an independent numerical reference.** Representative 10 s slow pour; shading marks the pouring interval. A terminal event delays all transfer until completion, while fine-window and per-tick updates preserve the delivery history. This panel is taken from the supplied temporal-coupling figure; the associated aggregate errors are reported in Table S3.

C Evaluation, Observations, and External Validity

Tables S4, S5, and S6 provide the full results accompanying the main evaluation. Sensor variants use optical-projector RGB rather than camera pixels. Protocol-selection rows reuse the same test contexts across seeds; they are not 500 independent contexts. Extraction is evaluated open-loop because a mature legal closed-loop observation was unavailable.

Table S4: Grounding changes agent evaluation and downstream protocol selection.

Experiment	Setting / supervision	N	Success	95% CI	Mean regret
Agent evaluation	Symbolic / open-loop	200	100.0%	[98.1%, 100.0%]	–
Agent evaluation	Embodied / open-loop	200	54.5%	[47.6%, 61.3%]	–
Agent evaluation	Embodied / closed-loop	200	66.5%	[59.7%, 72.7%]	–
Protocol selection	Symbolic labels	500	41.0%	–	0.300
Protocol selection	Terminal-event labels	500	63.4%	–	0.130
Protocol selection	Flux-grounded labels	500	64.8%	–	0.123

Agent-evaluation intervals are Wilson 95% intervals over 200 paired contexts. Protocol selection reports 500 evaluation rows (five seeds on 100 grounded test contexts); regret is not defined for the agent-evaluation block. Terminal and flux-grounded label results should not be presented as significantly different without an additional paired test.

Table S5: Closed-loop robustness to sensor-model observation choices.

Observation	N	Success	95% CI	Spectrum error	Temp. error (K)	Corrections
Projected float RGB	200	57.5%	[50.6%, 64.1%]	0.0573	2.259	0.94
8-bit RGB	200	57.5%	[50.6%, 64.1%]	0.0573	2.236	3.00
8-bit RGB + noisy T	200	66.0%	[59.2%, 72.2%]	0.0576	0.391	3.42
Noisy spectrum + T	200	66.5%	[59.7%, 72.7%]	0.0576	0.376	1.25
RGB + spectrum + T noise	200	66.0%	[59.2%, 72.2%]	0.0576	0.376	1.24
Full sensor + 3-tick delay	200	67.0%	[60.2%, 73.1%]	0.0576	0.355	1.27

All variants use the same paired physical cases and shared noise seeds. RGB is optical-projector output, not camera pixels. Overlapping intervals do not support claims that the 3-tick delay improves performance.

Table S6: Held-out perturbation and second-process-family evaluation.

Evaluation	Setting A	Result A	Setting B	Result B	Difference	Paired cases
Seen single-factor	Embodied open-loop	58.0%	Embodied closed-loop	69.3%	+11.3 pp	development evaluation
Held-out combinations	Embodied open-loop	11.0%	Embodied closed-loop	31.5%	+20.5 pp	200
Extraction process	Symbolic open-loop	100.0%	Embodied open-loop	53.5%	-46.5 pp	200

Held-out combinations use no scenario labels and no threshold retuning. Extraction is open-loop only because no mature legal extraction observation was available; no privileged phase state is used by an agent.

D PROCESS BREADTH

The evidence levels remain distinct: pouring and geometric heating use live Isaac measurements; concurrent transfer and heating include Isaac-derived execution; extraction uses the ChemGymRL process model; distillation uses a semi-physical phase-transfer model. Table S7 summarizes the supplied conservative process evaluations. These do not constitute CFD, industrial mass-transfer, or quantitative wet-lab validation.

Table S7: Process breadth under the shared flux-integrator interface.

Process	Physical factor	N	Observed outcome range	Max conservation error
Concurrent reaction	Pour quality + heater contact	4	Product 5.017-14.562 mmol; T 313.34-341.33 K	1.80e-15
Liquid-liquid extraction	Mixing intensity + bottom drain	5	Collected target 6.096-8.029 mmol; purity 93.6-100.0%	2.11e-16
Distillation	Heater contact	4	Collected water 0.000-10.812 mmol; T 298.15-380.92 K	6.10e-16

These are conservative lumped-process evaluations, not CFD, industrial mass-transfer, or quantitative wet-lab validation. Distillation uses the documented semi-physical phase-flux model.

E BOUNDED FAILURE REACHABILITY

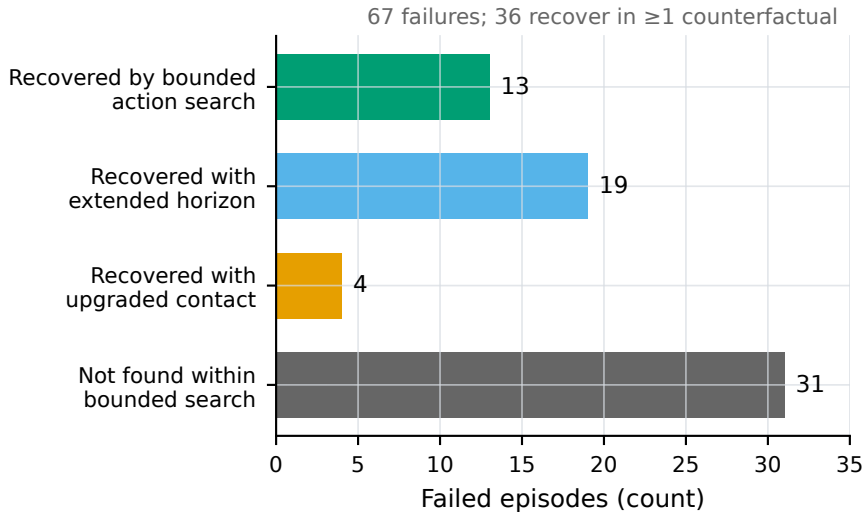


Figure S2: **Bounded counterfactual failure analysis.** Of 67 failures shared by the original closed-loop and privileged-heuristic baselines, 13 are recovered by bounded legal action search, 19 by extending the horizon, and four by improving heater contact; 31 remain unresolved within the search budget. This analysis is not a global oracle.