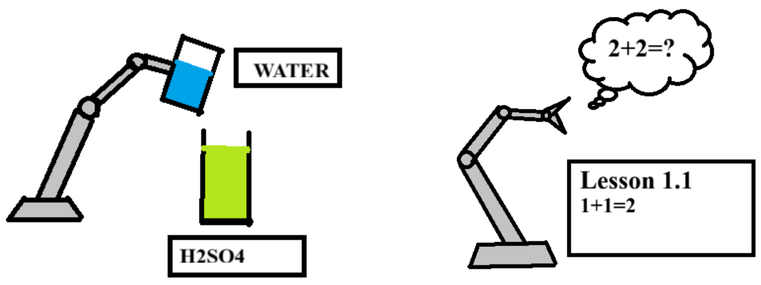
Provably Safe Robotics for Autonomous Science: Real-Time Formal Verification with the Verified Agentic Pipeline (VAP)

Saien Deng, David Scott Lewis, Enrique Zueco
AIXC, Spain

Introduction and motivation

The vision of autonomous “Robot scientists” – embodied agents capable of drafting, executing, and interpreting experiments-promises a paradigm shift in scientific discovery . Despite the development of Autonomous Generalist Scientist in constructing experiments, impendences still persists

1. **The Plausibility Trap (Lack of Grounding):** AI models generates plan that seems coherent but violates fundamental physical laws, posing safety and hallucination issues.
2. **Epistemic Inefficiency (Lack of Generalization)**: present models are limited on generalizing results from different experiment settings. Insights gained during one context were often stored or discarded that



Methodology

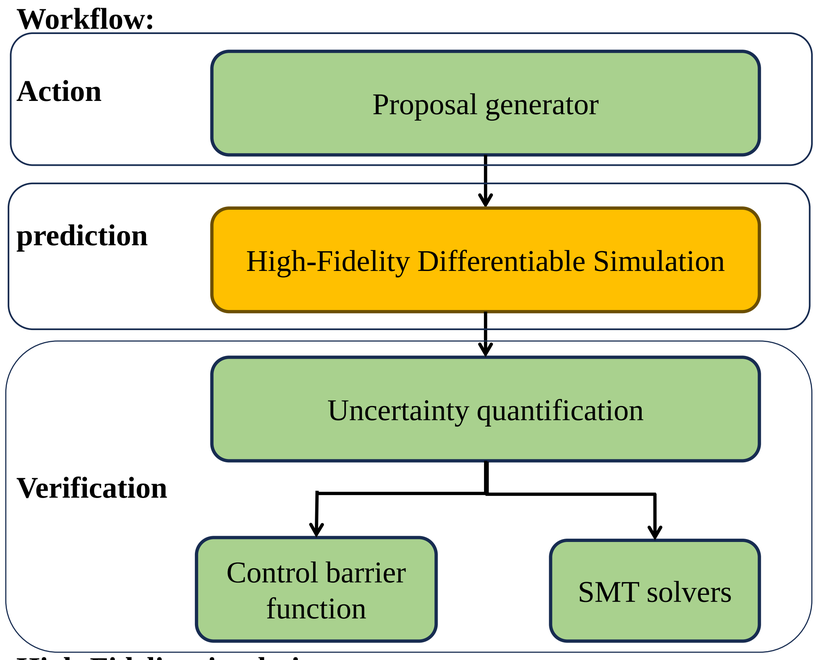
To overcome the plausibility barrier,. We hereby introduce the validated agentic pipeline(VAP), bridging the gap between automated planning with safety and physical validity

Foundation : Physically informed bases
Compared to black box emulation, VAP adopted models with physical grounding by introducing the physical informed machine learning (PIML) framework. This could allow us to establish known physical laws and other constraints by penalizing their violation in the loss function below

$$\mathcal{L}(\theta) = \lambda_{data}\mathcal{L}_{data}(\theta) + \lambda_{physics}\mathcal{L}_{physics}(\theta)$$

Beyond equations, VAP also Integrated structural inductive bias aimed at capturing the nuances of physical systems.

Workflow:



High-Fidelity simulation:
Enable virtual testing range for proposal test, reject unrealistic hypothesis and generate predicted outcome. Differentiable simulators were employed for policy refinement and loss function optimization in non-smooth condition
Uncertainty quantification:
Set worst-case bound from simulation prediction, provides safety barriers for ML’s uncertainty
Control barrier function:
Keep the set of actions in safe operational bounds with mathematical grounding
Refine or override action with the smallest change.
SMT solvers:
Keep complex action set logically obeying lab rules and other regulations.

Test outcomes

We generate two scenarios to assess safety, latency and multi-agent coordination

First scenario: simulated grasp scenario generated under the Hertzian contact model

G safety:
Test setup: 300 grasps, 150 safe (within force limits), 150 unsafe (exceed max force)
Comparison: VAP with verification (Montecarlo sampling + 95 percentile bound)
Reactive baseline (execute and measure force, abort if unsafe)
Neutral baseline (ML classifier trained on 1000 historical graph)
Hypothesis: VAP could guarantee $\geq 95\%$ safety with $\leq 5\%$ false rejection.

G latency:
Setup: 600 grasps, with varying monte Carlo sample size (100, 300, 600)
Time break down: action, prediction, verification
Device: Intel i7, 16GB RAM
Hypothesis: verification time remains below 15ms, making it capable of Integration into real-time control loops

Scenario 2: simulated dual- gripper corporative task (lifting a large beaker)
Comparation: VAP Distributed: Parallel verification with consensus protocol
Centralized Control: Single controller computes joint plan, sequential verification
Reactive Coordination: Force feedback-based adjustment (no pre-execution verification)
Independent Control: Grippers operate independently without coordination
Hypothesis: Distributed VAP verification perform <5% force imbalance (asymmetric loading) and >20% speedup over sequential verification.

Test result summary:

TABLE I
VAP EXPERIMENTAL VALIDATION RESULTS SUMMARY

Experiment	Metric	Result	Target
G-Safety	Safety Rate	100%	≥95%
	False Rejection	0.7%	≤5%
G-Latency	Verification Time	0.32ms	<15ms
G-Synergy	Force Imbalance	0.0%	<5%
	Speedup	75.9%	>20%

conclusions and future work

Our work presented a plausible idea for future integration of embodied agents in the field of scientific research. As presented in the experiment, Safety (100% rejection of unsafe contact) was ensured under high frequency Operating conditions(0.32ms verification time) by our pipeline, demonstrating its potential capabilities in real-time lab environments. Moreover,
The inclusion of LabMemo made knowledge generalization and further interpretability possible by focusing On the abstraction of reusable knowledge, reducing training time at the same time by allowing leaning in tests without the need of retraining.

Future works:
1. Automated constrain discovery: emphasize on developing VAP’s ability on Finding such constraints autonomously through data and experimentations.
2. Hierarchical abstraction: Scaling LabMemo on generalizing convoluted tasks require hierarchical Abstraction mechanisms, in which high level tasks are composed of low-level Skills,
3. Fluent Human-AI collaboration:
Improve agents' ability to explain reasoning, accept human Feed back on constraints (VAP) and knowledge(LabMemo), and collaboratively explore the Knowledge space.