# LANGUAGE-CONDITIONED EXECUTABLE WORLD HYPOTHESES FOR UAVS: CAUSAL EVALUATION AND CERTIFIED RESOLUTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Language-conditioned UAV systems can emit syntactically valid, executable actions whose behavior, referents, or evidential support are nevertheless wrong. Existing evaluations usually isolate interface validity, semantic accuracy, or trajectory error, obscuring whether a model's complete decision is both justified and executable. We formalize the *executable world hypothesis*, a contract-bound claim that couples an intended operation with typed grounding, epistemic support, and deterministic rollout, and define Executable-World-Hypothesis Correctness (EWHC) as its end-to-end criterion. We instantiate this problem in AeroWorld-32K, a benchmark independent of model family with 4,096 executable UAV scenes and 32,768 causally paired queries spanning supported and open world conditions. We further introduce *Certificate-guided Epistemic Resolution for Executable States* (CERES), a deterministic resolver that requires no training and audits one frozen model completion against a typed public evidence representation, certifying a uniquely grounded and supported interpretation or compiling an explicit abstention to the protocol defined safe hold. Across ten frozen language-model configurations and six deterministic comparators, unassisted models remain limited particularly at the open world support boundary, and a deterministic symbolic resolver outperforms every unassisted language model. Giving five representative models the same public information in a typed representation closes a mean 9.44% of their model-wise Direct-to-CERES gaps. CERES yields an average absolute EWHC gain of 35.07% across all model configurations, repairs 114,908 Direct errors, and causes no observed EWHC harm, harm on supported cases, or unsafe determinate output on the complete evaluation population. Preregistered closed loop simulations further show that small ADE/FDE can coexist with an incorrect executable world hypothesis. These results establish evidence grounded resolution as a practical complement to language conditioned action generation under open world uncertainty.

## 1 INTRODUCTION

Autonomous aerial systems increasingly support inspection, logistics, infrastructure monitoring, emergency response, and other operations conducted under changing airspace constraints and mission directives. Across these settings, UAV autonomy must combine structured state information with language that communicates route clearances, temporary restrictions, landing site availability, and operational contingencies. Recent world models conditioned on language world models and embodied agents show that linguistic context can support prediction, planning, and executable action (Driess et al., 2023; Jiang et al., 2023; Ghosh et al., 2024; Kim et al., 2025). Before a UAV acts, however, producing a plausible behavior or a compilable program is insufficient. The system must identify the requested operational transition, bind its arguments to the correct typed entities, and determine whether authenticated public evidence warrants a determinate commitment. Failure at any one of these levels can produce an executable trajectory that nevertheless represents the wrong world hypothesis.

Existing evaluation paradigms expose parts of this problem but not the complete decision chain. Learned world models are commonly assessed through predictive fidelity or downstream control,



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.

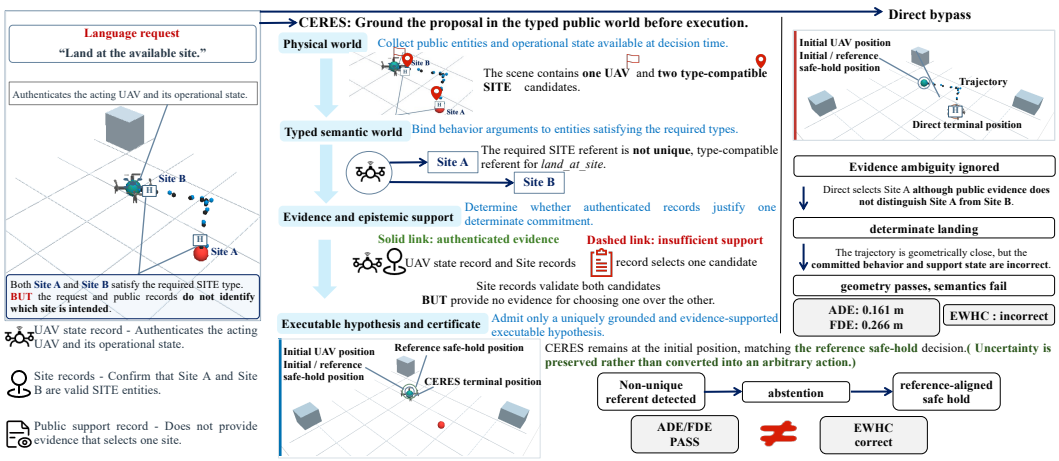


Figure 1: Motivating failure from a shared frozen proposal. Public evidence validates two sites but selects neither. CERES compiles a safe hold aligned with the reference; Direct commits to Site A, achieving low ADE/FDE but incorrect EWHC.

while embodied language systems often emphasize task success, planning, or action generation (Bruce et al., 2024; Hafner et al., 2025). Recent agent benchmarks broaden evaluation through interactive execution and detailed error taxonomies (Li et al., 2024; Xie et al., 2024). Yet output validity, semantic grounding, evidential support, deterministic compilation, and closed loop consequence remain difficult to observe jointly. This omission is consequential for UAV autonomy: unresolved linguistic ambiguity can be converted into a deterministic controller command while the system still receives credit for schema compliance, intervention sensitivity, or trajectory proximity. Once compiled, the hidden semantic or evidential error becomes an operational commitment, and downstream geometry cannot reveal whether the behavior, referent, and support state were justified. Figure 1 isolates this failure. The same frozen model proposal requests a landing at a candidate site, but the public world contains two sites of the required type and no authenticated record selects either one. Direct execution preserves the submitted commitment and lands at Site A. CERES audits the same proposal, preserves the unresolved evidence state, and compiles an auditable abstention to a safe hold aligned with the reference safe hold. Although the Direct trajectory remains geometrically close to the reference, its behavior and support state are incorrect. Reliable evaluation must therefore treat the complete executable world hypothesis, rather than its surface form or trajectory alone, as the unit of correctness.

We formalize the missing evaluation object as an *executable world hypothesis*: a claim governed by a contract over an intended operation, its typed entity and message grounding, and the epistemic support for committing that operation. Its consequence is determined by a frozen compiler, controller library, and simulator through a shared deterministic replay under common conditions. Executable-World-Hypothesis Correctness, abbreviated EWHC, evaluates the full chain by requiring agreement in behavior semantics, grounding, support state, and executable rollout. We instantiate this formulation in AeroWorld-32K, a benchmark independent of model family containing 4,096 executable UAV scenes and 32,768 queries conditioned on language. Each scene contributes paraphrase invariance pairs and interventions relevant to the decision, while explicit unknown conditions test the open world support boundary. AeroWorld-32K is generated from executable worlds to language realizations: canonical hypotheses and trajectories follow from the same frozen world state rather than from annotations added after observing model outputs.

We further introduce *Certificate-guided Epistemic Resolution for Executable States*, hereafter CERES, a deterministic resolver that requires no training and operates between a frozen language model and the shared execution backend. CERES consumes one frozen model completion, the public request, and fixed response contracts. A strict adapter preserves the submitted decision, a typed evidence graph exposes admissible referents and support conditions, and a resolver with two lanes either certifies one uniquely grounded, supported interpretation or emits an explicit abstention that compiles to the safe hold defined by protocol. CERES makes no additional model call, performs no

training or fine tuning, and receives no private label, Oracle state, role in a causal pair, or simulator feedback. Its certificate records why a hypothesis was admitted, canonicalized, or rejected.

We evaluate ten frozen language model configurations and six deterministic comparators on the complete AeroWorld-32K population. A deterministic symbolic resolver outperforms every unassisted language model, highlighting the role of explicit evidence and unknown state management. Across model families, the open world support boundary emerges as the dominant shared bottleneck. A five-model fairness control improves EWHC by 3.43% on average when the same public information is re-expressed as a typed representation, but this closes a mean 9.44% of the model-wise Direct-to-CERES gaps. CERES consistently improves EWHC in paired comparisons over scene clusters across all model configurations, with no observed EWHC harm, harm on supported cases, or unsafe determinate output on the complete evaluation population. Four preregistered PyBullet cases additionally show how EWHC errors can persist despite small trajectory deviations. Our contributions are threefold. (1) We formulate executable world hypotheses and EWHC, and introduce AeroWorld-32K, a causally paired benchmark of 4,096 UAV scenes and 32,768 supported and open world queries. (2) We propose CERES, a deterministic resolver that requires no training and certifies an interpretation grounded in evidence or abstains. (3) Across ten language model configurations and six deterministic comparators, we establish consistent CERES gains and examine how the resolved hypotheses propagate in preregistered closed loop cases.

## 2 Related Work

**Language-conditioned world models and embodied action.** World models support prediction, imagination, and planning; recent systems condition latent or visual futures on language and action (Lin et al., 2024; Bar et al., 2025; Zhu et al., 2025; Yang et al., 2024; Alonso et al., 2024; Wu et al., 2024b; Du et al., 2023). Robot-centric variants connect video prediction to manipulation or policy learning (Wu et al., 2024a; Zhen et al., 2024; Mazzaglia et al., 2024; Hu et al., 2023), while language-to-action systems ground proposals in skill affordances, generate executable programs, or emit robot actions (Ichter et al., 2023; Singh et al., 2023; Liang et al., 2023a; Zitkovich et al., 2023; Brohan et al., 2023; Chi et al., 2023; Ze et al., 2024). Language-mediated reasoning further decomposes, grounds, or revises behavior (Huang et al., 2023c;a;b; Yao et al., 2023; Shinn et al., 2023; Wang et al., 2024; Zeng et al., 2023), supported by large robot corpora (Open X-Embodiment Collaboration, 2024; Walke et al., 2023; Khazatsky et al., 2024). AeroWorld studies a complementary interface: the controller library and simulator are fixed, and a frozen text model must recover a typed, supported, executable world hypothesis from public operational evidence. EWHC therefore tests semantic resolution, compilation, and replay rather than learned visual dynamics or continuous control fidelity (Tang et al., 2024; Vafa et al., 2024).

**Causal, open-world, and executable evaluation.** Behavioral tests, contrast sets, and intervenable simulators examine whether decisions respond correctly to controlled changes (Ribeiro et al., 2020; Gardner et al., 2020; Ahmed et al., 2021). Multidimensional and interactive benchmarks broaden evaluation to agents whose outcomes can be externally checked (Liang et al., 2023b; Liu et al., 2024; Ruan et al., 2024; BIG-Bench Collaboration, 2023; Suzgun et al., 2023; Wang et al., 2023; Mialon et al., 2024; Zhou et al., 2024; Koh et al., 2024; Deng et al., 2023; Jimenez et al., 2024; Ma et al., 2024; Qin et al., 2024; Yao et al., 2025). Selective prediction and uncertainty-aware methods add rejection for risky or unsupported outputs (Geifman & El-Yaniv, 2019; Farquhar et al., 2024; Ren et al., 2023). AeroWorld unifies these concerns in one executable population through invariance pairs, decision-relevant interventions, explicit unknown conditions, and matched rollouts. Its paired metrics distinguish output change from joint correctness, while CERES uses fixed public contracts and evidence checks rather than a learned rejector or confidence threshold to certify one interpretation or emit an auditable safe hold.

## 3 Problem Formulation

Before an instruction-conditioned UAV can act, it must determine three coupled quantities: the requested operational change, the entities and message evidence that ground it, and whether that evidence warrants a determinate action at all. Action-only evaluation cannot distinguish a correctly grounded decision from an unsupported commitment, while intervention sensitivity alone does not

establish correctness. We therefore define an *executable world hypothesis* as a contract-bound claim about operation, grounding, and support whose consequence is obtained through shared frozen compilation and deterministic replay. A base scenario $s$ contains a shared pre-event history $H_s$, an event-time state $x_s$ of a target UAV, a catalog of environment entities $\mathcal{E}_s$, a target identifier $\iota_s$, and an event time $t_s$. The state $x_s = (p_s^N, p_s^E, p_s^U, v_s, \psi_s, \rho_s)$ records north, east, and up position, speed, heading, and remaining energy. The catalog $\mathcal{E}_s$ contains typed operational entities, including routes, destinations, bases, landing sites, restricted zones, and emergency corridors. A query associates the scenario with one authenticated natural-language message $m_{s,k}$:

$$q_{s,k} = P(B_s, m_{s,k}) = \big(x_s, \mathcal{E}_s, \iota_s, t_s, m_{s,k}\big), \qquad B_s = \big(H_s, x_s, \mathcal{E}_s, \iota_s, t_s\big), \tag{1}$$

where $P$ is the frozen public-request projection. The evaluated text model receives the event-time state $x_s$, not the private history $H_s$. The latter is retained only to construct a common pre-intervention world, generate counterfactual branches, and audit provenance. Within a paired comparison, the state, environment, target, event time, and message authority are fixed; only the controlled language condition changes.

**Executable hypotheses and resolution.** Under the frozen contract, the supported behavior set $\mathcal{B}$ comprises `route_follow` for route following, `hold` for holding, `return_to_base` for returning to base, `land_at_site` for landing at a designated site, `reroute_around_zone` for rerouting around a restricted zone, and `emergency_corridor` for using an emergency corridor. Given $q_{s,k}$, a frozen text model $f_\theta$ returns a seven-field structured hypothesis

$$\widehat{z}_{s,k} = f_\theta(q_{s,k}) = \big(\widehat{\sigma}, \widehat{\iota}, \widehat{b}, \widehat{\alpha}, \widehat{u}, \widehat{c}, \widehat{\mathcal{I}}\big). \tag{2}$$

where field indices $(s, k)$ are suppressed for readability. The fields specify the schema, subject, behavior, typed arguments, unknown reason, diagnostic confidence, and supporting message. A supported hypothesis must bind an allowed behavior to contract-valid typed arguments; an abstention must emit `unknown`, no arguments, and one frozen protocol reason. The tuple consequently separates operational semantics, evidence grounding, and epistemic support. Confidence never gates execution or changes a score. Direct submits the frozen model output unchanged, whereas CERES applies a frozen resolver that requires no training to the same public request and completion before both tracks enter the same compiler, controller, simulator, and scorecards. The resolver receives no private label, Oracle state, pair role, or simulator feedback, so the comparison isolates resolution before execution.

**Compilation and executable-world correctness.** The text model does not predict trajectories. The frozen compiler maps a submitted hypothesis to an executable controller or a reason-preserving safe hold; it never repairs schema or semantics. From the same event state, the simulator executes model and canonical programs for 30 seconds under zero and shared-fixed disturbances. A private deterministic Oracle constructs the canonical hypothesis from the complete synthetic world; the evaluated model never observes that world, its labels, or its trace. Compilation status, rollout equations, and Oracle adjudication use the same frozen contract for every evaluated method. Let $V_{s,k}$ indicate that a uniquely extracted payload satisfies the frozen seven-field schema and all public contracts, including exact subject and evidence binding. Full semantic exactness is

$$M_{s,k} = V_{s,k}\, \mathbf{1}\Big[(\widehat{b}, \widehat{\alpha}, \widehat{u}) = (b^*_{s,k}, \alpha^*_{s,k}, u^*_{s,k})\Big]. \tag{3}$$

Thus, subject and evidence errors are already excluded through $V_{s,k}$, while self-reported confidence is absent. Let $\mathrm{ADE}^0_{30}$ and $\mathrm{FDE}^0_{30}$ denote the mean and terminal 3D displacement from the canonical 30-second zero-disturbance rollout. The primary endpoint, Executable-World-Hypothesis Correctness or EWHC, is

$$\mathrm{EWHC}_{s,k} = M_{s,k}\, \mathbf{1}[\mathrm{ADE}^0_{30}(s,k) \le 5\ \mathrm{m}]\, \mathbf{1}[\mathrm{FDE}^0_{30}(s,k) \le 5\ \mathrm{m}]. \tag{4}$$

By construction, a contract-valid hypothesis that matches the canonical semantics compiles to the same deterministic program and rollout as the Oracle. EWHC therefore audits the complete sequence from generation through compilation and rollout; separate diagnostics characterize how non-exact hypotheses propagate.

DATASET SCALE

| Module | Scenes | Queries | Inv. | Int. |
|---|---|---|---|---|
| Reference | 1,536 | 12,288 | 3,072 | 3,072 |
| Information | 1,024 | 8,192 | 2,048 | 2,048 |
| Open world | 1,536 | 12,288 | 3,072 | 3,072 |
| **Total** | **4,096** | **32,768** | **8,192** | **8,192** |

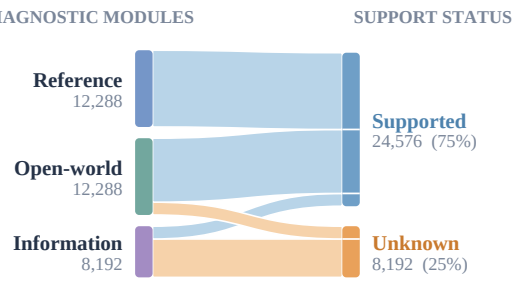


Figure 2: Composition of AeroWorld-32K. Left: module scale and numbers of semantic invariance and intervention pairs, abbreviated Inv. and Int. Module colors match the left nodes in the right panel. Right: flows from modules to support states; ribbon width shows query count and color shows supported versus expected unknown status.

**Paired causal correctness and scope.** Each causal group contains two messages attached to one base scene. An invariance pair changes surface form while preserving canonical semantics and future; an intervention pair changes one decision-relevant condition and must alter both. Consistency or sensitivity compares complete compiled decision signatures, whereas joint correctness requires both members to be semantically exact. An output change therefore does not imply a correct causal response. Exact pair constraints and metrics follow the frozen causal group contract. This formulation asks whether a frozen model recovers a grounded operational hypothesis, abstains when evidence is insufficient, and remains executable under a common backend. Appendix A collects the complete response, compilation, Oracle, extraction, and causal contracts.

## 4 AeroWorld-32K

Section 3 defines the object to be recovered; AeroWorld-32K operationalizes it as a benchmark independent of model family and requires no benchmark specific training. Each example is derived from an executable low-altitude UAV world before its language realization is rendered, so its canonical hypothesis and future trajectory are consequences of the same underlying state rather than annotations added after generation. Deterministic compilation and replay admit only executable instances. The resulting design evaluates pointwise semantic and execution correctness together with stability under meaning-preserving language and responsiveness to decision-relevant interventions. Simulator-backed embodied benchmarks demonstrate the value of explicit task interfaces, controlled worlds, and executable outcomes (Li et al., 2023; Gu et al., 2023; Deitke et al., 2022; Fan et al., 2022; Mees et al., 2022; Szot et al., 2021); AeroWorld uses these principles to isolate language-conditioned resolution under one frozen UAV execution backend.

### 4.1 Executable World Construction

Each base world fixes the event-time UAV state, typed entities, message authority, evidence support, and one of six behaviors across eight environment families. Private executable semantics are assigned before eight authenticated views form two invariance and two intervention pairs under fixed physical state and public context. Every admitted hypothesis compiles and replays for 30 s under both disturbances. Models receive only the public projection in Eq. 1, while labels, pair roles, traces, and executable branches remain private.

### 4.2 Diagnostic Modules and Causal Composition

The benchmark contains 4,096 base scenes and 32,768 queries: 24,576 supported instances and 8,192 expected-unknown instances. Each scene contributes eight queries arranged into four disjoint binary causal groups. Figure 2 reports absolute scale and the support profile of the three diagnostic modules. Reference grounding contains 12,288 supported queries that test whether a behavior binds to the unique typed entity denoted by explicit, relational, attribute, or event anchored language. Information integrity contrasts 6,144 requests missing a required argument or public referent with

2,048 supported controls. Open world support combines 10,240 supported queries with 2,048 categorically unsupported requests and tests whether abstention changes only when language crosses the frozen behavior boundary.

Each of the six supported behaviors and eight environment families contributes 4,096 queries, rendered through 96 module-local template families. Every scene contains two invariance and two intervention pairs, yielding 8,192 pairs of each type. Invariance changes surface realization while preserving canonical semantics and trajectories; intervention changes one decision-relevant condition and separates both. A frozen mapping within each scene adds one semantically incompatible partner with a separable trajectory for each query as a substitution diagnostic.

All 32,768 hypotheses pass semantic replay, and every causal pair satisfies its prescribed invariance or intervention relation. Exact duplicates and cross-family structural collisions are absent. Appendix A details construction and integrity audits.

# 5 CERES

CERES is a deterministic resolver that requires no training and operates between a frozen text model and the common execution backend. It operates on one frozen model completion and either certifies one publicly grounded executable interpretation or emits an explicit safe abstention. Its purpose is not to search for a better action, but to separate a uniquely admissible transition from unsupported, under-specified, or contradictory public evidence. CERES makes no additional model call and receives no private label, Oracle state, causal-pair role, or simulator feedback.

Figure 3 summarizes the governed dataflow. The frozen LLM contributes only the initial, untrusted hypothesis. A strict adapter, public evidence graph, two-lane resolver, and auditable certificate then operate under fixed contracts before the resolved hypothesis enters the shared backend.

## 5.1 Interface

Using a strict, label-free adapter $\mathcal{A}$, CERES first maps the public seven-field decision to an internal typed proposal:

$$p_{s,k} = \mathcal{A}(q_{s,k}, \widehat{z}^{\mathrm{dir}}_{s,k}) \in \mathcal{P} \cup \{\bot\}. \tag{5}$$

Acceptance requires the fixed response schema, valid public identities, and behavior-specific typed arguments. It canonicalizes only representation while preserving behavior, arguments, identities, and unknown cause; rejection maps to $\bot$ and cannot release a determinate action.

CERES next constructs a typed evidence graph $G_{s,k}$ from the public message, event state, and entity catalog. Behavior witnesses, argument types, candidate referents, and binding support expose a unique grounding, ambiguity, missing support, or a type/reference conflict without private propositions or Oracle trace.

## 5.2 Resolution

The first lane performs *contract-preserving canonicalization*. Let $b(p)$ be the proposed high-level behavior and $\mathcal{N}(q,p)$ its canonical typed representation. Lane A is eligible only if

$$\begin{aligned} g_A(q,p) = \mathbf{1}\big[b(p) \in \mathcal{B} \ \wedge\ \mathrm{Unique}(G) \ \wedge\ \mathrm{Bound}(q,p) \\ \wedge\ \mathrm{Canonicalizable}(p) \ \wedge\ \mathrm{Contract}(\mathcal{N}(q,p))\big] = 1. \end{aligned} \tag{6}$$

Here, $\mathrm{Bound}$ verifies subject and evidence identity and $\mathrm{Unique}$ excludes multiple candidates. Lane A may normalize schema fields or replace a unique selector with its typed identifier, but it preserves the proposed behavior, searches no alternative program, and compiles exactly once. A contradiction with public evidence yields safe abstention.

If Lane A cannot certify a supported transition, Lane B performs *causal unknown-state discrimination* with a frozen priority rule. A request outside $\mathcal{B}$ is unsupported behavior; a required referent that conflicts with the graph is an invalid reference; all remaining non-admissible cases are insufficient information. Lane B emits an unknown hypothesis that compiles to the existing safe hold. It never substitutes another supported behavior or promotes an unknown proposal to a determinate action.

**Certificate.** Both lanes emit a certificate $\kappa_{s,k}$ with the selected lane, verdict, proposed and final behavior, grounding and evidence status, unknown cause, and contradictions. CERES returns $\widehat{z}^{\mathrm{cer}}_{s,k}$ and $\kappa_{s,k}$ to the same compiler and rollout as Direct, so the methods differ only upstream of execution. Full contracts appear in Appendix B.

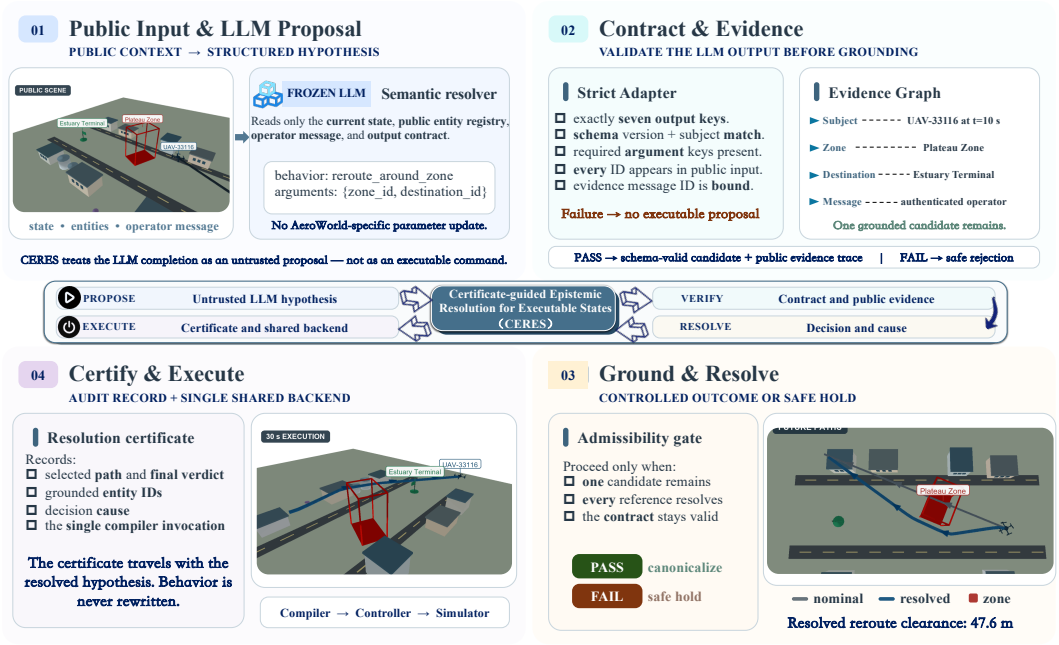


Figure 3: Overview of CERES. A frozen LLM provides an untrusted structured hypothesis. Using only fixed contracts and public evidence, CERES either preserves a certified supported behavior or emits an explicit safe hold, records an auditable certificate, and invokes the shared execution backend. The LLM receives neither Oracle state nor simulator feedback.

# 6 Experiment

## 6.1 Experimental Setup

**Models and baselines.** We evaluate ten frozen semantic resolvers: Llama-3.2-1B/3B, Llama-3.1-8B, Qwen3-4B/8B/14B, Phi-4-mini/14B, and Gemma-3-4B/12B, without AeroWorld training or parameter updates. Six deterministic comparators are Random-valid, Always-Unknown, State-only, Rule-parser, Keyword+Nearest, and a deterministic symbolic resolver. The latter is a stateless per-request method that uses fixed lexical, type, and public-geometry rules. Every method receives the same public request; Oracle state is used only for implementation self-checks.

**Protocol.** All 32,768 requests share one execution, disturbance, and scoring protocol. Direct, strict-adapter, and CERES reuse each frozen completion without another model call or response rewrite. A separate five-model fairness control instead makes one fresh frozen inference call after deterministically re-expressing the same public entities, behavior signatures, spatial indices, and authenticated records in typed form; it excludes private labels, Oracle state, CERES outputs, pair roles, and trajectories. EWHC is primary; diagnostics cover support, trajectory, interface, and paired causality, with uncertainty from paired resampling of 4,096 scene clusters. Appendices C and D provide the complete protocol and diagnostics.

## 6.2 Benchmark Results

Table 1 reports the complete Direct and deterministic baselines. Interface diagnostics are reported separately. Bold and underline denote the best and second best comparable performance values; ties receive the same mark. The support gap is diagnostic and is therefore not ranked.

The deterministic symbolic resolver leads Direct EWHC at 81.67%, exceeding every unassisted LLM, and is the only nontrivial method whose unknown EWHC exceeds its supported score. Phi-4 exceeds Qwen3-14B by 3.14% overall despite nearly identical supported accuracy; the difference is instead concentrated on unknown requests. Across noncollapsed LLMs, the positive S−U gaps range from 27.98% to 78.58%, exposing evidence support rather than executable formatting as the shared bottleneck. The nonmonotonic Qwen results and the differing Llama and Gemma progressions further rule out a parameter-only account of performance.

Table 1: Direct AeroWorld-32K results (%). Gap is supported EWHC minus unknown EWHC; Int. joint is intervention pair joint correctness.

| Class | Family | Model | EWHC | | | | Causal | Exec. |
|---|---|---|---|---|---|---|---|---|
| | | | *All* | *Supp.* | *Unk.* | *Gap* | *Int. joint* | *Exact* |
| Control | N/A | Always-Unknown | 18.75 | 0.00 | 75.00 | -75.00 | 0.00 | 37.50 |
| | | Random-valid | 6.90 | 7.93 | 3.80 | 4.13 | 0.38 | 15.15 |
| Non-LLM | N/A | State-only | 6.26 | 8.34 | 0.00 | 8.34 | 0.00 | 6.26 |
| | | Rule-parser | 35.99 | 47.99 | 0.00 | 47.99 | 14.15 | 47.24 |
| | | Keyword+Nearest | 63.84 | **85.12** | 0.00 | 85.12 | 40.19 | 64.03 |
| | | Symbolic resolver | **81.67** | 80.47 | **85.27** | -4.80 | **67.50** | **85.50** |
| LLM | Llama | Llama-3.2-1B | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 37.50 |
| | | Llama-3.2-3B | 21.42 | 28.41 | 0.43 | 27.98 | 2.93 | 43.75 |
| | | Llama-3.1-8B | 51.71 | 68.75 | 0.60 | 68.15 | 21.77 | 59.75 |
| | Qwen | Qwen3-4B | 58.72 | 74.86 | 10.29 | 64.57 | 26.98 | 63.17 |
| | | Qwen3-8B | 55.12 | 72.36 | 3.42 | 68.94 | 26.51 | 60.72 |
| | | Qwen3-14B | 62.43 | 82.08 | 3.50 | 78.58 | 29.85 | 69.06 |
| | Phi | Phi-4-mini | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 37.50 |
| | | Phi-4 | 65.57 | 82.06 | 16.11 | 65.95 | 37.28 | 71.83 |
| | Gemma | Gemma-3-4B | 45.33 | 60.40 | 0.12 | 60.28 | 13.54 | 51.91 |
| | | Gemma-3-12B | 56.33 | 74.80 | 0.90 | 73.90 | 26.61 | 61.03 |

**Causal and execution diagnostics.** Int.-J. requires both intervention members to be correct. The deterministic symbolic resolver reaches 67.50%, versus 40.19% for the next method and at most 37.28% among LLMs. Qwen3-14B and Phi-4 nevertheless change on 99.65% and 99.43% of pairs, showing that sensitivity alone overstates causal competence. Appendix D.1 reports continuous ADE/FDE and invariance diagnostics.

### 6.3 CERES Result

For all ten configurations, each frozen completion enters the Direct, strict-adapter, and CERES tracks, isolating resolution while interface scores remain fixed. In Table 2, D→C and Ctrl.→C are absolute EWHC gains; Cause is CERES exact unknown-cause classification.

Table 2: Complete AeroWorld-32K CERES and strict-control results (%).

| Family | Model | EWHC | | | | | Epist. | Exact trajectory match | |
|---|---|---|---|---|---|---|---|---|---|
| | | *Direct* | *Ctrl.* | *CERES* | *D→C* | *Ctrl.→C* | *Cause* | *Direct* | *CERES* |
| Llama | Llama-3.2-1B | 0.00 | 18.75 | 25.00 | +25.00 | +6.25 | **100.00** | 37.50 | 37.50 |
| | Llama-3.2-3B | 21.42 | 48.28 | 55.78 | +34.36 | **+7.50** | 94.24 | 43.75 | 65.31 |
| | Llama-3.1-8B | 51.71 | 85.22 | 89.04 | +37.33 | +3.82 | 93.73 | 59.75 | 90.62 |
| Qwen | Qwen3-4B | 58.72 | 90.80 | 93.96 | +35.24 | +3.17 | 96.34 | 63.17 | 94.89 |
| | Qwen3-8B | 55.12 | 90.26 | 94.88 | +39.76 | +4.63 | 99.83 | 60.72 | 94.93 |
| | Qwen3-14B | 62.43 | 93.75 | 98.38 | +35.95 | +4.64 | 99.66 | 69.06 | 98.47 |
| Phi | Phi-4-mini | 0.00 | 18.75 | 25.00 | +25.00 | +6.25 | **100.00** | 37.50 | 37.50 |
| | Phi-4 | **65.57** | 94.35 | 97.21 | +31.64 | +2.86 | 95.13 | **71.83** | 98.45 |
| Gemma | Gemma-3-4B | 45.33 | 86.76 | 88.89 | **+43.55** | +2.13 | 83.75 | 51.91 | 93.60 |
| | Gemma-3-12B | 56.33 | **96.76** | **99.16** | +42.83 | +2.40 | 98.65 | 61.03 | **99.50** |

CERES improves every model, with EWHC gains from 25.00% to 43.55%, and all paired effects over scene clusters are positive; mean and median gains are 35.07% and 35.60%. Phi-4 leads Direct at 65.57%, whereas Gemma-3-12B reaches the highest resolved score at 99.16%, showing that Direct ranking does not determine performance after evidence grounded resolution. The audit records

114,908 repairs, no EWHC or supported-case harm, and no unsafe determinate output. Gains beyond the strict adapter range from 2.13% to 7.50%, while cause exactness ranges from 83.75% to 100.00%; both results isolate epistemic resolution. In a separate five-model fairness control, typed re-expression raises mean EWHC from 58.23% to 61.67%, whereas CERES reaches 95.74%. All intervals comparing Typed with Direct exclude zero, but typed access closes a mean 9.44% of the model-wise gap and CERES remains 34.07% higher, so representation access explains little of the deterministic resolution gain.

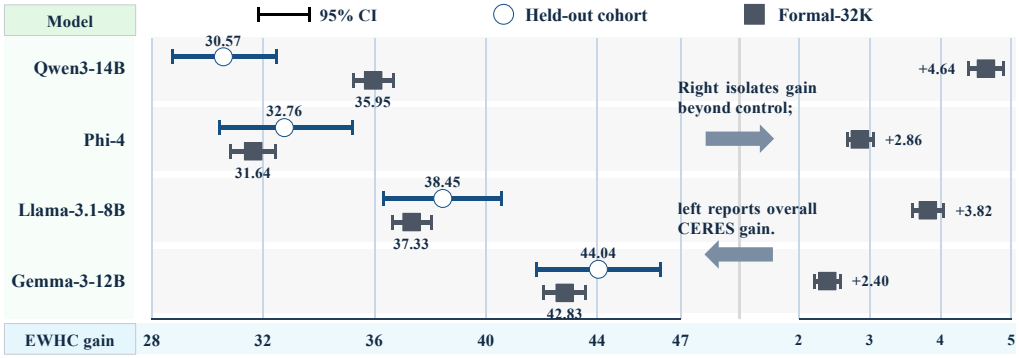


Figure 4: Cross-family CERES gains over Direct and the strict adapter; intervals are 95% paired confidence intervals.

Figure 4 separates confirmation from population evidence: gains remain positive across four families in both cohorts, while its right panel isolates gains from 2.40% to 4.64% beyond adapter normalization. CERES and the strict control have identical trajectories and release coverage, but all 14,299 incremental repairs correct causal unknown states. CERES releases 70.26% of supported requests with 100% EWHC and withholds 29.74%; it abstains on every unknown request, emits no unsafe determinate output, and identifies the exact cause in 96.13%. Sweeping ADE/FDE thresholds from 0 to 20 m changes no EWHC result, so the 5 m check does not drive the gain.

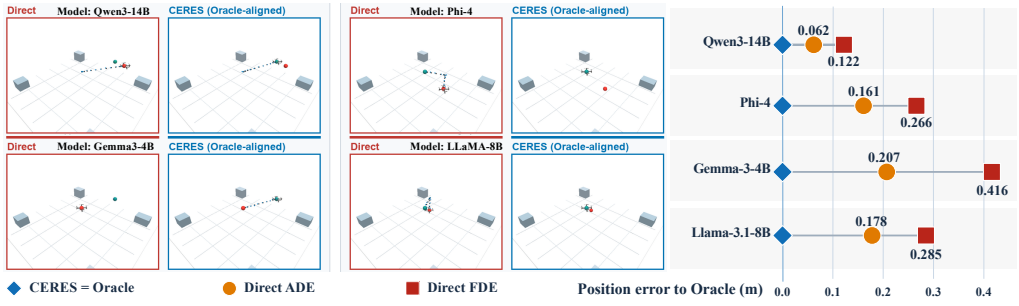


Figure 5: Four preregistered closed-loop cases. CERES matches the Oracle; Direct remains geometrically close while retaining an incorrect hypothesis.

Across four preregistered PyBullet cases, CERES matches the Oracle, whereas Direct remains EWHC-incorrect despite ADE from 0.062 to 0.207 m and FDE from 0.122 to 0.416 m, confirming that geometric proximity is insufficient. Under each frozen disturbance mode, all 155,000 CERES safe-hold rollouts satisfy the 30-second terminal contract.

## 7 Conclusion

Across ten models on AeroWorld-32K, CERES improves EWHC by an average 35.07% without training or another model call. Typed access closes only 9.44% of the model-wise gap, confirming the value of evidence grounded resolution for language conditioned UAV decisions.

## AI USE DISCLOSURE

Generative AI tools were used to provide feedback on benchmark and experimental design, assist with software implementation and debugging, and support language editing and scientific figure design. They were also used to stress-test candidate linguistic expressions and model interfaces. Generative AI did not generate or approve the formal semantic labels, Oracle decisions, or evaluation outcomes of AeroWorld-32K. It did not modify frozen model responses, determine benchmark membership, selectively remove samples, or decide which experimental outcomes were reported. The numerical results for CERES and all evaluated methods were produced by frozen programs and immutable experimental artifacts. AI-assisted code, text, and figures were reviewed by the authors, while experimental results were verified through complete query coverage, run manifests, checksums, deterministic replay, and prespecified statistical procedures. The authors take full responsibility for the final manuscript, its scientific claims, code, and research artifacts.

## REFERENCES

Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Yoshua Bengio, Bernhard Schölkopf, Stefan Bauer, and Manuel Wüthrich. CausalWorld: A robotic manipulation benchmark for causal structure and transfer learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=SK7A5pdrgov.

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in Atari. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1873. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6bdde0373d53d4a501249547084bed43-Abstract-Conference.html.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15791–15801, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Bar_Navigation_World_Models_CVPR_2025_paper.html.

BIG-Bench Collaboration. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=uyTL5Bvosj.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. RT-1: Robotics transformer for real-world control at scale. In *Proceedings of Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.025. URL https://roboticsproceedings.org/rss19/p025.html.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C. Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando De Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4603–4623, 2024. URL https://proceedings.mlr.press/v235/bruce24a.html.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin C. M. Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.026. URL https://roboticsproceedings.org/rss19/p026.html.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, Roozbeh Mottaghi, and Ali Farhadi. ProcTHOR: Large-scale embodied AI using procedural generation. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0433. URL

https://proceedings.neurips.cc/paper_files/paper/2022/hash/27c546ab1e4f1d7d638e6a8dfbad9a07-Abstract-Conference.html.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2Web: Towards a generalist agent for the web. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1220. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/5950bf290a1570ea401bf98882128160-Abstract-Datasets_and_Benchmarks.html.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8469–8488, 2023. URL https://proceedings.mlr.press/v202/driess23a.html.

Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0403. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1d5b9233ad716a43be5c0d3023cb82d0-Abstract-Conference.html.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. MineDojo: Building open-ended embodied agents with internet-scale knowledge. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1333. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/74a67268c5cc5910f64938cac4526a90-Abstract-Datasets_and_Benchmarks.html.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://www.nature.com/articles/s41586-024-07421-0.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020. doi: 10.18653/v1/2020.findings-emnlp.117.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Quan Vuong, Ted Xiao, Pannag R. Sanketi, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.090. URL https://www.roboticsproceedings.org/rss20/p090.html.

Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiaing Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yuan Yao, Xiaodi Yuan, Pengwei Xie, Zhiao Huang, Rui Chen, and Hao Su. ManiSkill2: A unified benchmark for generalizable manipulation skills. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=b_CQDy9vrD1.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. GAIA-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. URL https://arxiv.org/abs/2309.17080.

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. VoxPoser: Composable 3d value maps for robotic manipulation with language models. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 540–562, 2023a. URL https://proceedings.mlr.press/v229/huang23b.html.

Wenlong Huang, Fei Xia, Dhruv Shah, Danny Driess, Andy Zeng, Yao Lu, Pete Florence, Igor Mordatch, Sergey Levine, Karol Hausman, and Brian Ichter. Grounded decoding: Guiding text generation with grounded models for embodied agents. In *Advances in Neural Information Processing Systems*, volume 36, 2023b. doi: 10.52202/075280-2606. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/bb3cfcb0284642a973dd631ec9184f2f-Abstract-Conference.html.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Tomas Jackson, Noah Brown, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner monologue: Embodied reasoning through planning with language models. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 1769–1782, 2023c. URL https://proceedings.mlr.press/v205/huang23c.html.

Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 287–318, 2023. URL https://proceedings.mlr.press/v205/ichter23a.html.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. VIMA: Robot manipulation with multimodal prompts. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 14975–15022, 2023. URL https://proceedings.mlr.press/v202/jiang23b.html.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, et al. DROID: A large-scale in-the-wild robot manipulation dataset. In *Proceedings of Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.120. URL https://www.roboticsproceedings.org/rss20/p120.html.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713, 2025. URL https://proceedings.mlr.press/v270/kim25c.html.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. VisualWebArena: Evaluating multimodal agents on realistic visual web tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 881–905, 2024. doi: 10.18653/v1/2024.acl-long.50. URL https://aclanthology.org/2024.acl-long.50/.

Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, et al. BEHAVIOR-1K: A benchmark for embodied AI with 1,000 everyday activities and realistic simulation. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 80–93, 2023. URL https://proceedings.mlr.press/v205/li23a.html.

Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Li Erran Li, Ruohan Zhang, Weiyu Liu, Percy Liang, Li Fei-Fei, Jiayuan Mao, and Jiajun Wu. Embodied agent interface: Benchmarking LLMs for embodied decision making. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3188. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b631da756d1573c24c9ba9c702fde5a9-Abstract-Datasets_and_Benchmarks_Track.html.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation*, pp. 9493–9500, 2023a. doi: 10.1109/ICRA48891.2023.10160591.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023b. URL https://mlanthology.org/tmlr/2023/liang2023tmlr-holistic/.

Jessy Lin, Yuqing Du, Olivia Watkins, Danijar Hafner, Pieter Abbeel, Dan Klein, and Anca Dragan. Learning to model the world with language. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 29992–30017, 2024. URL https://proceedings.mlr.press/v235/lin24g.html.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zAdUB0aCTQ.

Chang Ma, Junlei Yang, Zhihao Gao, Jiaheng Liu, Quanyan Xie, Chenjun Yu, Jiawei Jin, Zhenzhong Li, Lei Wang, Yang Hu, Sen Shu, Jin Zhang, Xuelong Zhang, Yali Li, Siming Tian, Shaoting Tang, Mengke Zhong, Zhen Zhang, Huan Lin, Tao Zhang, Wenjun Liang, Weiming Liu, and Kaizhu Su. AgentBoard: An analytical evaluation board of multi-turn LLM agents. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2365. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/877b40688e330a0e2a3fc24084208dfa-Abstract-Datasets_and_Benchmarks_Track.html.

Pietro Mazzaglia, Tim Verbelen, Bart Dhoedt, Aaron Courville, and Sai Rajeswar. GenRL: Multimodal-foundation world models for generalization in embodied agents. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0865. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3076133f08b40607d00a8f48f6acd71c-Abstract-Conference.html.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. CALVIN: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022. doi: 10.1109/LRA.2022.3180108. URL https://doi.org/10.1109/LRA.2022.3180108.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: A benchmark for general AI assistants. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=fibxvahvs3.

Open X-Embodiment Collaboration. Open X-embodiment: Robotic learning datasets and RT-X models. In *2024 IEEE International Conference on Robotics and Automation*, pp. 6892–6903, 2024. doi: 10.1109/ICRA57147.2024.10611477. URL https://ieeexplore.ieee.org/document/10611477.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=dHng2O0Jjr.

Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 661–682, 2023. URL https://proceedings.mlr.press/v229/ren23a.html.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. Identifying the risks of LM agents with an LM-emulated sandbox. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=GEcwtMk1uA.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. ProgPrompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation*, pp. 11523–11530, 2023. doi: 10.1109/ICRA48891.2023.10161317.

Mirac Suzgun, Nathan Scales, Nathanael Sch"arli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, 2023. doi: 10.18653/v1/2023.findings-acl.824. URL https://aclanthology.org/2023.findings-acl.824/.

Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/021bbc7ee20b71134d53e20206bd6feb-Abstract.html.

Hao Tang, Darren Key, and Kevin Ellis. WorldCoder, a model-based LLM agent: Building world models by writing code and interacting with the environment. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2243. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/820c61a0cd419163ccbd2c33b268816e-Abstract-Conference.html.

Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0846. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2f6a6317bada76b26a4f61bb70a7db59-Abstract-Conference.html.

Homer Rich Walke, Kevin Black, Tony Z. Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, Abraham Lee, Kuan Fang, Chelsea Finn, and Sergey Levine. BridgeData V2: A dataset for robot learning at scale. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine*

*Learning Research*, pp. 1723–1736, 2023. URL https://proceedings.mlr.press/v229/walke23a.html.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. DecodingTrust: A comprehensive assessment of trustworthiness in GPT models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1361. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/63cb9921eecf51bfad27a99b2c53dd6d-Abstract-Datasets_and_Benchmarks.html.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. VOYAGER: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024. URL https://mlanthology.org/tmlr/2024/wang2024tmlr-voyager/.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *International Conference on Learning Representations*, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/2c37c5bcef24b9541550261dcd63261b-Abstract-Conference.html.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. iVideoGPT: Interactive VideoGPTs are scalable world models. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. doi: 10.52202/079017-2173. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7dbb5bfab324e3b86af9bd0df15498dd-Abstract-Conference.html.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1650. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5d413e48f84dc61244b6be550f1cd8f5-Abstract-Datasets_and_Benchmarks_Track.html.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/c4d66eae503694424123b93ac0fbaf17-Abstract-Conference.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. $\tau$-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=roNSXZpUDN.

Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In *Proceedings of Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.067. URL https://www.roboticsproceedings.org/rss20/p067.html.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael S. Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=G2Q2Mh3avow.

Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3D-VLA: A 3d vision-language-action generative world model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61229–61245, 2024. URL https://proceedings.mlr.press/v235/zhen24a.html.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abhishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oKn9c6ytLx.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. IRASim: A fine-grained world model for robot manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9834–9844, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhu_IRASim_A_Fine-Grained_World_Model_for_Robot_Manipulation_ICCV_2025_paper.html.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183, 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.