EMBODIEDCAUSALESCAPE: CAUSAL REASONING UNDER EMBODIED INTERACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal reasoning in embodied environments requires agents not only to recover relations from observation, but sometimes to discover unknown action–effect mechanisms through intervention. In practice, however, the evidence needed for such discovery may itself be distributed across viewpoints or revealed only through interaction, making it difficult to separate failures of mechanism discovery from failures of embodied evidence acquisition. We introduce EMBODIEDCAUSALESCAPE, a controlled 3D escape-room benchmark for studying causal mechanism discovery under varying embodied evidence conditions. It contains Association tasks, where task-relevant relations are encoded in observable evidence, and Intervention tasks, where an initially unknown action–effect mechanism must be identified through action. Matched diagnostic conditions preserve each underlying logical instance while varying mechanism information, spatial evidence access, and irrelevant interactions. Across four vision-language agents, Main-condition success ranges from 1/30 to 9/30, versus 25/30 for the human reference. On GPT-5.5, Intervention success rises from 4/15 to 14/15 when family-level mechanism structure is supplied, while spatial compaction raises overall success from 9/30 to 27/30. These results show that embodied task success depends not only on underlying task logic, but also on how mechanism-relevant evidence must be acquired and integrated.

1 INTRODUCTION

Reliable embodied behavior in unfamiliar environments requires more than perceiving objects and executing familiar actions: an agent must often infer the environment's task-relevant dynamics before it can act effectively. A central piece of this understanding is a predictive account of how actions change state—a *world model* used for prediction, planning, and control (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020; 2025; LeCun, 2022). We use the term in a task-level, operational sense: the action-conditioned transition structure an agent needs for the task at hand, not a separately trained generative simulator. The question we ask is whether a pretrained multimodal agent can recover such structure on the fly, within a single unfamiliar episode, and immediately put it to use for goal-directed behavior.

This differs from how world models are usually built. Canonical approaches fit a parametric dynamics model from accumulated experience and use it for planning or behavior learning (Hafner et al., 2019; 2020; 2025). Here we study a complementary, within-episode problem: an agent encounters a previously unseen local mechanism and must infer enough of its structure from a handful of interactions to act successfully, without any retraining. Operating an unfamiliar heating system illustrates the demand: reaching a target temperature may require determining which controls affect temperature or mode by acting on them, observing the consequences, and revising a hypothesis about how they work. Humans can acquire novel causal structure through intervention (Gopnik & Sobel, 2000; Gopnik et al., 2004; Edmonds et al., 2018), making on-the-fly world-model discovery a natural, if underexamined, demand on embodied intelligence.

Embodiment introduces an additional challenge: the evidence for building this task-relevant environmental structure does not arrive for free. Some task-relevant relations can already be recovered from observation alone, whereas unknown action-effect structure can only be identified through intervention. And in either case the required evidence must itself be acquired: relevant observations may be distributed across viewpoints, hidden behind ordinary interactions, or mixed in with irrelevant

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.

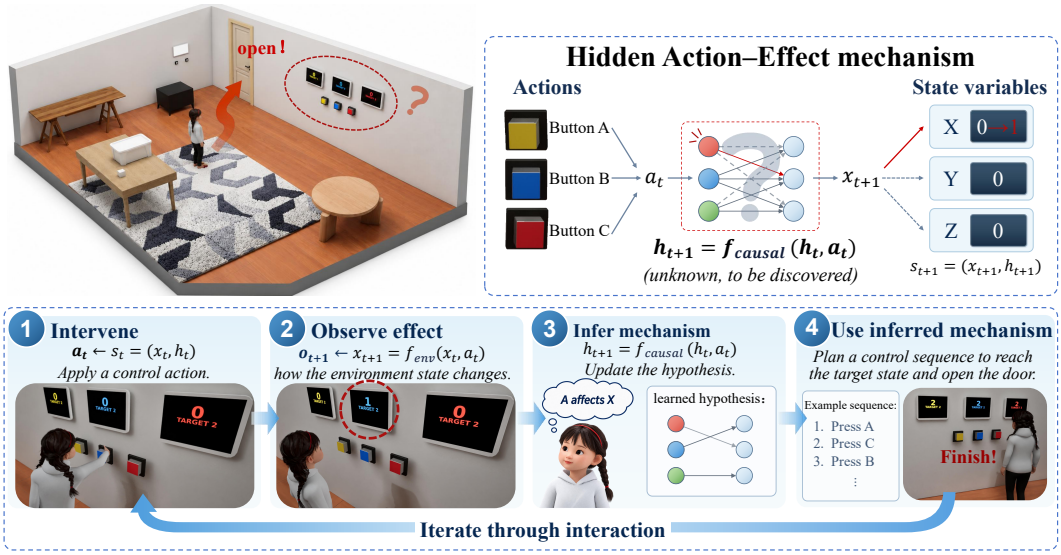


Figure 1: **Embodied causal mechanism discovery through interaction.** An agent must discover an initially unknown action-effect mechanism and use it to open the door. Through repeated interactions, it intervenes on controls, observes state changes, updates its mechanism hypothesis, and uses the inferred mechanism to reach the target state. The diagram illustrates task requirements rather than a prescribed agent pipeline; evaluated agents receive first-person simulator observations.

affordances, and identifying a hidden mechanism may require linking action outcomes across time. Interaction therefore plays two roles at once—executing behavior and generating the evidence needed to understand how the environment works (Fig. 1).

Existing embodied benchmarks emphasize navigation, grounded instruction following, manipulation, planning, and collaboration (Anderson et al., 2018; Shridhar et al., 2020; Yang et al., 2025; Chang et al., 2025), while causal-reasoning benchmarks typically supply the evidence to be reasoned over rather than requiring it to be gathered through interaction (Jin et al., 2023; Komanduri et al., 2025). Interactive causal learning has been studied more directly in OpenLock and ADAM (Edmonds et al., 2018; Yu & Lu, 2025). Yet end-to-end success or failure alone does not reveal whether an agent's failures come from the underlying task structure or from the embodied process of obtaining and using evidence about it.

This motivates our central question: *for a fixed underlying task, how strongly does embodied success depend on the way task-relevant evidence must be acquired and on whether the action–effect mechanism itself must be discovered?* To study this question, we introduce EMBODIEDCAUSALESCAPE, a controlled 3D escape-room benchmark for causal reasoning and mechanism discovery under embodied interaction (Fig. 2). The benchmark contains six task families across two information regimes: *Association* tasks require recovering relations already encoded in environmental evidence, whereas *Intervention* tasks require discovering initially unknown action-effect structure. Each logical instance is separated from its visual and spatial realization, enabling matched conditions that vary mechanism information, spatial evidence access, and irrelevant interactions while preserving the same logical problem.

Across four vision-language agents, Main-task success remains low, while matched GPT-5.5 diagnostics substantially recover performance when mechanism information is supplied or evidence access is compacted. Because the underlying logical instances are fixed, these recoveries show that difficulty depends strongly on how task-relevant structure must be acquired and used, rather than on task logic alone.

Our contributions are:

- **A benchmark for on-the-fly acquisition and use of task-relevant environmental structure and world models,** spanning observational relational inference and intervention-based causal mechanism discovery under first-person interaction.

Table 1: **Representative benchmarks for interactive and causal reasoning.** We compare the canonical task setting rather than auxiliary ablations or diagnostic experiments.

Benchmark	Env.	Interaction	Evidence	Mechanism	Goal
CausalVLBench (Komanduri et al., 2025)	Image	None	Supplied	Given	Predict
OpenLock (Edmonds et al., 2018)	2D	Sequential	Feedback	Hidden	Unlock
CausalWorld (Ahmed et al., 2021)	Robot3D	Control	Feedback	Latent	Control
Alchemy (Wang et al., 2021)	3D	Sequential	Feedback	Latent	Control
IVRE (Xu et al., 2023)	Visual	Experiment	Feedback	Hidden	Infer
DiscoveryWorld (Jansen et al., 2024)	2D	Experiment	Mixed	Hidden	Discover
EscapeBench (Qian et al., 2025)	Text	Sequential	Explore	None	Escape
MM-Escape (Wang et al., 2025)	FP3D	Embodied	Explore	None	Escape
VisEscape (Lim et al., 2025)	Views	Explore	Explore	None	Escape
Ours	**FP3D**	**Embodied**	**Mixed**	**Mixed**	**Escape**

Evidence: **Supplied** = provided with the input; **Explore** = acquired through environmental exploration; **Feedback** = acquired primarily from action outcomes; **Mixed** = combines environmental observation and action feedback. *Mechanism:* **Given** = relevant mechanism information is supplied; **Latent** = the task contains latent causal or dynamic structure; **Hidden** = action-effect structure must be identified through interaction; **Mixed** = Association tasks require no hidden action-effect mechanism, whereas Intervention tasks require action-effect structure to be identified; **None** = no explicit hidden action-effect mechanism is required.

- **A matched diagnostic design** that varies mechanism information, spatial evidence access, and irrelevant interactions while preserving root task logic.
- **Empirical evidence of condition-dependent difficulty,** showing low end-to-end success in the full setting and substantial increases in success on the same logical instances when embodied evidence-acquisition or mechanism-discovery demands are reduced.

2 Related Work

Embodied and interactive reasoning. Embodied benchmarks have expanded from navigation and instruction following (Anderson et al., 2018; Shridhar et al., 2020) to multimodal planning, manipulation, and collaboration (Yang et al., 2025; Chang et al., 2025). Escape-room benchmarks further couple exploration with long-horizon reasoning: EscapeBench emphasizes creative adaptation and tool use (Qian et al., 2025), MM-Escape evaluates multimodal reasoning in interactive 3D rooms (Wang et al., 2025), and VisEscape studies exploration-driven decision making (Lim et al., 2025). These settings motivate interactive evaluation, but our benchmark makes the *source of task-relevant relational and mechanistic information* an explicit design variable and holds root task logic fixed across matched evidence conditions.

Causal reasoning and mechanism discovery. Language and vision benchmarks such as CLadder, Corr2Cause, and CausalVLBench evaluate causal reasoning from supplied evidence (Jin et al., 2023; 2024; Komanduri et al., 2025). Interactive causal learning instead requires generating informative evidence: OpenLock studies causal transfer through sequential interaction (Edmonds et al., 2018), CausalWorld exposes controllable causal variables in robotic manipulation (Ahmed et al., 2021), and DiscoveryWorld evaluates sequential experimentation (Jansen et al., 2024); ADAM learns and uses causal knowledge in Minecraft (Yu & Lu, 2025). Related work on epistemic action and embodied question answering similarly emphasizes acting to obtain information (Kirsh & Maglio, 1994; Das et al., 2018; Gordon et al., 2018). We complement these directions with first-person tasks that distinguish acquiring existing relational evidence from discovering an unknown action-effect mechanism, together with matched variants that change evidence access without changing the underlying logical instance.

3 Preliminaries: Causal Reasoning under Interaction

We briefly review the association-intervention distinction for readers less familiar with causal inference before giving the operational definition used in our benchmark. Observational evidence concerns relations such as $P(Y \mid X = x)$, whereas interventional evidence concerns the consequences of

deliberately setting a variable, written $P(Y \mid \mathrm{do}(X = x))$ (Pearl, 2009). In the heating-system example from the Introduction, observing that a display value co-varies with a control differs from changing that control and using the resulting state change to infer what it does.

For embodied tasks, however, this distinction is not equivalent to "observing" versus "acting": many actions only reveal existing evidence. We call an action *mechanism-informative* when its observed outcome is needed to identify a previously unknown task-relevant action-effect relation. This yields two information regimes. In **Association** tasks, the relevant relation is already encoded in environmental evidence and must be acquired and used for control. In **Intervention** tasks, the action-effect mechanism is initially unknown. Writing h_t for hidden puzzle state and a_t for an action,

$$h_{t+1} = f_{\text{causal}}(h_t, a_t), \tag{1}$$

and the agent must infer the task-relevant structure of f_{causal} from interventions and their observed consequences.

Operational criterion for mechanism discovery. We classify a task as requiring mechanism discovery only when (i) the mechanism is ambiguous before intervention, (ii) action-outcome feedback can distinguish competing mechanisms, and (iii) the recovered mechanism is required for downstream control. These criteria exclude trial-and-error search in which action outcomes need not identify an underlying mechanism. For I1–I3, each instance makes the mechanism initially ambiguous, intervention feedback diagnostic, and subsequent control mechanism-dependent (see supplementary material).

Finally, the information regime specifies *what* must be inferred, whereas the evidence condition specifies *how* supporting evidence becomes available. We vary the latter through matched spatial, distractor, and mechanism-information conditions while preserving the same logical instance (Sec. 4.2).

4 Benchmark Design

4.1 Task Families and Logical Instances

EmbodiedCausalEscape evaluates causal and relational reasoning through interaction in unfamiliar first-person 3D environments. Each episode provides only the high-level objective `Open the door`; no symbolic problem description, family-specific instruction, or solution procedure is given.

We instantiate six task families, three per information regime. **Association** comprises **A1 Compositional relations**, which combines multiple visual relations; **A2 Object correspondence**, which recovers and maintains object-target bindings; and **A3 Selective matching**, which reveals candidate information and selects the subset relevant to the goal. **Intervention** comprises **I1 Factorized transitions**, where controls independently affect state variables; **I2 Coupled transitions**, where coupled controls jointly affect multiple state variables; and **I3 Hierarchical transitions**, where order-dependent effects and lower-level resets must be identified. These families represent distinct relational or mechanistic structures rather than calibrated levels of difficulty. In the original hidden-mechanism setting, mechanism-informative actions probe control—variable assignments in I1, coupled effects in I2, and order-dependent effects or resets in I3.

Each family contains five deterministic logical variants, yielding 30 logical instances. Variants change family-specific quantities—such as symbol or color mappings, object correspondences, candidate values, initial and target states, or hidden control assignments—while preserving the defining family-level structure. Thus, the five variants instantiate the same reasoning demand with different concrete logic rather than defining new task types.

This construction separates four levels that are otherwise easy to conflate. A *family* specifies the relational or transition structure to be reasoned about; a *logical instance* fixes the concrete parameters of that structure; an *embodied realization* determines how the instance is presented through objects, viewpoints, and spatial layout; and a *diagnostic condition* changes selected aspects of that realization or the supplied mechanism information while preserving the root instance. The matched analyses below rely on this separation.

The logical instances are realized in TongSIM (Sun et al., 2025) using a shared physical vocabulary of displays, controls, movable objects, target boards, containers, and doors. Low-level interaction primitives are shared across families, while visual content, object roles, and spatial arrangement vary.

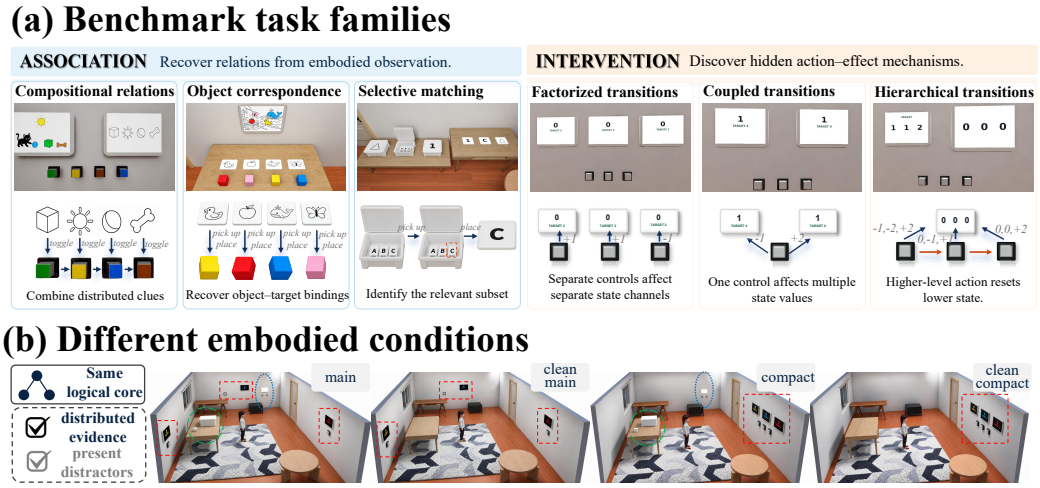


Figure 2: **Benchmark task families and matched diagnostic conditions. (a)** Six task families span Association and Intervention, each with five logical variants preserving its relational or transition structure. **(b)** Main and Compact vary spatial evidence access; Clean-Main and Clean-Compact remove irrelevant interactions from those layouts. For Intervention tasks, Given-Distributed and Given-Compact additionally supply family-level mechanism structure. All conditions preserve the root logical instance. Scenes are illustrative schematics, not simulator screenshots.

Task-relevant evidence may be distributed across viewpoints or revealed only through interaction, and canonical scenes contain task-irrelevant interactive objects so that affordance alone does not identify relevance. Each realization is validated for logical solvability, accessibility of required objects and controls, visual readability, and the absence of known shortcut solutions. Full task definitions and instance parameters are provided in the supplementary material.

4.2 Matched Diagnostic Conditions

Matched conditions preserve the same logical instance while changing how its evidence is made available. For all 30 instances, **Main** uses the canonical distributed scene with task-irrelevant interactions; **Clean-Main** removes those irrelevant interactions; **Compact** spatially concentrates task-relevant evidence and interaction objects while retaining distractors; and **Clean-Compact** combines compaction with distractor removal. These conditions alter embodied evidence-access demands without changing task logic.

For the 15 Intervention instances, we additionally manipulate whether the family-level mechanism structure is hidden or supplied, producing four matched conditions: **Hidden-Distributed** (the Main realization), **Hidden-Compact** (the Compact realization), **Given-Distributed**, and **Given-Compact**. Given conditions provide only the family-level causal structure rather than a concrete execution plan, with exact disclosures and remaining unknowns documented in the supplementary material. Clean-Main and Clean-Compact form a separate distractor-removal comparison over all 30 instances.

5 Experimental Setup

5.1 Models and Interaction Protocol

We evaluate GPT-5.5 (OpenAI, 2026), Claude Opus 4.7 (Anthropic, 2026), Gemini 3 Pro (Google DeepMind, 2025), and Qwen3.8-Max (Alibaba Cloud, 2026) on the same 30 Main instances under a shared interaction protocol.

At each decision step, an agent receives a concatenated image with first-person visual observation on the left and observed interactive object identifiers on the right, lightweight environment state, and accumulated interaction history, and outputs one structured action. The shared action space supports navigation and viewpoint changes, control activation, opening and closing objects, pickup and placement, and code entry. For object-directed actions, the agent specifies an exposed identifier and the executor handles the corresponding low-level simulator operations. These identifiers provide



Figure 3: **Agent-environment interface and interaction feedback.** At each step, the VLM maps first-person observations, object identifiers, lightweight state, and interaction history to one structured action. Subsequent observations and protocol feedback report the action outcome; hidden transition parameters and evaluator state are not exposed in the Main experiments.

stable references to observed objects but do not encode their puzzle roles or values. The model must therefore ground visual evidence to objects and maintain relevant relations or state changes across the interaction history; hidden transition parameters and evaluator state are not exposed in the Main experiments.

Each episode permits at most 70 model decisions and three final-door submissions. Invalid actions consume one decision but leave the environment unchanged. Figure 3 summarizes the interaction loop; full prompts, observation fields, action schemas, and feedback semantics are provided in the supplementary material.

5.2 Evaluation Protocol

The primary metric is *Escape Success*: an episode succeeds only when the required task state is achieved and the final door is physically opened. For the interaction-efficiency analysis, we additionally report *Decision Count*.

Human participants provide a reference for task solvability and interface difficulty on the Main experiments. The human evaluation contains 30 valid primary trials covering the same logical instances and uses the same limits of 70 interaction steps and three final-door submissions. Human performance is treated as a reference rather than an assumed upper bound or a direct measure of model-level causal reasoning ability. Detailed participant and trial-selection information is provided in the appendix.

6 Results and Analysis

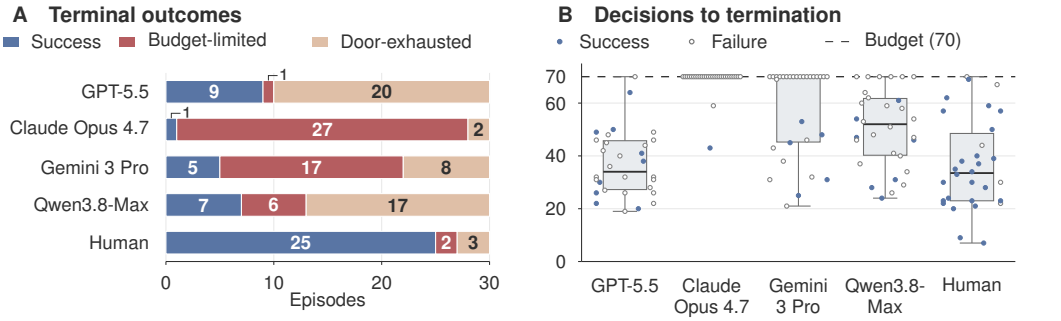
We first evaluate all four agents on Main, then use GPT-5.5 for matched diagnostics of mechanism information, evidence access, and irrelevant interactions.

6.1 Main Benchmark Performance

Table 2 reports instance-level escape success on the same 30 Main instances for all systems. End-to-end performance remains low across the four VLM agents: GPT-5.5 solves 9/30 instances, Gemini 3 Pro solves 5/30, Qwen3.8-Max solves 7/30, and Claude Opus 4.7 solves 1/30, compared with 25/30 for the human reference.

Table 2: **Main benchmark success at the instance level.** Within each family, the five symbols correspond to variants v1–v5 from left to right (●: success; ○: failure). Overall reports aggregate escape success.

	Association			Intervention			
System	**A1 Comp.**	**A2 Corr.**	**A3 Match**	**I1 Fact.**	**I2 Coupled**	**I3 Hier.**	**Overall**
GPT-5.5	● ○ ○ ● ○	● ○ ● ○ ○	○ ○ ● ○ ○	● ○ ○ ○ ○	○ ● ○ ○ ○	○ ● ● ○ ○	**9/30**
Claude Opus 4.7	○ ○ ○ ○ ○	○ ○ ○ ○ ○	○ ○ ○ ○ ○	● ○ ○ ○ ○	○ ○ ○ ○ ○	○ ○ ○ ○ ○	1/30
Gemini 3 Pro	○ ○ ○ ● ○	○ ○ ○ ○ ●	○ ○ ○ ○ ○	○ ○ ● ○ ○	○ ○ ○ ○ ○	● ● ○ ○ ○	5/30
Qwen3.8-Max	○ ○ ○ ○ ○	○ ○ ● ○ ○	● ○ ○ ○ ○	● ● ○ ● ○	● ○ ○ ○ ○	○ ● ○ ○ ○	7/30
Human	● ● ● ● ●	● ● ● ● ○	○ ● ● ● ●	● ● ● ○ ●	○ ● ○ ● ●	● ● ● ● ●	**25/30**



C Decision composition (%)

System	Navigation	Task interaction	Distractor interaction	Door attempts	Invalid / rejected
GPT-5.5	40.7	21.9	12.3	7.3	17.8
Claude Opus 4.7	64.3	11.3	7.3	1.3	15.9
Gemini 3 Pro	68.5	12.1	4.5	2.6	12.3
Qwen3.8-Max	49.3	14.1	7.9	4.5	24.2
Human	44.6	30.2	6.5	4.4	14.3

Percentages rounded to one decimal place; rows may not sum to 100%.

Figure 4: **Behavior on the Main benchmark. A**: terminal outcomes over 30 episodes/trials. **B**: decisions to termination (dashed: 70-decision budget; filled/open: success/failure). **C**: tabulated action-category shares.

The limited successes are broadly distributed across the benchmark. GPT-5.5 succeeds in all 6 task families, Qwen3.8-Max in 5, and Gemini 3 Pro in 4, while Claude Opus 4.7's single success occurs in I1. At the instance level, 14/30 tasks are unsolved by all four VLMs, 12 are solved by exactly one, and only four are solved by more than one; the human reference solves 13 of the 14 instances missed by every VLM. Together, these results show that low Main performance is neither confined to a single task family nor explained by a small shared set of uniformly difficult instances: different models succeed on substantially different subsets of the benchmark.

Failure endpoints are likewise heterogeneous (Fig. 4A). Among the 98 failed VLM Main episodes, 51 terminate after exhausting an interaction budget and 47 after exhausting the allowed final-door submissions. Claude Opus 4.7 and Gemini 3 Pro are predominantly budget-limited, whereas GPT-5.5 and Qwen3.8-Max more often exhaust their door submissions. Thus, similar aggregate success rates can mask different interaction-level failure profiles; these terminal categories describe where execution stops rather than the underlying cause of failure.

The behavioral panels show that success rates do not fully characterize Main behavior. Trajectory lengths differ across systems (Fig. 4B): Claude and Gemini frequently run to the 70-decision budget, whereas GPT-5.5 and Qwen3.8-Max more often terminate earlier or through repeated door attempts. The composition of decisions also differs (Fig. 4C). Gemini and Claude devote a larger share to navigation, GPT-5.5 has a larger share of task and distractor interactions, and Qwen3.8-Max has the highest share of invalid or rejected decisions. These are behavioral patterns in the recorded trajectories, not claims about the agents' internal reasoning.

6.2 Task Solvability under Reduced Embodied Demands

We evaluate GPT-5.5 on the same 30 root logical instances with reduced evidence-acquisition or filtering demands (Fig. 2), preserving family, variant, target, and task logic. Success rises from 9/30 in Main to 14/30 in Clean-Main, 27/30 in Compact, and 29/30 in Clean-Compact (Fig. 5A). Every one of the 21 Main failures is solved in at least one reduced-demand condition, and 20/21 are solved in Clean-Compact. Recoveries span all six task families (Fig. 5B).

These matched recoveries show that low Main success does not establish that the underlying logical problems are beyond the agent's capability. Solvability also depends on the embodied process through which task-relevant evidence is acquired and used.

The recovery is accompanied by lower behavioral friction rather than only more terminal successes (Fig. 5C). Pooling decisions within each condition, invalid or rejected actions fall from 17.8% to 11.0% of all recorded actions between Main and Clean-Compact, while premature door attempts fall from 6.5% to 1.8%. These summaries describe recorded interaction quality and do not imply that any one reduction is the sole cause of recovery.

6.3 Hidden Mechanism Structure and Embodied Evidence Access

To examine mechanism discovery and evidence access jointly, we compare the same 15 Intervention instances under Hidden/Given $\times$ Distributed/Compact (Fig. 2), varying family-level mechanism information and spatial evidence access while preserving task logic.

The resulting pattern is shown in Fig. 5D. GPT-5.5 solves 4/15 instances in Hidden-Distributed, 14/15 in Given-Distributed, and all 15/15 in both Hidden-Compact and Given-Compact. Among the 11 Hidden-Distributed failures, all 11 are recovered in Hidden-Compact and 10 are also recovered in Given-Distributed. Thus, the low performance in the full condition is not driven by a fixed subset of logical instances that remain difficult across matched realizations.

Providing mechanism structure raises success under Distributed access but not once evidence is Compact; conversely, compaction has a large effect when structure is Hidden but little additional effect when it is Given. The pattern does not support independent rankings of these demands: the same tasks are substantially harder when hidden mechanism structure and distributed evidence must be handled together.

6.4 Interaction Cost with Hidden Mechanism Structure

Keeping mechanism structure hidden also increases interaction cost in the Compact setting, where all 15 Intervention instances are completed under both conditions. Holding the logical instance and layout fixed, the median trajectory requires 29 model decisions in Hidden-Compact and 10 in Given-Compact (Fig. 5E); the median paired difference, defined as Hidden-Compact minus Given-Compact, is +17 decisions, and 13/15 matched instances require fewer decisions when mechanism is provided.

Equal success rates conceal differences in decision composition (Fig. 5F). Hidden-Compact has the same navigation share as Given-Compact but larger distractor and invalid/rejected shares, a smaller task-interaction share, and a comparable door-attempt share.

With layout and logical instance fixed, this comparison isolates the effect of withholding family-level mechanism structure. Given-Compact still requires instance-specific control grounding, effect identification, state tracking, planning, and execution; the paired difference measures interaction efficiency under the two information conditions, not internal reasoning quality.

7 Discussion and Limitations

The matched analyses show that embodied task difficulty depends on how reasoning is coupled with evidence acquisition and control, not on logical structure alone. Intervention success approaches ceiling when either mechanism structure is supplied or evidence access is compacted, yet withholding structure still increases interaction cost in the Compact setting.

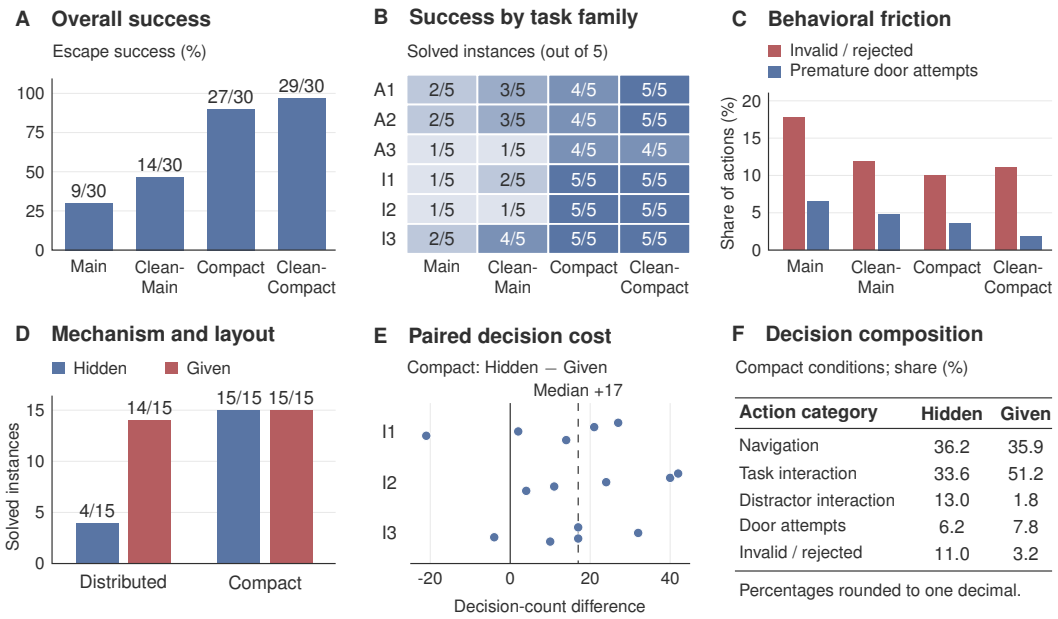


Figure 5: **Matched GPT-5.5 diagnostics. A–C**: success and behavioral friction across Main, Clean-Main, Compact, and Clean-Compact. **D**: Intervention success under Hidden/Given × Distributed/Compact. **E**: paired Hidden-Compact − Given-Compact decision counts. **F**: tabulated decision composition in Hidden-Compact and Given-Compact.

An end-to-end score from a single realization can confound the reasoning problem with where evidence is placed, which interactions reveal it, and which irrelevant affordances compete for attention. Matched-instance diagnostics make this dependence observable by testing whether the same logical problem changes solvability as its evidence conditions change.

At the same time, the matched conditions are not a decomposition into independent cognitive capacities. Spatial compaction simultaneously changes navigation, viewpoint switching, object re-identification, and cross-location evidence integration; distractor removal changes both available affordances and relevance filtering. Likewise, Given provides family-level mechanism structure but still requires instance-specific effect identification, visual grounding, state tracking, planning, and execution. Performance differences should therefore be interpreted as controlled changes to bundles of embodied demands, not as direct measurements of memory, attention, navigation, or causal reasoning in isolation. More targeted diagnostics could further separate these factors while preserving the interactive structure that makes the problem embodied.

Finally, EMBODIEDCAUSALESCAPE intentionally studies a controlled subset of embodied causal reasoning. The task structures are discrete and deterministic, the rooms share a simulator asset vocabulary, and the object-centric action interface abstracts away continuous motor control. More realistic settings may contain stochastic or changing mechanisms, noisy and partial observations, delayed effects, and richer physical interaction. We view EMBODIEDCAUSALESCAPE as a controlled testbed for studying these interactions rather than a complete decomposition of embodied intelligence.

8 CONCLUSION

We introduced EMBODIEDCAUSALESCAPE, a controlled benchmark for studying causal reasoning under embodied interaction. Current vision-language agents show low end-to-end success in the full setting, while matched diagnostics reveal that the same logical tasks often become solvable when mechanism ambiguity or embodied evidence-acquisition demands are reduced. The condition-dependent recoveries and interaction costs show why causal reasoning in embodied environments should be evaluated together with the process through which its evidence is acquired. More broadly, the benchmark provides a controlled way to study when an agent can discover how an unfamiliar environment works and turn that knowledge into goal-directed behavior.

Reproducibility Statement

The final artifact will include the frozen TongSIM task configurations, source-generated logical instance manifests, model-facing prompts, exact agent configurations, run manifests, evaluator code, and derived trajectory metrics. Logical instance, layout, appearance, and run identifiers will be versioned separately so that paired diagnostics can be reproduced without conflating new layouts with new logical tasks.

AI Use Statement

Generative AI tools assisted with literature retrieval, manuscript drafting, and editing. The authors remain responsible for verifying all references, benchmark implementations, empirical measurements, analyses, and final claims.

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 *ICLR*, 2021.

Alibaba Cloud. Qwen3.8-Max: Model information, 2026. URL https://www.alibabacloud.com/help/en/model-studio/qwen3-8-max.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, pp. 3674–3683, 2018.

Anthropic. Introducing Claude Opus 4.7, 2026. URL https://www.anthropic.com/news/claude-opus-4-7.

Matthew Chang, Gunjan Chhablani, Alexander Clegg, Mikael Dallaire Cote, Ruta Desai, Michal Hlavac, Vladimir Karashchuk, Jacob Krantz, Roozbeh Mottaghi, Priyam Parashar, Siddharth Patki, Ishita Prasad, Xavier Puig, Akshara Rai, Ram Ramrakhya, Daniel Tran, Joanne Truong, John M. Turner, Eric Undersander, and Tsung-Yen Yang. PARTNR: A benchmark for planning and reasoning in embodied multi-agent tasks. In *ICLR*, 2025.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *CVPR*, pp. 1–10, 2018.

Mark Edmonds, James Kubricht, Colin Summers, Yixin Zhu, Brandon Rothrock, Song-Chun Zhu, and Hongjing Lu. Human causal transfer: Challenges for deep reinforcement learning. In *CogSci*, 2018.

Google DeepMind. Gemini 3 Pro model card, 2025. URL https://deepmind.google/models/model-cards/gemini-3-pro/.

Alison Gopnik and David M. Sobel. Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development*, 71(5):1205–1222, 2000.

Alison Gopnik, Clark Glymour, David M. Sobel, Laura E. Schulz, Tamar Kushnir, and David Danks. A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111 (1):3–32, 2004.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. IQA: Visual question answering in interactive environments. In *CVPR*, pp. 4089–4098, 2018.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *ICLR*, 2020.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. DiscoveryWorld: A virtual environment for developing and evaluating automated scientific discovery agents. In *Advances in Neural Information Processing Systems*, volume 37, pp. 10088–10116, 2024.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. CLadder: Assessing causal reasoning in language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 31038–31065, 2023.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation? In *ICLR*, volume 2024, pp. 28663–28679, 2024.

David Kirsh and Paul Maglio. On distinguishing epistemic from pragmatic action. *Cognitive Science*, 18(4):513–549, 1994.

Aneesh Komanduri, Karuna Bhaila, and Xintao Wu. CausalVLBench: Benchmarking visual causal reasoning in large vision-language models. In *EMNLP*, pp. 30660–30680. Association for Computational Linguistics, 2025.

Yann LeCun. A path towards autonomous machine intelligence. OpenReview position paper, June 2022. URL https://openreview.net/forum?id=BZ5a1r-kVsf. Version 0.9.2, June 27, 2022.

Seungwon Lim, Sungwoong Kim, Jihwan Yu, Sungjae Lee, Jiwan Chung, and Youngjae Yu. VisEscape: A benchmark for evaluating exploration-driven decision-making in virtual escape rooms. In *EMNLP*, pp. 16020–16047. Association for Computational Linguistics, 2025.

OpenAI. GPT-5.5 system card, 2026. URL https://openai.com/index/gpt-5-5-system-card/.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009. ISBN 9780521895606.

Cheng Qian, Peixuan Han, Qinyu Luo, Bingxiang He, Xiusi Chen, Yuji Zhang, Hongyi Du, Jiarui Yao, Xiaocheng Yang, Denghui Zhang, Yunzhu Li, and Heng Ji. EscapeBench: Towards advancing creative intelligence of language model agents. In *ACL*, pp. 798–820. Association for Computational Linguistics, 2025.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10740–10749, 2020.

Zhe Sun, Kunlun Wu, Chuanjian Fu, Zeming Song, Langyong Shi, Zihe Xue, Bohan Jing, Ying Yang, Xiaomeng Gao, Aijia Li, Tianyu Guo, Huiying Li, Xueyuan Yang, Rongkai Liu, Xinyi He, Yuxi Wang, Yue Li, Mingyuan Liu, Yujie Lu, Hongzhao Xie, Shiyun Zhao, Bo Dai, Wei Wang, Tao Yuan, Song-Chun Zhu, Yujia Peng, and Zhenliang Zhang. TongSIM: A general platform for simulating intelligent machines. *arXiv preprint arXiv:2512.20206*, 2025.

Jane X. Wang, Michael King, Nicolas Porcel, Zeb Kurth-Nelson, Tina Zhu, Charlie Deck, Peter Choy, Mary Cassin, Malcolm Reynolds, Francis Song, Gavin Buttimore, David P. Reichert, Neil Rabinowitz, Loic Matthey, Demis Hassabis, Alexander Lerchner, and Matthew Botvinick. Alchemy: A benchmark and analysis toolkit for meta-reinforcement learning agents. In *NeurIPS Datasets and Benchmarks*, volume 1, 2021.

Ziyue Wang, Yurui Dong, Fuwen Luo, Minyuan Ruan, Zhili Cheng, Chi Chen, Peng Li, and Yang Liu. How do multimodal large language models handle complex multimodal reasoning? placing them in an extensible escape game. In *ICCV*, pp. 4807–4817, October 2025.

Manjie Xu, Guangyuan Jiang, Wei Liang, Chi Zhang, and Yixin Zhu. Interactive visual reasoning under uncertainty. In *Advances in Neural Information Processing Systems*, volume 36, pp. 42409–42432, 2023.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, Heng Ji, Huan Zhang, and Tong Zhang. EmbodiedBench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70576–70631, 2025.

Shu Yu and Chaochao Lu. ADAM: An embodied causal agent in open-world environments. In *ICLR*, 2025.