Grounding Before Generalizing: How AI Differs from Humans in Causal Transfer

Liangru Xiang[1,4,*], **Yuxi Ma**[2,3,4,5*], **Zhihao Cao**[1,4*], **Yixin Zhu**[3,2,4,5✉], **and Song-Chun Zhu**[1,2,4✉]

*Equal contributors Project Website: https://causal-openlock.github.io
[1] Department of Automation, Tsinghua University [2] Institute for Artificial Intelligence, Peking University
[3] School of Psychological and Cognitive Sciences, Peking University [4] State Key Laboratory of General Artificial Intelligence
[5] Beijing Key Laboratory of Behavior and Mental Health, Peking University

Abstract

Extracting abstract causal structures and applying them to novel situations is a hallmark of human intelligence (Griffiths & Tenenbaum, 2005; Holyoak & Cheng, 2011; Lake et al., 2017). While Large Language Models (LLMs) and Vision Language Models (VLMs) have shown strong performance on a wide range of reasoning tasks (Brown et al., 2020; Xu et al., 2025), their capacity for interactive causal learning—inducing latent structures through sequential exploration and transferring them across contexts—remains uncharacterized. Human learners accomplish such transfer after minimal exposure, whereas classical Reinforcement Learning (RL) agents fail catastrophically (Edmonds et al., 2018). Whether state-of-the-art Artificial Intelligence (AI) models possess human-like mechanisms for abstract causal structure transfer is an open question. Using the OpenLock paradigm (Edmonds et al., 2018) requiring sequential discovery of Common Cause (CC) and Common Effect (CE) structures, here we show that models exhibit fundamentally delayed or absent transfer: even successful models require initial environmental-specific mapping—what we term environmental grounding—before efficiency gains emerge, whereas humans leverage prior structural knowledge from the very first solution attempt. In the text-only condition, models matched or exceeded human discovery efficiency. In contrast, visual information—in both the image-only and text-and-image conditions—overall degraded rather than enhanced performance, revealing a broad reliance on symbolic processing rather than integrated multimodal reasoning. Models further exhibited systematic CC/CE asymmetries absent in humans, suggesting heuristic biases rather than direction-neutral causal abstraction. These findings reveal that large-scale statistical learning does not produce the decontextualized causal schemas underpinning human analogical reasoning, establishing grounding-dependent transfer as a fundamental limitation of current LLMs and VLMs.

Keywords: vision-language models; large language models; causal learning; structure transfer; active reasoning

Introduction

Large Language Models (LLMs) and Vision Language Models (VLMs) have achieved remarkable success across tasks ranging from natural language understanding to visual reasoning and mathematical problem-solving (Achiam et al., 2023; Anthropic, 2024; Brown et al., 2020; Gemini Team, 2023; Liu et al., 2025; OpenAI, 2023; Zhang et al., 2026). Yet a fundamental question remains: do these models engage in the kind of active causal learning and structural abstraction that characterizes human intelligence? While humans readily discover causal relationships through interaction and transfer this knowledge to new contexts, whether state-of-the-art AI models possess comparable capabilities remains largely unexplored.

Causal structure transfer—the ability to recognize and apply abstract relational patterns across different domains—represents a cornerstone of human cognition (Griffiths & Tenenbaum, 2005, 2009; Holyoak & Cheng, 2011; Holyoak & Thagard, 1996; Holyoak et al., 2010; Lu et al., 2008). Consider a smartphone unlockable via fingerprint, facial recognition, or passcode: this exemplifies a **many-to-one Common Effect (CE) structure**, where multiple independent causes converge on a single effect. Once grasped, the principle generalizes—a learner intuitively expects other secure systems to offer "multiple pathways to authorization." Conversely, a power strip failure that simultaneously cuts power to a lamp, laptop, and television instantiates a **one-to-many Common Cause (CC) structure**, where a single cause propagates to multiple effects. Recognizing such patterns guides efficient learning: one seeks a central breaker rather than inspecting each device individually. Crucially, such structural abstraction allows agents to navigate unfamiliar environments without relearning from scratch.

Empirical evidence confirms that humans excel at causal structure transfer in interactive settings, exhibiting marked efficiency gains when transitioning between structurally similar environments (Edmonds et al., 2018, 2019, 2020). Remarkably, this transfer occurs after minimal exposure—often a single episode. By contrast, traditional RL agents fail catastrophically on the same tasks despite orders of magnitude more training data, demonstrating that purely associative mechanisms cannot capture genuine structural abstraction.

Whether VLMs and LLMs can bridge this gap is a pressing open question. These models are trained on vast corpora encoding rich causal and relational knowledge: VLMs integrate visual perception with language understanding, potentially enabling causal pattern extraction from visual scenes (Anthropic, 2024; Gemini Team, 2023; OpenAI, 2023), while LLMs have shown promise in linguistic causal reasoning (Jin et al., 2023, 2024; Kiciman et al., 2023) and structural pattern extraction via in-context learning (Brown et al., 2020). Yet these capacities have not been tested in interactive settings that demand active exploration and discovery of latent causal structure through sequential decision-making—precisely the conditions under which human transfer is most striking.

We address this gap by adapting the OpenLock paradigm (Edmonds et al., 2018)—originally developed to benchmark human causal transfer against RL agents—to systematically probe four state-of-the-art AI models (GPT-5.2, Claude-4.5-Sonnet, Gemini-3-Flash, and DeepSeek-V3.2). The layout of the environment is shown in Fig. 1. This framework affords

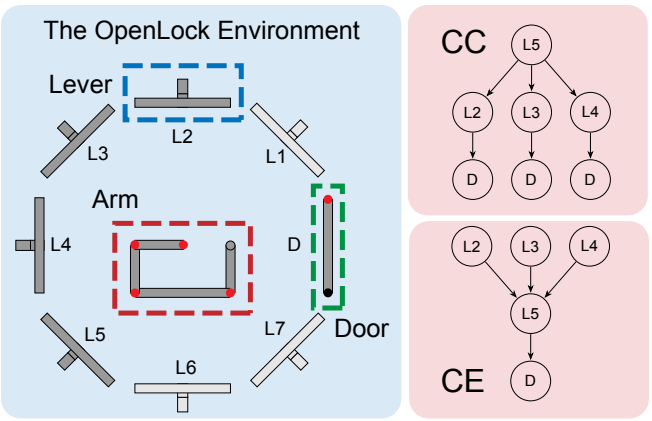

Figure 1: **OpenLock environment and causal structure schematics.** The virtual environment (*left*) contains seven levers and one door; agents discover sequential causal dependencies through active exploration, with lever positions, colors, and labels varying across environments while the underlying causal topology is preserved. The latent causal graphs (*right*) differ across the two experimental conditions: CC instantiates a *divergent* one-to-many structure in which a single first-stage lever (L_5) enables multiple independent second-stage paths to the door, whereas CE instantiates a *convergent* many-to-one structure in which multiple first-stage levers funnel through a shared bottleneck lever (L_5) before reaching the door.

a systematic dissociation between local causal discovery—finding solutions within a single environment—and genuine structural abstraction—transferring learned relational schemas to perceptually novel environments. By directly comparing model behavior against the human data reported in Edmonds et al. (2018), we evaluate models across three dimensions: (i) efficiency in causal discovery, (ii) the influence of input modality on active reasoning, and (iii) the transfer of learned structures to new environments.

Our findings reveal a fundamental divergence between human and model behavior. In the text-only condition, models matched or exceeded human discovery efficiency within a single environment. However, visual information—in both the image-only and text-and-image conditions—degraded rather than enhanced model performance, suggesting that current VLMs rely on symbolic processing rather than integrated visual-language reasoning. Most critically, while humans immediately exploited prior structural knowledge upon entering a new environment (Edmonds et al., 2018)—exhibiting strong positive transfer from the very first solution attempt—no model showed such *a priori* benefit. Instead, efficiency gains emerged only after models independently discovered an initial solution in the new context, a pattern consistent with post-hoc environmental grounding rather than genuine structural abstraction. We use "grounding' here strictly in the sense of environmental mapping between abstract structure and situational tokens, distinct from multimodal grounding. Models further exhibited systematic performance asymmetries between CC and CE configurations that were absent in human learners, suggesting reliance on heuristic biases encoded during training rather than on direction-neutral causal representations.

The OpenLock Paradigm

The OpenLock task (Edmonds et al., 2018) provides a controlled environment for studying interactive causal induction and structure transfer. Originally designed to compare human learners against RL agents—revealing that humans transfer causal structures after minimal exposure while RL agents fail to do so even with extensive training—the paradigm offers an ideal testbed for probing whether modern AI models exhibit similarly human-like abstraction. Each environment contains eight interactive components: seven levers and one door. Success requires discovering three unique solutions within a budget of 30 attempts, where each attempt is strictly limited to a three-action sequence: two lever manipulations followed by a door-opening attempt. A model is considered successful if it identifies all three solutions within this budget. Crucially, the underlying causal graph is latent—agents must infer it through active exploration rather than passive observation.

The two experimental variants instantiate distinct causal graph topologies over four active components (three levers and the door); the remaining four levers are inactive and serve as distractors:

- **Common Cause (CC)**: A *divergent* structure where a single first-stage lever (L_1) enables multiple second-stage options. Pushing L_1 unlocks L_2, L_3, or L_4, yielding three solutions: $L_1 \rightarrow \{L_2, L_3, L_4\} \rightarrow$ Door.
- **Common Effect (CE)**: A *convergent* structure where multiple first-stage levers funnel through a single enabler. Pushing any of L_1, L_2, or L_3 unlocks the same second-stage lever (L_4), yielding three solutions: $\{L_1, L_2, L_3\} \rightarrow L_4 \rightarrow$ Door.

This design affords a clean dissociation between surface-level and structural features: across environments, lever positions, colors, and labels change, but the underlying CC or CE topology is preserved. Genuine structure transfer therefore requires abstracting away perceptual details to recover the invariant relational schema.

We evaluated four state-of-the-art models—GPT-5.2, Claude-4.5-Sonnet, Gemini-3-Flash, and DeepSeek-V3.2—selected to represent a diverse range of architectural paradigms and reasoning capabilities. Human behavioral data from Edmonds et al. (2018) ($N = 80$, tested under equivalent task constraints) serve as the comparative reference throughout.

Experiment 1: Causal Structure Discovery

We first investigated whether modern AI models can discover all solutions within a *single* OpenLock environment through interactive exploration, extending the causal discovery benchmark of Edmonds et al. (2018) from RL agents to contemporary VLMs and LLMs. We examined how causal structure and presentation modality jointly influence discovery trajectories.

Experimental Design

Following the protocol of Edmonds et al. (2018), each model was given 30 attempts to find all solutions within a single OpenLock environment. To ensure performance stability across

random environment instantiations, we tested 30 independent agents per model on each of the two causal structures (CC and CE). GPT-5.2, Claude-4.5-Sonnet, and Gemini-3-Flash were evaluated across all three conditions; DeepSeek-V3.2 was evaluated under the text-only condition only, as it does not support visual input in the interactive setting used here.

Text-only (T) Condition Models interacted through a purely symbolic interface. Each prompt comprised: (i) high-level task objectives and operational constraints, explicitly requiring identification of all solutions; (ii) the initial state of all levers (position, color, and orientation) and door status in textual format; (iii) a sequential history of all executed actions and their outcomes; and (iv) a counter of remaining solutions. After each action, models received explicit feedback: either a null-change notification for unsuccessful attempts or a detailed state update for successful interactions (*e.g.*, "LOWERLEFT changes to GREY pushed"). Solution discovery was explicitly acknowledged (*e.g.*, "Solution found! 2 solutions remaining").

Image-only (I) Condition Models operated under a strictly visual paradigm in which all task-relevant information was conveyed through images alone. Specifically, (i) the initial environment configuration was presented solely via a representative image; (ii) lever and door state descriptions were conveyed exclusively through images; and (iii) post-action feedback consisted of dynamic visual sequences only. By removing all symbolic scaffolding, this condition isolates the contribution of pure visual input to causal discovery.

Text-and-Image (TI) Condition Models received both the textual interface of the text-only condition and supplemental visual inputs. Relative to the text-only condition, this condition additionally provided (i) an image of the initial environment state alongside the textual introduction, and (ii) dynamic visual feedback reflecting environment updates after each action. This condition was designed to test whether supplemental visual information facilitates causal discovery when symbolic information is already available.

Results

Overall Performance We compared model performance in the T condition against the human baseline from Edmonds et al. (2018) (see also Tab. 1), as this condition isolates logical inference from visual processing and thus provides the most direct comparison of causal reasoning capacity. Human participants achieved a 65.0% success rate across both CC and CE structures, requiring an average of 20.66 attempts ($SD = 9.09$, $N = 80$) to identify all three solutions. Gemini-3-Flash outperformed humans in both accuracy and efficiency, achieving a 100.0% success rate with significantly fewer attempts ($M = 9.08$, $SD = 2.76$, $N = 60$; $t(138) = 9.54$, $p < .001$). GPT-5.2 and Claude-4.5-Sonnet also exceeded the human baseline in efficiency: GPT-5.2 averaged 15.77 attempts ($SD = 7.95$, $N = 60$; $t(138) = 3.33$, $p = .001$) with success rates between 66.7% and 100.0% depending on structure, and Claude-4.5-Sonnet averaged 15.74 attempts ($SD = 9.53$, $N = 61$; $t(139) = 3.12$, $p = .002$). DeepSeek-V3.2, while achieving high success rates (96.7% for CC; 86.2% for CE), showed no statistically significant difference from humans in attempt count ($M = 19.35$, $SD = 6.63$, $N = 60$; $t(138) = 0.95$, $p = .346$).

Causal Structure Asymmetry Humans showed consistent success rates (65%) across both structures, with no significant difference in attempt counts between CC ($M = 19.4$, $SD = 9.66$, $N = 40$) and CE ($M = 22.0$, $SD = 8.40$, $N = 40$; $t(78) = -1.27$, $p = .207$). In contrast, all models exhibited systematic asymmetries between structures. GPT-5.2 showed a strong CC advantage in the T condition: 100.0% success with 11.8 attempts ($SD = 3.38$, $N = 30$) on CC, versus 66.7% success with 19.7 attempts on CE ($SD = 9.23$, $N = 30$; $t(58) = -4.38$, $p < .001$). Gemini-3-Flash similarly performed better on CC across all conditions: despite maintaining 100% success throughout, attempt counts were consistently lower for CC ($M = 8.13$, $SD = 3.44$, $N = 31$) than CE ($M = 10.03$, $SD = 2.63$, $N = 30$; $t(59) = -2.81$, $p = .0067$). Claude-4.5-Sonnet showed the opposite pattern, achieving a higher success rate on CE (86.7%) than CC (67.7%), though the difference in attempt counts was not significant ($M = 14.6$, $SD = 8.67$, $N = 30$ for CE *vs.* $M = 16.9$, $SD = 10.31$, $N = 31$ for CC; $t(59) = 0.94$, $p = .350$). These divergent asymmetry patterns—with GPT and Gemini favoring CC while Claude favors CE—were absent in human learners.

Impact of Modality on Causal Discovery Adding visual information degraded performance for most models. For GPT-5.2, the TI condition required significantly more attempts than the T condition ($M = 24.10$, $SD = 6.95$, $N = 60$ *vs.* $M = 15.77$, $SD = 7.95$, $N = 60$; $t(118) = -6.11$, $p < .001$). Gemini-3-Flash showed a smaller but significant efficiency drop from T to TI ($M = 9.08$, $SD = 2.76$, $N = 60$ *vs.* $M = 10.41$, $SD = 3.54$, $N = 70$; $t(128) = -2.36$, $p = .020$), though it maintained 100% success across all three conditions.

Table 1: **Experiment 1: Performance across causal structures and modalities.** Success rate (%) and average attempt counts for Common Cause (CC) and Common Effect (CE) structures under each presentation condition. T: text-only; I: image-only; TI: text-and-image. Human data reproduced from Edmonds et al. (2018).

Model	Condition	Success Rate (%)		Avg. Attempts	
		CC	CE	CC	CE
Human (Edmonds et al., 2018)	—	65.0	65.0	19.4	22.0
GPT	T	100.0	66.7	11.8	19.7
	I	38.7	10.3	26.1	29.1
	TI	66.7	50.0	22.8	25.4
Claude	T	67.7	86.7	16.9	14.6
	I	45.2	64.5	22.9	19.8
	TI	86.7	93.3	17.9	10.2
Gemini	T	100.0	100.0	8.1	10.0
	I	100.0	100.0	10.0	12.3
	TI	100.0	100.0	8.7	11.8
DeepSeek	T	96.7	86.2	18.6	20.1

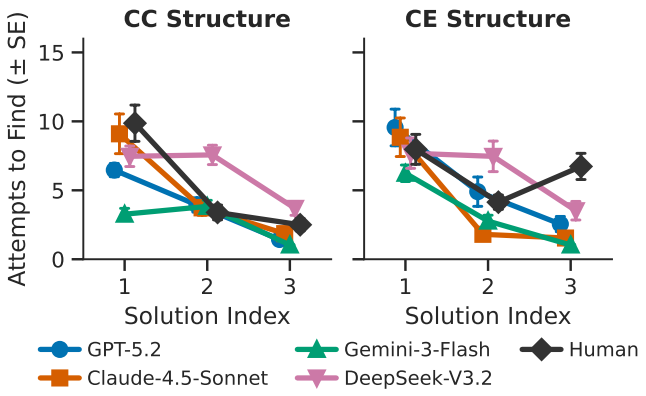

Figure 2: **Sequential discovery efficiency across causal structures.** Marginal discovery cost (attempts required to find each successive solution) within a single environment, shown separately for CC (left) and CE (right) structures. AI models are evaluated in the T condition. Humans and Claude-4.5-Sonnet exhibit sharp non-linear acceleration after the first solution, whereas GPT-5.2, Gemini-3-Flash, and DeepSeek-V3.2 show more gradual improvement. Error bars denote standard error.

Claude-4.5-Sonnet was the exception, showing no significant difference between T and TI ($M = 15.74$, $SD = 9.53$, $N = 61$ *vs.* $M = 14.05$, $SD = 8.12$, $N = 60$; $t(119) = 1.05$, $p = .297$). Removing symbolic scaffolding entirely in the I condition led to further significant performance declines for GPT-5.2 (I: $M = 27.55$, $SD = 5.01$, $N = 60$ *vs.* TI: $M = 24.10$, $SD = 6.95$, $N = 60$; $t(118) = -3.12$, $p = .002$) and Claude-4.5-Sonnet (I: $M = 21.37$, $SD = 9.03$, $N = 62$ *vs.* TI: $M = 14.05$, $SD = 8.12$, $N = 60$; $t(120) = -4.70$, $p < .001$). Taken together, these results indicate that models rely primarily on symbolic text for causal reasoning, with visual input acting as a distractor rather than a facilitative cue.

Sequential Discovery Patterns To characterize within-environment learning dynamics, we analyzed the marginal discovery cost—the number of attempts required to find each successive solution (see also Fig. 2). Human learners exhibited **non-linear acceleration**: discovery cost dropped sharply from the first solution ($M = 7.37$) to the second ($M = 3.65$; $t(51) = 4.31$, $p < .001$). Claude-4.5-Sonnet (T condition) closely mirrored this pattern, with discovery cost falling from $M = 7.19$ to $M = 2.62$ ($t(46) = 4.80$, $p < .001$). GPT-5.2 and Gemini-3-Flash, by contrast, showed only **gradual, incremental improvement**: for example, Gemini in the T condition decreased from $M_{L1} = 4.75$ to $M_{L2} = 3.30$ ($t(59) = 2.28$, $p = .026$), a substantially smaller reduction in magnitude than that observed in humans or Claude. These differences in learning dynamics were consistent across both CC and CE structures.

Experiment 2: Causal Structure Transfer

Having established baseline patterns in causal structure discovery, we investigated whether providing models with complete solutions from a structurally similar environment would facilitate discovery in a new environment—testing models' capacity for structure transfer.

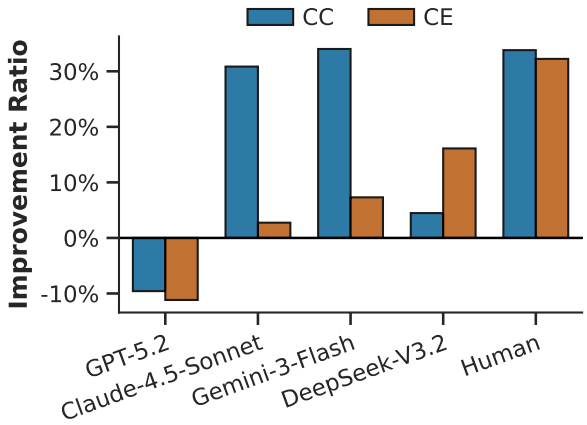

Figure 3: **Efficiency gain from causal structure transfer.** Improvement ratio in average attempt counts between Experiment 1 (Baseline) and Experiment 2 (Transfer), calculated as $(Attempts_{\text{base}} - Attempts_{\text{trans}})/Attempts_{\text{base}}$, shown separately for CC and CE structures. Positive values indicate positive transfer (fewer attempts required in the transfer condition).

Experimental Design

We modified the prompts for all models to include explicit textual descriptions of all three solutions from a previously completed environment with the same underlying causal structure (CC or CE). Each solution description specified the exact action sequence—first lever pushed, second lever pushed, and door-opening attempt (*e.g.*, "Solution 1: Push LOWERLEFT lever, then push UPPERLEFT lever, then try door"). The previous environment had a different spatial configuration of levers and potentially different color assignments, requiring models to abstract the structural principle beyond specific positions or visual attributes. We tested 30 agents per model per structure (240 in total: 4 models × 2 structures × 30 agents).

Results

Overall Transfer Effects The improvement ratios relative to Experiment 1 baselines are shown in Fig. 3 and summarized in Tab. 2. Human participants demonstrated robust structural transfer, significantly reducing average attempts from baseline ($M = 20.66$, $SD = 9.09$) to transfer ($M = 13.85$, $SD = 10.00$; $t(158) = 4.51$, $p < .001$, $d = 0.71$).

In contrast, models exhibited limited transfer capabilities. Gemini-3-Flash was the only model to achieve statistically significant overall transfer, reducing average attempts from $M_{\text{base}} = 9.08$ ($SD = 2.76$) to $M_{\text{trans}} = 7.33$ ($SD = 3.60$;

Table 2: **Experiment 2: Overall transfer effects.** Average attempt counts in the Baseline (Experiment 1, T condition) and Transfer (Experiment 2) conditions, with improvement ratio calculated as $(Attempts_{\text{base}} - Attempts_{\text{trans}})/Attempts_{\text{base}}$. Positive improvement ratios indicate positive transfer. Human data reproduced from Edmonds et al. (2018). $^{**}p < .01$; $^{***}p < .001$; unmarked: $p > .05$.

Model	Baseline M (SD)	Transfer M (SD)	Improv. (%)
Human (Edmonds et al., 2018)	20.66 (9.09)	13.85 (10.00)	$+33.0^{***}$
GPT-5.2	15.77 (7.95)	17.43 (8.61)	-10.5
Claude-4.5-Sonnet	15.74 (9.53)	12.92 (8.47)	$+17.9$
Gemini-3-Flash	9.08 (2.76)	7.33 (3.60)	$+19.3^{**}$
DeepSeek-V3.2	19.35 (6.63)	17.32 (5.89)	$+10.5$

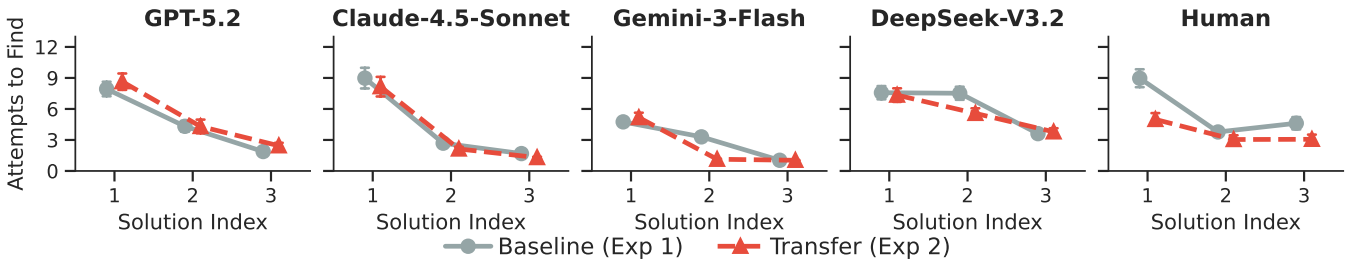

Figure 4: **Impact of causal structure transfer on sequential discovery dynamics.** Marginal discovery cost (attempts required to find each successive solution) for Baseline (Experiment 1, gray) and Transfer (Experiment 2, red) conditions, shown for each model and for human participants. Error bars denote standard error. Humans exhibit immediate transfer, with significantly lower first-solution cost under transfer. All models exhibit delayed transfer, with performance gains emerging only at the second solution (if at all), indicating that models require initial environmental grounding before leveraging prior structural knowledge.

$t(118) = 2.99$, $p = .003$, $d = 0.55$). Claude-4.5-Sonnet ($M_{\text{base}} = 15.74 \rightarrow M_{\text{trans}} = 12.92$; $t(119) = 1.71$, $p = .090$) and DeepSeek-V3.2 ($M_{\text{base}} = 19.35 \rightarrow M_{\text{trans}} = 17.32$; $t(118) = 1.72$, $p = .089$) showed numerical trends toward improvement that did not reach statistical significance. GPT-5.2 showed no positive transfer; its attempt count numerically increased in the transfer condition ($M_{\text{base}} = 15.77 \rightarrow M_{\text{trans}} = 17.43$; $t(118) = -1.12$, $p = .266$), though this decline was not statistically reliable.

Delayed Transfer Effects in Sequential Discovery To characterize *when* during the search process transfer effects emerge, we analyzed the marginal discovery cost for each successive solution (see also Fig. 4). Human participants showed immediate transfer: attempts to find the *first* solution were significantly reduced from baseline ($M = 8.97$, $SD = 7.24$) to transfer ($M = 4.99$, $SD = 5.36$; $t(145) = 3.82$, $p < .001$, $d = 0.63$). In stark contrast, none of the four models showed a significant reduction in first-solution discovery cost, indicating that prior structural knowledge did not guide initial exploration of the new environment.

Transfer effects in models emerged only at the *second* solution. Gemini-3-Flash showed a dramatic reduction in second-solution discovery cost from baseline ($M_{\text{base}} = 3.30$, $SD = 2.41$) to transfer ($M_{\text{trans}} = 1.13$, $SD = 0.39$; $t(118) = 6.88$, $p < .001$, $d = 1.26$). DeepSeek-V3.2 similarly showed significant acceleration ($M_{\text{base}} = 7.52$, $SD = 4.73$ to $M_{\text{trans}} = 5.57$, $SD = 3.66$; $t(110) = 2.44$, $p = .016$, $d = 0.46$). GPT-5.2 and Claude-4.5-Sonnet did not show statistically significant improvements at the second solution. Thus, models that did benefit from prior structural knowledge did so only after independently discovering an initial solution in the new environment, in direct contrast to humans who leveraged structural knowledge from the very first attempt.

Discussion

Immediate *vs*. Delayed Transfer

Humans immediately applied prior structural knowledge to first-solution discovery in new environments, whereas all AI models showed delayed or absent transfer—requiring initial environmental grounding before efficiency gains emerged (Fig. 4). This contrast suggests that humans construct decontextualized causal schemas that directly guide action in novel contexts, while models must first establish mappings between surface tokens and structural roles through direct environmental interaction before latent structural knowledge becomes operative. Rather than possessing a fully portable causal schema, current LLMs appear to exhibit a *grounding-dependent transfer mechanism*: structural knowledge acquired from prior experience remains latent until activated by concrete situational feedback. Delayed transfer could also partly reflect in-context learning effects (Min et al., 2022) rather than a grounding-specific mechanism. This implies that for current LLMs, the instantiation of abstract rules in novel contexts is a process remaining critically sensitive to situational grounding cues rather than an immediate byproduct of scale.

This pattern contrasts sharply with classic findings in human analogical reasoning, where successful transfer depends on recognizing structural similarity between source and target domains, independent of surface features (Gentner, 1983; Holyoak & Thagard, 1996). While humans readily form what Gick and Holyoak (1983) termed "problem schemas"—abstract relational representations that transfer across perceptually distinct instantiations—our results suggest that current AI systems remain bound to context-specific instantiations, requiring direct experience with a new environment before prior structural knowledge can be exploited. This is a meaningful distinction: human transfer is *prospective* (structural knowledge guides initial exploration), whereas model transfer is *retrospective*.

The CC/CE asymmetries across models further support this interpretation. GPT-5.2 consistently favored CC structures while Claude-4.5-Sonnet showed superior performance on CE configurations (Tab. 1). This pattern suggests that models encode directional statistical regularities from training corpora, where diagnostic reasoning (tracing effects back to causes) and predictive reasoning (projecting causes forward to effects) may differ in distributional frequency, rather than forming abstract causal representations that transcend directional preference. As Gentner (1983) emphasized, genuine structure mapping should enable transfer regardless of relational direction; the asymmetries indicate that current models lack this flexibility, further evidencing their reliance on surface-level statistical associations rather than genuine structural abstraction.

Insight *vs*. Gradual Optimization

Humans and Claude-4.5-Sonnet exhibited nonlinear acceleration—a reduction in discovery cost after the first solution—while GPT-5.2 and Gemini-3-Flash showed only gradual, incremental improvement (Fig. 2). This abrupt efficiency gain in humans is consistent with *representational change* (Ohlsson, 1992): discovering the first solution reveals the underlying causal structure, enabling the sudden elimination of entire classes of incorrect hypotheses and a qualitative reorganization of search strategy. By contrast, the smooth improvement curves of GPT-5.2 and Gemini suggest a process of iterative statistical refinement—narrowing a broad probability distribution over possible solutions rather than a discrete restructuring of the problem representation.

The in-context learning observed here thus functions more like iterative adjustment to prompt history than the discrete logical updates characteristic of human insight. Importantly, Claude-4.5-Sonnet's trajectory more closely resembles the human pattern, raising the question of whether this reflects architectural differences that better support flexible hypothesis revision, or an artifact of different search heuristics. Resolving this requires systematic investigation into how training objectives and architectural choices shape within-context learning dynamics (Nakkiran et al., 2020).

Multimodal Interference and the Abstraction Gap

The addition of visual information degraded performance for most models, with the image-only condition yielding the worst results overall. This finding reveals that while humans can selectively attend to task-relevant modalities and suppress irrelevant perceptual input (Shams & Seitz, 2008), current VLMs appear to lack the hierarchical control necessary to filter low-level visual features when abstract symbolic reasoning is required. In the OpenLock task, causal rules are fully determined by relational structure—lever positions, colors, and geometries are perceptually salient but causally irrelevant. Rather than providing useful abstraction support, these visual features appear to compete for processing capacity, obscuring the underlying symbolic structure.

This failure reveals a broader limitation of current multimodal architectures: an inability to distinguish between high-level causal invariants and low-level perceptual variance. Human causal reasoning relies on what might be termed *abstraction through subtraction*—the capacity to ignore specific visual appearances in order to isolate invariant relational rules. Current VLMs, by contrast, appear to perform undifferentiated fusion of visual and linguistic inputs, forcing the reasoning process to integrate perceptual noise into causal hypotheses. The contrast between Claude-4.5-Sonnet—which showed no significant performance cost from added visual input—and GPT-5.2 and Gemini-3-Flash—which showed degradation—suggests that models differ in their ability to de-weight irrelevant visual information, though none achieved positive visual facilitation. Effective multimodal causal reasoning may therefore require architectures in which symbolic abstraction governs primary reasoning, with visual input serving a secondary verification role.

Implications and Future Directions

Together, these three findings—delayed transfer, absence of insight-like restructuring, and multimodal interference—converge on a shared conclusion: large-scale statistical learning over text and image corpora does not, by itself, produce the flexible, decontextualized causal representations that underpin human structural abstraction. Current LLMs and VLMs excel at local causal search within a single context, but fail to apply structural knowledge prospectively when entering new environments. This gap is not a matter of scale or data, but appears qualitative, reflecting a fundamental difference in how humans and current AI systems represent and deploy abstract relational structure.

These findings point toward several concrete directions for future work. First, structure-mapping curricula—training regimes that explicitly reward transfer across perceptually distinct instantiations of the same relational schema—may help bridge the prospective/retrospective transfer gap identified here. Second, the modality interference results suggest that multimodal architectures may benefit from more explicit mechanisms for cross-modal attention control, allowing visual input to inform rather than distort symbolic reasoning. Third, the divergent learning dynamics across models (Claude *vs*. GPT and Gemini) suggest that architectural and training choices meaningfully shape within-context learning, warranting systematic study. More broadly, the OpenLock paradigm offers a reusable benchmark for probing structural generalization in interactive settings—one that dissociates genuine abstraction from context-bound pattern matching in a way that static benchmarks cannot.

Conclusion

By comparing human and AI performance in causal structure discovery and transfer using the OpenLock paradigm, we identified three fundamental differences in how current LLMs and VLMs differ from humans in abstract causal reasoning. Humans demonstrate immediate transfer of structural knowledge, show rapid nonlinear acceleration consistent with sudden representational insight, and leverage multimodal information more selectively than current AI systems. Despite strong performance on static reasoning benchmarks, none of these capacities were reliably present in state-of-the-art models. Our findings suggest that large-scale statistical learning does not inherently produce the flexible, decontextualized causal schemas that characterize human structural abstraction—the gap between humans and current AI is not merely quantitative, but qualitative. This work points toward concrete development paths—including structure-mapping curricula and architectures with explicit cross-modal attention control—for building AI systems capable of human-like causal abstraction.

Acknowledgment This work is supported in part by the National Science and Technology Major Project

(2022ZD0114900), National Natural Science Foundation of China (62376009), the PKU-BingJi Joint Laboratory for Artificial Intelligence, the Wuhan Major Scientific and Technological Special Program (2025060902020304), the Hubei Embodied Intelligence Foundation Model Research and Development Program, and the National Comprehensive Experimental Base for Governance of Intelligent Society, Wuhan East Lake High-Tech Development Zone.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (cit. on p. 1).

Anthropic. (2024). Claude 3 model card. *Anthropic Technical Report* (cit. on p. 1).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)* (cit. on p. 1).

Edmonds, M., Kubricht, J., Summers, C., Zhu, Y., Rothrock, B., Zhu, S.-C., & Lu, H. (2018). Human casual transfer: Challenges for deep reinforcement learning. *Annual Meeting of the Cognitive Science Society (CogSci)* (cit. on pp. 1–4).

Edmonds, M., Ma, X., Qi, S., Zhu, Y., Lu, H., & Zhu, S.-C. (2020). Theory-based causal transfer: Integrating instance-level induction and abstract-level structure learning. *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)* (cit. on p. 1).

Edmonds, M., Qi, S., Zhu, Y., Kubricht, J., Zhu, S.-C., & Lu, H. (2019). Decomposing human causal learning: Bottom-up associative learning and top-down schema reasoning. *Annual Meeting of the Cognitive Science Society (CogSci)* (cit. on p. 1).

Gemini Team. (2023). Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (cit. on p. 1).

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170 (cit. on p. 5).

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*(1), 1–38 (cit. on p. 5).

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384 (cit. on p. 1).

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661 (cit. on p. 1).

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology*, *62*(1), 135–163 (cit. on p. 1).

Holyoak, K. J., Lee, H. S., & Lu, H. (2010). Analogical and category-based inference: A theoretical integration with bayesian causal models. *Journal of Experimental Psychology: General*, *139*(4), 702 (cit. on p. 1).

Holyoak, K. J., & Thagard, P. (1996). *Mental leaps: Analogy in creative thought*. MIT press. (Cit. on pp. 1, 5).

Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Lyu, Z., Blin, K., Gonzalez Adauto, F., Kleiman-Weiner, M., Sachan, M., et al. (2023). Cladder: Assessing causal reasoning in language models. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)* (cit. on p. 1).

Jin, Z., Liu, J., Lyu, Z., Poff, S., Sachan, M., Mihalcea, R., Diab, M. T., & Schölkopf, B. (2024). Can large language models infer causation from correlation? *Proceedings of International Conference on Learning Representations (ICLR)* (cit. on p. 1).

Kiciman, E., Ness, R., Sharma, A., & Tan, C. (2023). Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research* (cit. on p. 1).

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40* (cit. on p. 1).

Liu, A., Mei, A., Lin, B., Xue, B., Wang, B., Xu, B., Wu, B., Zhang, B., Lin, C., Dong, C., et al. (2025). Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556* (cit. on p. 1).

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*(4), 955 (cit. on p. 1).

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work? *EMNLP* (cit. on p. 5).

Nakkiran, P., Neyshabur, B., & Sedghi, H. (2020). The deep bootstrap framework: Good online learners are good offline generalizers. *Proceedings of International Conference on Learning Representations (ICLR)* (cit. on p. 6).

Ohlsson, S. (1992). Information-processing explanations of insight and related phenomena. *Advances in the Psychology of Thinking*, 1–44 (cit. on p. 6).

OpenAI. (2023). Gpt-4v(ision) system card. *OpenAI Technical Report* (cit. on p. 1).

Shams, L., & Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in Cognitive Sciences*, *12*(11), 411–417 (cit. on p. 6).

Xu, F., Hao, Q., Zong, Z., Wang, J., Zhang, Y., Wang, J., Lan, X., Gong, J., Ouyang, T., Meng, F., et al. (2025). Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686* (cit. on p. 1).

Zhang, C., Song, J., Li, S., Liang, Y., Ma, Y., Wang, W., Zhu, Y., & Zhu, S.-C. (2026). Proposing and solving olympiad geometry with guided tree search. *Nature Machine Intelligence*, *8*, 84–95 (cit. on p. 1).