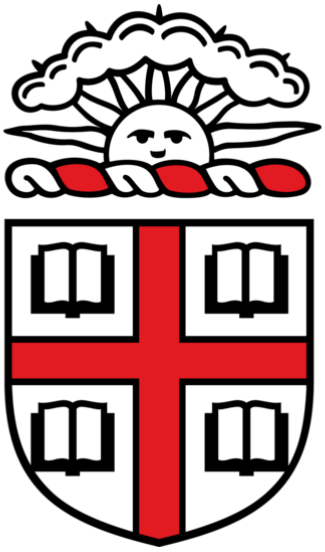
Structural vs. Functional Causal Models for Robotic Planning

Elicitation Methods and Task-Specific Performance in Assembly and Troubleshooting

Moon Hwan Kim



May 2025

Advisors:
Iris Bahar, Department of Computer Science
Steven Sloman, Department of Cognitive and Psychological Sciences

ABSTRACT

This study investigates the distinction between structural and functional causal models in the context of robotic planning. We examine whether humans naturally construct different types of causal models when reasoning about physical systems and how these representations affect performance in assembly and troubleshooting tasks. Our research compares three elicitation methods: a traditional intervention-based approach, a graphical drawing method, and an integrated hierarchical interface. Across diverse everyday objects (including bicycles, lamps, and sinks), we found that different elicitation methods significantly affect model fidelity, with the intervention-based approach consistently yielding higher hit rates at the cost of increased false alarms. When integrated into a Partially Observable Markov Decision Process (POMDP) planning framework, our ground truth models confirm that structural representations maximize assembly performance while functional models excel in troubleshooting scenarios. However, participant-elicited models frequently diverge from these theoretical expectations, with significant performance gaps between expert-generated and user-generated causal models in POMDP simulations. Despite these limitations, we demonstrate that different causal conditions (structural and functional) elicit statistically distinct causal representations, suggesting inherent differences in reasoning frameworks. Our findings indicate that adaptive robotic planning systems should maintain dual structural-functional representations while recognizing the challenges in accurately eliciting these models from human subjects. This work contributes to both cognitive theory by clarifying how humans conceptualize causal systems and practical robotics by highlighting important considerations for human-aligned planning approaches.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisors, Iris Bahar and Steven Sloman, for their invaluable guidance and support over the past two years, which have been instrumental in bringing this work to fruition. I'm also grateful to Tom Williams for his thoughtful insights and encouragement.

I wish to thank Semanti Basu and Semir Tatlidil for their mentorship and for inspiring me to continue my journey into the exciting world of human–robot interaction.

I'd also like to extend my gratitude to all the other members of our lab—past and present—whose collaboration, lively discussions, and unwavering support have greatly enriched this work.

This work was sponsored by the Office of Naval Research (ONR), under grant #N00014-22-12494. The views and conclusions contained herein are those of the authors only and should not be interpreted as representing those of ONR, the U.S. Navy or the U.S. Government.

Contents

1 Introduction **4**
- 1.1 Causal Reasoning: Theoretical Frameworks . . . 4
- 1.2 Causal Reasoning: Robotic Applications . . . 4
- 1.3 Study Overview . . . 4

2 Causal Elicitation Methods **5**
- 2.1 Prior Work . . . 5

3 Hierarchical Causal Elicitation: An Integrated Approach **7**
- 3.1 Methodology . . . 7
- 3.2 System Implementation and User Interface . . . 7
- 3.3 Stage-wise Experimental Workflow and Question Types . . . 10
- 3.4 Hierarchical Algorithm for Constructing Causal Graphs . . . 11
- 3.5 Findings and Comparative Insights . . . 13

4 Evaluating Distinct Causal Reasoning Modes: Structural vs Functional Causal Models **16**
- 4.1 Methodology . . . 16
- 4.2 Hierarchical Interface Method . . . 17
- 4.3 Intervention Method . . . 18

5 Causally Informed Planning under Uncertainty for Object Assembly and Troubleshooting **21**
- 5.1 Baseline Evaluation with Expert Ground Truth Models . . . 21
- 5.2 Assembly . . . 24
- 5.3 Pairwise Post-hoc Comparisons . . . 25
- 5.4 Troubleshooting . . . 26

6 Discussion **29**
- 6.1 Evidence for Distinct Causal Reasoning Modes . . . 29
- 6.2 Methodological Limitations in Structural Model Elicitation . . . 29
- 6.3 Refined Approaches to Structural Model Elicitation . . . 30
- 6.4 Task-Dependent Utility of Causal Models . . . 30
- 6.5 Object-Specific Effects and Context Sensitivity . . . 31
- 6.6 Implications for Robotic Planning . . . 31
- 6.7 Limitations and Future Directions . . . 31
- 6.8 Conclusion . . . 32

7 References **33**

8 Appendix **35**

1 Introduction

1.1 Causal Reasoning: Theoretical Frameworks

Humans use causal reasoning to organize events into coherent narratives by inferring mechanisms from observations (e.g., a car's failure to start suggests a dead battery or empty fuel tank) [8, 13, 11]. Formalizing this process, *Causal Bayesian networks* model variables as nodes in a directed acyclic graph with probabilistic dependencies that support both qualitative insights and quantitative inference [20, 21]. In contrast, *mental models* theory treats causal assertions as deterministic links, interpreting "A causes B as an unequivocal relation [13, 14]. Although empirical studies demonstrate all-or-nothing judgments under unspecified probabilities [9], this approach fails to capture real-world uncertainty [18].

The *force-dynamics* framework conceptualizes causation as interactions (enabling, preventing, resisting) that reflect nuanced distinctions in everyday language [23, 25, 12]. While *functional* causal models (e.g., Cheng's causal power theory) quantify how causes generate outcomes [7], no formal framework exists for *structural* causal models that capture component-level dependencies within physical systems. Existing elicitation methods represent causal relationships as a single directed graph, conflating structural and functional aspects and leaving this critical distinction unexplored [10].

1.2 Causal Reasoning: Robotic Applications

In robotics, distinguishing structural from functional knowledge is essential for robust decision-making under uncertainty. Early systems encoded causal rules for assembly and diagnosis, enabling error detection and recovery through probing component connections [5]. Subsequent research introduced adaptive models that update causal beliefs through intervention [19], demonstrated on platforms like Baxter where robots experimentally learn tool affordances and functional relations [6]. Logic-based planners further integrated both structural and functional constraints to support coordinated multi-robot tasks [17].

Recent advances in deep causal learning and counterfactual reasoning now allow systems to infer latent mechanisms from high-dimensional inputs and simulate hypothetical outcomes [22, 16]. Incorporating human-derived causal models into a POMDP framework has significantly improved assembly efficiency and interpretability [1, 2, 4, 3]. However, it remains unclear whether human reasoning adaptively shifts between structural and functional abstractions based on task context.

This distinction has profound practical implications: robots aligned with human causal intuitions can better anticipate errors, guide interventions, and collaborate effectively. Our work investigates whether different physical systems naturally elicit structural versus functional causal insights, and how these variations influence robotic planning and troubleshooting performance.

1.3 Study Overview

This study examines the structure of human causal reasoning in physical systems, focusing specifically on whether individuals spontaneously construct distinct *structural* models (emphasizing physical interconnections) and *functional* models (highlighting goal-directed mechanisms) independent of elicitation method. Motivated by challenges in robotic assembly and adaptive planning, we analyze how variations in elicited causal

structures relate to ground-truth models and influence performance when integrated into a Partially Observable Markov Decision Process (POMDP) framework.

We systematically compare two elicitation approaches: a traditional *intervention-based* method grounded in counterfactual reasoning, and a *drawing-based* graphical method. While intervention-based techniques typically yield higher-fidelity models, they are vulnerable to participant fatigue. Drawing-based methods offer greater scalability but potentially compromise structural richness. To address these limitations, we introduce a novel hierarchical elicitation interface that integrates graphical clustering with targeted counterfactual refinement, capturing structured causal representations while minimizing cognitive load.

Building on our previous work [1, 2, 4, 3], we integrate participant-elicited causal models into a POMDP-based planner to evaluate their effectiveness in robotic troubleshooting and assembly tasks. Specifically, we assess how causal representation fidelity shaped by reasoning mode and elicitation format influences task success rates under uncertainty. By connecting cognitive theory, elicitation methodology, and computational implementation, this study advances our understanding of how human causal abstraction can enhance autonomous robotic planning.

2 Causal Elicitation Methods

2.1 Prior Work

Note: The complete details of this work are presented in [24]. We provide only a brief summary of our methods and findings below.

In our work, we developed and compared two distinct approaches for eliciting causal structure from participants analyzing everyday physical artifacts. We evaluated how different elicitation procedures affect the completeness, granularity, and accuracy of the resulting causal models: the *Interventional Method* and the *Graphical (Drawing) Method*.

The Interventional Method decomposes an object into its constituent parts and presents participants with systematic counterfactual questions (e.g., If one removed [X part] from [object], would [Y part] still perform its function?"). Grounded in the make-a-difference view of causality, this method requires exhaustive, pairwise evaluation of all potential causal links. Participants view an object diagram alongside functional descriptions of each component, then evaluate every possible pairing. While this comprehensive approach yields more reported causal links including some false positives, it produces higher hit rates (HR, also known as true positive rates) and improved overall discriminability (d') between genuine and spurious causal relations.

The Graphical/Drawing Method employs an online interface (Loopy) where participants construct causal graphs by drawing directed arrows between nodes representing object parts. This approach allows participants to indicate causal relationships in a single, global step rather than through systematic pairwise queries. Though more intuitive and scalable, particularly for objects with numerous components, it yields lower HRs by not enforcing exhaustive consideration of every potential connection.

Our experiments provide compelling evidence for these trade-offs. In Experiment 1, which examined four light-producing objects (desk lamp, flashlight, kerosene lamp, and wall lamp), the Interventional Method significantly outperformed the Drawing Method in hit rates. The desk lamp's HR increased from 0.61 (Drawing) to 0.78 (Intervention), with similar improvements across all objects (see Table 1 and Table) 2. Although the In-

tervention condition produced somewhat higher false alarm rates (FAR), signal detection analysis confirmed superior overall discriminability (d') under the Interventional Method.

Experiment 2 extended these findings to a more diverse set of 10 objects (including a bicycle, cannon, electric mixer, hand mixer, paddle boat, pistol, scooter, sink, toilet, and tricycle), with component counts ranging from 5 to 12. Participants were randomly assigned 3 of the 10 objects, with average HR and FAR scores computed across these objects. Results again demonstrated a robust advantage for the Interventional Method: the average HR (true positive rate) was 0.54 compared to 0.29 in the Drawing condition, while the average FAR was 0.23 versus 0.09, respectively. Signal detection measures confirmed higher discriminability (d') and lower bias (c) in the Intervention condition, indicating that participants in this condition not only reported more causal links overall but also more accurately differentiated true causal relations from erroneous ones.

Tables 1 and 2 summarize the key findings from both experiments, illustrating performance differences across the full range of tested objects.

Table 1: Experiment 1: Means (with standard errors) for hit rates (HR) and false alarm rates (FAR) for four light-producing objects under the Intervention and Drawing conditions.

Object	**HR (Intervention)**	**HR (Drawing)**	**FAR (Intervention) / (Drawing)**
Desk lamp	0.78 (0.02)	0.61 (0.02)	0.14 (0.01) / 0.08 (0.01)
Flashlight	0.91 (0.02)	0.71 (0.03)	0.12 (0.02) / 0.10 (0.01)
Kerosene lamp	0.61 (0.02)	0.29 (0.02)	0.30 (0.01) / 0.17 (0.01)
Wall lamp	0.73 (0.02)	0.54 (0.02)	0.14 (0.01) / 0.15 (0.01)

Table 2: Experiment 2: Means (with standard errors) for hit rates (HR) and false alarm rates (FAR) for 10 objects under the Intervention and Drawing conditions.

Object	**HR (Drawing)**	**FAR (Drawing)**	**HR (Intervention)**	**FAR (Intervention)**
Bicycle	0.25 (0.05)	0.07 (0.01)	0.56 (0.02)	0.26 (0.02)
Cannon	0.26 (0.02)	0.10 (0.00)	0.50 (0.02)	0.30 (0.03)
Electric mixer	0.26 (0.02)	0.09 (0.00)	0.45 (0.03)	0.26 (0.03)
Hand mixer	0.42 (0.03)	0.13 (0.01)	0.56 (0.03)	0.25 (0.02)
Paddle boat	0.32 (0.04)	0.09 (0.01)	0.60 (0.03)	0.16 (0.01)
Pistol	0.25 (0.03)	0.06 (0.01)	0.52 (0.02)	0.20 (0.02)
Scooter	0.15 (0.03)	0.12 (0.01)	0.52 (0.03)	0.17 (0.02)
Sink	0.57 (0.07)	0.06 (0.01)	0.68 (0.03)	0.22 (0.02)
Toilet	0.24 (0.03)	0.07 (0.01)	0.50 (0.02)	0.23 (0.02)
Tricycle	0.18 (0.04)	0.12 (0.01)	0.47 (0.04)	0.26 (0.02)

Both experiments demonstrate that the Interventional Method elicits not only more reported causal relations but also yields more accurate and discriminating models, despite the increased false alarms, compared to the Graphical Method. These findings highlight a fundamental trade-off in causal elicitation: the systematic, counterfactual approach of the Interventional Method offers higher fidelity in capturing causal structure but faces scalability limitations due to its exhaustive requirements. Conversely, the Graphical Method, with its intuitive and efficient design, may be preferable for rapid, large-scale elicitation scenarios, even if some causal nuances remain uncaptured.

3 Hierarchical Causal Elicitation: An Integrated Approach

To address limitations in previous elicitation methods, we introduce a hierarchical framework that integrates the structured rigor of intervention-based techniques with the usability of graphical interfaces. Rather than treating these approaches as mutually exclusive, our method guides participants through sequential stages of causal reasoning, beginning with intuitive graphical mapping followed by targeted counterfactual prompts that refine and validate key connections.

This hybrid design aims to minimize cognitive overload by avoiding exhaustive pairwise queries upfront, focusing instead on areas of ambiguity or uncertainty. By streamlining the elicitation process, the framework preserves the discriminative benefits of interventional methods while enhancing engagement and scalability comparable to graphical approaches.

Implemented as an online interface, the system enables users to construct initial causal graphs and then interactively refine their models through context-sensitive prompts. This iterative structure improves both the accuracy and richness of elicited representations while providing a practical pathway for collecting structured causal knowledge in large-scale or real-time settings.

3.1 Methodology

This study examines whether a hierarchical elicitation interface can reduce cognitive demands typically associated with exhaustive causal questioning while still yielding accurate and interpretable causal models. Our primary research questions are: (1) Can a hierarchical interface produce causal graphs of comparable quality to those generated through traditional survey-based methods? and (2) Can complex objects be decomposed into hierarchical clusters to facilitate more efficient causal elicitation?

We implemented a controlled experimental design with quantitative evaluation metrics. The stimuli consisted of three mechanically distinct everyday objects (a bicycle, a sink, and a toilet) selected for their structural complexity and general familiarity. We recruited 300 participants, with each object assessed by 10 participants per condition. Participants were randomly assigned to one of three condition conditions: structural, functional, or causal (control), designed to subtly influence how they interpreted and grouped object components without introducing explicit bias.

Participants used the hierarchical interface to construct causal models for their assigned object. The interface guided them through a staged process of mapping component relationships, beginning with higher-level groupings and progressively refining connections through targeted prompts.

We evaluated the quality of elicited causal models using quantitative metrics, including hit rate (true positive rate) and false alarm rate (false positive rate), measured against expert ground truth graphs to assess both sensitivity and specificity.

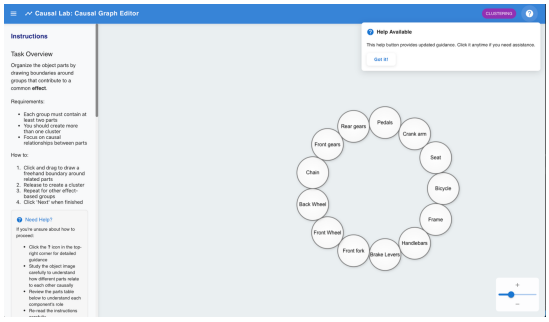
3.2 System Implementation and User Interface

The elicitation system was implemented with React for the frontend and Python Flask for the backend. The core feature is an interactive graph editor that enables participants

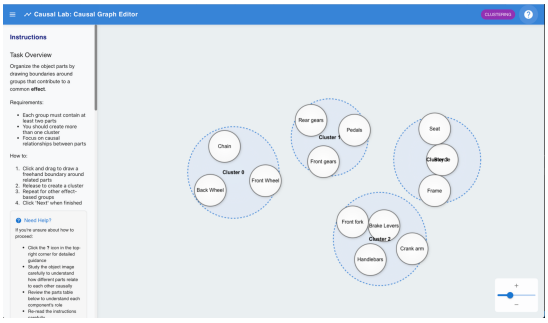
to construct causal models by clustering object parts and drawing directional causal connections. Participants organize components by drawing freehand boundaries that render as smooth elliptical shapes, with dynamic visual feedback provided through gradient fills, Bezier curves, and shadow effects.

The system continuously captures interaction data throughout the process. Every action, from clustering decisions to connection arrows, is recorded automatically. These data are processed to compute statistical measures such as hit rates, false alarm rates, and z-scores, and compared against expert benchmarks in post-experiment analysis. A persistent sidebar implemented via Material-UI's `Drawer` dynamically displays stage-specific instructions, contextual help, and supplementary visual aids to guide users during the experiment.

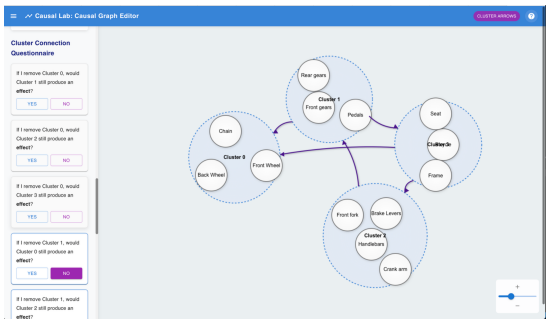
Figures 1(a)–(d) provide a visual overview of the system components, illustrating the editor help interface, initial counterfactual clustering demo, cluster-connection demo, and a sample causal map respectively. Figures 2(a)–(d) showcase the complete workflow from task entry introducing the interface, contextual help, and counterfactual questionnaire guidance/resources. **For more detailed versions of these images see Appendix 8.**



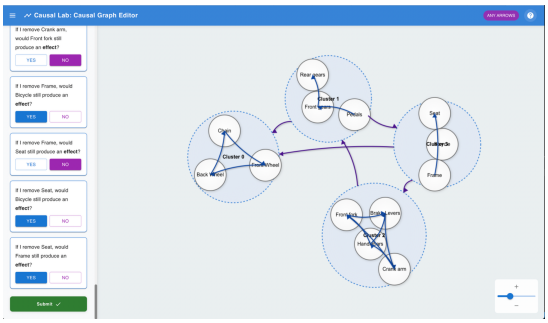
(a) Editor help panel showing keyboard shortcuts, bicycle-parts list, and an instructions sidebar.



(b) Initial clustering demonstration on the bicycle model.

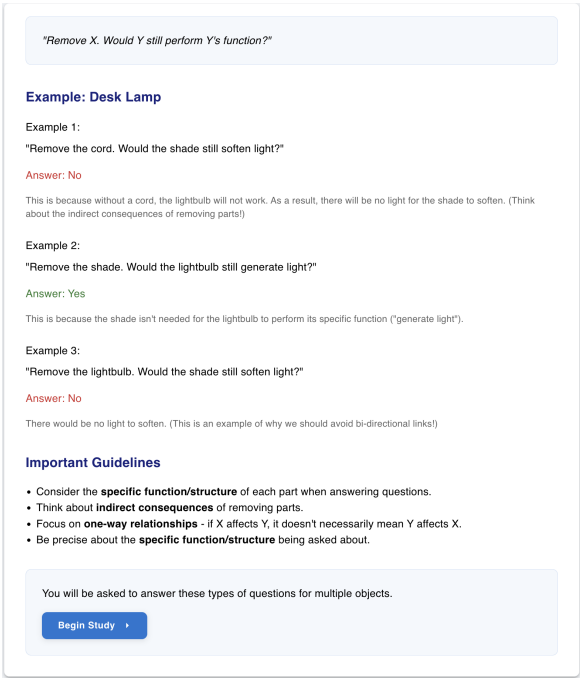


(c) Cluster-to-cluster connection tutorial on the bicycle model.

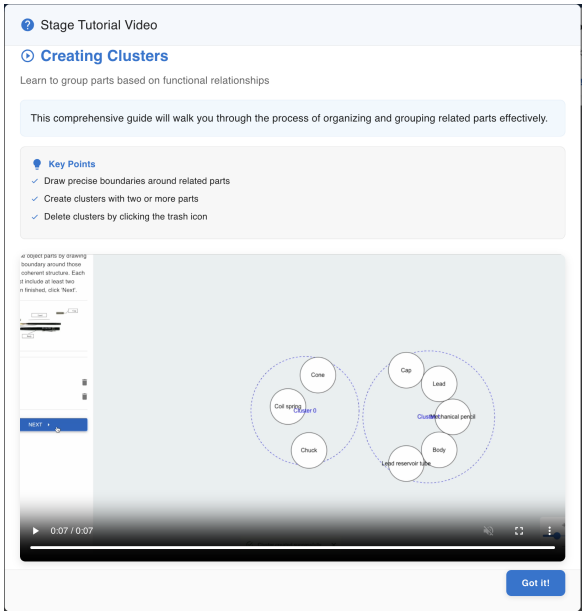


(d) Completed causal map displaying intra- and inter-cluster relationships.

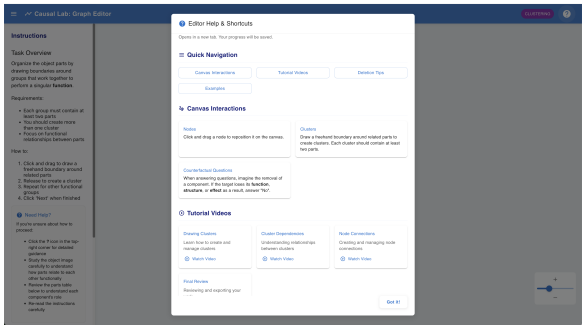
Figure 1: Visual overview of the application's workflow and interfaces: **(a)** Editor help and shortcuts, **(b)** Bicycle clustering demo, **(c)** Cluster-connection tutorial, **(d)** Full causal map with intra-/inter-cluster links.



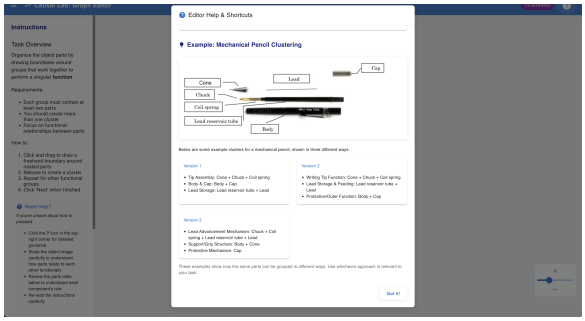
(a) Entry page introducing the interface and counterfactual reasoning task.



(b) Clustering interface with embedded video tutorial.



(c) Help popup with guidance text and resource links for causality.



(d) Help popup illustrating a mechanical pencil causal-map example.

Figure 2: Detailed screenshots of the counterfactual reasoning module: **(a)** Task entry page, **(b)** Clustering with video support, **(c)** Contextual help and resource links, **(d)** Example popup for a sample object - mechanical pencil.

3.3 Stage-wise Experimental Workflow and Question Types

Our experimental protocol comprised five sequential stages that methodically guided participants through causal model construction while minimizing cognitive burden. Table 3 presents this progression from orientation to final submission.

Table 3: Five-stage causal model elicitation workflow.

Stage	**Task Description**
General Instructions	Overview of study methodology and counterfactual reasoning examples
Clustering	Organization of object components based on condition-specific criteria
Cluster Arrows	Assessment of inter-cluster causal dependencies
Node Arrows	Evaluation of component-level relationships within clusters
Final Review	Validation and submission of the complete causal graph

In the **General Instructions** stage, participants received comprehensive orientation to both the interface mechanics and the conceptual foundations of counterfactual reasoning, establishing the cognitive framework for subsequent tasks.

The **Clustering** stage required participants to organize object components through freehand boundary-drawing, guided by condition-specific criteria: *functional* clusters emphasized components collaborating toward specific tasks; *structural* clusters prioritized physical cohesion and spatial arrangement; and *causal* clusters highlighted joint contribution toward particular outcomes, serving as a general causal effect control condition without specific structural or functional emphasis.

During the **Cluster Arrows** stage, participants evaluated inter-cluster dependencies through systematically varied counterfactual prompts:

- *Structural*: "If we remove [Cluster X], would [Cluster Y] maintain its structural integrity?"
- *Functional*: "If we remove [Cluster X], would [Cluster Y] still [perform its specific function]?" (with custom function descriptions)
- *Causal*: "If we remove [Cluster X], would [Cluster Y] still produce its effect?"

The **Node Arrows** stage then narrowed focus to component-level relationships within established clusters, employing parallel question structures across conditions:

- *Structural*: "If we remove [Part X], would [Part Y] maintain its structural integrity?"
- *Functional*: "If we remove [Part X], would [Part Y] still [perform its function]?" (tailored to each component)
- *Causal*: "If we remove [Part X], would [Part Y] still produce its effect?"

In both stages, negative responses generated directed arrows representing dependencies within the causal graph.

Each question type is formulated to evoke distinct cognitive frameworks: functional prompts highlight goal-oriented utility and purpose; structural prompts emphasize physical support relationships; and causal prompts target mechanisms through which components influence outcomes. The causal condition serves as our control baseline, enabling direct comparison with more specialized reasoning modes while maintaining consistent counterfactual structure.

This dual-level approach, applying parallel reasoning at both cluster and component levels, aims to create an effective cognitive scaffold. By first addressing abstract, higher-order relationships between clusters before examining granular component interactions, we seek to facilitate more coherent mental model construction. This hierarchical progression aligns with established principles from the learning sciences, where complex systems comprehension improves when learners begin with macro-level conceptual anchors before transitioning to micro-level analysis. The approach also aims to reduce cognitive load by compartmentalizing the reasoning task, potentially enabling participants to construct more consistent and structured system representations.

3.4 Hierarchical Algorithm for Constructing Causal Graphs

Traditional causal inference approaches often rely on labor-intensive methods involving exhaustive pairwise comparisons or extensive counterfactual queries. While theoretically robust, such approaches are cognitively demanding and frequently lead to participant fatigue and response inconsistencies. We propose a novel hierarchical algorithm that minimizes user burden while maximizing meaningful causal structure extraction from minimal inputs.

The algorithm assumes complex systems can be decomposed into semantically meaningful substructures or clusters representing functional component groupings. By utilizing predefined templates, the algorithm constructs a two-tiered representation: intra-cluster dependencies capture causal relationships within subsystems, while inter-cluster dependencies reflect interactions between subsystems.

Each cluster forms a directed acyclic graph (DAG) with nodes representing components and arrows indicating causal influence. Arrows within clusters culminate at terminal nodes, typically the final output of that subsystem. Inter-cluster connections link terminal nodes of one cluster to initiating nodes of another, forming a higher-order DAG encoding the overall causal architecture.

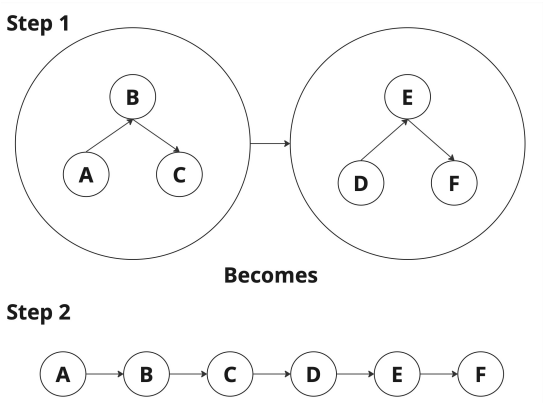


Figure 3: A high-level view of hierarchical causal graph construction. Each cluster contains internal directed flows culminating in a terminal node. Inter-cluster arrows define system-wide causal dependencies based on convergence and initiation points across clusters.

In the interface, participants begin by creating their own clusters, grouping nodes according to their system understanding. Once clusters are defined, they answer counterfactual questions to elicit pairwise causal dependencies between clusters. These generate directed relationships across all cluster pairs. Participants then evaluate pairwise causal relations between individual nodes across and within clusters.

Based on these responses, the algorithm constructs a candidate causal graph incorporating all confirmed intra-cluster and inter-cluster causal links. Unlike systems relying on predefined templates, this method leverages participant intuition to define structure from the bottom up. The final phase synthesizes all validated node-level and cluster-level connections into a unified hierarchical causal graph encoding both micro-level and macro-level causal dependencies.

Algorithm 1 Hierarchical Causal Graph Construction

Require: Object type O, component list C, participant responses R
Match O to template $\mathcal{T}$ with predefined clusters $\mathcal{C}$
for each cluster $c_i \in \mathcal{C}$ **do**
    Initialize DAG $G_i = (V_i, E_i)$ for components in c_i
    Add all possible intra-cluster edges to E_i
end for
Initialize inter-cluster edge set E_{inter}
for each participant response $(u, v, \text{label}) \in R$ **do**
    Update edge direction or weight between nodes u and v
    If $u \in c_i$ and $v \in c_j$ with $i \neq j$, update E_{inter}
end for
Refine E_{inter} based on cluster terminal and initiating nodes
Combine all G_i and E_{inter} into final graph $G = (V, E)$
return G =0

By reducing reliance on direct counterfactual questions and inferring connections from structured object decompositions, this algorithm aims to yield robust and user-friendly causal graphs that effectively integrate automated inference with targeted user input.

3.5 Findings and Comparative Insights

Hit rate (HR) and false alarm rate (FAR) values were computed by comparing participants' responses against expert-derived causal models, ensuring performance metrics reflected conceptually meaningful dependencies rather than superficial co-occurrences.

Across all three objects, participant performance under the *causal* and *functional* conditions followed similar trends in both HR and FAR. This similarity has theoretical grounding: both conditions emphasize goal-directed interactions among components. In mechanical systems, functional roles frequently align with causal relationships; a bicycle pedal's function to rotate the gear is also its causal effect, suggesting conceptual overlap between these reasoning modes.

The *structural* mode demonstrated greater variability, revealing distinct reasoning patterns focused on spatial arrangement and physical adjacency rather than dynamic interactions. Participants reasoning structurally often identified physically adjacent components correctly but also misattributed relevance to parts with no functional or causal role, resulting in higher false alarm rates.

These results suggest structural reasoning facilitates recognition of physical relationships but may obscure dynamic dependencies emphasized in causal and functional models. The convergence between causal and functional reasoning indicates that learners frequently conflate these modes, likely due to their alignment in many mechanical contexts.

Table 4: Means and standard deviations for hit rate (HR, true positive rate) and false alarm rate (FAR) per object and reasoning mode.

Object	**Mode**	**HR Mean**	**HR SD**	**FAR Mean**	**FAR SD**
Bicycle	Structural	0.362	0.155	0.100	0.044
	Causal	0.286	0.137	0.182	0.111
	Functional	0.314	0.123	0.139	0.079
	All Modes	**0.325**	**0.148**	**0.139**	**0.087**
Sink	Structural	0.487	0.250	0.172	0.069
	Causal	0.309	0.215	0.165	0.080
	Functional	0.349	0.241	0.164	0.093
	All Modes	**0.382**	**0.241**	**0.167**	**0.081**
Toilet	Structural	0.378	0.170	0.168	0.113
	Causal	0.368	0.139	0.206	0.120
	Functional	0.407	0.155	0.197	0.113
	All Modes	**0.383**	**0.157**	**0.190**	**0.115**

To evaluate how interaction modalities affect causal understanding, we compared our *hierarchical interface* with the *Drawing* and *Intervention* conditions from Experiment 2, benchmarking all against expert causal models.

Across all objects, the **hierarchical interface achieved intermediate performance**, higher than the unguided Drawing condition but lower than the Intervention condition. For the *Bicycle*, hit rates progressed from 0.25 (Drawing) to 0.325 (Hierarchical) to 0.56 (Intervention), indicating consistent stepwise improvement in causal identification. This trend replicated across all objects.

False alarm rates followed a parallel pattern. The Drawing condition yielded the lowest FARs (e.g., 0.06 for Sink) but with compromised sensitivity, missing many valid connections. The Intervention condition maximized hit rates but introduced more false alarms (e.g., 0.26 for Bicycle), likely due to overgeneralization from explicit cues. The hierarchical interface provided a balanced middle ground without imposing fixed causal schemas.

These results suggest the hierarchical interface scaffolds causal reasoning more effectively than freeform drawing while avoiding pitfalls of direct instructional intervention. By aligning with natural conceptual hierarchies, part-whole structures and flow-based dependencies, it enables coherent mental model construction without constraining exploratory reasoning. However, hit rates below 0.5 across all objects raise concerns about reliability for accurate model elicitation, questioning scalability to domains with greater abstraction or latent causal structures.

Table 5: Comparison of mean hit rates (HR) and false alarm rates (FAR) for three objects under three interfaces: Hierarchical (current study), Drawing, and Intervention (Experiment 2).

Object	**HR (Hier)**	**HR (Draw)**	**HR (Interv)**	**FAR (Hier)**	**FAR (Draw)**	**FAR (Interv)**
Bicycle	0.325	0.25	0.56	0.139	0.07	0.26
Sink	0.382	0.57	0.68	0.167	0.06	0.22
Toilet	0.383	0.24	0.50	0.190	0.07	0.23

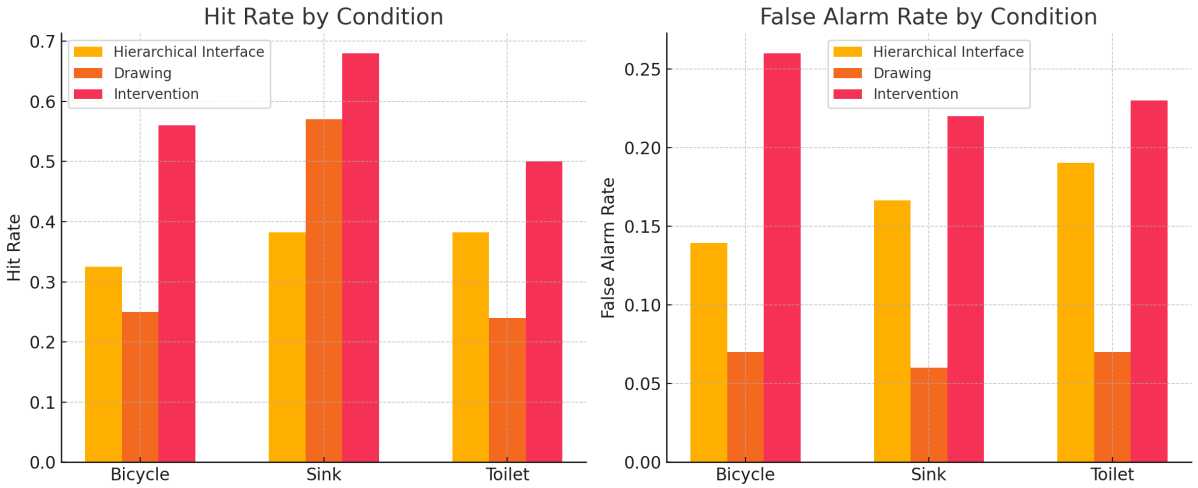


Figure 4: Side-by-side comparison of hit rate (left) and false alarm rate (right) across three interfaces. The hierarchical interface consistently falls between the Drawing and Intervention conditions. See Table 6 for corresponding numerical values.

Given that the Intervention method consistently achieved higher hit rates (HRs), and prior studies demonstrate a significant correlation between higher HRs and better rewards in POMDPs for both troubleshooting and assembly tasks [3, 2, 1, 4], we replicated the counterfactual condition from the Hierarchical interface in the Intervention condition. Participants were asked the same mode-specific counterfactual questions, allowing direct comparison between methods while isolating the effect of reasoning mode on performance.

To understand how reasoning mode interacts with interface type, we compared HR and FAR across Structural, Causal, and Functional conditions for both Hierarchical and Intervention methods (Table 6).

Across all reasoning modes, the Intervention method produced significantly higher HRs. Under causal reasoning with the *sink* object, HR was 0.783 with Intervention, more than double the 0.309 observed with the Hierarchical interface. Similar patterns emerged in other modes, suggesting that direct feedback and mode-specific condition enhance correct link identification and support robust causal reasoning.

This benefit comes with trade-offs. FARs under the Intervention method were substantially higher. In the bicycle-causal condition, FAR increased from 0.182 (Hierarchical) to 0.407 (Intervention), indicating liberal response patterns likely driven by overconfidence under guided conditions. These findings reinforce that interventions prioritizing performance gains may sacrifice precision.

The Hierarchical interface, emphasizing structured exploration and layered decision-making, promoted more conservative model construction. Though HRs were lower, FARs remained moderate, suggesting participants exercised greater caution when identifying links, valuing model accuracy over completeness.

These findings support the conclusion that the Intervention method is more effective for eliciting detailed, actionable causal models for robotic troubleshooting and assembly. While it introduces higher false positive risk, the increased identification of correct links may outweigh this cost in applications where reward maximization depends on successful link recognition, as will be explored in Section 5.

Table 6: Comparison of means and standard deviations for hit rate (HR, true positive rate) and false alarm rate (FAR) per object and reasoning mode between Hierarchical and Intervention interfaces.

Object	**Mode**	**Method**	**HR Mean**	**HR SD**	**FAR Mean**	**FAR SD**
Bicycle	Structural	Hierarchical	0.362	0.155	0.100	0.044
	Structural	Intervention	0.546	0.154	0.288	0.137
	Causal	Hierarchical	0.286	0.137	0.182	0.111
	Causal	Intervention	0.514	0.248	0.407	0.193
	Functional	Hierarchical	0.314	0.123	0.139	0.079
	Functional	Intervention	0.521	0.159	0.297	0.084
	All Modes	Hierarchical	**0.325**	**0.148**	**0.139**	**0.087**
	All Modes	Intervention	**0.527**	**0.189**	**0.329**	**0.152**
Sink	Structural	Hierarchical	0.487	0.250	0.172	0.069
	Structural	Intervention	0.506	0.161	0.397	0.148
	Causal	Hierarchical	0.309	0.215	0.165	0.080
	Causal	Intervention	0.783	0.161	0.329	0.094
	Functional	Hierarchical	0.349	0.241	0.164	0.093
	Functional	Intervention	0.796	0.125	0.339	0.084
	All Modes	Hierarchical	**0.382**	**0.241**	**0.167**	**0.081**
	All Modes	Intervention	**0.695**	**0.200**	**0.355**	**0.115**
Toilet	Structural	Hierarchical	0.378	0.170	0.168	0.113
	Structural	Intervention	0.573	0.147	0.300	0.117
	Causal	Hierarchical	0.368	0.139	0.206	0.120
	Causal	Intervention	0.605	0.102	0.336	0.100
	Functional	Hierarchical	0.407	0.155	0.197	0.113
	Functional	Intervention	0.590	0.127	0.317	0.104
	All Modes	Hierarchical	**0.383**	**0.157**	**0.190**	**0.115**
	All Modes	Intervention	**0.589**	**0.126**	**0.317**	**0.107**

4 Evaluating Distinct Causal Reasoning Modes: Structural vs Functional Causal Models

In previous sections, we investigated techniques to minimize exhaustive questioning while still capturing meaningful causal distinctions, hierarchical object decomposition for complex systems, and the comparative quality of hierarchical versus traditional elicitation approaches. Our findings indicated that the hierarchical method, while efficiently scalable and producing fewer false alarms than intervention-based approaches, still faced challenges in overall accuracy.

We now address a different set of questions: How do participants form clusters and causal links when explicitly prompted to consider structural versus functional relationships? Do different prompt framings influence the resulting causal maps, or do participants ultimately converge on similar representations regardless of phrasing?

4.1 Methodology

To investigate these questions, we compared the hierarchical interface with a modified intervention-based approach. This revised intervention method incorporated structured condition to guide participants toward reasoning in structural, functional, or unspecified causal terms.

We evaluated the resulting causal graphs across three experimental conditions (structural, functional, and unspecified prompting), focusing on whether graphs created within the same condition demonstrated greater similarity to each other than to graphs from different conditions. Our analysis employed four complementary graph comparison metrics:

- **Hamming Distance**: Measures the number of differing edges between graphs. High values indicate divergent causal attributions at the edge level, while low scores suggest convergence in specific causal links.
- **Jaccard Similarity**: Calculates the ratio of shared edges to the union of all edges, providing a normalized measure especially sensitive to sparse graph structures.
- **Spectral Distance**: Compares eigenvalue spectra of graph adjacency matrices, capturing global structural patterns including connectivity and clustering. This metric reveals whether participants formed similarly organized causal models beyond specific edge configurations.
- **Deltacon** [15]: Measures overall graph similarity using node affinity matrices, balancing sensitivity to both local and global structural changes. This metric particularly captures perceptual similarity as experienced by humans.

For each metric, we calculated average within-group similarity (comparing graphs from the same experimental condition) versus between-group similarity (comparing graphs from different conditions), using permutation tests to determine statistical significance. Significant differences would indicate that prompt framing systematically influences causal conceptualization strategies. Together, these metrics provide comprehensive assessment at both individual causal link and overall structural organization levels. Statistically significant differences would support the hypothesis that causal reasoning about physical systems can be decomposed into distinct structural and functional modes.

4.2 Hierarchical Interface Method

Across three systems (Bicycle, Sink, and Toilet), participants' graph constructions varied based on their assigned reasoning condition. Through hierarchical node clustering followed by counterfactual questioning both within and across clusters, we observed how participants both grouped components and inferred dependencies between them. The Sink object demonstrated the most pronounced differentiation, with significant differences across nearly all pairwise comparisons. In contrast, the Bicycle showed weaker differentiation, while the Toilet exhibited selective divergence primarily between structural and causal modes.

Table 7: Permutation Test p-values for Pairwise Mode Comparisons Across Metrics and Objects (Hierarchical Interface)

Object	**Comparison**	**Hamming**	**Jaccard**	**Spectral**	**DeltaCon**
Sink	Structural vs Functional	0.012	0.001	0.594	0.003
	Functional vs Causal	0.006	0.189	0.044	0.011
	Structural vs Causal	0.002	0.001	0.124	0.000
Bicycle	Structural vs Functional	0.521	0.087	0.264	0.208
	Functional vs Causal	0.702	0.785	0.648	0.578
	Structural vs Causal	0.682	0.072	0.713	0.447
Toilet	Structural vs Functional	0.258	0.169	0.406	0.258
	Functional vs Causal	0.080	0.859	0.617	0.130
	Structural vs Causal	0.227	0.070	0.140	0.126

Our results provide tentative evidence that structural, functional, and causal reasoning frames potentially reflect distinct inferential strategies corresponding to different types of causal reasoning. This pattern was most evident in the Sink object, where all three pairwise comparisons showed statistically significant differences in at least one metric. The structural versus causal comparison yielded significant Jaccard similarity difference ($p = 0.001$) and DeltaCon difference ($p = 0.000$), while functional versus causal differed significantly in spectral distance ($p = 0.044$). The consistent divergence between structural and functional reasoning (Jaccard $p = 0.001$) further supports that these frames lead to distinguishable system representations.

However, this separation was not uniformly observed across all objects. The Bicycle task showed no significant differences across any pairwise comparison, while the Toilet object exhibited only marginal trends, particularly between structural and causal reasoning (Jaccard $p = 0.070$). This variability suggests that the expression of distinct reasoning strategies may depend on object-specific features such as complexity, familiarity, or component role ambiguity.

These mixed findings complicate straightforward interpretation. Rather than a generalizable distinction among reasoning frames, the results point toward context-sensitive differences emerging under specific conditions. The Sink object, with its ambiguous affordances (such as the tailpiece being both spatially adjacent and functionally instrumental), may have encouraged broader variation in causal interpretation.

The task design itself may have contributed to these effects. The hierarchical interface requirement to first group components before reasoning about counterfactual changes may

have amplified or restricted certain reasoning strategies. Participants in structural and functional conditions might have anchored more rigidly to initial groupings, while those in the causal condition reasoned more fluidly across clusters. For complex objects like the Bicycle or Toilet, this interface could have introduced unnecessary cognitive load, potentially diminishing differences between conditions.

While the Sink results suggest that different reasoning frames can yield meaningfully distinct representations, this interpretation requires caution. The inconsistent effects across domains raise the possibility that they reflect task design artifacts rather than genuine causal ontology differences. Features of the hierarchical interface such as forced clustering sequence or spatial layout mechanics may have subtly influenced representational structure.

4.3 Intervention Method

Complex Objects

To disentangle potential confounds mentioned previously, we employed the intervention method as a complementary approach. By reexamining participants' representations through this alternative methodology, we aimed to determine whether distinctions among reasoning frames were robust or merely artifacts of the original interface design.

Across all three complex systems, participants' mental representations were systematically shaped by their assigned reasoning frame. As shown in Table 8, permutation tests yielded statistically significant differences ($p < 0.05$) for nearly all pairwise comparisons, especially when contrasting causal reasoning with structural or functional reasoning. These effects suggest that reasoning prompts elicit distinct representational strategies rather than merely influencing superficial graph features.

Table 8: Permutation Test p-values for Pairwise Mode Comparisons Across Metrics and Objects (Intervention Method)

Object	Comparison	Hamming	Jaccard	Spectral	DeltaCon
Sink	Structural vs Functional	0.000	0.000	0.009	0.000
	Functional vs Causal	0.097	0.035	0.606	0.103
	Structural vs Causal	0.000	0.000	0.061	0.000
Bicycle	Structural vs Functional	0.033	0.000	0.186	0.004
	Functional vs Causal	0.003	0.000	0.006	0.001
	Structural vs Causal	0.008	0.003	0.020	0.003
Toilet	Structural vs Functional	0.001	0.000	0.001	0.000
	Functional vs Causal	0.000	0.000	0.003	0.001
	Structural vs Causal	0.000	0.000	0.000	0.000

Causal reasoning yielded the most consistent and pronounced divergence. For the Toilet object, the structural versus causal comparison yielded a Spectral Distance of 3.095 (within-structural) and 3.007 (within-causal), with a between-group mean of 3.353 and an observed difference of 0.302 ($p = 0.000$). DeltaCon Similarity showed a corresponding pattern: 0.640 (structural), 0.622 (causal), and 0.620 between groups ($p = 0.000$). These results indicate that causal reasoning induced not only local changes in edge configuration

(reflected in Hamming and Jaccard) but also large-scale reorganization of graph topology, suggesting a shift toward representing dynamic processes and influence propagation.

This shift appears to reflect a qualitatively different cognitive process. While structural reasoning emphasizes spatial arrangement and component adjacency, and functional reasoning focuses on utility or goal-directed activity, causal reasoning promotes abstraction over temporal dependencies and inferred mechanisms. Participants engaged in causal reasoning appear to build models structured around latent, propagative relationships, redefining not just what connects to what, but why components matter to the system holistically.

Though more similar to each other than to causal reasoning, structural and functional reasoning also showed meaningful differences. In the Bicycle dataset, the structural versus functional comparison yielded a Hamming Distance of 0.280 (structural), 0.239 (functional), with a between-group mean of 0.269 and an observed difference of 0.007 ($p = 0.030$). DeltaCon Similarity for the same pair was 0.611 (structural), 0.625 (functional), and 0.610 between groups ($p = 0.004$). These results suggest that while both reasoning types may draw on observable features, their inferential emphases diverge: structural reasoning grounds itself in compositional integrity and containment, whereas functional reasoning directs attention toward interactive affordances and task-related mechanisms.

The Toilet object revealed strong and consistent separability among all three reasoning modes. For structural versus causal, the Jaccard Similarity was 0.269 (structural) versus 0.356 (causal), with a between-group mean of 0.283 and an observed difference of 0.030 ($p = 0.000$). This high separability may reflect the system's complexity or interpretive ambiguity, which likely increased reliance on the framing condition. When component functions are opaque or highly interdependent, participants appear especially sensitive to prompt guidance.

Spectral Distance and DeltaCon metrics were particularly informative in highlighting global topological effects. While Hamming and Jaccard captured edge-level discrepancies, Spectral Distance revealed shifts in graph modularity and connectivity, and DeltaCon reflected changes in perceived paths of information flow. The consistent elevation of Spectral Distance in structural versus causal comparisons (e.g., Sink, $p = 0.061$ and Bicycle, $p = 0.020$) supports the claim that causal reasoning leads to restructured conceptual models oriented around dependency chains and modular influence.

These results demonstrate that structural, functional, and causal reasoning modes all lead to significantly distinct graph representations, with causal reasoning exerting a particularly transformative influence on system modeling. These distinctions reflect different inferential commitments and mental modeling strategies prompted by each reasoning frame.

Extension: Lamp Objects

To assess whether reasoning frame effects generalize beyond large, complex systems, we extended our study to include four smaller-scale, more familiar household objects: desk lamp, wall lamp, kerosene lamp, and flashlight. These objects vary in complexity, internal mechanism visibility, and functional ambiguity, offering a valuable testbed for evaluating the robustness of frame-induced differences.

Across all four objects, participants' graph representations remained sensitive to assigned reasoning frames. Crucially, the same core pattern persisted: unspecified causal reasoning consistently produced graphs topologically distinct from both structural and

Table 9: Permutation Test p-values for Pairwise Mode Comparisons Across Metrics and Lamp Objects

Object	**Comparison**	**Hamming**	**Jaccard**	**Spectral**	**DeltaCon**
Desk Lamp	Structural vs Functional	0.000	0.000	0.000	0.000
	Functional vs Causal	0.001	0.000	0.012	0.000
	Structural vs Causal	0.000	0.000	0.000	0.000
Wall Lamp	Structural vs Functional	0.000	0.000	0.676	0.000
	Functional vs Causal	0.011	0.004	0.382	0.007
	Structural vs Causal	0.000	0.000	0.051	0.000
Kerosene Lamp	Structural vs Functional	0.000	0.000	0.070	0.000
	Functional vs Causal	0.069	0.051	0.447	0.050
	Structural vs Causal	0.000	0.000	0.011	0.000
Flashlight	Structural vs Functional	0.000	0.000	0.037	0.000
	Functional vs Causal	0.000	0.000	0.000	0.000
	Structural vs Causal	0.000	0.000	0.000	0.000

functional reasoning. However, new insights emerged regarding the relative separability of structural and functional reasoning, especially in simpler systems.

In less mechanically complex objects such as the desk lamp and flashlight, structural and functional reasoning led to significantly different graph structures across all metrics ($p < 0.001$). This highlights that differences between structural and functional representations are not merely a byproduct of task complexity but reflect fundamental differences in the inferential lens through which participants interpret systems.

Interestingly, object-specific patterns suggest interaction effects between object transparency and reasoning frame. For the wall lamp and kerosene lamp (objects with potentially hidden mechanisms), Spectral Distance did not significantly differ between structural and functional groups (e.g., Wall Lamp, $p = 0.676$). However, local differences captured by Hamming and Jaccard remained robust, indicating that even when global topology appears similar, fine-grained edge-level interpretations diverge based on prompt framing.

The causal condition again yielded the most pronounced reorganization. In the Desk Lamp dataset, the Spectral Distance between structural and causal groups showed striking contrast (0.992 vs. 1.971 within-group means; between-group mean = 2.036; $p = 0.000$), while DeltaCon reflected major shifts in perceived influence pathways. These patterns reiterate that causal reasoning reshapes mental models along abstract, dynamic dimensions, not simply by adding or removing connections but by altering how information and effects are expected to flow through the system.

The lamp object findings confirm and extend our previous results. Reasoning frames consistently shape mental representations across systems of all sizes, with even closely related modes (structural and functional) producing distinct graph structures. These effects persist even for intuitively graspable objects, highlighting framing's importance in shaping cognition. Notably, structural versus functional differences appear more pronounced in lamp objects, likely because these simpler mechanical systems have clearer gravitational influences, making structural relationships more easily distinguishable from functional processes.

5 Causally Informed Planning under Uncertainty for Object Assembly and Troubleshooting

We have seen evidence that conditioning users to think more structurally, functionally, or causally leads to different types of causal models, but are these models useful for robotic assembly and troubleshooting?

5.1 Baseline Evaluation with Expert Ground Truth Models

We begin by feeding our two expert ground truth models, *structural* and *functional*, directly into the POMDP to establish an empirical baseline against which all subsequent embedding-based experiments can be compared. Specifically, we instantiate two instances of the POMDP simulator, each configured with the complete list of true connections and a confidence parameter of 0.9. This baseline setup yields the maximum achievable planning returns and isolates the impact of model uncertainty.

In *assembly*, the agent views the object as a set of N parts with $\binom{N}{2}$ possible connections, of which E are true edges. Each probe incurs a cost of -1 (and -10 for any repeated probe), and upon discovering all E true edges the agent receives a completion bonus of $+\binom{N}{2}$. If the expert ground truth proposes a set P of connections, we define

$$\mathrm{HR} \;=\; \frac{\left|P \cap T\right|}{E}, \qquad \mathrm{FP} \;=\; \left|P \setminus T\right|,$$

where T is the true edge set. Since the total probes required is $E + \mathrm{FP}$, the assembly reward is

$$R_{\text{assembly}} \;=\; \binom{N}{2} \;-\; \left(E + \mathrm{FP}\right),$$

so that higher hit rate (i.e., smaller FP) increases reward linearly.

In *troubleshooting*, a known set of E_e erroneous connections T_e is injected; each flip of a connection costs -1 (and -10 if re-flipping), and a bonus of $+\binom{N}{2}$ is awarded once all E_e errors have been corrected. If the model identifies a set P_e of flips, we define

$$\mathrm{HR}_e \;=\; \frac{\left|P_e \cap T_e\right|}{E_e}, \qquad \mathrm{FP}_e \;=\; \left|P_e \setminus T_e\right|.$$

Since the total flips is $E_e + \mathrm{FP}_e$, the troubleshooting reward is

$$R_{\text{troubleshoot}} \;=\; \binom{N}{2} \;-\; \left(E_e + \mathrm{FP}_e\right),$$

so that higher hit rate (i.e., larger HR_e and smaller FP_e) yields proportionally higher returns.

Figure 5 shows that the structural ground truth attains near-perfect assembly performance—100.0% on *Wall* and *Flash*, 93.3% on *Kero*, 92.0% on *Desk*, 84.3% on *Sink*, 86.4% on *Toilet*, and 83.8% on *Bicycle*—for an overall mean of 91.4%. In contrast, the functional ground truth omits non-causal links and thus probes extra edges, yielding only 100.0% on *Wall*, but dropping to 50.5% on *Flash*, 28.3% on *Kero*, 42.5% on *Desk*, 84.4% on *Sink*, 19.1% on *Toilet*, and 19.2% on *Bicycle*, for a mean of 49.1%.

As seen in Figure 6a, the functional ground truth generally outperforms the structural in troubleshooting by precisely targeting causal failures: it achieves 86.2% vs. 68.5% on

Wall, 72.0% vs. 61.8% on *Flash*, 74.8% vs. 63.4% on *Kero*, 81.1% vs. 63.8% on *Desk*, and 86.9% vs. 81.6% on *Sink*, for an average of 78.6% compared to 71.6% for the structural ground truth. Notably, *Toilet* and *Bicycle* reverse this trend (69.7% vs. 81.4%, and 79.8% vs. 80.5%, respectively) because our error-location annotation included certain structurally plausible links—such as valve assemblies in the toilet and chain-pedal interactions on the bicycle—that were omitted from the purely causal model.

Figure 6b succinctly confirms that the structural ground truth maximizes assembly returns (91.4% vs. 49.1%), while the functional ground truth maximizes troubleshooting returns (78.6% vs. 71.6%). Occasional object-specific reversals reflect the interplay between physical plausibility and causal relevance in our unbiased error-location process, suggesting similar exceptions in participant models.

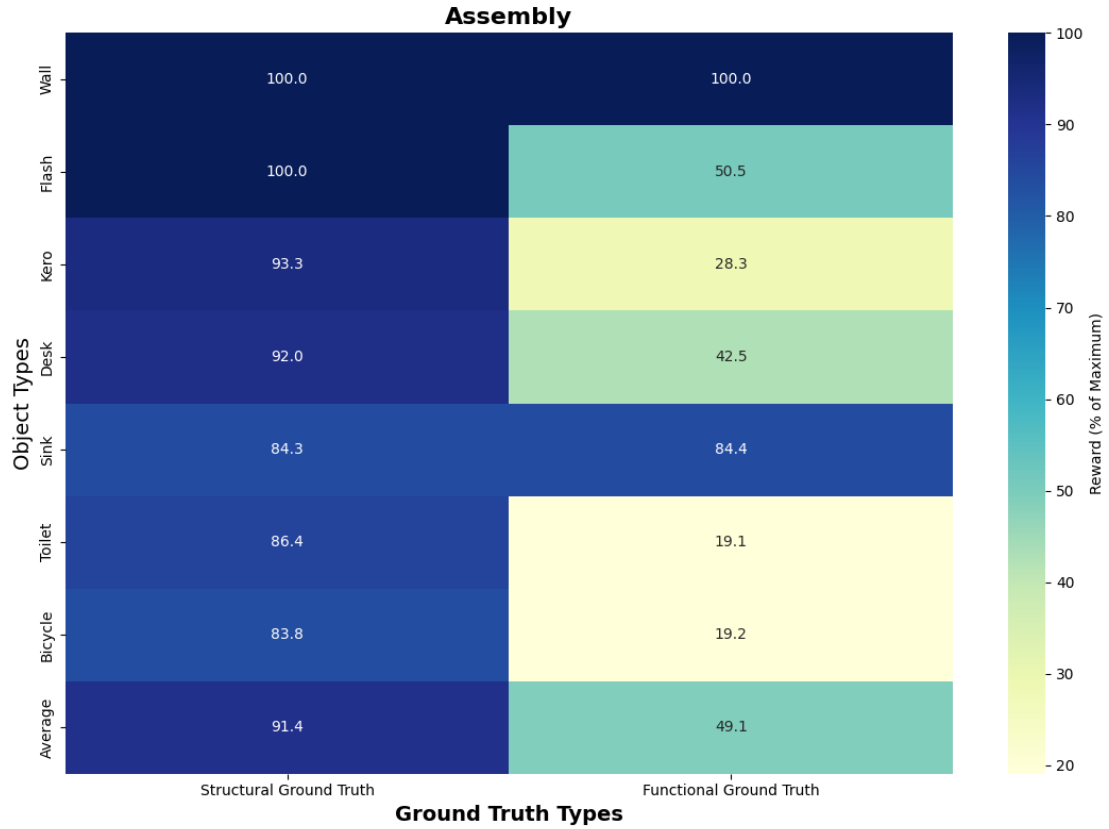
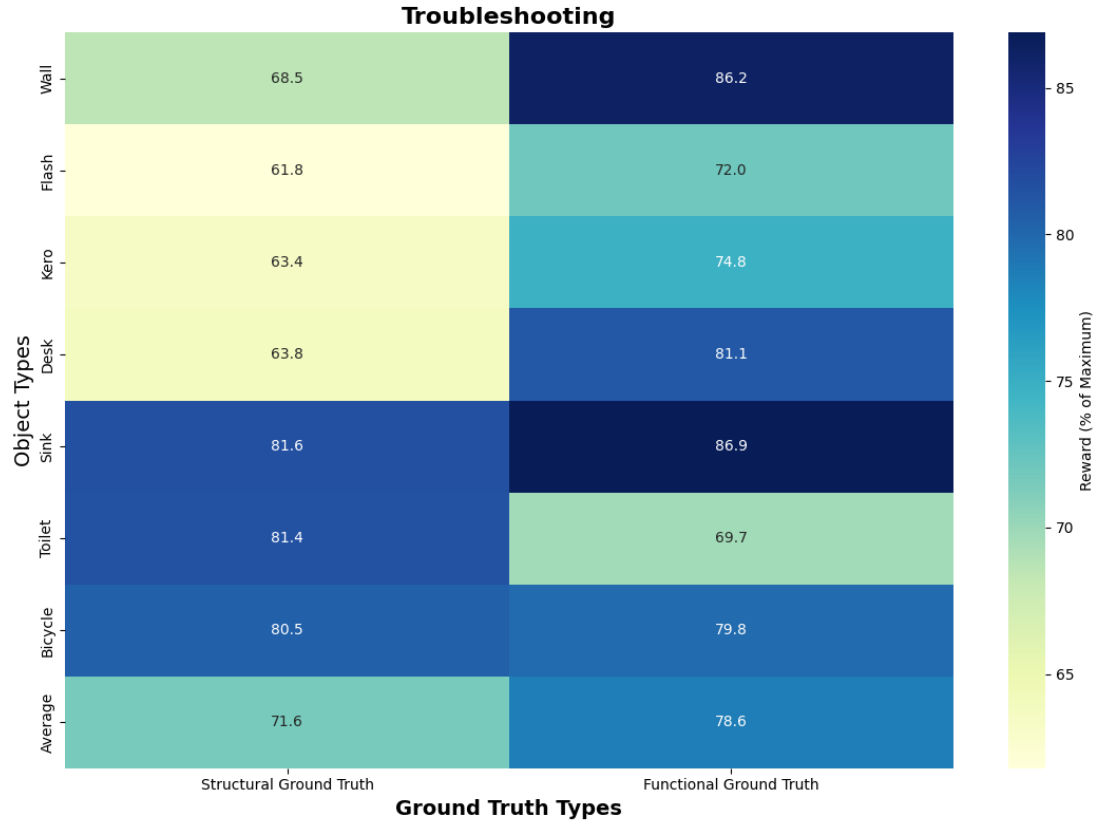
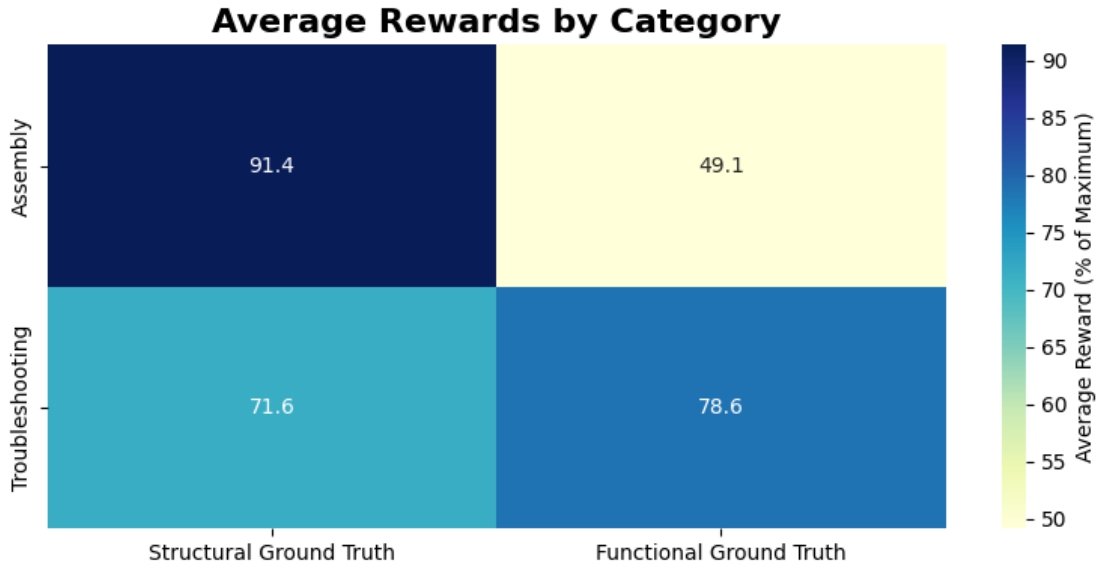


Figure 5: Normalized assembly rewards (% of maximum) under structural vs. functional expert ground truths.



(a) Normalized troubleshooting rewards (% of maximum) under structural vs. functional expert ground truths.



(b) Average rewards by category (% of maximum) for structural vs. functional expert ground truths.

Figure 6: Expert ground-truth performance for (a) troubleshooting and (b) average reward comparisons.

5.2 Assembly

Statistical Analysis of Reasoning Conditions Across Objects

The reward distributions under different reasoning conditions are illustrated in Figure 7. This heatmap compares normalized task-completion rewards across seven object categories (*Wall*, *Flash*, *Kero*, *Desk*, *Sink*, *Toilet*, and *Bicycle*) for each of the three conditions (*Structural*, *Functional*, and *Causal*), alongside the expert benchmark and the average user performance from a prior, control iteration.

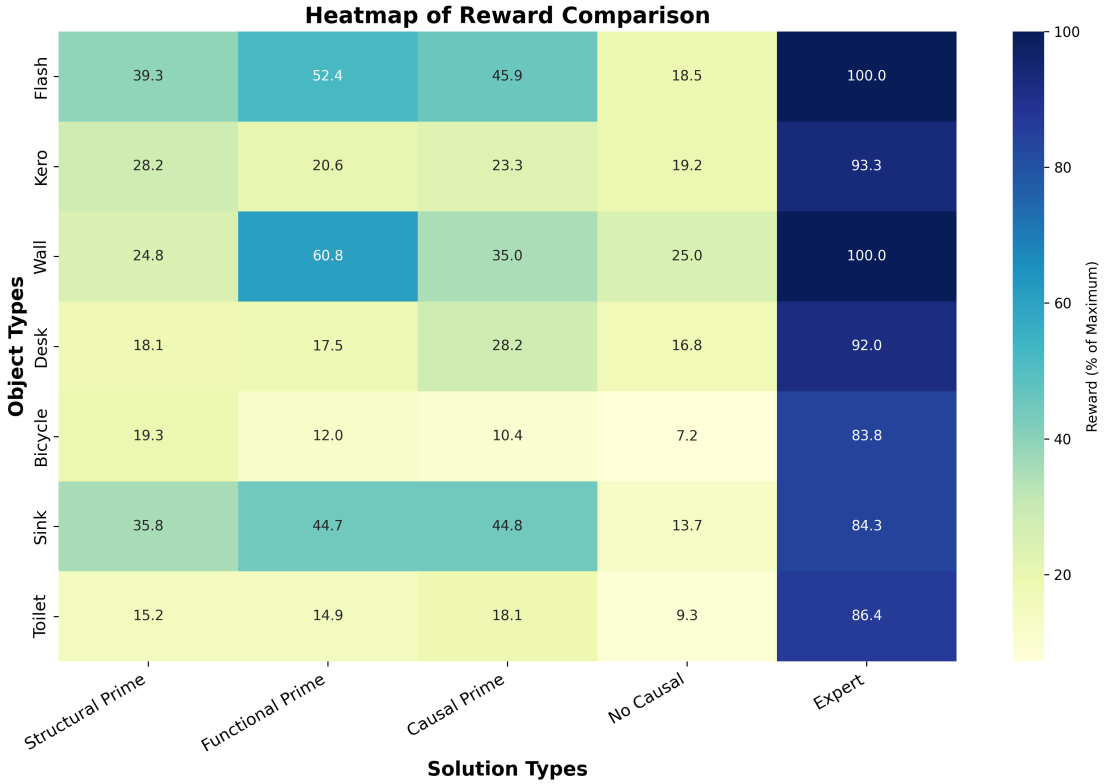


Figure 7: Normalized rewards for each object under different reasoning conditions compared to expert benchmarks and prior user averages. Higher values indicate better task performance.

Several patterns emerge. For complex, goal-directed objects, causal framing yields the largest gains. For example, the *Toilet* under the Causal condition (18.1) exceeds both the Structural (15.2) and Functional (14.9) conditions. Similarly, for the *Sink*, Causal (44.8) and Functional (44.7) both outperform Structural (35.8), suggesting that understanding component functions or causes aids troubleshooting of fluid systems.

In contrast, objects with more familiar, physical-assembly demands exhibit different trends. For the *Desk*, the Causal condition (28.2) substantially outperforms both Structural (18.1) and Functional (17.5), reflecting that causal insight drives efficient assembly of simple furniture. Meanwhile, the *Wall* shows the strongest benefit from Functional (60.8) over Structural (24.8) and Causal (35.0), indicating that component-affordance reasoning best supports wall-mount tasks.

Finally, the *Bicycle*, the most intricate object, yields low rewards across all modes (Structural: 19.3; Functional: 12.0; Causal: 10.4), with a modest advantage for Structural, reflecting reliance on physical part mapping when confronting novel mechanical

assemblies. Across all objects, expert rewards (scaled to 100) provide an upper-bound reference.

5.3 Pairwise Post-hoc Comparisons

To test these observations formally, we ran Kruskal–Wallis H tests for each object ($\alpha = 0.05$). Significant omnibus effects were found for Bicycle ($p < 0.0001$), Wall ($p = 0.0001$), Sink ($p = 0.0364$), and Desk ($p = 0.0002$), while Toilet ($p = 0.3631$), Kero ($p = 0.0799$), and Flash ($p = 0.0823$) showed no overall differences. We then performed Bonferroni-corrected Dunn's pairwise tests to locate specific contrasts. The most pronounced pairwise effects are summarized in Table 10. Notably:

- **Bicycle**: Structural differs from both Functional ($p = 0.0126$) and Causal ($p < 0.0001$), confirming that part-structure framing uniquely boosts this complex assembly.
- **Wall**: Functional exceeds Structural ($p = 0.0001$) and Causal ($p = 0.0205$), reflecting the advantage of component-affordance reasoning.
- **Desk**: Causal outperforms Functional ($p = 0.0002$) and Structural ($p = 0.0007$), aligning with the highest rewards under causal framing.
- **Sink**: Although the omnibus test is significant, no pairwise contrast survives correction, suggesting only marginal differences.
- **Toilet**, **Kero**, **Flash**: No significant pairwise differences, indicating reasoning mode had little effect.

Table 10: Bonferroni-corrected Dunn's Test p-values for Pairwise Mode Comparisons

Object	Structural vs Functional	Functional vs Causal	Structural vs Causal
Bicycle	**0.0126**	n.s.	< 0.0001
Toilet	n.s.	n.s.	n.s. ($p = 0.3631$)
Sink	n.s.	n.s.	n.s.
Flash	n.s.	n.s.	n.s. ($p = 0.0823$)
Desk	n.s.	**0.0002**	**0.0007**
Wall	**0.0001**	**0.0205**	n.s.
Kero	n.s.	n.s.	n.s. ($p = 0.0799$)

Structural yielded significantly higher rewards than functional for the *Bicycle* ($p = 0.0126$), indicating a clear structural advantage, whereas functional outperformed structural for the *Wall* ($p = 0.0001$). For the remaining objects (i.e., *Desk*, *Sink*, *Toilet*, *Flash*, and *Kero*), the structural versus functional comparisons were non-significant.

Alignment of Subject Rewards with Expert Ground Truths

To assess whether participants' elicited rewards under functional and structural conditions approach the corresponding expert benchmarks (Figure 5), we directly compare the mean normalized rewards R_{func} and R_{struct} against the functional and structural expert values. Because reward is driven by hit rate (HR), each condition should yield R_{func} nearer the functional-GT values and R_{struct} nearer the structural-GT values.

Table 11: Per-artifact means of hit rate (HR), false positives (FP), false negatives (FN), and normalized reward under functional vs. structural conditions ($N = 60$ per condition). The ΔR column has been omitted to focus on each condition's absolute alignment with its ground truth.

Object	HR_{func}	$\text{HR}_{\text{struct}}$	FP_{func}	$\text{FP}_{\text{struct}}$	FN_{func}	$\text{FN}_{\text{struct}}$	R_{func}	R_{struct}
Bicycle	0.613	0.721	29.800	29.133	5.033	3.633	0.120	0.193
Desk	0.733	0.274	5.167	4.148	1.333	3.630	0.175	0.181
Flash	0.800	0.556	1.467	2.778	0.800	1.778	0.524	0.393
Kero	0.700	0.602	5.567	4.111	1.200	1.593	0.206	0.282
Sink	0.614	0.562	9.900	10.933	2.700	3.067	0.447	0.358
Toilet	0.655	0.639	23.833	21.367	3.800	3.967	0.149	0.152
Wall	0.862	0.530	3.172	2.818	0.414	1.409	0.608	0.248

Across the seven artifacts, we observe:

- For **Bicycle**, $R_{\text{struct}} = 0.193 \ll 0.838$ (structural-GT) and $R_{\text{func}} = 0.120 \ll 0.192$ (functional-GT), but the ordering $0.193 > 0.120$ matches the expert ordering ($83.8 > 19.2$).
- For **Desk**, $0.181 \ll 0.920$ vs. $0.175 \ll 0.425$, with $0.181 > 0.175$ mirroring $92.0 > 42.5$.
- For **Kero**, $0.282 \ll 0.933$ vs. $0.206 \ll 0.283$, again $0.282 > 0.206$ as $93.3 > 28.3$.
- For **Sink**, $0.358 \ll 0.843$ vs. $0.447 \ll 0.844$, and $0.358 < 0.447$ follows $84.3 < 84.4$.
- For **Toilet**, $0.152 \ll 0.864$ vs. $0.149 \ll 0.191$, with $0.152 > 0.149$ matching $86.4 > 19.1$.
- For **Flash**, however, $R_{\text{struct}} = 0.393$ and $R_{\text{func}} = 0.524$ invert the expert ordering ($100 > 50.5$), so structural underperforms.
- For **Wall**, 0.248 vs. 0.608 likewise contradicts the equal expert returns ($100 = 100$), as functional dominates.

Further, every elicited reward is a small fraction of its expert benchmark (e.g., 0.282 vs. 0.933), reflecting that the subject hit rates (0.274–0.862) remain well below perfect link recovery. In sum, our user models generally align with the correct ordering expected in comparison to the expert ground truths for five of seven objects (though not all differences are statistically significant) but fail to approach expert performance and produce clear reversals for Flash and Wall.

5.4 Troubleshooting

To assess the impact of the different causal reasoning conditions on troubleshooting tasks, we analyzed participants' performance across three condition conditions (i.e., *structural*, *functional*, and *causal*) using normalized reward scores for each object, with participants' embeddings used as priors. Normalization was performed relative to each object's maximum achievable reward (e.g., 77 for *Bicycle*, 20 for *Desk*, and 9 for both *Wall* and *Flash*).

The results were visualized in a heatmap (Figure 8) and evaluated using Kruskal-Wallis tests followed by Bonferroni-corrected Dunn's post-hoc tests for pairwise mode comparisons (Table 12).

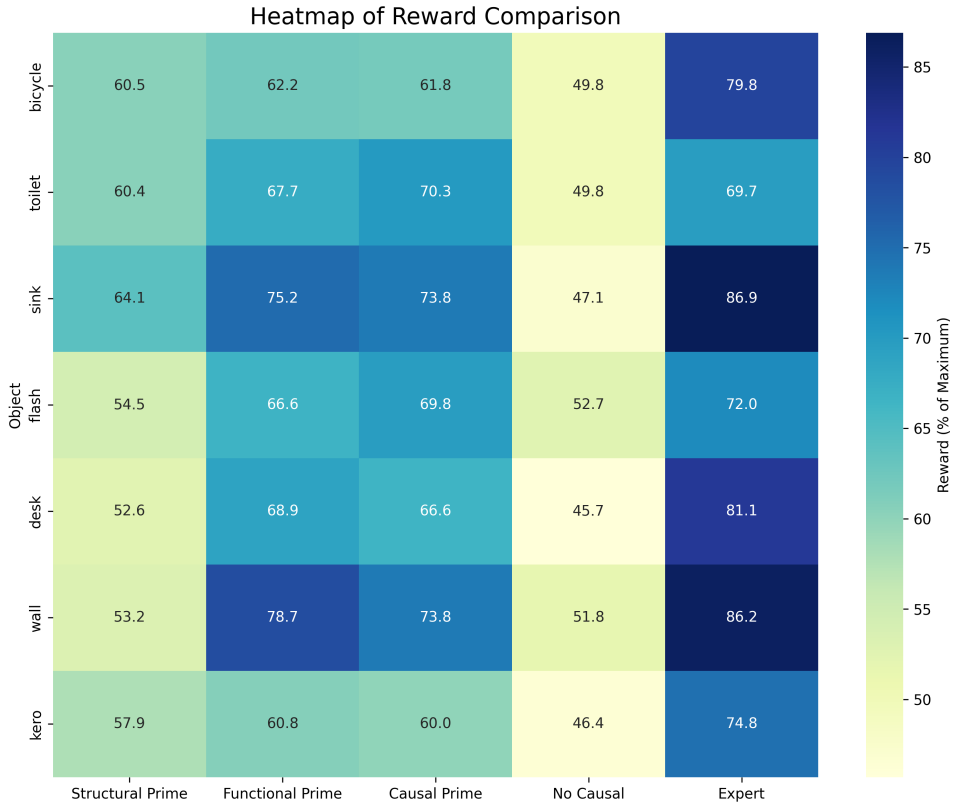


Figure 8: Heatmap of Average Normalized Reward (% of Maximum) Across Objects and Condition Conditions.

Table 12: Bonferroni-corrected Dunn's Test p-values for Pairwise Mode Comparisons (Structural, Functional, Causal). Non-significant comparisons are marked "n.s."

Object	**Structural vs Functional**	**Functional vs Causal**	**Structural vs Causal**
Desk	6.78×10^{-8}	7.73×10^{-1}	5.01×10^{-7}
Flash	1.47×10^{-6}	2.77×10^{-1}	1.24×10^{-8}
Wall	7.93×10^{-8}	9.52×10^{-1}	2.28×10^{-5}
Sink	1.94×10^{-4}	n.s. ($p = 1.00$)	2.34×10^{-4}
Toilet	1.57×10^{-3}	n.s. ($p = 1.00$)	3.40×10^{-5}
Kero	n.s. ($p = 0.0708$)	n.s.	n.s.
Bicycle	n.s. ($p = 0.4578$)	n.s.	n.s.

The results indicate that condition had a statistically significant effect for five of the seven objects (Desk, Flash, Wall, Sink, and Toilet). In each of these cases, **structural condition** yielded significantly lower normalized rewards compared to both **functional** and **causal condition**, suggesting that when participants were guided by purely structural framing, they omitted critical dependencies needed for efficient troubleshooting. By contrast, **functional and causal condition** produced comparable performance (all

functional vs. causal comparisons were non-significant), implying overlapping cognitive representations or flexible strategy shifts between these schemas.

For the *Desk* task, average normalized rewards were 52.6% (structural), 68.9% (functional), and 66.6% (causal), with highly significant differences between structural vs. functional ($p = 6.78 \times 10^{-8}$) and structural vs. causal ($p = 5.01 \times 10^{-7}$). Similarly, in the *Flash* condition, structural condition led to 54.5% reward compared to 66.6% and 69.8% under functional and causal modes, respectively (structural vs. functional $p = 1.47 \times 10^{-6}$; structural vs. causal $p = 1.24 \times 10^{-8}$), underscoring the importance of functional dependencies in circuit-style reasoning.

In contrast, *Kerosene* and *Bicycle* did not exhibit significant differences across condition conditions (Kruskal-Wallis $p = 0.0708$ and $p = 0.4578$, respectively). Although the descriptive mean reward for Bicycle under structural condition was higher ($\Delta R = 7.3$; see Table 13), this did not reach statistical significance, indicating that the observed gain likely reflects sampling variability rather than a systematic effect of the structural prior.

Finally, expert solutions achieved normalized rewards between 69.7% (*Toilet*) and 86.9% (*Sink*), substantially outperforming all participant conditions. Conversely, the no-condition baseline consistently produced the lowest scores (e.g., Desk 45.7%, Flash 52.7%), confirming the overall benefit of any structured cognitive scaffold over an unframed approach.

Alignment of Troubleshooting Rewards with Expert Ground Truths

Similar to how we compared our user models to the expert ground truth baselines for assembly, we directly compare the mean normalized rewards R_{func} and R_{struct} against the functional and structural expert ground truth returns for the troubleshooting task. Because reward is driven by hit rates, each condition should yield a normalized reward closer to its matching ground truth.

Table 13: Per-artifact means of hit rate (HR), false positive (FP), false negative (FN), and normalized reward under functional vs. structural condition (troubleshooting; $N = 60$ per condition).

Object	HR_{func}	$\text{HR}_{\text{struct}}$	FP_{func}	$\text{FP}_{\text{struct}}$	FN_{func}	$\text{FN}_{\text{struct}}$	R_{func}	R_{struct}	ΔR
Bicycle	0.613	0.721	29.800	29.133	5.033	3.633	12.0%	19.3%	+7.3%
Desk	0.733	0.274	5.167	4.148	1.333	3.630	17.5%	18.1%	+0.6%
Flash	0.800	0.556	1.467	2.778	0.800	1.778	52.4%	39.3%	−13.1%
Kero	0.700	0.602	5.567	4.111	1.200	1.593	20.6%	28.2%	+7.6%
Sink	0.614	0.562	9.900	10.933	2.700	3.067	44.7%	35.8%	−8.9%
Toilet	0.655	0.639	23.833	21.367	3.800	3.967	14.9%	15.2%	+0.3%
Wall	0.862	0.530	3.172	2.818	0.414	1.409	60.8%	24.8%	−36.0%

As shown in Table 13, structural versus functional condition yields mixed results across the seven artifacts. Four artifacts show better performance with structural condition: Bicycle (19.3% > 12.0%), Desk (18.1% > 17.5%), Kerosene (28.2% > 20.6%), and Toilet (15.2% > 14.9%). However, three artifacts show better performance with functional condition: Flash (52.4% > 39.3%), Sink (44.7% > 35.8%), and Wall (60.8% > 24.8%). The mixed pattern suggests that the effectiveness of each condition approach may be highly dependent on the specific characteristics of individual artifacts.

6 Discussion

Our research investigated whether humans construct distinct types of causal models (specifically structural and functional) when reasoning about physical systems, and how these different models influence performance in assembly and troubleshooting tasks. The results offer insights into causal reasoning and its applications in robotics, while revealing methodological challenges in eliciting accurate structural causal models.

6.1 Evidence for Distinct Causal Reasoning Modes

Our graph similarity analyses suggest that the way we frame counterfactual questions significantly affects the topology of elicited causal models. Using multiple graph-theoretic metrics (Hamming distance, Jaccard similarity, spectral distance, and DeltaCon), we observed that participants presented with structural, functional, or general causal reasoning frames constructed topologically different causal models of the same physical systems. The differences in graph structure were most pronounced when using the intervention-based elicitation method across the tested objects.

For simpler objects (desk lamps, flashlights), pairwise comparisons between the three conditions (structural, functional, and general causal) showed notable differences across multiple metrics. For more complex systems (bicycle, sink, toilet), we observed separation particularly between structural and general causal conditions, while functional condition responses sometimes resembled those from the general causal condition. These patterns suggest that counterfactual question framing may tap into potentially distinct modes of system conceptualization, though further research is needed to establish whether these represent fundamentally different cognitive processes rather than variations induced by our experimental framing.

The systematically different graph topologies across conditions point to the possibility that humans may possess multiple frameworks for reasoning about causal relationships in physical systems. However, we must interpret these differences cautiously, as they may also reflect methodological artifacts arising from our specific question formulations rather than inherent cognitive distinctions.

6.2 Methodological Limitations in Structural Model Elicitation

A critical limitation of our study lies in the formulation of counterfactual questions for structural reasoning. Our approach asking "If I remove part X, would part Y maintain its structural integrity?" inadvertently produced sparse structural models because physical parts often maintain integrity even when causally connected parts are removed. For example, removing a bicycle's chain doesn't compromise the structural integrity of the pedals, yet there exists a clear functional and assembly-relevant relationship between these components. This limitation is evidenced by the lower hit rates for most of the objects compared to the ground truth.

The bicycle showed less pronounced differences despite this limitation because its assembly process inherently emphasizes structural dependencies that align with our question framing. Many bicycle components genuinely do lose structural integrity when directly connected supporting elements are removed, making our structural counterfactual questions more effective for this particular object.

This limitation manifests in the consistently lower hit rates for structural models in

assembly tasks. For the bicycle task, although the structural condition yielded a higher hit rate (0.721) than the functional condition (0.613), it still fell far short of the ground-truth structural model, which captured 83.8% of the maximum possible reward compared to only 19.3% achieved by participant-elicited structural models. For the desk lamp, the structural condition hit rate was just 0.274, substantially lower than the functional condition hit rate of 0.733, despite the ground-truth structural model outperforming the functional model (92.0% vs. 42.5% of maximum reward).

This discrepancy suggests that our structural counterfactual questions failed to capture important connections critical for assembly that don't involve immediate structural integrity. The issue is compounded by the implicit presence of extraneous variables such as gravity. Removing a support might not compromise a part's integrity if it can rest on a surface, yet the connection remains essential for assembly.

6.3 Refined Approaches to Structural Model Elicitation

To address these limitations, we propose several methodological refinements for eliciting structural causal models.

Assembly-focused counterfactuals could be more effective, such as asking "If I remove part X during assembly, would this prevent part Y from being correctly positioned in the final assembled object?" This formulation directs attention to assembly relevance rather than immediate structural consequences.

Positional framing may prove beneficial: "If part X were missing or misaligned, would this necessitate repositioning part Y?" This approach avoids confounds from external supports by focusing explicitly on relative positioning.

Hierarchical decomposition presents a compelling alternative approach where participants first identify meaningful component clusters before establishing causal connections between them. This method more closely aligns with natural human conceptualization of complex systems. By applying our updated counterfactual questions to this hierarchical interface, we could assess whether it enhances the accuracy of user-generated structural causal models.

Future work should systematically compare these alternative formulations to determine which most effectively captures assembly-relevant structural causal knowledge while maintaining clear distinction from functional reasoning. Hybrid approaches combining multiple question types might yield more comprehensive structural models that better support robotic assembly planning.

6.4 Task-Dependent Utility of Causal Models

Despite methodological limitations, our findings support the hypothesis that different causal models exhibit differential effectiveness across task contexts. Our POMDP simulations demonstrated that ground-truth structural models significantly outperform functional models in assembly tasks (91.4% vs. 49.1% of maximum reward), while functional/causal models generally excel in troubleshooting scenarios (78.6% vs. 71.6%).

The gap between ground-truth performance and participant-elicited models suggests not that structural reasoning is inherently less useful for assembly, but rather that our elicitation method failed to capture the relevant structural dependencies. This interpretation is supported by the fact that for objects where structural dependencies are more

salient (such as the bicycle), structural condition still showed advantages over functional condition in assembly tasks ($p = 0.0126$).

6.5 Object-Specific Effects and Context Sensitivity

Our results reveal significant context-sensitivity in causal reasoning effectiveness across different physical systems, linked to the inherent properties and complexity of each object. For assembly tasks, structural condition significantly outperformed functional condition for the bicycle ($p = 0.0126$), suggesting that mechanically complex objects with visible linkages naturally evoke structural-spatial reasoning. Conversely, functional condition dramatically outperformed structural condition for the wall lamp ($p = 0.0001$), indicating that simpler systems with clear input-output relationships may better align with functional reasoning.

Particularly noteworthy, for the desk lamp, causal condition yielded the highest rewards, outperforming both functional ($p = 0.0002$) and structural ($p = 0.0007$) modes. This demonstrates that neutral causal framing can be optimal when an object's causal structure involves both structural dependencies and functional relationships that must be integrated for successful assembly.

These object-specific effects likely reflect inherent differences in how people conceptualize different systems. Objects with clear mechanical linkages (bicycles) naturally evoke structural thinking, while objects with less visible mechanisms but clear input-output relationships (lamps) better align with functional or causal reasoning. Understanding these object-specific tendencies could help tailor elicitation methods to particular domains.

6.6 Implications for Robotic Planning

Our findings have direct implications for robotic planning in assembly and troubleshooting contexts. Robots should maintain dual representations: structural for assembly and functional for troubleshooting, rather than relying on a single causal model. Furthermore, elicitation methods for capturing human causal knowledge must be carefully designed to match the specific reasoning mode being targeted.

For assembly tasks, robots might benefit from structural models that capture positional and sequential dependencies rather than simple integrity relationships. The consistent gap between human-elicited and ground-truth performance suggests robots should combine human input with structured domain knowledge and learning from interaction to build more comprehensive causal representations.

6.7 Limitations and Future Directions

Despite limitations in our structural counterfactual question framing, our results demonstrate the powerful effect of reasoning frames on causal model construction. The clear differentiation between causal (control), structural, and functional models confirms that question framing fundamentally shapes how people conceptualize system relationships. This is most apparent in the lamp objects study, where all three reasoning modes produced statistically distinct graph representations across multiple metrics, regardless of elicitation method.

The fact that even our control condition (general causal reasoning) yielded distinct representations from both structural and functional framings for most objects suggests

people possess flexible cognitive frameworks that can be selectively activated through appropriate prompting. This presents significant opportunities for tailored elicitation strategies in human-robot interaction. Through carefully crafted questions, we could potentially guide users toward the most task-appropriate reasoning mode without requiring explicit training in causal modeling.

Several directions for future research emerge. First, domain generalization should be explored by investigating how structural and functional reasoning modes manifest in domains with hidden components, abstract causal relationships, or emergent properties (software systems, ecological networks, financial markets). Second, alternative structural elicitation methods warrant systematic comparison, particularly question formulations targeting assembly relevance, positional relationships, or temporal sequences to better capture assembly-relevant causal knowledge. Third, longitudinal studies examining how causal models evolve with repeated interaction could yield insights into learning dynamics and expertise development. Novices might initially rely on structural representations based on visible spatial relationships, gradually developing more sophisticated functional or hybrid models through experience.

6.8 Conclusion

Our results provide empirical evidence that structural and functional reasoning frames elicit distinct causal representations, with each showing task-specific advantages. The intervention method demonstrated particular promise in eliciting high-fidelity models, though all methods revealed sensitivity to prompt framing. While our ground truth models confirmed theoretical expectations that structural representations should maximize assembly performance and functional models excel in troubleshooting, the actual participant-elicited models frequently diverged from these theoretical expectations. The significant performance gaps between expert-generated and user-generated causal models in POMDP simulations highlight the challenges in effectively eliciting these conceptual distinctions from human subjects. Despite these limitations, we demonstrated that different causal conditions through counterfactual questions can indeed elicit statistically distinct causal representations, suggesting inherent differences in reasoning frameworks. Future work should focus on developing more precise definitions of structural and functional reasoning and refining elicitation techniques to better align with cognitive predispositions. These improvements will help us more effectively harness human causal insights for adaptive, transparent robotic planning systems.

7 References

[1] Semanti Basu, Semir Tatlidi, Moon Hwan Kim, Steven A. Sloman, and R. Iris Bahar. Human causal reasoning guided autonomous object assembly under uncertainty. In *CAUSAL-HRI: Causal Learning for Human–Robot Interaction Workshop at HRI 2024*, March 2024.

[2] Semanti Basu, Semir Tatlidi, Moon Hwan Kim, Steven A. Sloman, and R. Iris Bahar. Can llms learn causal reasoning from humans and aid robots in assembly? In *Workshop on Human Interactive Robot Learning (HIRL) at HRI 2024*, March 2024.

[3] Semanti Basu, Semir Tatlidil, Moon Hwan Kim, Steven Sloman, and R. Iris Bahar. Using causal information to enable more efficient robot operation. In *Proceedings of the 61st ACM/IEEE Design Automation Conference*, pages 1–4. ACM/IEEE, jun 2024. Published on June 23, 2024.

[4] Semanti Basu, Semir Tatlidi, Moon Hwan Kim, Tiffany Tran, Serena Saxena, Tom Williams, Steven A. Sloman, and R. Iris Bahar. Robot planning under uncertainty for object assembly and troubleshooting using human causal models. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025.

[5] R. D. Borchelt and S. Alptekin. Error recovery in intelligent robotic workcells. *International Journal of Production Research*, 32(1):65–73, 1994. doi: 10.1080/00207549408956916.

[6] J. Brawer, M. Qin, and B. Scassellati. A causal approach to tool affordance learning. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8394–8399, Las Vegas, NV, USA, 2020. doi: 10.1109/IROS45743.2020.9341262.

[7] P. W. Cheng. From covariation to causation: A causal power theory. *Psychological Review*, 104(2):367–405, 1997.

[8] Kenneth Craik. *The Nature of Explanation*. Cambridge University Press, 1943.

[9] C. A. Frosch and Philip N. Johnson-Laird. Is everyday causation deterministic or probabilistic? *Acta Psychologica*, 137:280–291, 2011. doi: 10.1016/j.actpsy.2011.01.015.

[10] Eugenia Goldvarg and Philip N. Johnson-Laird. Naïve causality: A mental model theory of causal meaning and reasoning. *Cognitive Science*, 25:565–610, 2001. doi: 10.1207/S15516709COG2502_2.

[11] David Hume. *An Enquiry Concerning Human Understanding*. Open Court, 1748. Modern edition, 1988.

[12] P. N. Johnson-Laird and S. Khemlani. Causal reasoning with mental models. *Frontiers in Human Neuroscience*, 8:849, 2014.

[13] Philip N. Johnson-Laird. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press, 1983.

[14] Philip N. Johnson-Laird and Ruth M.J. Byrne. Conditionals: A theory of meaning, inference, and pragmatics. In *Advances in Psychology*, volume 78, pages 1–73. North-Holland, 1991.

[15] Danai Koutra, Joshua T. Vogelstein, and Christos Faloutsos. Deltacon: A principled massive-graph similarity function. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 162–170. SIAM, 2013. doi: 10.1137/1.9781611972832.18. Also available at arXiv:1304.4657 [cs.SI].

[16] Katharina Lagemann, Christian Lagemann, Benedikt Taschler, et al. Deep learning of causal structures in high dimensions under data limitations. *Nature Machine Intelligence*, 5:1306–1316, 2023. doi: 10.1038/s42256-023-00744-z.

[17] D. Meli, H. Nakawala, and P. Fiorini. Logic programming for deliberative robotic task planning. *Artificial Intelligence Review*, 56:9011–9049, Jan 2023. doi: 10.1007/s10462-022-10389-w.

[18] Michael Oaksford and Nick Chater. *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford University Press, 2007.

[19] L. Parker and R. Kannan. Adaptive causal models for fault recovery in multi-robot teams. *IEEE Transactions on Robotics*, 22(3):21–29, 2006. doi: 10.1109/TRO.2006.874574.

[20] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[21] Clark Glymour Peter Spirtes and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.

[22] S. C. Smith and S. Ramamoorthy. Counterfactual explanation and causal in service of robustness in robot control. In *Proceedings of the 2020 Joint IEEE 10th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 1–8, Virtual Conference, Chile, Oct 2020. doi: 10.1109/ICDL-EpiRob48136.2020.9278061.

[23] Leonard Talmy. Force dynamics in language and cognition. *Cognitive Science*, 12 (1):49–100, 1988. ISSN 0364-0213. doi: 10.1016/0364-0213(88)90008-0.

[24] Semir Tatlidil, Steven A. Sloman, Semanti Basu, Tiffany Tran, Serena Saxena, Moon Hwan Kim, and Iris Bahar. A comparison of methods to elicit causal structure. *Frontiers in Cognition*, 2025. To appear in the Section "Reason and Decision-Making" (Research Topic: "Causal Cognition in Humans and Machines – Volume II").

[25] Philip Wolff. Representing causation. *Journal of Experimental Psychology: General*, 136(1):3–21, 2007.

8 Appendix

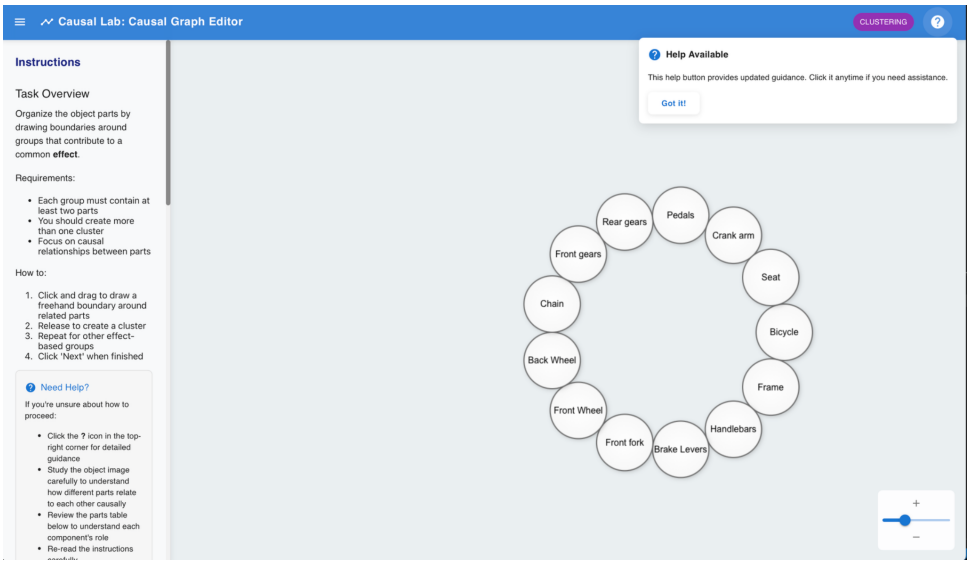


Figure 9: Editor Help & Shortcuts menu showing Quick Navigation, Canvas Interactions, for clustering task. Includes instructions for nodes, clusters, and counterfactual questions.

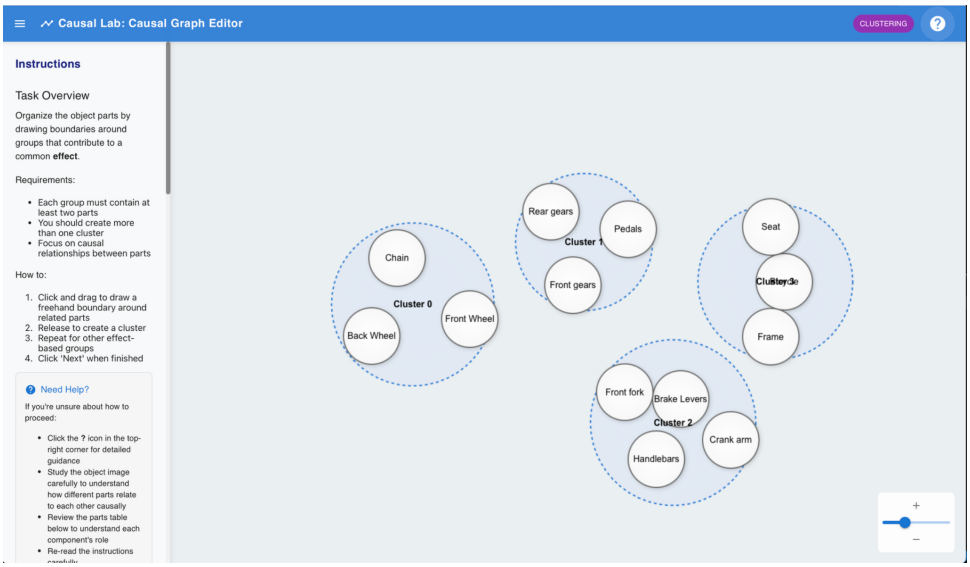


Figure 10: Example demonstration of clustering using a bicycle.

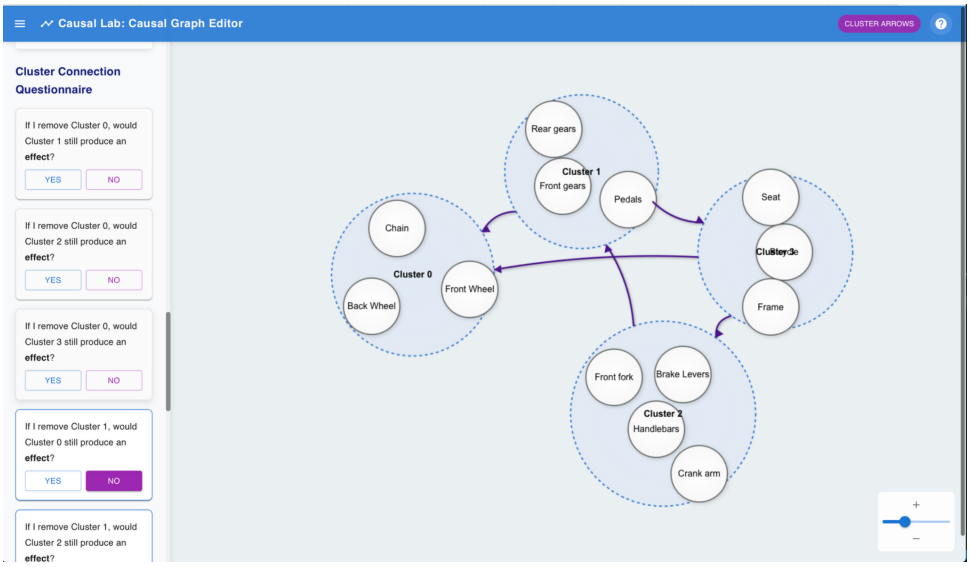


Figure 11: Example demonstration of cluster connections using a bicycle. Shows answering of counterfactual questions.

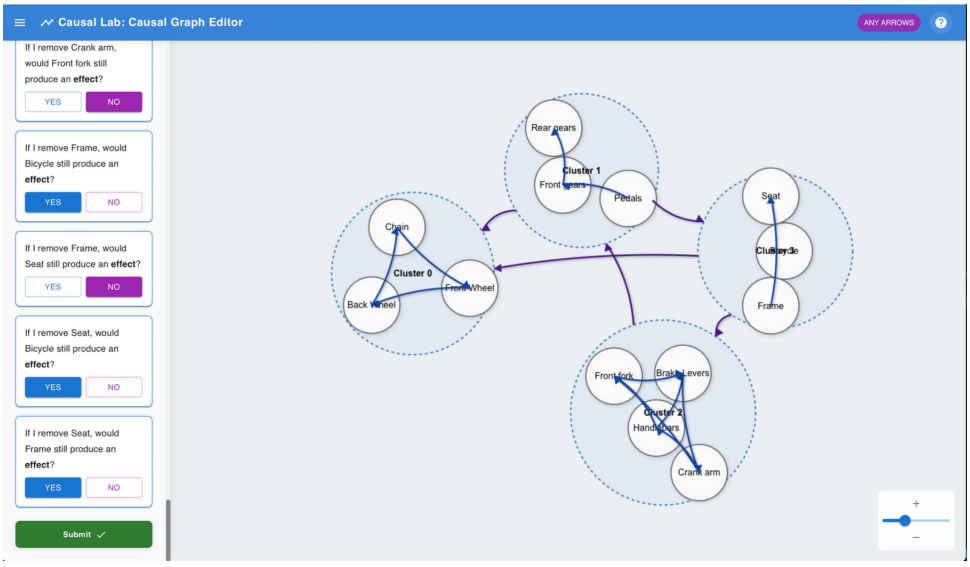


Figure 12: Example demonstration of complete causal map (intra/inter-cluster connections) for bicycle.

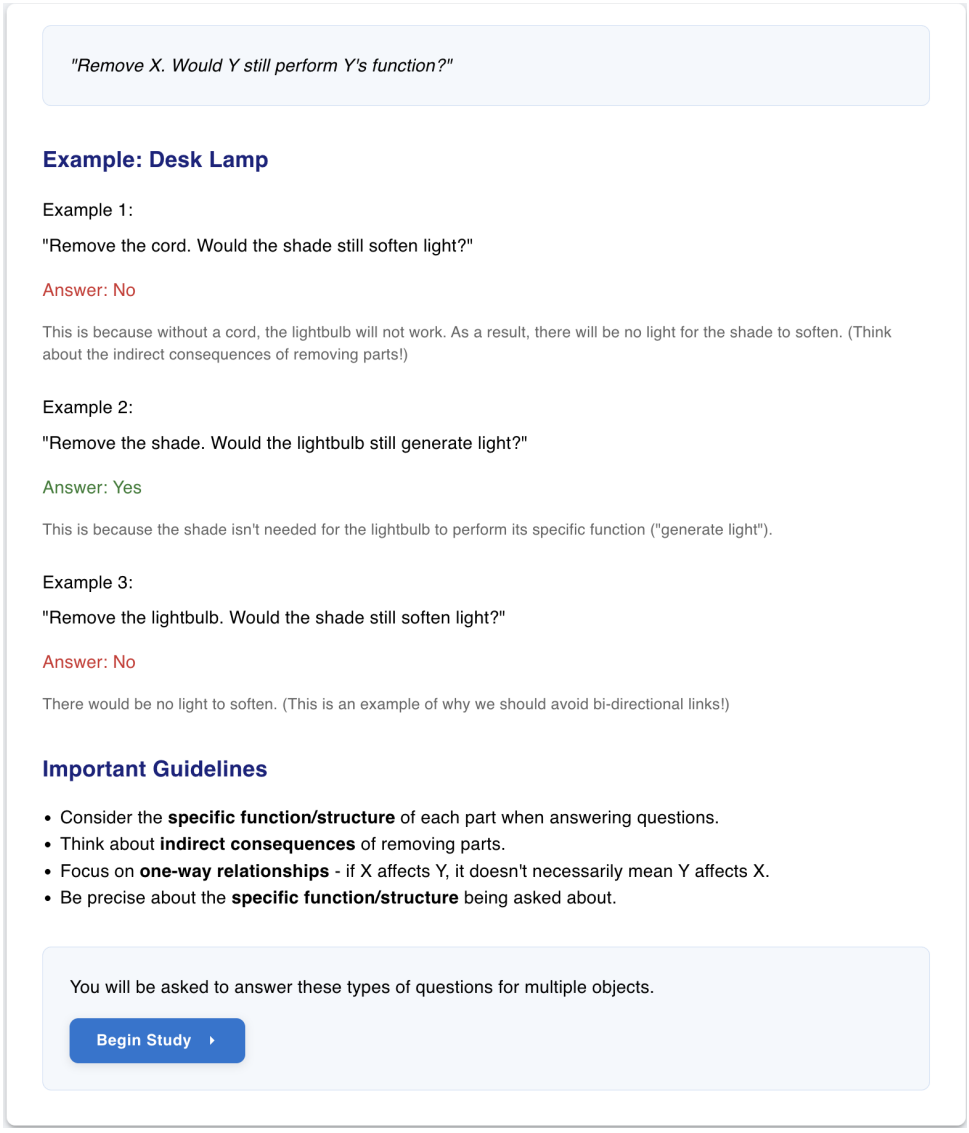


Figure 13: Entry page for participants to give an overview of counterfactual reasoning.

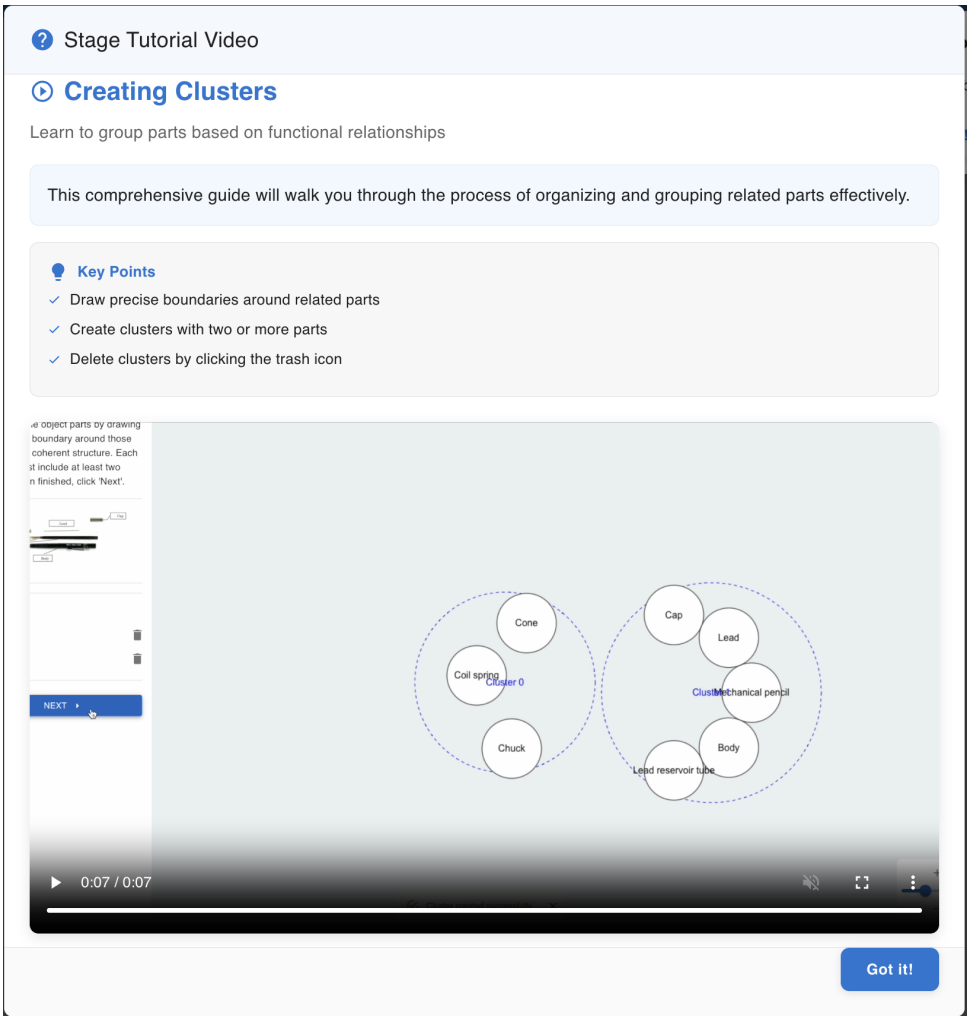


Figure 14: Clustering task interface with video tutorial on how to interact with the user interface.

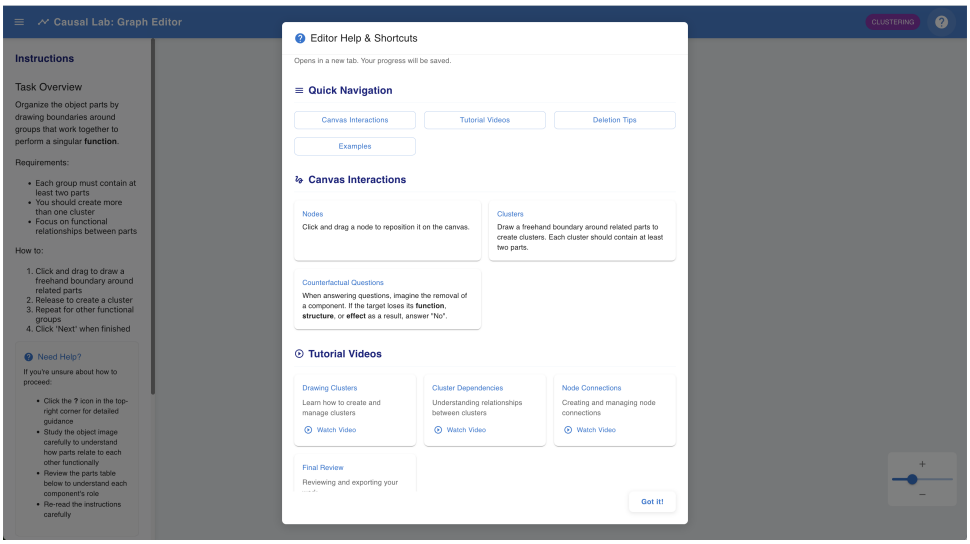


Figure 15: Help Available popup providing updated guidance to users.

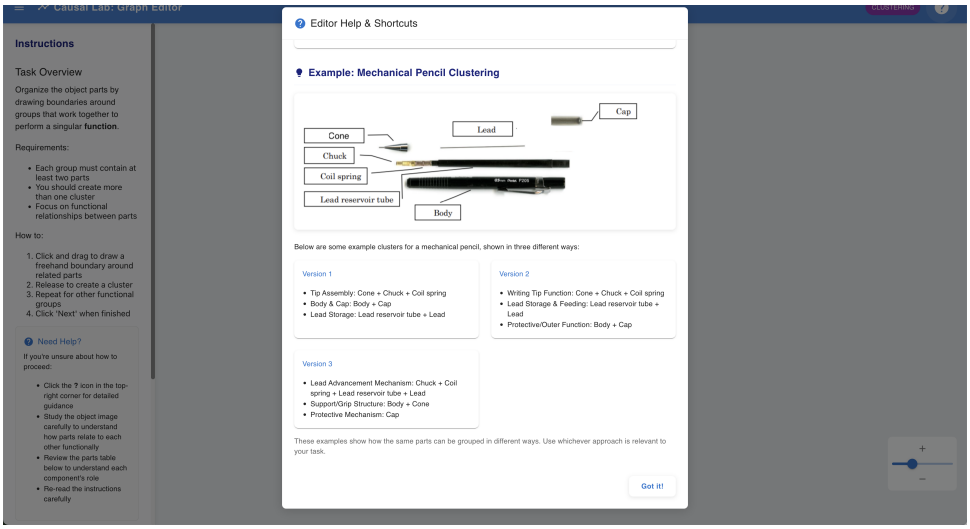


Figure 16: Help Available popup providing updated guidance to users with mechanical pencil example.