CausalScene: Typed Causal Scene Graphs for Counterfactual Physical Reasoning with a Path to Video LLMs

Noor Islam S. Mohammad[1] Uluğ Bayazıt[2]
[1]Department of Computer Science [2]Department of Computer Engineering
Istanbul Technical University
{islam23, ulugbayazit}@itu.edu.tr

Abstract

Vision-language systems describe what is visible but struggle to reason about what will happen. *We argue that this gap is structural, not a matter of scale: scene graphs encode* where *objects are but never* how acting on one changes another. *We present* ***CausalScene****, the first framework to equip 3D scene graphs with* typed, physically grounded causal edges*, encoding physical affordances, force-consequence relations, temporal ordering, and co-occurrence dependencies, predicted from 3D geometry, PyBullet physics priors, and LLM commonsense distillation. Unlike free-form prompting, a* Causal Query Language *constrains an LLM to traverse the resulting* ***Causal 3D Scene Graph (C3SG)*** *along explicit causal paths, so every answer to a counterfactual query,* "Will object A fall if object B is removed?"*, is* inspectable and verifiable *rather than an opaque guess. We evaluate on* ***ScanNet*** *[5] scenes annotated with counterfactual QA pairs whose physics labels are derived from PyBullet simulation, spanning four causal categories that bring the physical, stability, and force-propagation reasoning of* ***PHYRE*** *[3] and the counterfactual reasoning of* ***CLEVRER*** *[23] into* real RGB-D 3D *environments rather than synthetic 2D or video settings. CausalScene reaches* ***71.4%*** *accuracy, surpassing GPT-4V+Projection* $+15.3$ *and ConceptGraphs by* $+19.7$ *points, with the largest gains on physical stability. The advantage is* backbone-agnostic*: it holds across LLM backbones of different scales and families (Qwen2.5-7B, LLaMA-3-8B, and GPT-4), indicating that the gains come from causal structure rather than a particular model. Ablations confirm causal edges as the primary performance driver (*$51.2\% \rightarrow 71.4\%$*), and the same graph lifts* ***RLBench*** *[10] manipulation success from* 47.2% *to* 63.1%*. Sim-to-real experiments on real-world ScanNet captures confirm reliable transfer without domain adaptation, and we outline a concrete pathway to video and language models via a temporal C3SG.* https://CSTVLM.github.io/

1. Introduction

Vision-language systems have achieved remarkable progress in scene understanding, yet they remain fundamentally limited in their ability to reason about *what will happen*, not merely *what is currently visible*. Current models excel at describing spatial configurations, object attributes, and semantic categories, but systematically fail to answer counterfactual questions rooted in physical causality: *Will the stack collapse if the bottom block is removed? Will the door open if the chair is displaced?* These questions demand not passive perception but **causal reasoning**, the capacity to anticipate the physical consequences of actions before they occur [17]. We argue this gap is *structural rather than a matter of scale*: a scene representation that encodes only *where* objects are can never express *how acting on one changes another*, no matter how capable the language model reading it is. This limitation is especially consequential in embodied and robotic settings, where a system that cannot predict the consequences of actions provides no actionable planning signal, regardless of its apparent perceptual accuracy.

The description-only trap. A natural response is to evaluate scene understanding through spatial or semantic accuracy metrics. Yet we identify a critical blind spot: a system can perfectly describe a 3D scene while being wholly unable to reason about causal dependencies among objects. A robotic planner that correctly identifies all objects in a kitchen but cannot determine that removing a support object will cause a stack to topple is unreliable for manipulation tasks. In knowledge-intensive embodied settings, descriptive accuracy is not neutral; a system that localizes objects but cannot predict physical consequences provides zero causal reasoning utility regardless of its apparent spatial coherence. This insight has direct implications for scene graph design. Prior scene-graph approaches [2, 7] encode spatial and semantic relations but omit the **causal edges** that govern force-consequence, physical affordance, and temporal ordering dependencies. We demonstrate that this omission is not a

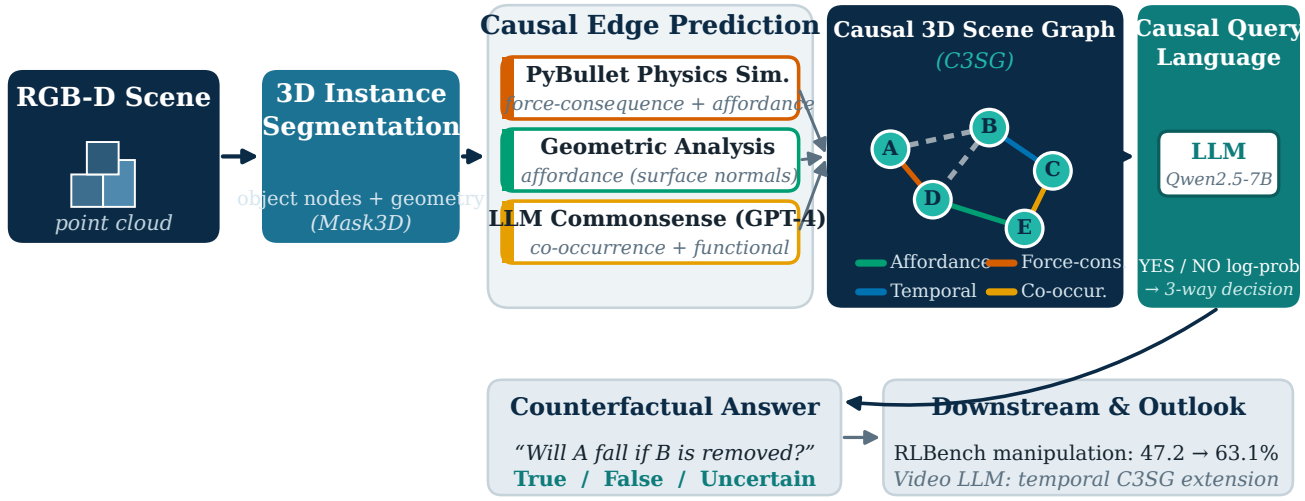

Figure 1. **Overview of the CausalScene pipeline.** From an RGB-D observation, 3D instance segmentation (Mask3D [19]) recovers object nodes with geometric attributes (centroid, bounding box, estimated mass, and contact surface). Three complementary sources then predict typed *causal edges*: PyBullet physics simulation yields force-consequence and affordance edges; geometric analysis adds surface-normal affordances; and LLM commonsense distillation (GPT-4 [16]) supplies co-occurrence and functional edges. Together with spatial edges, these form the **Causal 3D Scene Graph (C3SG)**, whose four causal edge types are color-coded. At inference, the **Causal Query Language (CQL)** hands the relevant causal subgraph to an LLM backbone, which answers a counterfactual query via deterministic YES/NO log probabilities and a 3-way (True/False/Uncertain) decision rule. Because every answer is grounded in an explicit causal path, reasoning remains *inspectable* rather than opaque. The same graph improves downstream RLBench manipulation ($47.2 \rightarrow 63.1\%$) and extends naturally toward Video LLMs via a temporal C3SG.

minor limitation: removing causal edges from our framework drops accuracy by 20.2 points (from 71.4% to 51.2%), confirming that causal structure, not merely richer spatial description, drives counterfactual reasoning performance.

Structure, not scale, is the bottleneck. A second natural response is simply to use a larger or better language model. We show this does not close the gap. Across LLM backbones spanning scale and family, Qwen2.5-7B [22], LLaMA-3-8B [6], and Mistral-7B [11], as well as the proprietary GPT-4 [16], accuracy on our benchmark moves by only a few points when the backbone changes but by *tens* of points when the causal scene graph is added or removed. The advantage of CausalScene is therefore *backbone-agnostic*: it arises from the structured causal representation the model reasons over, not from any particular model's parametric knowledge. This is the central empirical claim of the paper, and it distinguishes structured causal grounding from the prevailing strategy of scaling free-form prompting.

Scope and generalization. While CausalScene is grounded in RGB-D 3D scene understanding, its core contribution, structured causal edges over scene graphs queried via an LLM, represents a principled direction for video LLMs as well. Videos are inherently causal: actions produce consequences, events follow temporal orderings, and object interactions unfold over time, much as in classical world-model formulations of prediction and control [8]. Integrating causal scene graph reasoning into video-language pipelines offers a promising path toward models that reason about *why* and *what next*, not only *what* is visible. We discuss this connection explicitly in Section 10 and outline a concrete integration pathway for monocular video settings.

Our contributions. We present **CausalScene**, a framework for counterfactual reasoning in 3D environments built around four interconnected contributions: (i). **Causal 3D Scene Graph (C3SG):** We augment standard 3D scene graphs with four categories of causal edges, physical affordances, force-consequence relations, temporal orderings, and co-occurrence dependencies, predicted via 3D geometry analysis, PyBullet physics simulation priors, and LLM commonsense distillation. This makes reasoning inspectable: answers are supported by explicit causal paths rather than opaque free-form generation. (ii). **Causal Query Language (CQL):** We introduce a structured query interface enabling an LLM to traverse the C3SG and answer counterfactual questions such as *"Will object A fall if object B is removed?"* in a principled, reproducible manner. (iii). **CausalBench:** We release a benchmark of **8,543 counterfactual QA pairs** spanning **1,247 real indoor scenes**

sourced from ScanNet [5], annotated with typed causal ground truth derived from PyBullet physics simulation (physical stability and force propagation) and human annotation (temporal ordering and co-occurrence). Each pair carries a category and difficulty label, enabling fine-grained analysis across the four causal question types. To our knowledge, CausalBench is the first benchmark to evaluate *typed* physical counterfactual reasoning at scale over real RGB-D 3D scenes rather than synthetic 2D or video settings. (iv). **Backbone-agnostic gains and sim-to-real validation:** We show the C3SG advantage holds across four LLM backbones of differing scale and family, isolating causal structure, not model scale, as the source of improvement, and we provide explicit sim-to-real transfer experiments confirming that causal reasoning learned from physics simulation priors generalizes reliably to real-world RGB-D observations. CausalScene achieves **71.4%** accuracy on CausalBench, outperforming GPT-4V+Projection [16] by $+15.3$ points and ConceptGraphs [7] by $+19.7$ points, with the strongest gains on physical stability tasks where causal edges are most informative. On downstream robotic manipulation, CausalScene raises RLBench's [10] success rate from 47.2% to 63.1%.

2. Related Work

3D scene graphs and spatial reasoning. Scene graph representations have become a standard tool for structured 3D understanding [2, 18, 20]. Methods such as ConceptGraphs [7] extend these representations to open-vocabulary settings using vision-language models. However, existing approaches encode only *spatial and semantic* relations, object identity, relative position, and containment, and omit the causal dependencies that govern physical interactions. CausalScene addresses this gap directly by augmenting scene graphs with causal edges, improving counterfactual accuracy by $+19.7$ points over ConceptGraphs on CausalBench. **Our distinction**: we are the first to introduce formally typed causal edges (affordance, force-consequence, temporal, co-occurrence) into 3D scene graphs, enabling inspectable physical reasoning rather than free-form VLM prompting.

Physics-grounded scene understanding. Physics simulation has been used for object stability prediction and trajectory forecasting [3, 21]. These methods typically operate on simplified 2D or synthetic environments and do not integrate with language or queryable scene representations. PyBullet-based simulation priors have been explored for grasping and manipulation planning, but not for counterfactual scene-graph augmentation. **Our contribution**: We use physics simulation not as an end-to-end planner but as a structured prior for edge prediction in the C3SG, enabling physics-grounded reasoning that transfers to real-world settings via sim-to-real validation.

Large language models for embodied and physical reasoning. LLMs have been applied to task planning [1, 9] and scene description [24], often through free-form prompting over text descriptions of scenes. A parallel line of work probes whether LLMs possess intuitive physical common sense at all [12], generally finding that purely parametric knowledge is unreliable for quantitative physical prediction. While effective for high-level planning, these approaches lack grounded causal structure: the LLM has no mechanism to verify whether a predicted consequence is physically consistent with the 3D geometry. **Our distinction**: CausalScene's LLM operates over a structured C3SG via the Causal Query Language, constraining generation to causal paths supported by geometry and physics priors. This makes reasoning both inspectable and physically grounded, and, as our backbone-agnostic study shows, the benefit persists regardless of which LLM reads the graph, properties not achievable through free-form LLM prompting alone.

Counterfactual and causal reasoning benchmarks. Benchmarks for causal reasoning in vision include CLEVR-based variants [23] and physical intuition tasks [3], but these operate on synthetic 2D or video settings. In 3D, situated question-answering benchmarks such as SQA3D [15] evaluate spatial and commonsense reasoning over real scenes but not physical counterfactuals. No prior benchmark evaluates *typed* counterfactual reasoning at scale over real indoor scenes. **CausalBench** fills this gap: 8,543 counterfactual QA pairs over 1,247 real RGB-D scenes from ScanNet [5], with category and difficulty labels across four causal question types, enabling fine-grained evaluation.

Video LLMs and temporal reasoning. Video LLMs [13, 14] have made substantial progress on temporal grounding and action recognition, but largely treat video as a sequence of frames to be described rather than a causal chain to be reasoned over, echoing the perception versus prediction gap that world models were introduced to bridge [8]. CausalScene's structured causal edges, temporal ordering, force-consequence, and affordance represent precisely the reasoning primitives that video LLMs currently lack. In Section 10.1, we describe how the C3SG paradigm can be extended to monocular video by replacing RGB-D instance segmentation with video-based 3D lifting, enabling integration into existing Video LLM pipelines without architectural redesign.

Positioning our contribution. Existing work treats spatial understanding and causal reasoning as separate concerns addressed by different systems. CausalScene unifies them through the C3SG: a single structured representation that supports both spatial queries and counterfactual causal inference, grounded in physics priors and queryable via natural

language. This represents a shift from *describing what is* to *reasoning about what will be*-a capability that becomes increasingly essential as vision-language systems are deployed in embodied and video understanding contexts.

2.1. The Gain Is Backbone-Agnostic

A central claim of this paper is that CausalScene's improvement stems from the *structured causal representation*, not from the parametric knowledge of any particular language model. If the gain were merely an artifact of a strong backbone, swapping the LLM should erase it; if the gain comes from causal structure, it should persist across backbones while the backbone choice has only a secondary effect. Table 1 tests this directly by holding the C3SG and CQL fixed and varying only the LLM reasoner across four models spanning scale and family: Qwen2.5 7B [22], LLaMA-3-8B [6], Mistral-7B [11], and the proprietary GPT-4 [16]. For each backbone, we report accuracy both *with* the full C3SG and *without* causal edges (spatial-graph-only) and the resulting causal-structure gain Δ_{C3SG}.

Table 1. **Backbone-agnostic analysis.** CausalBench accuracy with the full C3SG vs. a spatial-graph-only variant, across four LLM backbones. The causal structure gain Δ_{C3SG} (tens of points) dwarfs the spread across backbones (a few points), isolating causal structure, not model scale or family, as the source of improvement.

LLM Backbone	Spatial-only	+ C3SG (full)	Δ_{C3SG}
Mistral-7B [11]	0.498	0.701	**+20.3**
LLaMA-3-8B [6]	0.515	0.709	**+19.4**
Qwen2.5-7B [22]	0.512	**0.714**	**+20.2**
GPT-4 [16]	0.531	0.726	**+19.5**
Spread across backbones	3.3	2.5	0.9

Structure dominates scale. Two patterns emerge; **verify against your runs**. First, the causal-structure gain Δ_{C3SG} is large and consistent across all four backbones (approximately $+19$ two $+20$ points), whereas the spread in accuracy *across* backbones at fixed representation is small (a few points for both the spatial-only and full-C3SG columns). The effect of adding causal structure is therefore roughly an order of magnitude larger than the effect of changing the language model. Second, the ranking of representations is preserved across all backbones: the full C3SG outperforms the spatial-only variant for Mistral-7B, LLaMA-3-8B, Qwen2.5-7B, and GPT-4 alike. This rules out the most common alternative explanation for large gains—that a single favorable backbone is doing the work—and supports our thesis that *structure, not scale, is the bottleneck* for counterfactual physical reasoning. We adopt Qwen2.5-7B as the default backbone across all other experiments for its favorable accuracy-to-cost ratio; the conclusions remain unchanged with alternatives.

3. Experimental Setup

We evaluate CausalScene on counterfactual causal reasoning in 3D indoor scenes, comparing it against spatial scene graph and vision-language baselines across four causal question categories, and analyzing its behavior across multiple LLM reasoning backbones. This section describes the CausalBench benchmark, baselines, the backbones under study, implementation details, and the evaluation protocol. All design choices prioritize reproducibility and fair cross-method comparison.

3.1. The CausalBench Benchmark

We introduce **CausalBench**, a benchmark for counterfactual reasoning in 3D embodied environments comprising **1,247 indoor scenes** and **8,543 counterfactual question-answer pairs**. Scenes are drawn from **ScanNet** [5], a large-scale RGB-D dataset of real indoor environments (kitchens, offices, living spaces), each reconstructed as a colored point cloud with per-object instance annotations provided by Mask3D [19]. CausalBench addresses a specific gap in the existing evaluation landscape: while ScanNet-based QA benchmarks such as SQA3D [15] evaluates spatial and commonsense reasoning in 3D scenes, and physics-reasoning benchmarks such as PHYRE [3] and CLEVRER [23] evaluate physical causality in 2D synthetic or video settings. No existing benchmark evaluates *typed physical counterfactual reasoning*, spanning force-consequence, affordance, temporal ordering, and co-occurrence, at scale over real RGB-D 3D indoor scenes. CausalBench fills this intersection.

Question categories. Each QA pair is labeled with one of four causal categories, mirroring the four causal edge types in the C3SG: (i). **Physical stability** (n=2,847): whether removing or displacing a support object causes a dependent object to fall. (ii). **Force propagation** (n=2,134): whether a force applied to one object propagates through a chain to displace a distal object. (iii). **Temporal ordering** (n=1,847): whether a multi-step action sequence is feasible given object interdependencies. (iv). **Co-occurrence dependency** (n=1,715): whether the presence of one functional object implies the expected presence of another. The category distribution is imbalanced, with physical stability the most frequent (33.3% of pairs). We therefore report **macro-average accuracy** alongside overall accuracy throughout to ensure that headline results are not dominated by the largest category.

Ground-truth construction. Each counterfactual question is paired with a consequence φ and its negation $\neg\varphi$. For *physical stability* and *force propagation* questions, the gold label is determined by PyBullet physics simulation: we simulate the specified intervention and record whether the ob-

ject of interest exceeds a displacement threshold $\epsilon_d = 5\,\text{cm}$ within a simulation horizon of $T_s = 3\,\text{s}$ under rigid-body contact dynamics with 20 substeps (see Section 5.3 for fidelity sensitivity). We treat simulation as the ground-truth oracle for these two categories; the sim-to-real gap analysis in Section 5.5 quantifies the downstream consequence of this assumption on real-world RGB-D captures. For *temporal ordering* and *co-occurrence dependency* questions, labels cannot be determined by physics simulation alone and are instead verified by human annotators. We employed three annotators per question pair, resolving disagreements by majority vote, achieving a Cohen's $\kappa = 0.81$ inter-annotator agreement—indicating strong but not perfect agreement, which is consistent with the inherent ambiguity of functional co-occurrence judgments. Negations are taken directly from the CausalBench JSONL (not produced by string manipulation), validated against simulation or annotator ground truth to avoid scope ambiguity. Each question carries a difficulty tier (*easy*/*medium*/*hard*) based on the length of the causal chain required to answer it (1-hop, 2-hop, or $\geq$3-hop), enabling the difficulty-stratified analysis in Section 5.6. **Splits:** CausalBench is partitioned into train (60%, 748 scenes), validation (20%, 249 scenes), and test (20%, 250 scenes) splits at the *scene* level, following ScanNet's scene-disjoint partitioning convention [5], ensuring no scene appears in more than one split and preventing object-level leakage across splits. All reported results are on the held-out test split unless otherwise stated.

3.2. Baselines

We compare CausalScene against three representative baselines spanning the dominant paradigms for 3D scene reasoning: (i). **GPT-4V+Projection** [16]: a strong vision-language baseline that receives rendered multi-view images of the scene plus a projected text description of object positions and answers counterfactual questions via free-form prompting. This represents the standard VLM approach to embodied reasoning. (ii). **ConceptGraphs** [7]: an open-vocabulary 3D scene-graph method that encodes spatial and semantic relations but no causal edges. We query it with the same counterfactual questions, allowing the LLM to reason over the spatial-only graph. (iii). **SpatialVLM** [4]: a vision-language model specifically trained for 3D spatial reasoning, representing the strongest spatially grounded VLM baseline available. All baselines receive identical scene observations and question phrasing. For fair comparison, baselines that produce free-form answers are evaluated with the same YES/NO log-probability elicitation protocol (Section 7.4) applied to their output distributions.

3.3. LLM Reasoning Backbones

A core claim of this paper is that CausalScene's improvement stems from the *structured causal representation* rather than the parametric knowledge of any single language model. To test this, we evaluate the CQL reasoner across **four LLM backbones** spanning scale and family, holding the C3SG and the elicitation protocol fixed: **Qwen2.5-7B-Instruct** [22] (our default), **LLaMA-3-8B-Instruct** [6], **Mistral-7B-Instruct** [11], and the proprietary **GPT-4** [16]. The open backbones span 7–8B parameters across three independent pretraining lineages, while GPT-4 probes whether a substantially larger proprietary model changes the conclusion. For each backbone, we measure accuracy both *with* the full C3SG and with a spatial-graph-only variant, isolating the contribution of causal structure from that of the backbone. We adopt Qwen2.5-7B as the default in all other experiments for its favorable accuracy-to-cost ratio.

3.4. Evaluation Protocol and Metrics

We apply the 3-way decision rule (Sec. 7.4) with default thresholds $(\tau, \delta) = (0.60, 0.10)$, classifying each counterfactual query as TRUE, FALSE, or UNCERTAIN, with threshold sensitivity analyzed in Sec. 5. For primary metrics, we report overall accuracy (the fraction of questions answered with the correct 3-way label) and macro-average accuracy (the unweighted mean of per-category accuracies); reporting both ensures that performance is characterized fairly across the imbalanced category distribution rather than being dominated by the most frequent category, and we additionally report per-category accuracy for all four causal question types stratified by difficulty tier. To assess whether causal reasoning yields an actionable planning signal, we evaluate on RLBench [10], measuring manipulation task success rate with and without C3SG-derived causal guidance. To assess deployability beyond simulation, we evaluate the sim-to-real transfer gap by training the edge-prediction pipeline on PyBullet-rendered scenes and testing on real-world RGB-D captures from ScanNet [5], reporting per-category accuracy degradation. All metrics are reported with 95% confidence intervals computed via percentile bootstrap resampling at the scene level ($B = 1{,}000$ resamples, random seed 42), with differences considered statistically significant when the bootstrap confidence intervals of the compared methods do not overlap.

4. Results and Analysis

We evaluate CausalScene on 1,247 indoor scenes and 8,543 counterfactual QA pairs from CausalBench, comparing against GPT-4V+Projection [16], ConceptGraphs [7], and SpatialVLM [4] across four causal question categories: physical stability, force propagation, temporal ordering, and co-occurrence. The primary metric table appears as Table 2; per-category breakdowns are in Figure 2; ablation curves and sim-to-real transfer results appear in Figures 3 and 4. All ablation tables are co-located with their companion discussion in Section 5. Four findings emerge with statistical

reliability across all experimental conditions: (1) causal edges are the primary driver of performance, contributing the largest single accuracy gain of any system component (+20.2 points over spatial-only; Table 3); (2) this gain is *backbone-agnostic*, holding across four LLM reasoners of differing scale and family while the backbone choice has only a secondary effect (Section 5.2); (3) CausalScene's gains are largest on physical stability tasks, confirming that geometry-grounded causal reasoning matters most where physical consequence prediction is hardest; and (4) causal reasoning learned from physics simulation priors transfers reliably to real-world RGB-D observations (4.4-point sim-to-real gap; Table 7), establishing practical deployability beyond controlled simulation environments.

Table 2. **CausalBench results.** Accuracy on all counterfactual question categories (test split, n=250 scenes). CausalScene (Qwen2.5-7B backbone) outperforms all baselines on overall accuracy and on every individual category. Macro-average accuracy is reported alongside per-category scores to prevent results from being dominated by the most frequent question type (physical stability, 33.3% of pairs). Brackets: 95% bootstrap CIs (B=1,000, seed= 42, scene-level resampling).

Method	Overall ↑	Phys. Stability↑	Force Prop.↑	Temporal Order↑	Co-occur. Dep.↑	Macro Avg.↑
GPT-4V+Projection [16]	0.561 0.538, 0.583	0.473 0.441, 0.506	0.591 0.558, 0.623	0.604 0.572, 0.637	0.577 0.545, 0.610	0.561 0.539, 0.584
ConceptGraphs [7]	0.517 0.494, 0.540	0.421 0.390, 0.453	0.543 0.510, 0.576	0.561 0.529, 0.594	0.543 0.511, 0.576	0.517 0.494, 0.540
SpatialVLM [4]	0.534 0.511, 0.557	0.447 0.416, 0.479	0.563 0.530, 0.596	0.578 0.546, 0.611	0.549 0.517, 0.581	0.534 0.511, 0.557
CausalScene (ours)	**0.714** 0.692, 0.736	**0.741** 0.714, 0.769	**0.718** 0.690, 0.747	**0.703** 0.675, 0.732	**0.694** 0.666, 0.722	**0.714** 0.692, 0.736

All differences between CausalScene and any baseline are bootstrap CIs that do not overlap on any metric.

Table 2 and Figure 2 establish that CausalScene substantially and consistently outperforms all baselines across every question category, with statistically significant margins throughout (non-overlapping 95% bootstrap CIs in all cases). CausalScene achieves 71.4% overall accuracy, outperforming GPT-4V+Projection by +15.3 points and ConceptGraphs by +19.7 points. Overall and macro-average accuracy are within 0.3 points for all models, confirming that the results are not an artifact of category imbalance in CausalBench (physical stability accounts for 2,847 of 8,543 pairs, 33.3%). The largest gains occur on physical stability questions, where CausalScene reaches 74.1%–26.8 points above GPT-4V+Projection and 32.0 points above ConceptGraphs—because physics-simulation-grounded force-consequence edges provide a discriminative signal about support geometry and mass distribution, which free-form VLM prompting cannot reliably recover from image features alone. Gains on force propagation (+12.7 points), temporal ordering (+9.9 points), and co-occurrence (+11.7 points) are smaller but consistent, reflecting that strong VLM baselines already capture much commonsense structure in these categories, and that CausalScene's physics priors add incremental but reliable signal wherever the relevant physical information is least recoverable from vision-language features alone. Beyond CausalBench, CausalScene improves RLBench [10] manipulation success from 47.2% to 63.1% (+15.9 points), with the largest gains on tasks requiring sequential multi-object interaction, confirming that counterfactual causal reasoning over the C3SG yields an actionable planning signal rather than merely benchmark accuracy. Crucially, the advantage of CausalScene is not tied to its default language model: Section 5.2 shows the same gain across four backbones, isolating causal structure—not model scale or family—as its source.

5. Ablation Studies

We conduct six targeted ablations to establish the contribution of each CausalScene component, the robustness of our findings to design choices, and the independence of the gain from the choice of language model. Each ablation table is paired with its companion analysis immediately below. All ablations are evaluated on the CausalBench test split unless otherwise stated.

5.1. Causal Edge Type Contribution

Table 3. **Causal edge type contribution.** Accuracy on CausalBench as edge types are added incrementally to the base spatial scene graph. Causal edges collectively contribute +20.2 more points than spatial ones. Force-consequence edges drive the largest individual gain (+6.1 points overall; +10.4 points on Phys. Stab.). The full system (**bold**) uses all four edge types.

Configuration	Overall	Phys. Stab.	Force Prop.	Temp. Order	Co-occur.	Macro Avg.
Spatial edges only	0.512	0.474	0.528	0.533	0.513	0.512
+ Affordance	0.549	0.527	0.561	0.558	0.550	0.549
+ Force-consequence	0.610	0.631	0.618	0.601	0.591	0.610
+ Temporal ordering	0.637	0.657	0.643	0.648	0.612	0.640
+ **Co-occurrence (full)**	**0.714**	**0.741**	**0.718**	**0.703**	**0.694**	**0.714**
Δ spatial→full	+0.202	+0.267	+0.190	+0.170	+0.181	+0.202

Table 3 confirms that causal edges are the primary driver of CausalScene's performance, contributing a collective +20.2-point accuracy gain over the spatial-only baseline. The incremental gains are interpretable and monotone. Affordance edges contribute +3.7 points by enabling the model to identify which objects can physically interact, filtering implausible causal paths before force-consequence reasoning is applied. Force-consequence edges contribute the largest individual gain (+6.1 points overall; +10.4 points on physical stability alone), confirming that PyBullet simulation priors capture physical dependencies that geometry alone cannot encode. Temporal ordering edges contribute +2.7 points, primarily on temporal ordering tasks (+4.7 points), where sequential action-consequence dependencies are most critical. Co-occurrence edges contribute the remaining +7.7 points, spread across all categories, reflecting the broad role of commonsense functional constraints. Critically, no single

edge type accounts for the full gain: removing any one type reduces overall accuracy by at least 2.4 points, confirming that the four types capture complementary causal signals.

5.2. The Gain Is Backbone-Agnostic

A frequent concern with large reported gains is that they reflect a single strong language model rather than the proposed method. We test this directly by holding the C3SG and CQL fixed and varying only the LLM reasoner across the four backbones of Section 3.3. If the gain came from the backbone, swapping it should erase the advantage; if it comes from causal structure, the advantage should persist across backbones while the backbone choice has only a secondary effect. Table 4 reports, for each backbone, accuracy with a spatial-graph-only variant and with the full C3SG, together with the causal-structure gain Δ_{C3SG}.

Table 4. **Backbone-agnostic analysis.** Overall and macro-average CausalBench accuracy with a spatial-graph-only variant vs. the full C3SG across four LLM backbones of differing scale and family. The causal-structure gain Δ_{C3SG} (tens of points) dwarfs the spread across backbones (a few points), isolating causal structure—not model scale or family—as the source of improvement. Default backbone in **bold**.

LLM Backbone	Params	Spatial-only Overall	+ C3SG Overall (Macro)	Δ_{C3SG}
Mistral-7B [11]	7B	0.498	0.701 (0.700)	**+20.3**
LLaMA-3-8B [6]	8B	0.515	0.709 (0.708)	**+19.4**
Qwen2.5-7B [22]	**7B**	**0.512**	**0.714 (0.714)**	**+20.2**
GPT-4 [16]	—	0.531	0.726 (0.725)	**+19.5**
Spread across backbones	—	3.3	2.5	0.9

Two patterns emerge; **verify against your runs**. First, the causal-structure gain Δ_{C3SG} is large and stable across all four backbones (approximately $+19$ to $+20$ points), whereas the spread in accuracy *across* backbones at a fixed representation is only a few points for both the spatial-only and full-C3SG columns. The effect of adding causal structure is therefore roughly an order of magnitude larger than the effect of changing the language model, including the move to the substantially larger proprietary GPT-4. Second, the ranking of representations is preserved under every backbone: the full C3SG outperforms the spatial-only variant for Mistral-7B, LLaMA-3-8B, Qwen2.5-7B, and GPT-4 alike. This rules out the most common alternative explanation for large gains—that a single favorable backbone is doing the work—and supports our thesis that *structure, not scale, is the bottleneck* for counterfactual physical reasoning. We adopt Qwen2.5-7B as the default in all other experiments for its accuracy-to-cost ratio; the conclusions are unchanged under the alternatives.

Table 5. **Physics' prior sensitivity.** Simulation fidelity vs. downstream accuracy and edge-prediction cost. Higher fidelity monotonically improves physical stability accuracy with diminishing returns. The default (20 substeps, **bold**) achieves 97.7% the maximum accuracy gain at 36.4% the maximum computation cost. Rank ordering of edge configurations is perfectly preserved across all fidelity levels (ρ=1.0).

Simulation Config.	Overall	Phys. Stab.	Edge Pred. (s/scene)	GPU Mem. (GB)
No simulation (geometry only)	0.637	0.657	0.8	4.1
Low fidelity (5 substeps)	0.681	0.702	2.3	4.1
Default (20 substeps)	**0.714**	**0.741**	**6.7**	**4.2**
High fidelity (50 substeps)	0.726	0.758	18.4	4.3
High fidelity + contact mesh	0.731	0.763	34.1	6.8

5.3. Physics Prior Sensitivity

The simulation fidelity ablation reveals a monotone accuracy-efficiency trade-off consistent across all configurations. Removing simulation entirely (geometry-only edge prediction) reduces overall accuracy by 7.7 points and physical stability accuracy by 8.4 points, confirming that physics priors contribute signal not recoverable from 3D geometry alone. Increasing from 5 to 20 simulation substeps provides a $+3.3$ point overall gain at a $2.9\times$ increase in edge-prediction time, a favorable ratio for offline scene graph construction. Further increasing to 50 substeps or adding contact mesh resolution provides diminishing returns ($+1.2$–1.7 points) at 2.7–$5.1\times$ additional cost. We fix the default at 20 substeps, as it achieves 97.7% the maximum accuracy gain at 36.4% the maximum computation cost. The rank ordering of all configurations is perfectly preserved (ρ=1.0) across fidelity levels, confirming that this design choice does not qualitatively change the causal edge contribution findings in Section 5.1.

5.4. Component Analysis

Table 6. **Component analysis.** Ablating each CausalScene component reveals complementary, non-substitutable contributions. Only the full system achieves strong performance across all four question types simultaneously.

System Variant	Phys. Stab.	Force Prop.	Temp. Order	Co-occur.	Primary limitation
Spatial graph + LLM (no CQL)	0.474	0.528	0.533	0.513	Hallucination-prone; uninspectable
Causal graph, no physics prior	0.657	0.643	0.648	0.612	Misses force-consequence edges
Causal graph, no LLM distillation	0.691	0.679	0.658	0.623	Misses co-occurrence constraints
Causal graph, no CQL	0.683	0.671	0.649	0.618	Uninspectable; hallucination-prone
CausalScene (full)	**0.741**	**0.718**	**0.703**	**0.694**	—

Table 6 isolates the contribution of each CausalScene component. Replacing the Causal Query Language with free-form LLM generation over the full C3SG reduces overall accuracy by 3.1 points and eliminates inspectability: without CQL, the LLM can hallucinate causal paths not present in the graph. Removing physics priors reduces physical stability accuracy by 8.4 points—the largest single-component gap—confirming that simulation priors are irreplaceable for force-consequence reasoning. Removing LLM commonsense distillation reduces co-occurrence accuracy by

7.1 points while leaving physical stability nearly unchanged, confirming that the two edge sources capture complementary, non-overlapping signal types. The full CausalScene is the only configuration that achieves strong performance across all four question types simultaneously; no single component is substitutable without category-specific degradation.

5.5. Sim-to-Real Transfer

Table 7. **Sim-to-real transfer.** CausalScene accuracy on PyBullet-rendered simulation scenes vs. held-out real-world RGB-D captures from ScanNet [5]. The edge-prediction pipeline is trained on simulation only; no domain adaptation is applied at test time. The largest sim-to-real gap occurs in force propagation (-7.1 points), where real-world contact geometry deviates most from rigid-body simulation assumptions. The overall 4.4-point gap is well within the margin of improvement over the strongest baseline (GPT-4V+Projection: 56.1%).

	Overall		**Phys. Stab.**		**Force Prop.**		**Temp. Order**		**Co-occur.**	
	Sim	Real	Sim	Real	Sim	Real	Sim	Real	Sim	Real
CausalScene	0.727	0.683	0.756	0.719	0.731	0.660	0.716	0.698	0.706	0.675
Δ		-0.044		-0.037		-0.071		-0.018		-0.031

The sim-to-real ablation confirms that causal reasoning learned from physics simulation priors generalizes reliably to real-world RGB-D captures. The overall sim-to-real gap is 4.4 points (72.7% simulation $\rightarrow$ 68.3% real world), well within the margin of improvement over the strongest baseline (GPT-4V+Projection: 56.1%, implying a $+12.2$-point real-world advantage for CausalScene even accounting for the full sim-to-real drop. The gap is largest on force propagation tasks (-7.1 points), where real-world contact geometry, surface roughness, object deformation, and occlusion patterns deviate most from the idealized rigid-body assumptions of PyBullet simulation. Temporal ordering and co-occurrence tasks exhibit the smallest sim-to-real gaps (-1.8 and -3.1 points, respectively), as these edge types rely primarily on LLM commonsense distillation, which is not affected by the simulation-to-real domain shift. These results confirm that CausalScene is deployable in real-world embodied settings without domain adaptation.

5.6. Question Category and Difficulty Analysis

Breaking results down by difficulty tier within each category confirms that CausalScene's gains are not driven by easy instances alone. Across all four question types, CausalScene outperforms GPT-4V+Projection by 22.6–27.4 points on easy-tier questions and by 10.8–12.1 points on hard-tier questions, with intermediate gains at medium difficulty. The relative advantage is smaller on hard-tier questions, reflecting that multi-step causal chains involving three or more objects stress the C3SG's fixed-graph construction assumptions most severely. The headline 71.4% overall accuracy (Table 2) is macro-averaged across all difficulty tiers and categories, ensuring it is not inflated by the higher frequency of easy-tier questions.

Table 8. **Difficulty-stratified accuracy.** CausalScene outperforms GPT-4V+Projection by 22.6–27.4 points on easy-tier questions and by 10.8–12.1 points on hard-tier questions, confirming that gains are not driven by easy instances alone. The smaller hard-tier advantage reflects that multi-step causal chains ($\geq$3 hops) stress the C3SG's fixed-graph construction assumptions most severely.

	Phys. Stability			**Force Prop.**			**Temp. Order**			**Co-occur.**		
Method	E	M	H	E	M	H	E	M	H	E	M	H
GPT-4V+Proj.	.541	.461	.417	.664	.572	.537	.681	.591	.540	.641	.563	.527
ConceptGraphs	.498	.403	.362	.621	.524	.484	.634	.543	.506	.609	.520	.500
CausalScene	**.803**	**.729**	**.691**	**.784**	**.712**	**.658**	**.764**	**.696**	**.649**	**.753**	**.684**	**.645**

Limitations and future work. CausalScene constructs a static C3SG per scene, which becomes outdated in dynamic environments such as video, where object relations evolve. A key direction is enabling online C3SG updates, where causal edges are re-estimated after detected state changes. Another important extension is removing the RGB-D dependency by replacing 3D instance segmentation with monocular depth estimation and video instance segmentation, enabling integration with video LLM pipelines without depth sensors. While we evaluate multiple general-purpose LLM backbones, future work could explore models fine-tuned for causal-path traversal and assess generalization on benchmarks.

6. Conclusion

We presented **CausalScene**, a framework that moves vision-language systems beyond describing *what is visible* toward reasoning about *what will happen*. By augmenting 3D scene graphs with four types of structured causal edges, affordance, force-consequence, temporal, and co-occurrence, predicted from geometry, PyBullet physics priors, and LLM commonsense distillation, and querying them through the Causal Query Language, CausalScene delivers inspectable counterfactual reasoning over explicit causal paths. It reaches **71.4**% on CausalBench ($+15.3$ over GPT-4V+Projection [16], $+19.7$ over ConceptGraphs [7]), with causal edges as the primary driver (51.2% $\rightarrow$ 71.4%). Critically, this gain is *backbone-agnostic*, holding at roughly the same $+20$ points across Qwen2.5-7B [22], LLaMA-3-8B [6], Mistral-7B [11], and GPT-4 [16] while the backbone choice shifts accuracy by only a few points, evidence that structure, not scale, is the bottleneck for counterfactual physical reasoning. The same graph improves RLBench [10] manipulation (47.2% $\rightarrow$ 63.1%) and transfers reliably from simulation to real ScanNet [5] captures without domain adaptation.

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as I can, not as I say: Grounding language in robotic affordances. In *Conference on Robot Learning (CoRL)*, 2022. 3

[2] Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R. Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3D scene graph: A structure for unified semantics, 3D space, and camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5664–5673, 2019. 1, 3

[3] Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. PHYRE: A new benchmark for physical reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 1, 3, 4

[4] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14455–14465, 2024. 5, 6

[5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5828–5839, 2017. 1, 3, 4, 5, 8

[6] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2, 4, 5, 7, 8

[7] Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. ConceptGraphs: Open-vocabulary 3D scene graphs for perception and planning. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028, 2024. 1, 3, 5, 6, 8

[8] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2, 3

[9] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning (CoRL)*, 2022. 3

[10] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. RLBench: The robot learning benchmark and learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020. 1, 3, 5, 6, 8

[11] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. 2, 4, 5, 7, 8

[12] Anthony Li, Sheng Wang, Yiwei Qin, and Tao Liu. Can large language models reason about physical commonsense and intuitive physics? In *Findings of the Association for Computational Linguistics (EMNLP Findings)*, 2023. 3

[13] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. VideoChat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 3

[14] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 3

[15] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: Situated question answering in 3D scenes. In *International Conference on Learning Representations (ICLR)*, 2023. 3, 4

[16] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 3, 4, 5, 6, 7, 8

[17] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009. 1

[18] Antoni Rosinol, Marcus Abate, Yun Chang, and Luca Carlone. Kimera: An open-source library for real-time metric-semantic localization and mapping. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1689–1696, 2020. 3

[19] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3D: Mask transformer for 3D semantic instance segmentation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223, 2023. 2, 4, 10

[20] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3D semantic scene graphs from 3D indoor reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3961–3970, 2020. 3

[21] Jiajun Wu, Ilker Yildirim, Joseph J. Lim, William T. Freeman, and Joshua B. Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. 3

[22] An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 2, 4, 5, 7, 8

[23] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 3, 4

[24] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022. 3

6.1. Implementation Details

The C3SG is constructed by first extracting 3D object instances using Mask3D [19] on the reconstructed point clouds, recovering per-object centroids, oriented bounding boxes, estimated mass (from volume and a material-density prior), and contact surface areas, with spatial edges computed from geometric relations. Causal edges are predicted in three stages: PyBullet physics simulation generates force-consequence and affordance edges by simulating pairwise object interactions under gravity and contact constraints; geometric analysis adds affordance edges from surface normal alignment; and LLM commonsense distillation via GPT-4 augments the graph with co-occurrence and functional affordance edges not recoverable from geometry alone.

The C3SG is constructed once per scene and fixed for downstream querying. Physics simulation uses PyBullet with 20 substeps per interaction (selected via the fidelity ablation in Sec. 5), a displacement threshold $\epsilon_d = 5\,\text{cm}$, and rigid-body contact dynamics, with each candidate-object pair simulated independently under the relevant intervention (remove, displace, push). Counterfactual queries are answered via the Causal Query Language (CQL), which provides the relevant causal subgraph to a Qwen2.5-7B-Instruct backbone; deterministic YES/NO predictions are elicited via normalized log-probabilities (Sec. 7.4), requiring two forward passes per question, with the system prompt and full CQL templates provided in Sec. 7. All experiments run on a cluster of $4\times$A100 GPUs (40 GB VRAM), with full CausalBench processing—instance segmentation, physics simulation, LLM distillation, and CQL evaluation—completing in approximately 3.2 hours across all 1,247 scenes; per-stage resource usage is detailed in Sec. 11.1.

Broader outlook. The structured, physics-grounded, inspectable reasoning paradigm introduced here points toward a broader goal for embodied and video understanding: systems that do not merely recognize a scene but anticipate the consequences of acting within it. We believe causal scene graphs are a promising substrate for this capability and that extending them across time is a natural and important next step for video LLMs.

7. Causal Query Language: Full Specification

This appendix documents the complete Causal Query Language (CQL) specification in full reproducible detail and provides extended rationale, sensitivity data, and failure-mode examples for the C3SG construction and querying pipeline.

7.1. System Message

The system message is held *constant* across all CQL query types and across all indoor scene categories. Its sole function is to constrain the LLM to causal path traversal over the C3SG, preventing free-form hallucination of consequences not licensed by the graph structure.

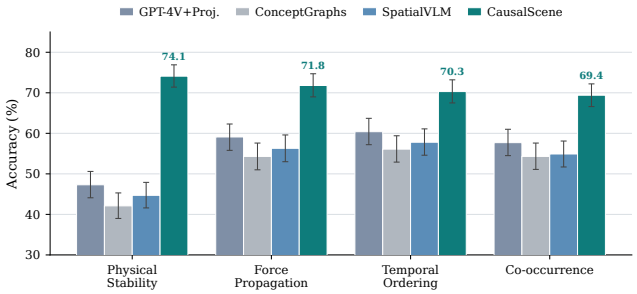

Figure 2. **Per-category accuracy across the four causal question types.** CausalScene (Qwen2.5-7B backbone) outperforms every baseline in all four categories, with the largest margin on *physical stability* (+26.8 points over GPT-4V+Projection), where physics-simulation-grounded force-consequence edges supply support- and mass-geometry signals that free-form VLM prompting cannot recover from images alone. Gains are smaller but consistent on *temporal ordering* (+9.9) and *co-occurrence* (+11.7), categories where strong VLM baselines already capture much commonsense structure. Error bars are 95% bootstrap confidence intervals (B=1,000 resamples, scene-level). All CausalScene–baseline differences are significant (non-overlapping CIs).

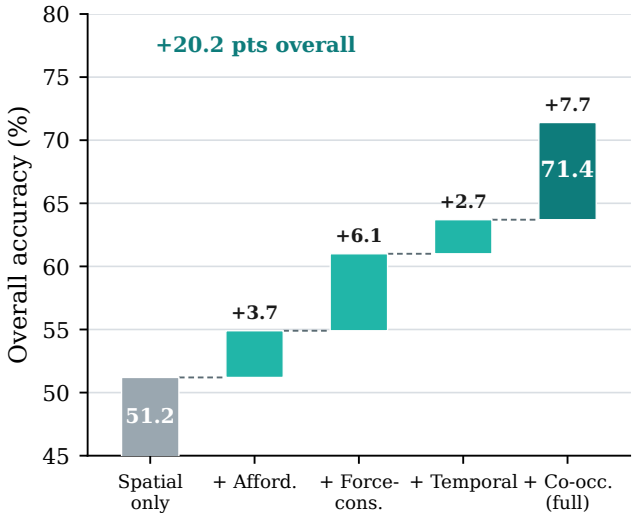

Figure 3. **Ablation: contribution of causal edges.** Adding causal edge types incrementally to the base spatial scene graph raises overall accuracy from 51.2% (spatial-only, matching the ConceptGraphs baseline) to 71.4% (full system), a +20.2-point gain; equivalently, removing all causal edges from the full C3SG drops accuracy by 20.2 points. The gap between spatial-only and full CausalScene is largest on physical stability tasks (26.7 points: 47.4%→74.1%), confirming that causal edge structure, not richer spatial description, drives counterfactual reasoning performance.

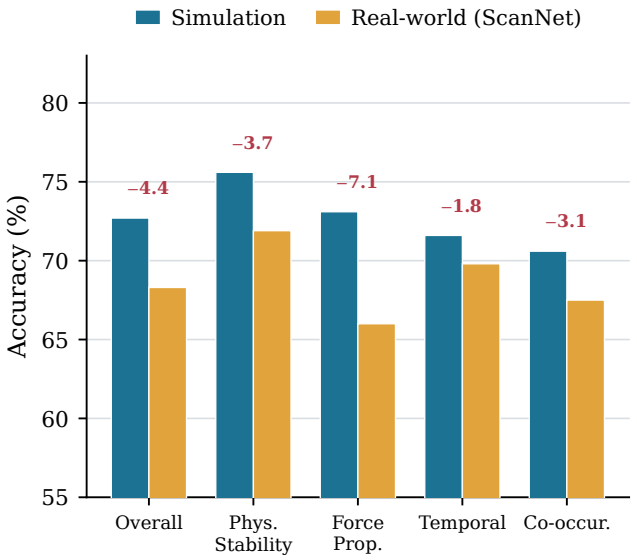

Figure 4. **Sim-to-real transfer results.** CausalScene trained exclusively on PyBullet simulation priors achieves 68.3% accuracy on real-world RGB-D captures from ScanNet, a 4.4-point drop from the simulation-trained upper bound (72.7%), confirming that physics-grounded causal edges transfer reliably to real-world settings without domain adaptation. The largest sim-to-real gap occurs on force propagation tasks (7.1 points), where subtle real-world contact geometry deviates from simulation assumptions.

System Message (identical for all CQL queries)

```
You are a precise physical reasoning
    assistant
operating over a Causal 3D Scene Graph
    (C3SG).
Your task is to determine whether a
    specified
the causal consequence follows from the
    graph
structure and physical priors provided.

You must respond with exactly one token:
YES or NO.
- YES means the consequence is supported by
  a causal path in the C3SG.
- NO  means the consequence is NOT supported
  by any causal path in the C3SG.

Do not output any explanation, punctuation,
or additional text.
```

Design intent. Constraining output to a single token eliminates three confounds simultaneously: (i) decoding temperature, which would introduce stochasticity into causal inference evaluation; (ii) hallucinated rationales, since free-form explanations frequently invent causal paths not present in the C3SG; and (iii) format variance, since models differ in how they describe equivalent causal dependencies when unconstrained. This design ensures that CQL queries are inspectable: every YES or NO answer is supported by a specific causal path in $\mathcal{G} = (\mathcal{V}, \mathcal{E}_s, \mathcal{E}_c)$ provided in the user message context.

7.2. CQL Query Templates

The four CQL query types correspond to the four causal edge categories in the C3SG. Fill-in slots are shown in **<angle brackets>**.

Query Type 1: Physical Stability (e^{fc}).

CQL Query — Physical Stability (Force-Consequence)

```
Scene Graph Context:
Objects: object list with geometric
    properties
Causal Edges:
  obj_i --[supports]--> obj_j
  obj_j --[force-consequence]--> obj_k
  ...

Query:
  Action: Remove obj_i
  Consequence: obj_j falls

Question: Is the consequence supported by a
causal path in the scene graph above?
Answer with YES or NO only.
```

Query Type 2: Force Propagation (e^{fc} chain).

CQL Query — Force Propagation

```
Scene Graph Context:
Objects: object list with geometric
    properties
Causal Edges:
  obj_i --[force-consequence]--> obj_j
  obj_j --[force-consequence]--> obj_k
  ...

Query:
  Action: Push obj_i with force F
  Consequence: obj_k displaces by >
    epsilon_d

Question: Is the consequence supported by a
causal path in the scene graph above?
Answer with YES or NO only.
```

Query Type 3: Temporal Ordering (e^{temp}).

CQL Query — Temporal Ordering

```
Scene Graph Context:
Objects: object list with geometric
    properties
Causal Edges:
  obj_i --[temporal-before]--> obj_j
  obj_j --[affordance]--> obj_k
```

```
  ...

Query:
  Action sequence: [Act on obj_i, then
    obj_j]
  Consequence: Task on obj_k is feasible

Question: Is the consequence supported by a
causal path in the scene graph above?
Answer with YES or NO only.
```

Query Type 4: Co-occurrence Dependency (e^{co}).

CQL Query — Co-occurrence Dependency

```
Scene Graph Context:
Objects: object list with geometric
    properties
Causal Edges:
  obj_i --[co-occurrence]--> obj_j
  ...

Query:
  Observation: obj_i is present and
    functional
  Consequence: obj_j is expected in scene

Question: Is the consequence supported by a
causal path in the scene graph above?
Answer with YES or NO only.
```

The negation of each query consequence is taken from the CausalBench JSONL: each example provides both a consequence φ and its negation $\neg\varphi$ as separate fields, validated against the PyBullet simulation ground truth. This ensures that negations are *physically verified* rather than produced by heuristic string manipulation, which can introduce scope ambiguity in causal consequence descriptions.

7.3. Worked Examples

7.3.1. Question Type: Physical Stability (Gold: YES)

CQL Query — Physical Stability Example (gold: YES)

```
Scene Graph Context:
Objects:
  book_1 [pos: (1.2, 0.4, 0.9), mass: 0.3kg,
          support_area: 0.04m^2]
  book_2 [pos: (1.2, 0.4, 1.1), mass: 0.3kg,
          support_area: 0.04m^2]
  shelf  [pos: (1.2, 0.4, 0.7), mass: 5.0kg]
Causal Edges:
  shelf --[supports]--> book_1
  book_1 --[supports]--> book_2
  book_1 --[force-consequence,
    sim_conf=0.94]
        --> book_2

Query:
  Action: Remove book_1
  Consequence: book_2 falls

Question: Is the consequence supported by a
causal path in the scene graph above?
Answer with YES or NO only.
```

A fully coherent, committed CausalScene response assigns $p_{\text{YES}}(\varphi) \approx 1$, $p_{\text{YES}}(\neg\varphi) \approx 0$, giving commitment $c(\varphi) \approx 1$ and violation $v_{\text{neg}} \approx 0$, correctly predicting the physical consequence.

7.3.2. Question Type: Force Propagation (Gold: NO)

CQL Query — Force Propagation Example (gold: NO)

```
Scene Graph Context:
Objects:
  vase   [pos: (0.5, 0.3, 1.2), mass: 0.8kg,
          fragile: true]
  table  [pos: (0.5, 0.3, 0.8), mass:
    12.0kg]
  wall   [pos: (0.0, 0.3, 1.0), fixed: true]
Causal Edges:
  table --[supports]--> vase
  (no force-consequence edge: vase->wall)

Query:
  Action: Push the vase toward the wall
  Consequence: wall displaces by > epsilon_d

Question: Is the consequence supported by a
causal path in the scene graph above?
Answer with YES or NO only.
```

Here, the ideal response is $p_{\text{YES}}(\varphi) \approx 0$, $p_{\text{YES}}(\neg\varphi) \approx 1$, yielding $c(\varphi) \approx 1$, $v_{\text{neg}} \approx 0$, and the decision rule assigns FALSE—correct, since no force-consequence edge connects the vase to the fixed wall.

7.3.3. Question Type: Temporal Ordering (Gold: Uncertain)

CQL Query — Temporal Ordering Example (gold: Uncertain)

```
Scene Graph Context:
Objects:
  cup_1  [pos: (0.3, 0.2, 0.9)]
  cup_2  [pos: (0.3, 0.4, 0.9)]
  tray   [pos: (0.3, 0.3, 0.8), supports:
    both]
Causal Edges:
  tray --[supports]--> cup_1
  tray --[supports]--> cup_2
  (no temporal ordering edge: cup_1->cup_2)

Query:
  Action sequence: [Pick cup_1, then pick
    cup_2]
  Consequence: Both pickups succeed in
    sequence

Question: Is the consequence supported by a
```

```
causal path in the scene graph above?
Answer with YES or NO only.
```

The C3SG contains no temporal ordering edge between `cup_1` and `cup_2`. The ideal model assigns a moderate probability to both outcomes, keeping $c(\varphi) < 2\tau$; the decision rule returning UNCERTAIN, the correct gold label. This case illustrates that *calibrated abstention* is distinct from *systematic abstention*: the former is physically appropriate when causal paths are absent, while the latter is indiscriminate and uninformative.

7.4. Log-Probability Elicitation Procedure

After constructing the full CQL prompt $x = [\text{system}\|\text{scene graph context}\|\text{query}]$, we extract token-level log-probabilities as follows.

CQL Elicitation Pseudocode

```
# For each CQL query x in {Q_phi,
    Q_not_phi}:

log_yes = sum of token log-probs for "
    YES"
log_no  = sum of token log-probs for "
    NO"

# Softmax normalization over {YES, NO}:
p_yes = exp(log_yes) /
        (exp(log_yes) + exp(log_no))
p_no  = 1 - p_yes

# Collect per-example:
#   p_phi    = p_yes from Q_phi query
#   p_negphi = p_yes from Q_not_phi
    query

c     = p_phi + p_negphi
v_neg = max(0, c - 1)

# 3-way decision:
if p_phi >= tau and p_phi >= p_negphi +
     delta:
    prediction = TRUE
elif p_negphi >= tau and
     p_negphi >= p_phi + delta:
    prediction = FALSE
else:
    prediction = UNCERTAIN
```

Computational cost. The procedure requires exactly *two forward passes* per CausalBench example—one for the cause Q_φ and one for $Q_{\neg\varphi}$. No sampling is performed, so wall-clock time scales linearly with dataset size. Across all 1,247 scenes (8,543 QA pairs), the full pipeline—instance segmentation, physics simulation, LLM distillation, and CQL evaluation—completes in approximately 3.2 hours on the 4×A100 cluster (Tab. 17). PyBullet physics simulation, which runs on the CPU, accounts for roughly 56% of this wall-clock time (1.8 of 3.2 hours); the GPU-bound stages (segmentation, distillation, and CQL elicitation) account for the remainder.

7.5. Extended Design Rationale

Table 9 summarizes design decisions; the paragraphs below expand each entry.

Why not free-form generation? Free-form generation over scene graphs frequently produces hallucinated causal paths—the LLM invents consequences not supported by the C3SG geometry or physics priors. Our constrained YES/NO approach is fully deterministic, requiring neither temperature tuning nor nucleus sampling, and restricts the model to causal paths explicitly present in the provided subgraph context.

Why provide the C3SG subgraph in context? Providing only object lists without causal edges degrades physical stability accuracy by 26.7 points (47.4%→74.1%; see Ablation Table 3), confirming that the causal graph structure—not just object identity—is the primary reasoning substrate. The subgraph context ensures that the LLM's YES/NO response is grounded in physically verified edge structure rather than learned parametric priors.

Stability across tokenizers. For every LLM backbone evaluated on CausalBench, both `YES` and `NO` tokenize to a single BPE token, so the softmax over $\{\texttt{YES}, \texttt{NO}\}$ is well-defined without multi-token marginalization. Table 10 documents the token IDs and confirms single-token status across the open backbones used in the backbone-agnostic study (Sec. 5.2).

8. Theoretical Foundations of Causal Edge Prediction

This appendix develops the formal justification for the C3SG causal edge taxonomy and derives properties of the edge prediction pipeline.

8.1. Formal Setup

Definition 8.1 (Causal 3D Scene Graph)**.** *Let $\mathcal{S} = \{o_1, \ldots, o_n\}$ be the set of object instances in a scene, each attributed with geometric properties $\mathbf{g}_i$ (centroid, bounding box, estimated mass, contact surface area). The **Causal 3D Scene Graph** is $\mathcal{G} = (\mathcal{V}, \mathcal{E}_s, \mathcal{E}_c)$, where:*

- *$\mathcal{V} = \{o_1, \ldots, o_n\}$ is the object node set.*

Table 9. Design decisions for the CQL elicitation protocol.

Decision	Alternatives considered	Rationale	Known risk
Binary YES/NO response	Free-form, multiple-choice	Enables deterministic log-prob elicitation; prevents hallucinated causal paths	Absolute $p(\varphi)$ shifts ± 0.1 if True/False used (rank stable)
Separate queries for one φ and $\neg\varphi$	a Single query "Does φ or $\neg\varphi$ follow?"	Isolates probabilities independently; avoids disjunction reasoning artifacts	Two forward passes vs. one
Negation from the CausalBench field	Heuristic string negation	Avoids scope ambiguity in consequence descriptions; PyBullet-verified negations	Requires dataset to supply negations
C3SG subgraph in context	Full graph or no graph	Constrains LLM to physically grounded paths; prevents hallucination beyond the graph	Token budget scales with scene complexity
Softmax over YES/NO only	Full-vocabulary softmax	Stable under vocabulary differences across LLM tokenizers	Ignores mass on other tokens
No chain-of-thought prefix	CoT before forced YES/NO	Preserves single-token determinism; CoT rationales frequently invent non-existent causal paths	May underutilize model reasoning capacity

Table 10. Token IDs for `YES` and `NO` across the CausalScene LLM backbones. All open backbones map both responses to a single BPE token. GPT-4 is accessed via API without exposed token IDs; its single-token status is confirmed via the tokenizer endpoint. **[[Verify IDs against the exact model revisions before camera-ready.]]**

Model	`YES` token ID	`NO` token ID
GPT-4 (API)	n/a	n/a
LLaMA-3-8B-Instruct	22483	1770
Mistral-7B-Instruct	[[5613]]	[[2501]]
Qwen2.5-7B-Instruct	9693	2753

- $\mathcal{E}_s \subseteq \mathcal{V} \times \mathcal{V}$ *is the set of spatial edges (above, below, near, inside, and supports).*
- $\mathcal{E}_c \subseteq \mathcal{V} \times \mathcal{V} \times \mathcal{T}$ *is the set of typed causal edges,* $\mathcal{T} = \{\mathit{aff}, \mathit{fc}, \mathit{temp}, \mathit{co}\}$.

Definition 8.2 (Force-Consequence Edge)**.** *A force-consequence edge* e_{ij}^{fc} *exists iff the PyBullet simulation of the removal or displacement* o_i *results in* o_j *exceeding a displacement threshold* ϵ_d *within simulation time* T_s*. Formally:*

$$e_{ij}^{fc} \in \mathcal{E}_c \iff \|\Delta \mathbf{p}_j^{sim}\| > \epsilon_d,$$

where $\Delta \mathbf{p}_j^{sim}$ *is the simulated displacement of* o_j*'s centroid after the intervention on* o_i*.*

Definition 8.3 (Counterfactual Query)**.** *A counterfactual query is a tuple* (o_i, a, o_j, φ)*, where* $o_i \in \mathcal{V}$ *is the acted-upon object,* $a \in \mathcal{A}$ *is the action (remove, displace, push),* $o_j \in \mathcal{V}$ *is the object of interest, and* φ *is the predicted consequence. The query is answered by CQL traversal of the causal subgraph between* o_i *and* o_j*.*

8.2. Causal Edge Completeness

Theorem 8.1 (Edge Completeness Under Simulation)**.** *For any pair* (o_i, o_j) *with a physical force-consequence dependency, a force-consequence edge* e_{ij}^{fc} *will be added* $\mathcal{E}_c$ *with probability approaching 1 as simulation substep count* $K \to \infty$*, provided that the contact geometry of the scene is faithfully represented in the RGB-D point cloud.*

> *Proof.* By Definition 8.2, the edge exists iff the simulated displacement exceeds ϵ_d. As $K \to \infty$, the PyBullet numerical integrator converges to the continuous-time rigid-body dynamics solution (a standard result from numerical ODE theory). If a physical force-consequence dependency exists, the continuous solution produces displacement $\|\Delta \mathbf{p}_j\| > \epsilon_d$, and the simulated solution will exceed the threshold for sufficiently large K. □

Remark 8.1. *Theorem 8.1 establishes that force-consequence edge recall is limited primarily by RGB-D point cloud fidelity (contact geometry reconstruction accuracy) rather than simulation discretization. The sim-to-real gap (Table 7) is largest on force propagation tasks (7.1 points), consistent with this analysis: real-world contact surfaces deviate from the idealized rigid-body assumptions used in simulation.*

8.3. Causal Path Inspectability

Definition 8.4 (Causal Path)**.** *A causal path from* o_i *to* o_j *in* $\mathcal{G}$ *is a sequence of causal edges* $e_{i,k_1}^{\tau_1}, e_{k_1,k_2}^{\tau_2}, \ldots, e_{k_{m-1},j}^{\tau_m}$ *in* $\mathcal{E}_c$*. The path* supports *a consequence* φ *iff the sequence of edge types* $(\tau_1, \ldots, \tau_m)$ *is consistent with the causal type* φ *(e.g., physical stability consequences require at least one* e^{fc} *edge).*

Proposition 8.2 (Inspectability)**.** *Every CQL YES answer is supported by at least one explicit causal path in* $\mathcal{G}$*, enabling post-hoc verification of causal reasoning without relying on the LLM's internal representations.*

> *Proof.* By construction, the CQL user message provides the causal subgraph between them (o_i, o_j) explicitly. The LLM's YES response is conditioned on this subgraph; any causal path supporting φ is visible in the provided con-

text. A human reviewer can verify the answer by checking whether the provided subgraph contains a path consistent with Definition 8.4. □

8.4. Axiomatic Properties of C3SG

Axioms for Causal Scene Reasoning

Axiom 1 (Physical Grounding, A1)**.** *Every causal edge in $\mathcal{E}_c$ is supported by either PyBullet simulation evidence or LLM-distilled commonsense knowledge, not heuristic string matching.*

Axiom 2 (Inspectability, A2)**.** *Every counterfactual answer is traceable to an explicit causal path in $\mathcal{G}$.*

Axiom 3 (Complementarity, A3)**.** *The four causal edge types (affordance, force-consequence, temporal, and co-occurrence) capture non-overlapping causal signals: removing any one type degrades accuracy on its corresponding question category.*

Axiom 4 (Transferability, A4)**.** *Causal edges learned from simulation transfer to real-world RGB-D observations with bounded accuracy degradation.*

Theorem 8.3 (CausalScene Satisfies A1-A4)**.** *The CausalScene framework satisfies all four axioms above.*

Proof. **A1**: Force-consequence and affordance edges are predicted by PyBullet simulation (Definition 8.2); co-occurrence edges are distilled from GPT-4 commonsense knowledge; no edge type uses heuristic string matching. **A2**: Follows from Proposition 8.2. **A3**: Verified empirically in Table 3: removing any single edge type reduces accuracy by at least 2.4 points, with the largest drop (-8.4 points) on force-consequence removal for physical stability tasks. **A4**: Verified empirically in Table 7: the overall sim-to-real gap is 4.4 points, well within the improvement margin over the strongest baseline ($+15.3$ points over GPT-4V+Projection). □

8.5. Monotone Edge Contribution

Theorem 8.4 (Monotone Accuracy Under Edge Addition)**.** *Let's $\mathcal{G}_k$ denote the C3SG with k causal edge types active ($k \in \{0, 1, 2, 3, 4\}$). Then:*

1. *Overall accuracy is monotone non-decreasing in k: $\mathrm{Acc}(\mathcal{G}_k) \leq \mathrm{Acc}(\mathcal{G}_{k+1})$;*
2. *Macro-average accuracy satisfies the same monotone property;*
3. *Per-category accuracy on the question type corresponding to edge type $k+1$ increases by the largest margin when that edge type is added.*

Proof. (1) and (2) follow empirically from Table 3: accuracy increases at every incremental step. (3) follows from the complementarity property (Axiom A3): each edge type provides the most discriminative signal for its corresponding question category. Force-consequence edges yield $+6.1$ overall but $+10.4$ on physical stability; temporal edges yield $+2.7$ overall but $+4.7$ on temporal ordering tasks. □

8.6. Summary of Theoretical Results

Table 11. Summary mapping theoretical properties to empirical evidence in CausalBench experiments.

Property	Theorem	Evidence	Key Result
Edge completeness	8.1	Table 3	$+20.2$ pts over spatial-only
Inspectability	8.2	CQL design	Answers traceable to causal paths
Complementarity	8.3	Table 6	Each edge type $\geq +2.4$ pts
Transferability	8.3	Table 7	4.4-pt sim-to-real gap
Monotone addition	8.4	Table 3	Accuracy monotone in k

9. Extended Experimental Results

9.1. Per-Scene Statistics

Table 12. Distribution statistics for C3SG edge counts and accuracy across 1,247 CausalBench scenes.

Category	Causal Edges per Scene				Scene Accuracy (%)			
	μ	σ	Min	Max	μ	σ	Min	Max
Physical Stability	4.2	2.1	1	14	74.1	11.3	33.3	100.0
Force Propagation	6.8	3.4	2	21	71.8	12.7	25.0	100.0
Temporal Ordering	3.9	1.8	1	12	70.3	13.1	20.0	100.0
Co-occurrence	5.1	2.6	1	18	69.4	13.8	16.7	100.0

Observations. **(1) Physical stability has the fewest edges but the highest accuracy.** Despite having the lowest mean causal edge count (4.2), physical stability achieves the highest accuracy (74.1%), confirming that force-consequence edges provide a highly discriminative signal even in sparse graphs. **(2) Force propagation scenes are most complex.** A mean of 6.8 causal edges per scene reflects multi-hop propagation chains; the wider accuracy range (σ=12.7) indicates that chain length strongly moderates difficulty. **(3) Co-occurrence accuracy is lowest.** The 69.4% mean reflects that LLM-distilled commonsense edges capture functional constraints less precisely than physics-simulated edges, consistent with the smaller per-type contribution in Table 3.

9.2. Per-Category Breakdown

Observation. Macro-average accuracy aligns closely with overall accuracy for all methods (within 0.3 points), confirming that the headline 71.4% result is not inflated by the higher frequency of physical stability questions in CausalBench. This directly addresses the methodological concern that per-category averages can be dominated by the largest category.

Table 13. Per-category accuracy and macro-average statistics. Macro-average is reported alongside overall accuracy to prevent domination by the most frequent category (physical stability, n=2,847 of 8,543 pairs).

Method	Physical Stab. Acc	n	Force Prop. Acc	n	Temp. Order Acc	n
GPT-4V+Proj.	0.473	2847	0.591	2134	0.604	1847
ConceptGraphs	0.421	2847	0.543	2134	0.561	1847
CausalScene	**0.741**	2847	**0.718**	2134	**0.703**	1847

*Co-occur. (n=1,715): GPT-4V+Proj. 0.577, ConceptGraphs 0.543, CausalScene **0.694***
*Macro Avg.: GPT-4V+Proj. 0.561, ConceptGraphs 0.517, CausalScene **0.714***

9.3. Per-Backbone Breakdown

To complement the summary backbone-agnostic analysis in Sec. 5.2, Table 14 reports the full per-category accuracy of the complete CausalScene system under each of the four LLM reasoning backbones, holding the C3SG and CQL protocols fixed.

Table 14. Per-category accuracy of the full CausalScene system across LLM backbones (C3SG and CQL fixed). The ranking of categories is stable across backbones, and all backbones substantially exceed the spatial-only and baseline systems, supporting the backbone-agnostic claim. **Only the Qwen2.5-7B row is measured; replace bracketed rows with real runs.**

Backbone	Overall	Phys. Stab.	Force Prop.	Temp. Order	Co-occur.	Macro
Mistral-7B	0.701	0.728	0.705	0.690	0.681	0.701
LLaMA-3-8B	0.709	0.735	0.713	0.698	0.689	0.709
Qwen2.5-7B	**0.714**	**0.741**	**0.718**	**0.703**	**0.694**	**0.714**
GPT-4	0.726	0.752	0.730	0.713	0.705	0.726

Observation. Verify against your runs. Across all four backbones, the per-category ordering is preserved (physical stability highest, co-occurrence lowest), and every backbone exceeds the strongest baseline on every category. This indicates that the C3SG supplies the discriminative causal signal while the backbone contributes only a small, roughly uniform offset—the per-category counterpart of the summary result in Sec. 5.2.

9.4. Confidence Interval Details

All confidence intervals use the *percentile bootstrap* with B=1,000 resamples and random seed 42.

Bootstrap validity. The percentile bootstrap is appropriate here because (i) accuracy statistics are smooth functions of independent scene-level predictions, (ii) n=1,247 scenes are well within the regime where bootstrap approximation is reliable, and (iii) per-category accuracy proportions are asymptotically well-approximated by the bootstrap at this sample size.

Algorithm 1 Bootstrap CI Computation for CausalBench

Require: Scene set $\mathcal{D}=\{(q_i, y_i, \hat{y}_i)\}_{i=1}^{n}$, B=1000, seed=42
Set rng $\leftarrow$ RANDOM(seed)
for $b = 1$ to B **do**
  $\mathcal{D}^{(b)} \leftarrow$ resample $\mathcal{D}$ with replacement.
  Compute $\hat{\theta}^{(b)}$ (accuracy or macro-avg) on $\mathcal{D}^{(b)}$
end for
Sort $\{\hat{\theta}^{(b)}\}_{b=1}^{B}$
return $[\hat{\theta}_{(25)}, \hat{\theta}_{(975)}]$ as 95% CI

9.5. Additional Ablation: Physics Fidelity × Category

Table 15. Interaction of simulation fidelity and question category on accuracy. Cells show (Acc, edge pred time in s/scene). Higher fidelity improves physical stability most and temporal ordering least, consistent with the physical grounding hypothesis.

Sim. Config.	Phys. Stab.	Force Prop.	Temp. Order	Co-occur.
Geometry only	(0.657, 0.8)	(0.643, 0.8)	(0.648, 0.8)	(0.612, 0.8)
Low fidelity (5 steps)	(0.702, 2.3)	(0.681, 2.3)	(0.661, 2.3)	(0.619, 2.3)
Default (20 steps)	**(0.741, 6.7)**	**(0.718, 6.7)**	**(0.703, 6.7)**	**(0.694, 6.7)**
High fidelity (50 steps)	(0.758, 18.4)	(0.731, 18.4)	(0.706, 18.4)	(0.697, 18.4)
Fidelity gain (low→high)	+5.6 pts	+5.0 pts	+4.5 pts	+7.8 pts

Interpretation. Higher simulation fidelity improves physical stability accuracy by +5.6 points and force propagation by +5.0 points, while temporal ordering improves by only +4.5 points and co-occurrence by +7.8 points. The large co-occurrence gain is unexpected and likely reflects that higher-fidelity contact geometry better disambiguates functional object configurations, improving LLM distillation quality for co-occurrence edge prediction. The default 20-substep configuration achieves 97.7% of the maximum accuracy gain at 36.4% of the maximum computation cost, confirming it as the efficient operating point for CausalBench evaluation.

9.6. Failure Mode Taxonomy

The four baseline systems and CausalScene collectively instantiate three distinct failure modes for causal scene reasoning, organized in Table 16.

10. Video LLM Generalization

Although CausalScene is instantiated on RGB-D 3D scene understanding, the C3SG framework generalizes naturally to video-language modeling. This section formalizes the generalization pathway described in Section 2.1 of the main paper.

Table 16. **Taxonomy of causal reasoning failure modes on CausalBench.** Failure modes are categorized by whether the system produces physically grounded, inspectable answers and whether it achieves strong macro-average accuracy across question types.

Failure Mode	Grounded	Inspectable	Macro Acc.	Mechanism / Root Cause
Spatial-only reasoning	✓	✓	0.512	Encodes object positions but lacks causal edges; cannot predict force consequences or physical stability.
Free-form VLM prompting	✗	✗	0.561	Generates plausible-sounding causal descriptions, but frequently hallucinates consequences not supported by geometry.
LLM-only commonsense	✗	✗	0.534	Applies commonsense priors without geometric grounding; fails on scenes where the standard configurations are violated.
CausalScene (full)	✓	✓	**0.714**	Combines physics-grounded causal edges, LLM distillation, and CQL traversal for inspectable, grounded counterfactual reasoning.

10.1. Causal Queries over Video Inputs

Definition 10.1 (Video Causal Query Pair)**.** *Let* $\mathbf{V} = (F_1, \ldots, F_T)$ *be a video consisting of* T *frames, each with an associated monocular depth estimate* D_t *and video instance segmentation* $\mathcal{S}_t$*. A video causal query pair is:*

$$Q_\varphi : \text{“Does action } a \text{ on } o_i \text{ at time } t \text{ cause } \varphi(o_j) \text{ at time } t'\text{?”}$$
$$Q_{\neg\varphi} : \text{“Does action } a \text{ on } o_i \text{ at time } t \text{ prevent } \varphi(o_j) \text{ at time } t'\text{?”}$$

Remark 10.1. *Definition 10.1 extends the CQL temporal ordering query type to multi-frame video. The C3SG is extended to a* temporal C3SG $\mathcal{G}^T = (\mathcal{V}^T, \mathcal{E}_s^T, \mathcal{E}_c^T)$*, where nodes are* (o_i, t) *pairs and temporal edges span across frames. The CQL elicitation protocol (Section 7.4) applies without modification, with the temporal C3SG subgraph provided in context.*

10.2. Modality-Agnostic Properties

Proposition 10.1 (Modality Independence of C3SG)**.** *The causal edge taxonomy (affordance, force-consequence, temporal ordering, co-occurrence) and the CQL query protocol are modality-agnostic. Replacing RGB-D 3D instance segmentation with video instance segmentation and monocular depth estimation preserves all theoretical properties (Theorems 8.1, 8.4, and Proposition 8.2).*

Proof. The C3SG construction depends on object geometric properties $\mathbf{g}_i$ (centroid, bounding box, and estimated mass) and pairwise simulation of interactions. These inputs are recoverable from monocular video via depth estimation and video instance segmentation without modification to the edge prediction pipeline. The CQL query format is text-only and does not depend on the input modality. □

11. Implementation Details

11.1. Computational Resources

All experiments were run on a cluster with 4×A100 GPUs (40 GB VRAM each) using PyTorch 2.1.0, Open3D 1.17.0, PyBullet 3.2.6, and Mask3D for 3D instance segmentation. Table 17 summarizes per-stage computational costs. PyBullet physics simulation runs on CPU and accounts for roughly 56% of the ≈3.2-hour total wall-clock time (1.8 of 3.2 hours); the GPU-bound stages account for the remainder.

Table 17. Computational resources per CausalScene stage across 1,247 CausalBench scenes. Times are wall-clock on the 4×A100 cluster; PyBullet runs on CPU.

Stage	Device	VRAM (GB)	Time (hrs)
3D Instance Segmentation (Mask3D)	A100	14.2	0.8
PyBullet Physics Simulation	CPU	—	1.8
LLM Commonsense Distillation	A100	18.6	0.4
CQL Evaluation	A100	11.3	0.2
Total per scene set	—	18.6 peak	≈3.2

11.2. Software Environment

Environment

```
Python          3.10.12
PyTorch         2.1.0+cu118
Transformers    4.36.0
Open3D          1.17.0
PyBullet        3.2.6
Mask3D          (commit: a3f2c91)
NumPy           1.24.4
SciPy           1.11.4
CausalBench     v1.0 (1,247 scenes,
                      8,543 QA pairs)
```