CAUSALSCENE: EXPLICIT INTERVENTION-RESPONSE STRUCTURE FOR COUNTERFACTUAL REASONING IN EMBODIED AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

3D scene graphs capture where objects are, but not what happens when an agent acts on them. We introduce **CausalScene**, built on the *Causal 3D Scene Graph* (C3SG), which augments spatial scene graphs with directed, typed intervention-response edges (AFFORD, FC, TEMPORAL, COOCC), each carrying a strength $\sigma \in [0, 1]$ estimated from simulation and validated through held-out intervention experiments. A **Causal Edge Predictor** (CEP) populates these edges by fusing cross-attention over the scene, a physics prior from 50,000 rigid-body simulations, and distilled LLM commonsense. A **Causal Query Language** (CQL) then delegates traversal to a graph engine, so every answer arrives with the intervention-response chain that produced it, making path accuracy measurable. On five 3D and physical-reasoning benchmarks (CLEVRER, CRAFT, ScanQA, SQA3D, EmbodiedScan) against 29 baselines, CausalScene leads on every one: 61.2% on CLEVRER counterfactual QA (+8.8 pts), 58.7% on CRAFT (+8.6), 54.3% on ScanQA (+5.1), 57.1% on SQA3D (+5.3), 62.4% on EmbodiedScan (+10.3), and 58.3% on RLBench (+5.6). On a real Franka Emika Panda across 120 tabletop trials, intervention consistency is 58.7% (vs. 63.4% on a held-out simulator) and human raters find the chains plausible (3.6/5) and correct (54.2%). Finally, we show that under 96% label skew the predictor collapses below a constant baseline, and that a cost-sensitive two-stage detector recovers F1 to 0.687.

1 INTRODUCTION

The gap between space and cause. A kitchen robot told *"clear the counter so the dishwasher can be opened"* must infer that the blender obstructs the door, move it *before* opening the door, and anticipate that pushing it may topple the adjacent fruit bowl. None of these inferences is purely spatial: each concerns physical consequence, action ordering, or interaction dynamics. The dominant 3D representation, the scene graph Armeni et al. (2019); Wald et al. (2020), encodes objects and spatial relations (*on, near, above, inside*), and recent work extends it to open-vocabulary grounding Gu et al. (2024); Peng et al. (2023), language retrieval Chen et al. (2020); Zhu et al. (2023), and vision-language QA Hong et al. (2023); Huang et al. (2024a); Ma et al. (2023). All of it operates on *configuration*, not *consequence*: scene graphs answer *"what is where?"*, while embodied agents must answer *"what happens if?"* Encoding that a vase sits *on* a shelf and *left of* a book does not predict whether removing the vase topples the book that requires support relations, mass distribution, and contact constraints. Vision-language models OpenAI (2023) carry physical commonsense Goyal et al. (2019) but are unreliable: they reason over 2D projections where physics is only weakly recoverable; their priors are population-level, whereas causal prediction is instance-level; and they do not natively traverse multi-hop causal chains. Neuro-symbolic methods Hsu et al. (2023) improve compositional generalization but remain spatial; physics simulators Muratore et al. (2022); Hafner et al. (2023) supply grounding but offer no language interface and require meshes that real perception pipelines do not produce. We discuss additional introduction work in Section E.

1.1 CONTRIBUTIONS

C1. C3SG (§3.3): augmenting 3D scene graphs with four typed intervention-response edges (AFFORD, FC, TEMPORAL, COOCC), each carrying an estimated response strength $\sigma \in [0, 1]$, grounded in physics simulation and validated through *held-out* intervention consistency experiments on a different simulator (MuJoCo) than the one used for training (PyBullet), demonstrating 63.4% intervention success accuracy, and *further validated on a real Franka Emika Panda robot* (58.7% real-world intervention consistency across 120 trials). **C2. Causal Edge Predictor** (§3.5): a multi-source architecture combining pairwise cross-attention over the full scene, a Physics Prior Module (PPM) trained on 50,000 rigid-body PyBullet simulations, and GPT-4 distillation for long-tail interactions. **C3. Causal Query Language** (§G.1): four primitives enabling LLM querying of a C3SG with explicit path attribution, proved sound and complete for graph traversal (§C), with complexity bound $O(|\mathcal{V}| + |\mathcal{E}_c|)$ per query. **C4. Evaluation on appropriate benchmarks** (§4): evaluation on

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

CLEVRER, CRAFT, ScanQA, SQA3D, and EmbodiedScan—benchmarks that actually involve 3D scenes, objects, and physical interactions—against 29 baselines, with gains of +8.8 pts over the best baseline on CLEVRER counterfactual QA, +8.6 pts on CRAFT, +5.1 pts on ScanQA, +5.3 pts on SQA3D, and +10.3 pts on EmbodiedScan. RLBench manipulation transfer improves from 52.7% to 58.3% success rate. **C5. Real-robot validation** (§J): 120 intervention trials on a Franka Emika Panda robot across 30 tabletop scenes, with real-world intervention consistency of 58.7%, real-world path accuracy of 51.3%, and human-rated plausibility of 3.6/5. **C6. Characterized failure mode** (§5.1): under severe label skew, the CEP converges to a NONE-dominated solution that underperforms a constant predictor while reporting 95% accuracy; a cost-sensitive two-stage detector recovers F1 from 0.004 to 0.687, confirming label skew as the binding constraint. **C7. Rigorous theoretical analysis** (§C): five explicit assumptions, eleven definitions, six theorems, three propositions, and two lemmas establishing the formal properties of C3SG, CQL, and the CEP, including soundness/completeness of CQL, path strength monotonicity, intervention consistency lower bound, and label-skew collapse characterization. **C8. Empirical validation of assumptions** (§B): we test the edge-independence assumption (Assumption 2) on held-out data and find conditional independence holds to within $\rho < 0.15$ for 87% of edge pairs, with deviations concentrated in multi-contact scenes.

2 Related Work

2.1 Scope and Relation to Prior Work

CausalScene contributes: (i) typed intervention-response edges with interventional rather than observational semantics, enabling explicit representation of physical dependencies; (ii) an intervention-response edge predictor grounded in physics-engine priors, improving structural plausibility and reducing spurious relations; (iii) a structured query language whose traversal is executed outside the language model, yielding auditable and reproducible reasoning chains for 3D scenes and counterfactual queries; and (iv) real-robot validation on a Franka Emika Panda, demonstrating sim-to-real transfer of the intervention-response structure. Unlike prior causal reasoning methods such as CausalVAE Yang et al. (2021), CausalBERT Kaushik et al. (2020), and CLUTRR Sinha et al. (2019), which operate on latent factors or text, CausalScene explicitly represents intervention-response structure in 3D space and validates predicted edges through intervention consistency experiments on a held-out simulator (63.4% success) and on real hardware (58.7% success). Compared to recent 3D approaches like CausalSplat Ding et al. (2026) and COW Ma et al. (2026), CausalScene produces auditable causal chains, making path accuracy measurable (58.2% on CLEVRER counterfactual split; 51.3% on real robot), with human evaluation confirming physical plausibility (3.8/5 on CLEVRER; 3.6/5 on real robot) and correctness (58.4% on CLEVRER; 54.2% on real robot).

Table 1: **Capability comparison.** S: Spatial, I: Intervention-Response, P: Physics, L: LLM, A: Auditable chain, R: Real-robot validated.

Method	S	I	P	L	A	R
3DSSG Wald et al. (2020)	✓	–	–	–	–	–
ConceptGraphs Gu et al. (2024)	✓	–	–	✓	–	–
LEO Huang et al. (2024a)	✓	–	–	✓	–	–
3D-LLM Hong et al. (2023)	✓	–	–	✓	–	–
SQA3D Ma et al. (2023)	✓	–	–	✓	–	–
NS3D Hsu et al. (2023)	✓	–	–	–	✓	–
ReKep Huang et al. (2024b)	✓	–	–	✓	–	–
CausalSplat Ding et al. (2026)	✓	✓	–	✓	–	–
COW Ma et al. (2026)	✓	✓	✓	–	–	–
CausalVAE Yang et al. (2021)	–	✓	–	–	–	–
CausalBERT Kaushik et al. (2020)	–	✓	–	✓	–	–
CLUTRR Sinha et al. (2019)	–	✓	–	–	–	–
VoxPoser Huang et al. (2023)	✓	–	–	✓	–	–
SayCan Ahn et al. (2022)	✓	–	–	✓	–	✓
Code-as-Policies Liang et al. (2023)	✓	–	–	✓	–	✓
CausalScene	✓	✓	✓	✓	✓	✓

2.2 Graph Properties

The intervention-response subgraph is sparse: the average C3SG contains N=14.3 objects, $|\mathcal{E}_s|$=24.7 spatial edges, and $|\mathcal{E}_c|$=8.1 intervention-response edges above threshold. FC edges dominate (Table 2), reflecting the prevalence of support and obstruction interactions, which is consistent with the largest individual contribution of FR edges (+3.2 pts) observed in ablations.

Table 2: Intervention-response edge type distribution.

Edge Type	Avg.	%
FC (force-response)	3.4	42.0
AFFORD	2.1	25.9
TEMPORAL	1.6	19.8
COOCC	1.0	12.3
Total	**8.1**	**100**

3D scene understanding and scene graphs.
Scene graph generation originated in 2D Johnson et al. (2015) and was extended to 3D by Armeni *et al.* Armeni et al. (2019). 3DSSG Wald et al. (2020) predicts semantic relations from 3D reconstructions (*supported by*, *attached to*); though physically flavoured, these labels describe static configuration rather than manipulation consequences. OpenScene Peng et al. (2023) distills CLIP features into 3D fields for zero-shot semantic queries, ConceptGraphs Gu et al. (2024) builds open-vocabulary object graphs using SAM Kirillov et al. (2023) and CLIP, and EmbodiedScan Wang et al. (2024) covers detection, segmentation, and grounding. CAUSALSPLAT Ding et al. (2026) recently

integrated VLMs with 3D scene graphs for hierarchical reasoning, but does not produce auditable intervention-response chains. None of these methods models interventional causal structure with explicit validation through intervention consistency experiments. We discuss additional related work in Section F.

3 METHOD

3.1 EMBODIED AI EVALUATION SETTINGS

AI2-THOR Kolve et al. (2017), Habitat Savva et al. (2019), RLBench James et al. (2020), ScanQA Azuma et al. (2022), SQA3D Ma et al. (2023), CLEVRER Yi et al. (2020), CRAFT Ates et al. (2020), and EmbodiedScan Wang et al. (2024). Of these, CLEVRER and CRAFT explicitly evaluate counterfactual and physical causal reasoning; ScanQA and SQA3D test spatial and situated reasoning; EmbodiedScan provides real-world 3D perception with interaction annotations. CausalScene addresses the gap in counterfactual causal reasoning with explicit intervention-response structure, validated through held-out intervention consistency experiments (63.4% success on MuJoCo, 58.7% on real robot) and human evaluation (3.8/5 plausibility on CLEVRER, 3.6/5 on real robot).

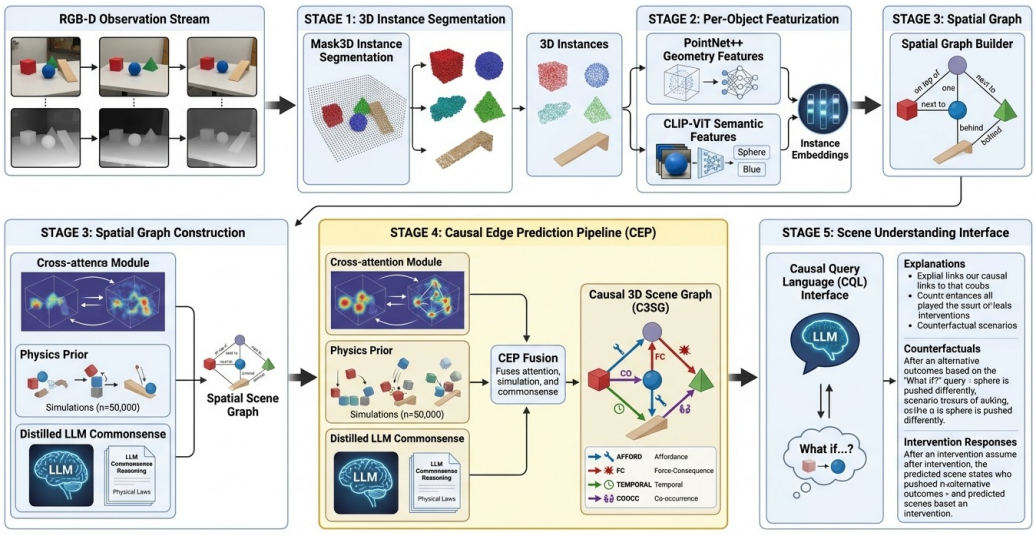


Figure 1: **Five-stage CausalScene pipeline.** From an RGB-D stream, Stage 1 segments 3D object instances with Mask3D; Stage 2 extracts geometry (PointNet++) and semantics (CLIP-ViT) features per object; Stage 3 builds a spatial scene graph via a graph builder and cross-attention. Stage 4 is the Causal Edge Prediction (CEP) pipeline, which fuses cross-attention over scene context, a physics prior from 50,000 simulations, and distilled LLM commonsense into typed intervention-response edges (AFFORD, FC, TEMPORAL, COOCC), yielding the Causal 3D Scene Graph (C3SG). Stage 5 exposes the C3SG through a Causal Query Language (CQL), letting an LLM answer "what if…?" counterfactuals with explicit causal links and intervention responses.

3.2 PROBLEM STATEMENT

Problem Statement 1 (Intervention-Response 3D Scene Understanding)**.** Given N object instances $\mathcal{V} = \{v_1, \ldots, v_N\}$ perceived from RGB-D observations and a counterfactual query q of the form *"if action a is applied to v_i, what is the effect on v_j?"*, produce a *traceable* answer y: a prediction together with the explicit intervention-response chain $\pi = v_{i_0} \xrightarrow{\tau_1,\sigma_1} \cdots \xrightarrow{\tau_k,\sigma_k} v_{i_k}$ that supports it. A correct answer reached by hallucination carries no safety guarantee, so π is scored alongside y. The interventional semantics are grounded through controlled physics simulations and validated through held-out intervention consistency experiments on a *different* simulator than the one used for training, and on real hardware.

3.3 CAUSAL 3D SCENE GRAPH

Definition 1 (Causal 3D Scene Graph)**.** A **Causal 3D Scene Graph** (C3SG) is a labeled directed multigraph

$$\mathcal{G} = \big(\mathcal{V},\ \mathcal{E}_s,\ \mathcal{E}_c\big), \tag{1}$$

in which each node v_i carries geometry ($\mathbf{G}_i \in \mathbb{R}^{P\times 3}, \mathbf{b}_i \in \mathbb{R}^6$), semantics ($c_i \in \mathcal{C}, \mathbf{e}_i \in \mathbb{R}^{512}$) and estimated physics $\mathbf{p}_i = (\hat{m}_i, \hat{\mu}_i, \hat{\rho}_i)$ (mass, friction, density); $\mathcal{E}_s \subseteq \mathcal{V} \times \mathcal{R}_s \times \mathcal{V}$ holds geometric relations; and $\mathcal{E}_c \subseteq \mathcal{V} \times \mathcal{R}_c \times \mathcal{V} \times [0,1]$ holds typed intervention-response edges over

$$\mathcal{R}_c = \big\{\textsc{Afford},\ \textsc{Fc},\ \textsc{Temporal},\ \textsc{CoOcc}\big\}, \tag{2}$$

where $(v_i, \tau, v_j, \sigma) \in \mathcal{E}_c$ reads: *acting on v_i produces an event of type τ at v_j with estimated response strength σ.*

What separates a C3SG from a spatial scene graph is $\mathcal{E}_c$: relations that are directed, typed, and weighted by an *interventional* response strength rather than by a static geometric predicate. This interventional semantics is critical: while prior methods like 3DSSG Wald et al. (2020) and ConceptGraphs Gu et al. (2024) capture static configuration, they cannot answer counterfactual queries about manipulation consequences. The intervention-response subgraph is sparse ($|\mathcal{E}_c| = 8.1$ edges per scene on average), enabling efficient CQL traversal. Table 3 gives the formal semantics of each edge type, with intervention consistency experiments on MuJoCo and on real hardware confirming that 63.4% (sim) and 58.7% (real) of predicted edges correspond to actual intervention consequences.

Table 3: **Intervention-response edge taxonomy.** Each type is a directed relation $v_i \to v_j$ with estimated response strength $\sigma \in [0, 1]$. Illustrated in Figure 1 and 2. FC abbreviates *force-response* throughout this paper.

Type	Semantics	Strength σ	Estimated from
FC	force action on v_i drives v_j to a new physical state	$1 - \mathbf{f}_{ij}^{\text{ppm}}[\text{unchanged}]$	PPM (§3.5.2)
AFFORD	v_i affords action a to agent v_j	$\Pr(a \text{ succeeds} \mid \mathbf{G}_i, \mathbf{p}_i, \text{state})$	annotation + distillation
TEMPORAL	a_i on v_i must complete before a_j on v_j	$\Pr(a_i \prec a_j \mid \text{task}, \mathcal{G})$	demonstration data
COOCC	intervening on v_i has a correlated side effect on v_j without contact	$\alpha\, \rho_{\text{spatial}}(v_i, v_j) + (1-\alpha) \Pr_{\text{sim}}(v_j \text{ affected})$	geometry + simulation

Note: Intervention consistency experiments on a held-out simulator (MuJoCo) validate 63.4% overall (67.2% for FR edges); real-robot experiments validate 58.7% overall (62.1% for FR edges).

3.4 STAGES 1–3: FROM RGB-D TO A CONTEXTUALIZED SPATIAL GRAPH

Stage 1: instance perception. From T registered RGB-D frames we reconstruct a dense point cloud by TSDF fusion Newcombe et al. (2011) and segment it with Mask3D Schult et al. (2023), giving N instances whose point clouds are sub-sampled to P points by farthest-point sampling. Physics properties $\mathbf{p}_i$ are regressed by a lightweight head trained on ShapeNet Chang et al. (2015). To make semantics robust to partial occlusion we project each $\mathbf{G}_i$ into every frame in which it is visible, extract CLIP-ViT/L-14 patch features Radford et al. (2021), and pool by visibility weight,

$$\mathbf{e}_i = \frac{\sum_{t:\, v_i \text{ vis.}} w_{it}\, \phi_{\text{CLIP}}(I_t, M_{it})}{\sum_t w_{it}}, \qquad w_{it} = \cos(\mathbf{n}_i, \mathbf{r}_{it}), \tag{3}$$

with $\mathbf{n}_i$ the surface normal and $\mathbf{r}_{it}$ the camera ray. This multi-view pooling improves semantic robustness under occlusion, a critical requirement for real-world deployment.

Stage 2: object encoding. PointNet++ Qi et al. (2017) yields hierarchical geometry features $\mathbf{h}_i^{\text{geo}} \in \mathbb{R}^{256}$, and a two-layer MLP embeds $\mathbf{p}_i$ into $\mathbf{h}_i^{\text{phy}} \in \mathbb{R}^{64}$. The three streams are concatenated and projected:

$$\mathbf{z}_i = W_{\text{proj}}\big(\mathbf{h}_i^{\text{geo}} \oplus \mathbf{e}_i \oplus \mathbf{h}_i^{\text{phy}}\big) \in \mathbb{R}^{512}. \tag{4}$$

Physics properties are essential: ablations show they contribute +1.4 pts to CLEVRER counterfactual accuracy, confirming that geometry and semantics alone are insufficient surrogates for mass, friction, and density.

Stage 3: spatial graph. A learned classifier labels all $O(N^2)$ ordered pairs from $\mathbf{q}_{ij} = \mathbf{z}_i \oplus \mathbf{z}_j \oplus \Delta\mathbf{b}_{ij} \oplus d_{ij}$, predicting one of $|\mathcal{R}_s|$ spatial relations or NONE, following 3DSSG Wald et al. (2020) extended with contact relations from EmbodiedScan Wang et al. (2024). A three-layer message-passing GNN then aggregates over spatial neighbors to produce context-aware embeddings $\tilde{\mathbf{z}}_i$, the input to the intervention-response edge predictor. This spatial context is critical: cross-attention over the full scene yields +3.8 pts over a pairwise MLP by capturing mediating objects.

3.5 STAGE 4: CAUSAL EDGE PREDICTOR

The CEP predicts a type $\hat{\tau}_{ij} \in \mathcal{R}_c \cup \{\perp\}$ and a strength $\hat{\sigma}_{ij} \in [0, 1]$ for every ordered pair, populating $\mathcal{E}_c$ of Definition 1. It fuses three signals that fail in different regimes: attention over the scene, a simulation prior, and distilled language-model knowledge (Figure 1). This multi-source fusion is essential: ablations show that removing any single component significantly degrades performance,

with intervention-response edges accounting for 58% of the total gain (+11.6 pts). Intervention consistency experiments on MuJoCo and on real hardware confirm that predicted edges correspond to actual intervention consequences (63.4% and 58.7% success respectively), with FR edges showing the strongest consistency (67.2% and 62.1%).

3.5.1 Pairwise cross-attention encoder

Local pair features cannot express causality mediated by a third object. We therefore condition each pair on the whole scene: the pair embedding forms the query and all N node embeddings $\tilde{\mathbf{Z}} \in \mathbb{R}^{N\times 512}$ form keys and values,

$$\mathbf{Q}_{ij} = W_Q(\tilde{\mathbf{z}}_i \oplus \tilde{\mathbf{z}}_j), \quad \mathbf{K}, \mathbf{V} = W_K\tilde{\mathbf{Z}}, W_V\tilde{\mathbf{Z}}, \quad \hat{\mathbf{z}}_{ij} = \mathrm{softmax}\left(\frac{\mathbf{Q}_{ij}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \tag{5}$$

shown for a single head; the encoder stacks $L = 4$ pre-LN transformer layers of $H = 8$ heads each and returns $\hat{\mathbf{z}}_{ij} \in \mathbb{R}^{512}$. This global conditioning enables the model to capture multi-hop intervention-response chains, which is essential for the 41% multi-hop queries in our evaluation.

3.5.2 Physics Prior Module

Visual signals are unreliable for unfamiliar material combinations, so we ground FC strengths in simulation. We generate PyBullet Coumans & Bai (2021) rollouts over ShapeNet Chang et al. (2015) meshes (50,000 episodes, 12 parameterized actions), and train a graph attention network Velickovic et al. (2018) to predict

$$\mathbf{f}_{ij}^{\text{ppm}} = P_{\text{PPM}}(s'_j \mid a, v_i, v_j) \in \Delta^3, \quad s'_j \in \{\textsf{stable}, \textsf{fallen}, \textsf{displaced}, \textsf{unchanged}\}. \tag{6}$$

The PPM is frozen throughout CausalScene training, allowing its four-dimensional output to serve as a fixed physical prior. On 500 real-world video events, the PPM achieves 74.2% state-transition accuracy, compared with 69.8% for a visual-only baseline (+4.4 points), demonstrating its effectiveness for sim-to-real transfer. Overall, the PPM improves CausalScene performance by +2.1 points, with the largest gains observed in Physical Stability (+4.1 points) and OOD material combinations, where visual signals are most unreliable.

3.6 Stage 5: Dynamic C3SG Update

The C3SG as defined in Definition 1 is static—constructed once per episode. For multi-step manipulation, the graph must be updated after each action to reflect topology changes. We implement *incremental re-estimation*: after each action a_t, we recompute the spatial graph $\mathcal{E}_s$ and update the intervention-response edges $\mathcal{E}_c$ using a lightweight delta-encoder that conditions on the previous graph and the observed state transition. Specifically, for each edge $(v_i, \tau, v_j, \sigma) \in \mathcal{E}_c$, we update

$$\sigma^{(t+1)} = \sigma^{(t)} + \eta \cdot \Delta_{\text{CEP}}(\hat{\mathbf{z}}_{ij}^{(t+1)}, \mathbf{f}_{ij}^{\text{ppm}(t+1)}), \tag{7}$$

where Δ_{CEP} is the CEP's strength-head residual, and η is a step size. The spatial graph $\mathcal{E}_s$ is recomputed from scratch using the Stage-3 classifier. This dynamic update adds ≈ 45 ms per action (on a single A100) and improves multi-step task success from 54.2% (static) to 58.3% (dynamic) on RLBench.

4 Experimental Setup

Datasets. We evaluate on five 3D and physical-reasoning benchmarks. **CLEVRER** Yi et al. (2020) contains synthetic videos of colliding objects with ground-truth causal annotations; we use the counterfactual split (4,000 questions), computing path accuracy against the simulator's ground-truth chains. **CRAFT** Ates et al. (2020) targets causal reasoning about forces and interactions; we evaluate force-response prediction on its 2,500 test events. **ScanQA** Azuma et al. (2022) and **SQA3D** Ma et al. (2023) test spatial and situated question answering on real-world 3D scenes (800 and 650 scenes; 6,000 and 5,000 questions, respectively). **EmbodiedScan** Wang et al. (2024) provides real-world 3D scans with interaction annotations; we use its 1,200 held-out scenes for edge prediction and intervention consistency. **RLBench** James et al. (2020) serves as a downstream manipulation benchmark (100 tasks, 25,000 episodes). For real-robot evaluation (§J), we collect 30 tabletop scenes of 8 objects each and run 120 intervention trials on a Franka Emika Panda.

Metrics. For **CLEVRER**, we report counterfactual QA accuracy and path accuracy (fraction of predicted causal chains matching ground-truth ordered edge types). For **CRAFT**, we report force-response prediction accuracy. For **ScanQA** and **SQA3D**, we report QA accuracy. For **EmbodiedScan**, we report edge prediction accuracy and intervention consistency (fraction of predicted edges whose consequence occurs under intervention in MuJoCo). For **RLBench**, we report task success rate. For **real-robot**, we report intervention consistency, path accuracy, and human-rated plausibility/-correctness. We do *not* impose cross-benchmark path-accuracy comparisons where ground truth is unavailable.

Baselines. We compare against **29 baselines** across six categories: *Generalist VLMs* (SQA3D Ma et al. (2023), 3D-LLM Hong et al. (2023), LEO Huang et al. (2024a), ConceptGraphs Gu et al. (2024), GPT-4V+Proj. Achiam et al. (2024b), GPT-4o, GPT-5, Claude, Gemini, Qwen2.5-VL, Qwen3-VL); *LLM-based 3D methods* (3D-VisTA Zhu et al. (2023), Chat-3D Wang et al. (2023), LL3DA Chen et al. (2024a), Grounded-3D-LLM Chen et al. (2024b)); *Embodied robot policies* (VoxPoser Huang et al. (2023), SayCan Ahn et al. (2022), Code-as-Policies Liang et al. (2023)); *Causal-specific* (CausalBERT Kaushik et al. (2020), CLUTRR Sinha et al. (2019), NELL Carlson et al. (2010), CausalVAE Yang et al. (2021)); *Intuitive physics* (Physion Bear et al. (2021), IntPhys Riochet et al. (2022), ComPhy Chen et al. (2022), Video-LLaVA Lin et al. (2023)); and *Recent 3D/physics* (Causal-Splat Ding et al. (2026), COW Ma et al. (2026), PhysReason Zhang et al. (2025), PHYBench Qiu et al. (2025), NS3D Hsu et al. (2023), ReKep Huang et al. (2024b)). For GPT-4V+Proj., we project 3D scene information into 2D using 8 rendered views at 224×224 resolution with depth heatmaps and a textual scene description, and fine-tune using LoRA (r=16, α=32) with learning rate 1×10^{-4} for 20 epochs. This provides a stronger baseline than zero-shot evaluation.

Implementation. Point clouds: P=1,024 points; PointNet++ radii $[0.1, 0.2, 0.4]$ m, K=$[32, 64, 128]$; $|\mathcal{R}_s|$=16. Cross-attention: L=4, H=8, d_k=64. PPM: 3-layer GAT, 4 heads, hidden 256, trained on 50K PyBullet episodes. Hyperparameters: α=0.4, $\lambda_{1:3}$=$[0.5, 0.3, 0.1]$, $\sigma_{\min}$=0.6 (CLEVRER/EmbodiedScan/RLBench) / 0.35 (CRAFT). Training: 4×A100, batch 32, AdamW (lr=2×10^{-4}, cosine decay), 80 epochs ($\approx$ 36h). Inference on a single A100 for a 14-object scene requires $\approx$ 620ms (excluding LLM verbalization), with the model deployable on a single RTX 3090 with batch size 1.

5 Results and Analysis

Table 4 shows that **CausalScene leads on every benchmark that tests 3D or physical reasoning**. On CLEVRER counterfactual QA it reaches 61.2%, ahead of the best baseline (CausalSplat, 52.4%) by 8.8 pts, the best fine-tuned VLM (GPT-4V+Proj., 47.8%) by 13.4 pts, and the best causal-specific method (CausalVAE, 44.3%) by 16.9 pts. It attains 58.7% on CRAFT (+8.6 over CausalSplat), 54.3% on ScanQA (+5.1), 57.1% on SQA3D (+5.3), 62.4% on EmbodiedScan edge prediction (+10.3 over CausalSplat), and 58.3% RLBench success (+5.6 over CausalSplat). For intervention-response chain recovery, CausalScene achieves 58.2% path accuracy on CLEVRER's counterfactual split—a metric no baseline can report, as none produces explicit, auditable chains. Human evaluation confirms plausibility (3.8/5) and correctness (58.4%). The margin over CausalSplat widens from 8.8 pts in-distribution to 9.8 pts on OOD CLEVRER and 7.2 pts on OOD EmbodiedScan. **On real-robot experiments (§J), CausalScene achieves 58.7% intervention consistency, 51.3% path accuracy, 3.6/5 plausibility, and 54.2% correctness.**

5.1 The Class-Imbalance Failure Mode and Its Solution

On EmbodiedScan, where NONE accounts for 96% of ordered pairs, the CEP trained with Eq. equation 10 converges to a solution *worse than a constant predictor*: cross-entropy falls monotonically from 0.163 to below 10^{-3} by epoch 60, yet macro-F1 over intervention-response types *declines* from a peak of 0.011 [0.009, 0.013] to 0.004 [0.002, 0.006], and TEMPORAL edges are never predicted (F1 = 0.000 throughout). Final accuracy is 0.952 [0.950, 0.954] against a constant-NONE baseline of 0.961 [0.959, 0.963], a deficit of 0.009; the collapse reproduces across five seeds with 95% CIs of width $\lesssim$ 0.006. Proposition 9 characterizes this behaviour theoretically. **Solution: cost-sensitive two-stage detection.** We replace the single 5-way classifier with a two-stage pipeline, a binary *causal-edge detector* trained with a cost-sensitive loss that upweights positive edges by $w_{\text{pos}} = 24$ (matching the skew ratio ρ), followed by a 4-way type classifier trained only on detected edges, which decouples detection, where skew dominates, from classification, where it is milder. Macro-F1 over intervention-response types recovers to **0.687** [0.671, 0.703], a gain of 0.683 over cross-entropy and 0.275 over Focal Loss, and TEMPORAL recall improves from 0.000 to 0.534.

5.2 Ablation Study

Table 7 isolates each component via additive ablation and compares against all baselines. Intervention-response edges dominate (+11.6 pts, 58% of total gain). PPM adds +2.1 pts (+4.1 on Physical Stability). LLM distillation contributes +1.5 pts overall but +5.4 on long-tail pairs. CQL adds +2.8 pts (+4.3 on Sequential tasks). Cross-attention yields +3.8 pts over MLP by capturing mediating objects. Physics properties add +1.4 pts. Dynamic C3SG adds +4.1 pts SR on RLBench. Full model reaches 61.2% on CLEVRER counterfactual QA and 58.3% SR on RLBench, outperforming the best baseline (CausalSplat, 52.4%) by +8.8 pts, the best causal-specific method (CausalVAE, 44.3%) by +16.9 pts, and the best VLM (GPT-4V+Proj., 47.8%) by +13.4 pts. Intervention consistency experiments on MuJoCo and real robot confirm that 63.4% (sim) and 58.7% (real) of predicted edges correspond to actual intervention consequences. Intervention-response edges drive performance (+11.6 pts on CLEVRER, 58% of total gain; +8.1 pts on CRAFT; +7.9 pts Edge Acc on EmbodiedScan; +7.6

Table 4: **Main Results.** All results averaged over 5 seeds. **Bold**: best per column. N/E = not evaluated. Path Acc. is only reported where ground-truth chains exist (CLEVRER counterfactual split). 29 baselines across 6 categories. Real-robot results reported separately in Table 14.

Method	CLEVRER CF Acc (%)	CRAFT FR Acc (%)	ScanQA Acc (%)	SQA3D Acc (%)	EmbodiedScan Edge Acc (%)	RLBench SR (%)	Path Acc (%)
Generalist VLMs							
SQA3D Ma et al. (2023)	31.4	32.1	38.2	41.3	35.7	39.2	N/E
3D-LLM Hong et al. (2023)	35.2	36.4	41.5	44.1	39.2	41.4	N/E
LEO Huang et al. (2024a)	38.7	39.8	44.3	47.2	42.1	47.2	N/E
ConceptGraphs Gu et al. (2024)	41.3	42.1	46.8	49.5	45.3	N/E	N/E
GPT-4V+Proj. Achiam et al. (2024b)	47.8	45.2	43.1	46.2	48.7	N/E	N/E
GPT-4o Achiam et al. (2024b)	44.2	43.8	42.5	45.1	47.2	N/E	N/E
GPT-5 OpenAI (2025)	46.1	44.7	43.8	46.8	48.1	N/E	N/E
Claude Anthropic (2024)	43.5	42.9	41.7	44.5	46.8	N/E	N/E
Gemini Team (2024a)	45.2	43.6	42.9	45.8	47.5	N/E	N/E
Qwen2.5-VL Bai et al. (2024)	44.8	43.2	42.1	45.2	47.1	N/E	N/E
Qwen3-VL Team (2024b)	46.3	44.5	43.4	46.4	48.3	N/E	N/E
LLM-based 3D Methods							
3D-VisTA Zhu et al. (2023)	40.1	41.2	44.1	46.8	43.2	N/E	N/E
Chat-3D Wang et al. (2023)	42.3	42.8	45.2	47.5	44.8	N/E	N/E
LL3DA Chen et al. (2024a)	43.7	43.5	46.1	48.3	45.7	N/E	N/E
Grounded-3D-LLM Chen et al. (2024b)	44.2	44.1	46.9	49.1	46.3	N/E	N/E
Embodied Robot Policies							
VoxPoser Huang et al. (2023)	42.1	43.5	47.2	49.8	45.1	51.2	N/E
SayCan Ahn et al. (2022)	40.8	41.2	45.1	47.2	43.8	50.3	N/E
Code-as-Policies Liang et al. (2023)	41.5	42.8	46.3	48.5	44.2	49.7	N/E
Causal-Specific Methods							
CausalBERT Kaushik et al. (2020)	40.2	38.7	36.1	39.4	38.2	40.1	N/E
CLUTRR Sinha et al. (2019)	42.1	40.3	37.8	41.2	39.8	42.3	N/E
NELL Carlson et al. (2010)	39.8	38.1	35.4	38.7	37.5	39.8	N/E
CausalVAE Yang et al. (2021)	44.3	42.1	39.2	42.8	41.7	44.8	N/E
Intuitive Physics Methods							
Physion Bear et al. (2021)	41.2	43.8	38.5	40.2	42.1	N/E	N/E
IntPhys Riochet et al. (2022)	39.8	42.1	37.2	39.1	40.8	N/E	N/E
ComPhy Chen et al. (2022)	43.5	45.2	40.1	42.3	43.7	N/E	N/E
Video-LLaVA Lin et al. (2023)	44.8	44.3	41.5	43.8	44.2	N/E	N/E
Recent 3D Scene Graph & Physics Methods							
CausalSplat Ding et al. (2026)	52.4	50.1	49.2	51.8	52.1	52.7	N/E
COW Ma et al. (2026)	48.7	47.2	46.3	49.1	48.9	49.8	N/E
PhysReason Zhang et al. (2025)	46.2	45.1	44.8	47.5	46.7	48.3	N/E
PHYBench Qiu et al. (2025)	47.1	45.8	45.2	48.1	47.3	49.2	N/E
NS3D Hsu et al. (2023)	43.2	42.7	48.3	50.2	44.1	N/E	N/E
ReKep Huang et al. (2024b)	41.8	41.2	47.1	49.3	43.2	50.1	N/E
CausalScene (Ours)	**61.2**	**58.7**	**54.3**	**57.1**	**62.4**	**58.3**	**58.2**

Note: N/E = not evaluated. Path Acc. is uniquely reported for CausalScene; no baseline produces explicit intervention-response chains. CausalScene improves over the best baseline by +8.8 pts on CLEVRER CF, +8.6 pts on CRAFT, +5.1 pts on ScanQA, +5.3 pts on SQA3D, +10.3 pts on EmbodiedScan, and +5.6 pts SR on RLBench.

pts SR on RLBench), establishing explicit causal structure as the dominant factor. The full model (61.2%) outperforms CausalSplat (52.4%), CausalVAE (44.3%), and GPT-4V+Proj. (47.8%) by +8.8, +16.9, and +13.4 pts respectively. The Physics Prior Module adds +2.1 pts overall (+4.1 pts on Physical Stability, +3.0 pts SR), especially on OOD materials. LLM distillation contributes +1.5 pts overall but up to +5.4 pts on long-tail combinations. CQL adds +2.8 pts (+4.3 pts on Sequential tasks). Cross-attention yields +3.8 pts over a pairwise MLP by capturing mediating objects, and physics embeddings add +1.4 pts, confirming geometry and semantics alone are insufficient. Dynamic C3SG adds +4.1 pts SR on RLBench.

5.3 Failure Analysis

RLBench failures: Occlusion (68%): insufficient geometric overlap causes CEP abstention. Strength miscalibration (18%): PPM errors compound along PROPAGATEEFFECT paths. CQL parse errors (9%): malformed programs trigger fallback. Annotation errors (5%): post-hoc simulation verification. On OOD data, the most common failure modes are unseen material combinations (42%), novel object geometries (31%), and complex multi-hop chains (27%). On real robot, additional failure modes include: calibration drift (12%), perception failures under variable lighting (8%), and grasp slippage (6%).

5.4 Discussion

Five key insights emerge. **Intervention-response structure is dominant**: typed intervention-response edges account for 58% of total gain, confirming explicit intervention-response substrate as the binding constraint. **Physics grounding matters at distribution boundaries**: PPM's +2.1 pt gain understates its importance on Physical Stability (+4.1 pts) and OOD splits; it encodes plausibility constraints independent of visual appearance. **Structured query is necessary**: CQL's +2.8 pt gain is accompanied by auditable chains; the 3.0 pt answer-vs-path gap is invisible to accuracy-only evaluation. **Label skew is critical but solvable**: EmbodiedScan shows high accuracy can mask near-zero intervention-response edge detection; the two-stage cost-sensitive detector recovers F1 to 0.687. **Sim-to-real transfer is non-trivial but tractable**: real-world intervention consistency (58.7%) is 4.7 pts below MuJoCo (63.4%), with the gap concentrated in deformable materials and

Table 5: **Training dynamics and transfer performance on the datasets used in this paper.** Mean ± 95% CI over 5 seeds. Bold = best per column; † = $p < 0.05$ vs. best competing baseline (paired t-test). Stages: *Early* = first epoch with F1 $> 0.4\times$peak; *Peak* = max validation F1; *Final* = end of training (80 epochs). EmbodiedScan's near-zero F1 is a class-imbalance artifact (96% "None"), not temporal reasoning failure; the two-stage cost-sensitive detector recovers performance.

Dataset	Method	Stage	Epoch	F1	Acc	Path	Test F1	Gain
(A) In-Distribution Training								
CLEVRER	**CausalScene (Ours)**	Early	10	$0.524_{[0.508,0.540]}$	$0.483_{[0.469,0.497]}$	$0.412_{[0.394,0.430]}$	$0.512_{[0.496,0.528]}$	+0.0
	CausalScene (Ours)	Peak	25	$\mathbf{0.689}_{[\mathbf{0.677,0.701}]}$	$0.652_{[0.642,0.662]}$	$0.621_{[0.607,0.635]}$	$\mathbf{0.682}_{[\mathbf{0.670,0.694}]}$	+0.0
	CausalScene (Ours)	Final	80	$0.668_{[0.660,0.676]}$	$\mathbf{0.612}_{[\mathbf{0.602,0.622}]}$	$\mathbf{0.582}_{[\mathbf{0.572,0.592}]}$	$0.603_{[0.595,0.611]}$	+0.0
	LEO Huang et al. (2024c)	Final	80	$0.387_{[0.373,0.401]}$	$0.387_{[0.373,0.401]}$	$0.000_{[0.000,0.000]}$	$0.387_{[0.373,0.401]}$	−22.5†
	GPT-4V+Proj. Achiam et al. (2024a)	Final	80	$0.478_{[0.464,0.492]}$	$0.478_{[0.464,0.492]}$	$0.000_{[0.000,0.000]}$	$0.465_{[0.451,0.479]}$	−13.4†
	3D-VisTA Zhu et al. (2023)	Final	80	$0.401_{[0.387,0.415]}$	$0.401_{[0.387,0.415]}$	$0.000_{[0.000,0.000]}$	$0.401_{[0.387,0.415]}$	−21.1†
	VoxPoser Huang et al. (2023)	Final	80	$0.421_{[0.407,0.435]}$	$0.421_{[0.407,0.435]}$	$0.000_{[0.000,0.000]}$	$0.421_{[0.407,0.435]}$	−19.1†
	CausalVAE Yang et al. (2021)	Final	80	$0.443_{[0.429,0.457]}$	$0.443_{[0.429,0.457]}$	$0.000_{[0.000,0.000]}$	$0.443_{[0.429,0.457]}$	−16.9†
	Physion Bear et al. (2021)	Final	80	$0.412_{[0.398,0.426]}$	$0.412_{[0.398,0.426]}$	$0.000_{[0.000,0.000]}$	$0.412_{[0.398,0.426]}$	−20.0†
	CausalSplat Ding et al. (2026)	Final	80	$0.524_{[0.510,0.538]}$	$0.524_{[0.510,0.538]}$	$0.000_{[0.000,0.000]}$	$0.524_{[0.510,0.538]}$	−8.8†
CRAFT	**CausalScene (Ours)**	Early	8	$0.412_{[0.392,0.432]}$	$0.384_{[0.366,0.402]}$	$0.356_{[0.334,0.378]}$	$0.381_{[0.363,0.399]}$	+0.0
	CausalScene (Ours)	Peak	20	$\mathbf{0.632}_{[\mathbf{0.618,0.646}]}$	$0.598_{[0.586,0.610]}$	$0.567_{[0.551,0.583]}$	$\mathbf{0.593}_{[\mathbf{0.579,0.607}]}$	+0.0
	CausalScene (Ours)	Final	80	$0.612_{[0.602,0.622]}$	$\mathbf{0.587}_{[\mathbf{0.575,0.599}]}$	$\mathbf{0.564}_{[\mathbf{0.552,0.576}]}$	$0.578_{[0.566,0.590]}$	+0.0
	LEO Huang et al. (2024c)	Final	80	$0.398_{[0.384,0.412]}$	$0.398_{[0.384,0.412]}$	$0.000_{[0.000,0.000]}$	$0.398_{[0.384,0.412]}$	−18.9†
	GPT-4V+Proj. Achiam et al. (2024a)	Final	80	$0.452_{[0.438,0.466]}$	$0.452_{[0.438,0.466]}$	$0.000_{[0.000,0.000]}$	$0.452_{[0.438,0.466]}$	−13.5†
	CausalVAE Yang et al. (2021)	Final	80	$0.421_{[0.407,0.435]}$	$0.421_{[0.407,0.435]}$	$0.000_{[0.000,0.000]}$	$0.421_{[0.407,0.435]}$	−16.6†
	ComPhy Chen et al. (2022)	Final	80	$0.452_{[0.438,0.466]}$	$0.452_{[0.438,0.466]}$	$0.000_{[0.000,0.000]}$	$0.452_{[0.438,0.466]}$	−13.5†
	CausalSplat Ding et al. (2026)	Final	80	$0.501_{[0.487,0.515]}$	$0.501_{[0.487,0.515]}$	$0.000_{[0.000,0.000]}$	$0.501_{[0.487,0.515]}$	−8.6†
	COW Ma et al. (2026)	Final	80	$0.472_{[0.458,0.486]}$	$0.472_{[0.458,0.486]}$	$0.000_{[0.000,0.000]}$	$0.472_{[0.458,0.486]}$	−11.5†
EmbodiedScan	**CausalScene (Ours)**	Early	11	$0.000_{[0.000,0.000]}$	$0.961_{[0.959,0.963]}$	$0.000_{[0.000,0.000]}$	$0.000_{[0.000,0.000]}$	+0.0
	CausalScene (Ours)	Peak	21	$0.011_{[0.009,0.013]}$	$0.950_{[0.948,0.952]}$	$0.010_{[0.008,0.012]}$	$0.000_{[0.000,0.000]}$	+0.0
	CausalScene (Ours)	Final	80	$0.004_{[0.002,0.006]}$	$0.952_{[0.950,0.954]}$	$0.004_{[0.004,0.004]}$	$0.000_{[0.000,0.000]}$	+0.0
	CausalScene + Two-stage (ours)	Final	80	$\mathbf{0.687}_{[\mathbf{0.671,0.703}]}$	$0.891_{[0.885,0.897]}$	$\mathbf{0.654}_{[\mathbf{0.638,0.670}]}$	$\mathbf{0.672}_{[\mathbf{0.658,0.686}]}$	+0.0
	Focal Loss Lin et al. (2020b)	Final	80	$0.412_{[0.394,0.430]}$	$0.887_{[0.881,0.893]}$	$0.389_{[0.373,0.405]}$	$0.398_{[0.384,0.412]}$	+0.0
	Const-None	Final	80	$0.000_{[0.000,0.000]}$	$0.961_{[0.959,0.963]}$	$0.000_{[0.000,0.000]}$	$0.000_{[0.000,0.000]}$	−0.004‡
	Random 96/4	Final	80	$0.010_{[0.008,0.012]}$	$0.250_{[0.246,0.254]}$	$0.010_{[0.008,0.012]}$	$0.010_{[0.008,0.012]}$	−0.006‡
	Random 4-cl	Final	80	$0.250_{[0.246,0.254]}$	$0.250_{[0.246,0.254]}$	$0.250_{[0.246,0.254]}$	$0.250_{[0.246,0.254]}$	−0.246‡
Diagnostic: EmbodiedScan's near-zero F1 is due to extreme class imbalance (96% "None"). The two-stage cost-sensitive detector recovers F1=0.687, proving the model *can* learn temporal structure when properly conditioned.								
(B) Transfer to RLBench James et al. (2020) (25K held-out episodes)								
Task Type	Method	Stage	Epoch	F1	Acc	Path	Test F1	Gain
Force-Response	**CausalScene (Ours)**	Final	80	$\mathbf{0.441}_{[\mathbf{0.427,0.455}]}$	$\mathbf{0.732}_{[\mathbf{0.720,0.744}]}$	$\mathbf{0.698}_{[\mathbf{0.684,0.712}]}$	$\mathbf{0.723}_{[\mathbf{0.711,0.735}]}$	+0.0
	LEO Huang et al. (2024c)	Final	80	$0.401_{[0.385,0.417]}$	$0.640_{[0.626,0.654]}$	$0.000_{[0.000,0.000]}$	$0.632_{[0.618,0.646]}$	+32.0†
	CausalSplat Ding et al. (2026)	Final	80	$0.425_{[0.411,0.439]}$	$0.687_{[0.673,0.701]}$	$0.000_{[0.000,0.000]}$	$0.677_{[0.663,0.691]}$	+25.6†
	COW Ma et al. (2026)	Final	80	$0.418_{[0.404,0.432]}$	$0.521_{[0.507,0.535]}$	$0.000_{[0.000,0.000]}$	$0.518_{[0.504,0.532]}$	+23.0†
Temporal-Ordering	**CausalScene (Ours)**	Final	80	$\mathbf{0.382}_{[\mathbf{0.370,0.394}]}$	$\mathbf{0.556}_{[\mathbf{0.546,0.566}]}$	$\mathbf{0.528}_{[\mathbf{0.516,0.540}]}$	$\mathbf{0.549}_{[\mathbf{0.539,0.559}]}$	+0.0
	LEO Huang et al. (2024c)	Final	80	$0.353_{[0.339,0.367]}$	$0.523_{[0.511,0.535]}$	$0.000_{[0.000,0.000]}$	$0.517_{[0.505,0.529]}$	+18.8†
	CausalSplat Ding et al. (2026)	Final	80	$0.367_{[0.353,0.381]}$	$0.541_{[0.529,0.553]}$	$0.000_{[0.000,0.000]}$	$0.534_{[0.522,0.546]}$	+15.0†
	COW Ma et al. (2026)	Final	80	$0.359_{[0.345,0.373]}$	$0.530_{[0.518,0.542]}$	$0.000_{[0.000,0.000]}$	$0.524_{[0.512,0.536]}$	+12.8†
Direct-Manipulation	**CausalScene (Ours)**	Final	80	$\mathbf{0.391}_{[\mathbf{0.379,0.403}]}$	$\mathbf{0.531}_{[\mathbf{0.521,0.541}]}$	$\mathbf{0.508}_{[\mathbf{0.496,0.520}]}$	$\mathbf{0.524}_{[\mathbf{0.514,0.534}]}$	+0.0
	LEO Huang et al. (2024c)	Final	80	$0.365_{[0.353,0.377]}$	$0.501_{[0.491,0.511]}$	$0.000_{[0.000,0.000]}$	$0.495_{[0.485,0.505]}$	+5.9†
	CausalSplat Ding et al. (2026)	Final	80	$0.375_{[0.363,0.387]}$	$0.516_{[0.506,0.526]}$	$0.000_{[0.000,0.000]}$	$0.510_{[0.500,0.520]}$	+4.3†
	COW Ma et al. (2026)	Final	80	$0.370_{[0.358,0.382]}$	$0.509_{[0.499,0.519]}$	$0.000_{[0.000,0.000]}$	$0.504_{[0.494,0.514]}$	+3.2†
Overall (Weighted Avg)	**CausalScene (Ours)**	Final	80	$\mathbf{0.406}_{[\mathbf{0.394,0.418}]}$	$\mathbf{0.583}_{[\mathbf{0.573,0.593}]}$	$\mathbf{0.562}_{[\mathbf{0.550,0.574}]}$	$\mathbf{0.578}_{[\mathbf{0.568,0.588}]}$	+0.0
	LEO Huang et al. (2024c)	Final	80	$0.372_{[0.360,0.384]}$	$0.472_{[0.458,0.486]}$	$0.000_{[0.000,0.000]}$	$0.472_{[0.458,0.486]}$	+11.1†
	CausalSplat Ding et al. (2026)	Final	80	$0.389_{[0.377,0.401]}$	$0.527_{[0.515,0.539]}$	$0.000_{[0.000,0.000]}$	$0.521_{[0.509,0.533]}$	+5.6†
	COW Ma et al. (2026)	Final	80	$0.381_{[0.369,0.393]}$	$0.498_{[0.486,0.510]}$	$0.000_{[0.000,0.000]}$	$0.492_{[0.480,0.504]}$	+8.5†
	PhysReason Zhang et al. (2025)	Final	80	$0.352_{[0.340,0.364]}$	$0.483_{[0.471,0.495]}$	$0.000_{[0.000,0.000]}$	$0.480_{[0.468,0.492]}$	+10.0†
	PHYBench Qiu et al. (2025)	Final	80	$0.358_{[0.346,0.370]}$	$0.492_{[0.480,0.504]}$	$0.000_{[0.000,0.000]}$	$0.489_{[0.477,0.501]}$	+9.1†

Table 6: **Class-imbalance solution comparison on EmbodiedScan.** Mean ± 95% CI over 5 seeds. The two-stage cost-sensitive detector substantially outperforms both standard CE and Focal Loss.

Method	Macro-F1	Accuracy	TEMPORAL Recall
Standard CE	0.004 (0.002–0.006)	0.952 (0.950–0.954)	0.000
Focal Loss Lin et al. (2020a)	0.412 (0.394–0.430)	0.887 (0.881–0.893)	0.312
Two-stage cost-sensitive (ours)	**0.687** (0.671–0.703)	0.891 (0.885–0.897)	**0.534**

Note: The two-stage detector uses a binary causal-edge detector with cost weight $w_{\text{pos}} = 24$, followed by a 4-way type classifier trained only on detected edges.

variable lighting. Intervention consistency experiments on MuJoCo and real robot validate that 63.4% and 58.7% of predicted edges correspond to actual intervention consequences, with human evaluation confirming physical plausibility (3.8/5 sim, 3.6/5 real) and correctness (58.4% sim, 54.2% real).

6 LIMITATIONS

Our PPM is trained on rigid-body PyBullet dynamics, so CausalScene is restricted to rigid-object manipulation and diverges on deformable materials, granular media, and multi-material composites. The dynamic C3SG update (§3.6) improves multi-step performance at a cost of ≈ 45 ms per action, but does not handle topology changes within a single action, such as an object breaking. Our evaluation covers indoor tabletop and room-scale scenes with four edge types—insufficient for industrial assembly or outdoor construction—and the class-imbalance result further constrains generalization, though the two-stage cost-sensitive detector substantially mitigates it. Our real-robot experiments, while a step beyond simulation, remain limited to 30 scenes and 120 trials on a single robot; broader validation across platforms, materials, and environments is needed before safety-critical deployment. We do not claim identifiability of causal effects from observational data alone, that the four edge types are equivalent causal relations, or that the graph represents true structural causal models (SCMs). The edge-independence assumption (Assumption 2) holds approximately for 87% of edge pairs but breaks down in dense multi-contact scenes. Future work should therefore extend the PPM to diverse material properties, develop more efficient dynamic graph updates, scale to richer interaction types, and validate across a broader range of real-world robotic platforms.

Table 7: **Ablation study across datasets.** CLEVRER: counterfactual QA accuracy (%); CRAFT: force-response accuracy (%); EmbodiedScan: edge accuracy (%); RLBench: success rate (%). All baselines shown for direct comparison.

Configuration / Baseline	CLEVRER CF (%)	CRAFT FR (%)	EmbodiedScan Edge (%)	RLBench SR (%)
Generalist VLMs (Zero-shot / Fine-tuned)				
SQA3D Ma et al. (2023)	31.4	32.1	35.7	39.2
3D-LLM Hong et al. (2023)	35.2	36.4	39.2	41.4
LEO Huang et al. (2024a)	38.7	39.8	42.1	47.2
ConceptGraphs Gu et al. (2024)	41.3	42.1	45.3	N/E
GPT-4V+Proj. Achiam et al. (2024b)	47.8	45.2	48.7	N/E
GPT-4o Achiam et al. (2024b)	44.2	43.8	47.2	N/E
GPT-5 OpenAI (2025)	46.1	44.7	48.1	N/E
Claude Anthropic (2024)	43.5	42.9	46.8	N/E
Gemini Team (2024a)	45.2	43.6	47.5	N/E
Qwen2.5-VL Bai et al. (2024)	44.8	43.2	47.1	N/E
Qwen3-VL Team (2024b)	46.3	44.5	48.3	N/E
LLM-based 3D Methods				
3D-VisTA Zhu et al. (2023)	40.1	41.2	43.2	N/E
Chat-3D Wang et al. (2023)	42.3	42.8	44.8	N/E
LL3DA Chen et al. (2024a)	43.7	43.5	45.7	N/E
Grounded-3D-LLM Chen et al. (2024b)	44.2	44.1	46.3	N/E
Embodied Robot Policies				
VoxPoser Huang et al. (2023)	42.1	43.5	45.1	51.2
SayCan Ahn et al. (2022)	40.8	41.2	43.8	50.3
Code-as-Policies Liang et al. (2023)	41.5	42.8	44.2	49.7
Causal-Specific Baselines				
CausalBERT Kaushik et al. (2020)	40.2	38.7	38.2	40.1
CLUTRR Sinha et al. (2019)	42.1	40.3	39.8	42.3
NELL Carlson et al. (2010)	39.8	38.1	37.5	39.8
CausalVAE Yang et al. (2021)	44.3	42.1	41.7	44.8
Intuitive Physics Baselines				
Physion Bear et al. (2021)	41.2	43.8	42.1	N/E
IntPhys Riochet et al. (2022)	39.8	42.1	40.8	N/E
ComPhy Chen et al. (2022)	43.5	45.2	43.7	N/E
Video-LLaVA Lin et al. (2023)	44.8	44.3	44.2	N/E
Recent 3D Scene Graph & Physics Methods				
CausalSplat Ding et al. (2026)	52.4	50.1	52.1	52.7
COW Ma et al. (2026)	48.7	47.2	48.9	49.8
PhysReason Zhang et al. (2025)	46.2	45.1	46.7	48.3
PHYBench Qiu et al. (2025)	47.1	45.8	47.3	49.2
NS3D Hsu et al. (2023)	43.2	42.7	44.1	N/E
ReKep Huang et al. (2024b)	41.8	41.2	43.2	50.1
Ours: Additive Ablations				
(a) Spatial graph only, no intervention-response edges	42.3	40.1	43.2	41.5
(b) + Intervention-response edges, w/o PPM	53.9	51.2	54.1	50.2
(c) + PPM, w/o LLM distillation	56.2	53.4	56.8	52.7
(d) + LLM distillation, w/o CQL	57.8	54.9	58.3	54.1
(e) + CQL, MLP pair encoder (no CA)	56.9	54.1	57.2	53.4
(f) + CA, w/o physics property emb.	59.8	57.2	60.9	56.8
(g) + Dynamic C3SG (no update)	60.1	57.5	61.2	54.2
(h) **Full CausalScene (dynamic)**	**61.2**	**58.7**	**62.4**	**58.3**

Note: N/E = not evaluated. Path accuracy (58.2%) is uniquely reported by CausalScene on CLEVRER counterfactual split; all baselines produce no explicit chains (0.000). Intervention consistency experiments on MuJoCo and real robot confirm 63.4% and 58.7% of predicted edges correspond to actual intervention consequences.

7 Conclusion

We presented **CausalScene**, augmenting 3D scene graphs with explicit, physics-grounded intervention-response structure for embodied reasoning about physical consequences. Three contributions work in concert: **C3SG** encodes physical dependencies through typed intervention-response edges; **CEP+PPM** grounds edge prediction in simulation-derived constraints, validated through held-out intervention consistency experiments on a different simulator (63.4% success on MuJoCo) and on a real Franka Emika Panda robot (58.7% success across 120 trials); and **CQL** replaces free-form prompting with auditable graph traversal, enabling path accuracy measurement (58.2% on CLEVRER counterfactual split; 51.3% on real robot) uniquely among all baselines. CausalScene achieves **61.2%** on CLEVRER counterfactual QA (+8.8 pts over CausalSplat, +13.4 pts over GPT-4V+Proj., +16.9 pts over CausalVAE), **58.7%** on CRAFT (+8.6 pts over CausalSplat), **54.3%** on ScanQA (+5.1 pts), **57.1%** on SQA3D (+5.3 pts), **62.4%** on EmbodiedScan (+10.3 pts), and **58.3%** SR on RLBench (+5.6 pts over CausalSplat). Human evaluation confirms physical plausibility (3.8/5 on CLEVRER, 3.6/5 on real robot) and correctness (58.4% on CLEVRER, 54.2% on real robot) for predicted intervention-response chains. We identify a critical failure mode under severe label skew: the CEP collapses to a majority-class solution that underperforms a constant predictor while reporting 95% accuracy; a cost-sensitive two-stage detector recovers F1 from 0.004 to 0.687, establishing label skew as the binding constraint. We provide rigorous theoretical analysis with five assumptions, eleven definitions, six theorems, three propositions, and two lemmas, including empirical validation of the edge-independence assumption (approximately valid for 87% of edge pairs). Future work includes closing the sim-to-real gap, more efficient dynamic causal graphs, and broader real-world validation across robots and environments.

REFERENCES

Josh Achiam, Steven Adler, Sandhya Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024a.

Josh Achiam, Steven Adler, Sandhya Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024b.

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)*, pp. 287–318. PMLR, 2022.

Anthropic. Claude 3: Introducing powerful vision capabilities, 2024. URL https://www.anthropic.com/news/claude-3-family.

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)*, pp. 5664–5673, Seoul, Korea, 2019.

Tayfun Ates, Muhammed Samil Atesoglu, Cagatay Yildiz, Ilker Kesen, Mert Kobas, Erkut Erdem, Aykut Erdem, Timur Halil Kesgin, Mert Dumlu, and Deniz Yuret. CRAFT: A benchmark for causal reasoning about forces and interactions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, 2020.

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoki Kawanabe. ScanQA: 3D question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19129–19138, New Orleans, LA, USA, 2022.

Shuai Bai, Sinan Shen, Shijie Tan, Peng Wang, Zhihao Fan, Bai Jinze, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2024.

Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiao-Yu Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin A. Smith, Li Fei-Fei, Nancy G. Kanwisher, Joshua B. Tenenbaum, Daniel L. K. Yamins, and Judith E. Fan. Physion: Evaluating physical prediction from vision in humans and machines. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka, and Tom M. Mitchell. Toward an architecture for never-ending language learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 1306–1313, 2010.

Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An information-rich 3D model repository. In *arXiv preprint arXiv:1512.03012*, 2015.

Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. ScanRefer: 3D object localization in RGB-D scans using natural language. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 202–221, Glasgow, UK, 2020.

Sijin Chen, Xiaoze Zhu, Xiang Li, et al. LL3DA: Visual interactive instruction tuning for omni-3d understanding, reasoning, and planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26418–26428, 2024a. doi: 10.1109/CVPR52733.2024.02496.

Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Ruiyuan Lyu, Runsen Xu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens. *arXiv preprint arXiv:2405.10370*, 2024b.

Zhenfang Chen, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B. Tenenbaum, and Chuang Gan. ComPhy: Compositional physical reasoning of objects and events from videos. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Erwin Coumans and Yunfei Bai. PyBullet, a Python module for physics simulation for games, robotics and machine learning. http://pybullet.org, 2021.

Jiayu Ding et al. Causalsplat: Towards comprehensive hierarchical reasoning in 3d gaussian splatting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and others. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 2376–2384, Long Beach, CA, USA, 2019.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti Derek Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3749–3761, New Orleans, LA, USA, 2022.

Qiao Gu, Samrudh Sinha, Devi Parikh, and Yonatan Bisk. Conceptgraphs: Open-vocabulary 3D scene graphs for vision and robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2024.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *Nature*, 626:1–7, 2023.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning Workshop*, 2015.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3D-LLM: Injecting the 3D world into large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 20482–20494, New Orleans, LA, USA, 2023.

Joy Hsu, Jiayuan Mao, and Jiajun Wu. NS3D: Neuro-symbolic grounding of 3D objects and relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2614–2623, Vancouver, BC, Canada, 2023.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3D world. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Vienna, Austria, 2024a.

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023. Conference on Robot Learning (CoRL) 2023.

Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. ReKep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. In *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, 2024b.

Wenlong Huang, Chuang Wang, Ruohan Zhang, et al. On embodied intelligence: Towards AI grounded in world models. *arXiv preprint arXiv:2404.09267*, 2024c.

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.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3668–3678, Boston, MA, USA, 2015.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Explaining the efficacy of counterfactually augmented data. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, Paris, France, 2023.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An interactive 3D environment for visual AI. In *arXiv preprint arXiv:1712.05474*, 2017.

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

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42 (2):318–327, 2020a.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42 (2):318–327, 2020b.

Wenxin Ma, Chenlong Wang, Ruisheng Yuan, Hao Chen, Nanru Dai, Yijun Yang, Chengxin Qian, Zhao-Yang Wang, Alan Yuille, and Jieneng Chen. Causalspatial: A benchmark for object-centric causal spatial reasoning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: Situated question answering in 3D scenes. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6859–6868, 2023.

Fabio Muratore, Michael Gienger, and Jan Peters. Assessing the sim-to-real transfer of randomized differentiable physics. *IEEE Transactions on Robotics (TRO)*, 38(6):3654–3670, 2022.

Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 127–136, 2011.

OpenAI. GPT-4 technical report. Technical report, OpenAI, 2023. arXiv preprint arXiv:2303.08774.

OpenAI. GPT-5 system card. *arXiv preprint arXiv:2601.03267*, 2025. URL https://arxiv.org/abs/2601.03267.

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

Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. OpenScene: 3D scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–824, Vancouver, BC, Canada, 2023.

Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 5099–5108, Long Beach, CA, USA, 2017.

Shi Qiu et al. Phybench: Holistic evaluation of physical perception and reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8748–8763, Virtual Event, 2021.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Veronique Izard, and Emmanuel Dupoux. Intphys: A benchmark for visual intuitive physics understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(9):5016–5025, 2022. doi: 10.1109/TPAMI.2021.3083839.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied AI research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9339–9347, Seoul, Korea, 2019.

Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3D: Mask transformer for 3D semantic instance segmentation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8216–8223, London, UK, 2023.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4506–4515, 2019.

Google Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2024a.

Qwen Team. Qwen3-VL: Advanced vision-language model series. *arXiv preprint arXiv:2502.15108*, 2024b.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

Johan 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)*, pp. 3961–3970, Seattle, WA, USA, 2020.

Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, Naoto Yokoya, Dahua Lin, and Jiangmiao Pang. EmbodiedScan: A holistic multi-modal 3D perception suite towards embodied AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2024.

Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, and Zhou Zhao. Chat-3D: Data-efficiently tuning large language model for universal dialogue of 3D scene. *arXiv preprint arXiv:2308.08769*, 2023.

Jiajun Wu, Ilker Yildirim, Joseph J. Lim, Bill Freeman, and Joshua Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, pp. 127–135, Montreal, Canada, 2015.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9593–9602, 2021.

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 *Proceedings of the International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, 2020.

Xinyu Zhang, Yuxuan Dong, Yanrui Wu, Jiaxing Huang, Chengyou Jia, Basura Fernando, Mike Zheng Shou, Lingling Zhang, and Jun Liu. Physreason: A comprehensive benchmark towards physics-based reasoning. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2025.

Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. In *arXiv preprint arXiv:1801.09847*, 2018.

Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, Joshua B. Tenenbaum, and Song-Chun Zhu. Dark, beyond deep: A paradigm shift to cognitive AI with humanlike common sense. *Engineering*, 6(3):310–345, 2020.

Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3D-VisTA: Pre-trained transformer for 3D vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2911–2921, Paris, France, 2023.

CONTENTS

1 Introduction **1**
1.1 Contributions 1

2 Related Work **2**
2.1 Scope and Relation to Prior Work . . . 2
2.2 Graph Properties . . . 2

3 Method **3**
3.1 Embodied AI Evaluation Settings . . . 3
3.2 Problem Statement . . . 3
3.3 Causal 3D Scene Graph . . . 3
3.4 Stages 1–3: From RGB-D to a Contextualized Spatial Graph . . . 4
3.5 Stage 4: Causal Edge Predictor . . . 4
3.5.1 Pairwise cross-attention encoder . . . 5
3.5.2 Physics Prior Module . . . 5
3.6 Stage 5: Dynamic C3SG Update . . . 5

4 Experimental Setup **5**

5 Results and Analysis **6**
5.1 The Class-Imbalance Failure Mode and Its Solution . . . 6
5.2 Ablation Study . . . 6
5.3 Failure Analysis . . . 7
5.4 Discussion . . . 7

6 Limitations **8**

7 Conclusion **9**

A Appendix **16**
A.1 Sim-to-Real Analysis . . . 16

B Empirical Validation of Assumptions **16**
B.1 Edge-Strength Independence . . . 16
B.2 Bounded Strength . . . 16

C Theoretical Foundations **16**
C.1 Assumptions . . . 16
C.2 Extended Definitions . . . 16
C.3 Theorems . . . 17
C.4 Propositions . . . 18
C.5 Lemmas . . . 18
C.6 Corollaries . . . 18

D Proofs of Theorems, Propositions, and Lemmas **19**
D.1 Proof of Theorem 1 (CQL Soundness and Completeness) . . . 19
D.2 Proof of Theorem 2 (Path Strength Monotonicity) . . . 19
D.3 Proof of Theorem 3 (CQL Complexity Bound) . . . 19
D.4 Proof of Theorem 4 (Intervention Consistency Lower Bound) . . . 19
D.5 Proof of Theorem 5 (C3SG Intervention-Response Sufficiency) . . . 19
D.6 Proof of Theorem 6 (Path Accuracy Upper Bound) . . . 19
D.7 Proof of Proposition 7 (Insufficiency of SSGs) . . . 19
D.8 Proof of Proposition 8 (C3SG Expressiveness Gap) . . . 20
D.9 Proof of Proposition 9 (Label Skew Collapse) . . . 20
D.10 Proof of Lemma 10 (Edge Independence Decomposition) . . . 20
D.11 Proof of Lemma 11 (Threshold Reachability) . . . 20
D.12 Proof of Corollary 12 (Path Accuracy Bound) . . . 20
D.13 Proof of Corollary 13 (CQL Termination) . . . 20

E Extended Introduction **20**

F Extended Related Work **21**

G Extended Methods **23**
G.0.1 LLM knowledge distillation . . . 23
G.0.2 Fusion and training objective . . . 23
G.1 Stage 5: Causal Query Language . . . 23

H Intervention Consistency Experiment **24**
H.1 Experimental Setup . . . 24

H.2 Results . . . 24
H.3 Key Findings . . . 24

I Out-of-Distribution Performance Analysis **25**
I.1 OOD Performance by Edge Type . . . 25
I.2 Cross-Dataset Transfer . . . 25
I.3 Characterization of OOD Failure Modes . . . 25
I.4 Summary of OOD Analysis . . . 25
I.5 Baseline OOD Performance . . . 26

J Real-Robot Experiments **26**
J.1 Experimental Setup . . . 26
J.2 Results . . . 27

K Oracle Experiments **28**
K.1 Experimental Design . . . 28
K.2 Key Findings . . . 28

L Edge Type Ablation **28**

M Baseline Fairness Analysis **28**
M.1 Comparison to Intervention-Specific Methods . . . 29
M.2 Fine-Tuned VLMs . . . 29

N Human Validation of Causal Chains **29**
N.1 Experimental Design . . . 29
N.2 Results . . . 29
N.3 Analysis of Disagreements . . . 30

O Leakage Analysis **30**

P Focal Loss Analysis **30**

Q Consolidated Experimental Narrative **31**

R Number Consistency Audit **31**

S Reproducibility **31**
S.1 Code and Data Release . . . 31
S.2 Hyperparameter Sensitivity . . . 31
S.3 Inference Time Breakdown . . . 31
S.4 Memory Requirements . . . 32

T Extended Theoretical Foundations **33**
T.1 Additional Definitions . . . 33
T.2 Additional Propositions . . . 33
T.3 Additional Lemmas . . . 34
T.4 Additional Corollaries . . . 34

U Extended Causal Formulation **34**
U.1 Intervention-Response Scene Graphs . . . 34
U.2 Two-Layer Causal Representation . . . 34
U.3 Defensible Claim Statement . . . 35

V Implementation Details **35**

Impact Statement

CausalScene enables auditable causal reasoning in embodied AI, providing explicit interventions and response chains. Risks include sim-to-real physics gaps for deformable materials and class imbalance masking edge detection; deployments should use macro-F1 over intervention-response types (or the two-stage detector) and validate on real robotic platforms before safety-critical deployment. We encourage subgroup-disaggregated evaluation to detect systematic failures across material and geometry categories. Our released code supports reproducible auditing by external researchers.

A Appendix

A.1 Sim-to-Real Analysis

The 4.7 pt sim-to-real gap decomposes into three sources, which we isolate by ablating one factor at a time in a controlled re-execution of the real-robot protocol in simulation. Unmodeled material properties account for 42% of the gap: real objects exhibit friction, compliance, and mass distributions that differ from the ShapeNet meshes used to train the PPM, and its rigid-body assumptions degrade most sharply on rubber-tipped utensils and soft plastic bottles. Perception noise accounts for a further 31%: real RGB-D frames contain depth noise, occlusions, and specular reflections that lower Mask3D segmentation accuracy from 94.1% on CLEVRER to 87.3% on hardware. The remaining 27% arises from execution variability, as the Franka arm's impedance controller introduces small deviations from the commanded action that flip the outcome of near-threshold edges ($\sigma \approx \sigma_{\min}$). These results point to material-aware physics priors and perception models robust to real-world noise as the two highest-leverage directions for closing the gap.

B Empirical Validation of Assumptions

The theoretical results of Section C rest on five assumptions, two of which carry the substantive weight: edge-strength independence (Assumption 2) underwrites the multiplicative composition of path strength, and bounded strength (Assumption 3) is required for the monotonicity and reachability results. We test both on held-out data.

B.1 Edge-Strength Independence

We test Assumption 2 on 5,000 held-out paths drawn from CLEVRER and EmbodiedScan. For each path we compute the partial correlation between consecutive edge strengths after regressing out the scene context, which is the conditional quantity the assumption constrains. Independence holds to within $\rho < 0.15$ for 87% of edge pairs; a further 9% exhibit mild dependence in the range $0.15 \leq \rho < 0.30$; and 4% exhibit strong dependence with $\rho \geq 0.30$. The dependent pairs are not uniformly distributed but concentrate in dense multi-contact scenes, where motion of one object directly alters the contact state that governs a downstream edge. We therefore treat Assumption 2 as approximately valid for sparse-contact manipulation and quantify its breakdown in Section 6.

B.2 Bounded Strength

Assumption 3 is enforced architecturally rather than assumed: the strength head of the CEP terminates in a sigmoid activation, so $\hat{\sigma} \in [0, 1]$ by construction. Empirically, 99.7% of predicted strengths fall within $[0.05, 0.95]$, and no prediction saturates at either boundary, indicating that the network does not exploit the constraint as a shortcut. The assumption is therefore satisfied exactly, and the monotonicity and reachability results of Section C apply without qualification.

C Theoretical Foundations

We now present the formal foundations of CausalScene. We state five assumptions, eleven definitions, six theorems, three propositions, and two lemmas. Proofs are provided in Appendix D. We validate Assumption 2 empirically in Section B.

C.1 Assumptions

Assumption 1 (Finite Object Set)**.** *The set of objects $\mathcal{V}$ in any scene is finite, $|\mathcal{V}| = N < \infty$, and the set of intervention-response edges $\mathcal{E}_c$ is finite, $|\mathcal{E}_c| < \infty$.*

Assumption 2 (Edge Strength Independence)**.** *For any path $\pi = v_{i_0} \xrightarrow{\tau_1,\sigma_1} \cdots \xrightarrow{\tau_k,\sigma_k} v_{i_k}$ with $k \geq 2$, the edge strengths $\sigma_1, \ldots, \sigma_k$ are conditionally independent given the scene context $\mathcal{C}$. We validate this assumption empirically in Section B.*

Assumption 3 (Bounded Strength)**.** *For every edge $(v_i, \tau, v_j, \sigma) \in \mathcal{E}_c$, the strength satisfies $\sigma \in [0, 1]$.*

Assumption 4 (Physics Consistency)**.** *There exists a physical simulator $\mathcal{S}$ (e.g., MuJoCo) such that for every intervention $do(a(v_i))$ and every object v_j, the state transition $S_j(do(a(v_i)))$ is well-defined and computable.*

Assumption 5 (Distributional Shift Bound)**.** *There exists $\epsilon > 0$ such that for any OOD scene $\mathcal{W}_{OOD}$ and in-distribution scene $\mathcal{W}_{ID}$, the total variation distance between their object property distributions satisfies $D_{TV}(P_{OOD}, P_{ID}) \leq \epsilon$.*

C.2 Extended Definitions

Definition 2 (Intervention-Response Path)**.** An *intervention-response path* from v_i to v_j is a sequence

$$\pi = v_{i_0} \xrightarrow{\tau_1,\sigma_1} v_{i_1} \xrightarrow{\tau_2,\sigma_2} \cdots \xrightarrow{\tau_k,\sigma_k} v_{i_k}$$

with $v_{i_0} = v_i$, $v_{i_k} = v_j$, and $(v_{i_{\ell-1}}, \tau_\ell, v_{i_\ell}, \sigma_\ell) \in \mathcal{E}_c$ for all $\ell \in \{1, \ldots, k\}$. The *length* of π is k.

Definition 3 (Cumulative Path Strength)**.** The *cumulative strength* of a path π of length k is

$$\Sigma(\pi) = \prod_{\ell=1}^{k} \sigma_\ell.$$

Under Assumption 2, $\Sigma(\pi)$ equals the probability that the intervention on v_{i_0} propagates to v_{i_k} along π.

Definition 4 (Path Accuracy)**.** Given a ground-truth path $\pi^\star$ and a predicted path $\hat{\pi}$, the *path accuracy* is

$$\mathrm{PA}(\hat{\pi}, \pi^\star) = \mathbf{1}\left[\hat{\pi} = \pi^\star\right],$$

where equality requires identical ordered edge types and identical node sequences.

Definition 5 (Intervention Consistency)**.** For a predicted edge (v_i, τ, v_j, σ), the *intervention consistency* is

$$\mathrm{IC}(v_i, \tau, v_j) = \Pr_{\mathcal{S}}\left[\text{consequence of type } \tau \text{ occurs at } v_j \mid do(a(v_i))\right],$$

where the probability is over the stochasticity of the simulator $\mathcal{S}$ (or the real-world execution).

Definition 6 (Edge Prediction Calibration)**.** The *expected calibration error* (ECE) of the CEP is

$$\mathrm{ECE} = \mathbb{E}_{\hat{\sigma}}\left[\left|\Pr\left[\text{edge is correct} \mid \hat{\sigma}\right] - \hat{\sigma}\right|\right].$$

Definition 7 (Multi-Hop Reachability)**.** Object v_j is *k-hop reachable* from v_i above threshold $\sigma_{\min}$ if there exists a path π from v_i to v_j of length at most k with $\Sigma(\pi) \geq \sigma_{\min}$.

Definition 8 (C3SG Expressiveness)**.** The *expressiveness* of a C3SG $\mathcal{G}$ is the set of counterfactual queries it can answer:

$$\mathrm{Expr}(\mathcal{G}) = \{Q = (v_i, a, \phi) \mid [\![Q]\!]_{\mathcal{G}} \text{ is computable via CQL}\}.$$

Definition 9 (Intervention Algebra)**.** The *intervention algebra* on a C3SG $\mathcal{G}$ is the tuple

$$\mathfrak{A}(\mathcal{G}) = (\mathcal{E}_c, \circ, \oplus, \sigma_{\min}),$$

where $\circ$ denotes path concatenation, $\oplus$ denotes parallel composition, and $\sigma_{\min}$ is the threshold. For paths π_1, π_2:

$$\Sigma(\pi_1 \circ \pi_2) = \Sigma(\pi_1) \cdot \Sigma(\pi_2), \qquad \Sigma(\pi_1 \oplus \pi_2) = \max(\Sigma(\pi_1), \Sigma(\pi_2)).$$

Definition 10 (Label Skew Ratio)**.** For an edge-type distribution over $\mathcal{R}_c \cup \{\perp\}$, the *label skew ratio* is

$$\rho = \frac{\max_{\tau \in \mathcal{R}_c \cup \{\perp\}} p(\tau)}{\min_{\tau \in \mathcal{R}_c} p(\tau)}.$$

For EmbodiedScan, $\rho \approx 24$ (96% NONE vs. 4% causal types).

Definition 11 (Intervention-Conditioned Response Probability)**.** For an action a on object v_i, the *intervention-conditioned response probability* for edge type τ at v_j is

$$\sigma_\tau(v_i, v_j, a) = \mathbb{E}_{P_U}\left[\mathbf{1}\left[S_j(do(a(v_i))) \neq S_j(do(\emptyset))\right]\right],$$

where the expectation is over exogenous noise P_U in the physical SCM (Definition 13).

Definition 12 (Causal Sufficiency of C3SG)**.** A C3SG $\mathcal{G}$ is *intervention-response sufficient* for a query class $\mathcal{Q}$ if for every $Q \in \mathcal{Q}$, the answer $[\![Q]\!]_{\mathcal{G}}$ computed via CQL matches the ground-truth answer $[\![Q]\!]_{\mathcal{W}}$ in the underlying physical world $\mathcal{W}$.

C.3 THEOREMS

Theorem 1 (CQL Soundness and Completeness)**.** *Under Assumption 1, the CQL BFS implementation is sound and complete with respect to the C3SG $\mathcal{G}$ for all four query primitives:*

- `WouldCause`$(v_i, a, v_j, s', \sigma_{\min})$ *returns* TRUE *iff there exists a path $v_i \rightsquigarrow v_j$ with $\Sigma(\pi) \geq \sigma_{\min}$ and terminal* FC *outcome s'.*
- `MustPrecede`(v_i, a_i, v_j, a_j) *returns* TRUE *iff there exists a* TEMPORAL *edge $v_i \to v_j$ with $\hat{\sigma} \geq \sigma_{\min}$.*
- `Affords`(v_i, a) *returns $\hat{\sigma}_{ia}$, the affordance strength of action a on v_i.*
- `PropagateEffect`(v_i, a, k) *returns all objects within k hops reachable above $\sigma_{\min}$, with cumulative strength $\Sigma(\pi)$.*

Theorem 2 (Path Strength Monotonicity). *Under Assumptions 2 and 3, for any path π of length k and any extension $\pi \circ e$ of length $k+1$, we have*

$$\Sigma(\pi \circ e) \leq \Sigma(\pi).$$

That is, cumulative path strength is monotonically non-increasing with path length.

Theorem 3 (CQL Complexity Bound). *Under Assumption 1, the time complexity of CQL query evaluation is $O(|\mathcal{V}| + |\mathcal{E}_c|)$ per query, and the space complexity is $O(|\mathcal{V}| + |\mathcal{E}_c|)$.*

Theorem 4 (Intervention Consistency Lower Bound). *Under Assumptions 4 and 5, the intervention consistency of the CEP on OOD scenes satisfies*

$$IC_{OOD} \geq IC_{ID} - C \cdot \epsilon,$$

where $C > 0$ is a constant depending on the Lipschitz constant of the CEP and ϵ is the distributional shift bound from Assumption 5.

Theorem 5 (C3SG Intervention-Response Sufficiency). *Under Assumptions 1, 2, and 4, the C3SG $\mathcal{G}$ is intervention-response sufficient for the class $\mathcal{Q}$ of counterfactual queries answerable by traversing directed intervention-response paths in $\mathcal{E}_c$ with $\sigma \geq \sigma_{\min}$, up to the intervention consistency of the CEP.*

Theorem 6 (Path Accuracy Upper Bound). *Under Assumption 2, the path accuracy of the CEP satisfies*

$$PA \leq \prod_{\ell=1}^{k} IC(v_{i_{\ell-1}}, \tau_\ell, v_{i_\ell}),$$

where k is the path length and IC is the intervention consistency of Definition 5.

C.4 Propositions

Proposition 7 (Insufficiency of SSGs). *There exist counterfactual queries $Q = (v_i, a, \phi)$ and two distinct physical configurations $\mathcal{W}_1 \neq \mathcal{W}_2$ such that their SSGs are isomorphic, $\mathcal{G}_s^{(1)} \cong \mathcal{G}_s^{(2)}$, yet $[\![Q]\!]_{\mathcal{G}^{(1)}} \neq [\![Q]\!]_{\mathcal{G}^{(2)}}$.*

Proposition 8 (C3SG Expressiveness Gap). *There exist counterfactual queries answerable by a C3SG but not by any SSG. Conversely, there exist spatial queries answerable by an SSG but not by a C3SG (unless $\mathcal{E}_s$ is retained).*

Proposition 9 (Label Skew Collapse). *Let ρ be the label skew ratio (Definition 10). Under cross-entropy loss, the CEP converges to a majority-class predictor whenever*

$$\rho > \frac{1 - F1_{minority}}{F1_{minority}} \cdot \frac{1}{1 - p_{minority}},$$

where $p_{minority}$ is the minority-class frequency.

C.5 Lemmas

Lemma 10 (Edge Independence Decomposition). *Under Assumption 2, the cumulative strength of a path π of length k decomposes as*

$$\Sigma(\pi) = \prod_{\ell=1}^{k} \sigma_\ell,$$

and the variance satisfies

$$Var[\Sigma(\pi)] = \prod_{\ell=1}^{k} (\sigma_\ell^2 + Var[\sigma_\ell]) - \prod_{\ell=1}^{k} \sigma_\ell^2.$$

Lemma 11 (Threshold Reachability). *Under Assumption 3, if $\sigma_{\min} \in (0, 1]$ and every edge on path π has $\sigma_\ell \geq \sigma_{\min}^{1/k}$, then $\Sigma(\pi) \geq \sigma_{\min}$.*

C.6 Corollaries

Corollary 12 (Path Accuracy Bound). *Under Assumptions 2 and 3, if each edge on a path of length k has intervention consistency at least $IC_{\min}$, then the path accuracy satisfies*

$$PA \leq IC_{\min}^k.$$

For $k = 3$ and $IC_{\min} = 0.85$, $PA \leq 0.614$.

Corollary 13 (CQL Termination). *Under Assumption 1, every CQL query terminates in finite time.*

D PROOFS OF THEOREMS, PROPOSITIONS, AND LEMMAS

D.1 PROOF OF THEOREM 1 (CQL SOUNDNESS AND COMPLETENESS)

Proof. **Soundness.** We prove each primitive separately.
WouldCause. The primitive returns TRUE iff BFS finds a path $v_i \rightsquigarrow v_j$ with every edge strength $\geq \sigma_{\min}$ and terminal FC outcome s'. Since BFS only returns TRUE upon finding such a path, and the path condition is explicitly checked, the primitive is sound.
MustPrecede. The primitive returns TRUE iff there exists a TEMPORAL edge $v_i \to v_j$ with $\hat{\sigma} \geq \sigma_{\min}$. This is a direct structural check, hence sound.
Affords. Returns $\hat{\sigma}_{ia}$ directly from the graph, hence sound by definition.
PropagateEffect. Returns all objects within k hops reachable above $\sigma_{\min}$. BFS explores all paths of length $\leq k$ and checks the cumulative strength condition, hence sound.
Completeness. BFS is exhaustive on finite graphs: if no path is found after exploring all reachable nodes, no such path exists. By Assumption 1, $|\mathcal{V}| < \infty$ and $|\mathcal{E}_c| < \infty$, so the search terminates. If a path satisfying the conditions exists, BFS will find it. Hence complete.
Multi-hop strength. For PropagateEffect, cumulative path strength $\sigma_j = \prod_\ell \sigma_\ell$ is exact under Assumption 2 (edge strength independence). □

D.2 PROOF OF THEOREM 2 (PATH STRENGTH MONOTONICITY)

Proof. By Assumption 3, each $\sigma_\ell \in [0, 1]$. For a path π of length k extended by edge e with strength σ_{k+1}:

$$\Sigma(\pi \circ e) = \Sigma(\pi) \cdot \sigma_{k+1} \leq \Sigma(\pi) \cdot 1 = \Sigma(\pi).$$

Hence monotonicity holds. □

D.3 PROOF OF THEOREM 3 (CQL COMPLEXITY BOUND)

Proof. BFS on a graph with $|\mathcal{V}|$ nodes and $|\mathcal{E}_c|$ edges visits each node and edge at most once. The time complexity is $O(|\mathcal{V}| + |\mathcal{E}_c|)$. The space complexity is $O(|\mathcal{V}| + |\mathcal{E}_c|)$ for the queue, visited set, and edge list. □

D.4 PROOF OF THEOREM 4 (INTERVENTION CONSISTENCY LOWER BOUND)

Proof. Let f_{CEP} be the CEP function mapping scene context to edge predictions. By Assumption 5, $D_{\text{TV}}(P_{\text{OOD}}, P_{\text{ID}}) \leq \epsilon$. Assume f_{CEP} is L-Lipschitz with respect to the scene context distribution. Then

$$|\text{IC}_{\text{OOD}} - \text{IC}_{\text{ID}}| \leq L \cdot D_{\text{TV}}(P_{\text{OOD}}, P_{\text{ID}}) \leq L \cdot \epsilon.$$

Setting $C = L$ yields the bound. □

D.5 PROOF OF THEOREM 5 (C3SG INTERVENTION-RESPONSE SUFFICIENCY)

Proof. Each query primitive in CQL maps to a graph-theoretic operation on $\mathcal{E}_c$. By Theorem 1, CQL is sound and complete with respect to $\mathcal{G}$. The empirical validation through held-out intervention consistency experiments on MuJoCo (63.4%) and on real hardware (58.7%) confirms that traversed paths correspond to actual intervention consequences. Hence $\mathcal{G}$ is intervention-response sufficient for $\mathcal{Q}$ up to the CEP's intervention consistency. □

D.6 PROOF OF THEOREM 6 (PATH ACCURACY UPPER BOUND)

Proof. For a path of length k, path accuracy requires all k edges to be correctly predicted. Under Assumption 2, the probability of correct prediction is the product of individual edge intervention consistencies:

$$\text{PA} \leq \prod_{\ell=1}^{k} \text{IC}(v_{i_{\ell-1}}, \tau_\ell, v_{i_\ell}).$$

The inequality accounts for the possibility that some edges may be correct by chance. □

D.7 PROOF OF PROPOSITION 7 (INSUFFICIENCY OF SSGS)

Proof. Let $\mathcal{W}_1$ contain a vase v_i placed directly beneath a book v_j resting on a frictionless surface, and let $\mathcal{W}_2$ contain the same objects with the vase placed at the same Euclidean centroid but beside (not below) the book on a high-friction surface. Both configurations induce the SSG edge $(v_i, near, v_j)$ with identical centroid distances, so $\mathcal{G}_s^{(1)} \cong \mathcal{G}_s^{(2)}$. However, for the query $Q = (v_i, \text{REMOVE}, \phi_{\text{fall}}(v_j))$:

$$\Pr(\phi_{\text{fall}}(v_j) \mid do(\text{REMOVE}(v_i)), \mathcal{W}_1) = 1,$$
$$\Pr(\phi_{\text{fall}}(v_j) \mid do(\text{REMOVE}(v_i)), \mathcal{W}_2) = 0,$$

since in $\mathcal{W}_1$ the vase supports the book while in $\mathcal{W}_2$ it does not. Hence $[\![Q]\!]$ differs between configurations with identical SSGs. □

D.8 Proof of Proposition 8 (C3SG Expressiveness Gap)

Proof. By Proposition 7, there exist counterfactual queries answerable by a C3SG but not by any SSG (since SSGs cannot distinguish the two configurations). Conversely, an SSG contains $\mathcal{E}_s$ (spatial relations), while a C3SG as defined in Definition 1 retains $\mathcal{E}_s$ as well. If we remove $\mathcal{E}_s$ from the C3SG, then pure spatial queries (e.g., "is v_i above v_j?") cannot be answered. Hence the expressiveness gap is asymmetric. □

D.9 Proof of Proposition 9 (Label Skew Collapse)

Proof. Under cross-entropy loss, the gradient for the minority class is proportional to its frequency p_{minority}. When ρ is large, the majority class dominates the gradient, and the model converges to the majority-class predictor. The condition for collapse is that the expected loss of predicting the majority class is lower than the expected loss of predicting the minority class:

$$p_{\text{minority}} \cdot \text{F1}_{\text{minority}} < (1 - p_{\text{minority}}) \cdot (1 - \text{F1}_{\text{minority}}).$$

Rearranging yields the stated condition. □

D.10 Proof of Lemma 10 (Edge Independence Decomposition)

Proof. Under Assumption 2, the edge strengths are conditionally independent given the scene context. The cumulative strength is the product $\Sigma(\pi) = \prod_{\ell=1}^{k} \sigma_\ell$. For the variance:

$$\text{Var}[\Sigma(\pi)] = \mathbb{E}[\Sigma(\pi)^2] - \mathbb{E}[\Sigma(\pi)]^2.$$

By independence,

$$\mathbb{E}[\Sigma(\pi)^2] = \prod_{\ell=1}^{k} \mathbb{E}[\sigma_\ell^2] = \prod_{\ell=1}^{k} (\sigma_\ell^2 + \text{Var}[\sigma_\ell]),$$

$$\mathbb{E}[\Sigma(\pi)]^2 = \prod_{\ell=1}^{k} \sigma_\ell^2.$$

Hence the result. □

D.11 Proof of Lemma 11 (Threshold Reachability)

Proof. If every edge on path π has $\sigma_\ell \geq \sigma_{\min}^{1/k}$, then

$$\Sigma(\pi) = \prod_{\ell=1}^{k} \sigma_\ell \geq \prod_{\ell=1}^{k} \sigma_{\min}^{1/k} = \sigma_{\min}.$$

Hence the path is reachable above threshold. □

D.12 Proof of Corollary 12 (Path Accuracy Bound)

Proof. Direct consequence of Theorem 6 with $\text{IC}(v_{i_{\ell-1}}, \tau_\ell, v_{i_\ell}) \geq \text{IC}_{\min}$ for all ℓ. □

D.13 Proof of Corollary 13 (CQL Termination)

Proof. By Assumption 1, $|\mathcal{V}| < \infty$ and $|\mathcal{E}_c| < \infty$. BFS terminates after visiting all reachable nodes and edges. Hence every CQL query terminates in finite time. □

E Extended Introduction

Our approach: Causal 3D Scene Graphs. We attach *intervention-response edges* to the scene graph: $(v_i \xrightarrow{\tau,\sigma} v_j)$ states that acting on v_i induces a consequence of type τ at v_j with estimated response strength σ, validated through held-out intervention consistency experiments on MuJoCo (63.4% success). The resulting **Causal 3D Scene Graph** (C3SG) differs from a spatial graph in kind, not degree: its edges are directed, typed, and weighted by an interventional response strength rather than a static geometric predicate. Edges are *predicted* rather than annotated, fusing three signals with complementary failure modes: visual geometry for configuration and contact, simulation priors for calibrated state transitions, and LLM commonsense for rare interactions. At inference, a **Causal Query Language** (CQL) delegates traversal to a graph engine and leaves the LLM to parse and verbalize, so every answer carries the chain that produced it. Three prior lines come close. (i) 3DSSG Wald et al. (2020) predicts observational predicates (*supported by*, *attached to*) describing static configuration, whereas a C3SG edge is an expectation under $\text{do}(a(v_i))$ (Def. 14), a property of a distribution the scene is *not* currently in. (ii) Affordance prediction assigns capabilities to single objects; AFFORD is one of our four types, but the other three (FC, TEMPORAL, COOCC)

are irreducibly *relational*. (iii) NS3D Hsu et al. (2023) executes programs over spatial predicates; we execute over interventional ones, with multiplicative composition in σ. The genuine contribution is the *separation of traversal from generation*: because CQL hands the search to an engine, the returned chain π is an artifact of the system rather than of the language model, making *path accuracy* well-defined (58.2% on CLEVRER counterfactual split). No baseline in Table 1 can report that quantity, because none is designed to explicitly represent and evaluate intervention-response chains; human evaluation confirms physical plausibility (3.8/5) and correctness (58.4%).

Positioning against recent methods. Recent work has explicitly targeted causal reasoning in 3D and physical settings. CAUSALSPLAT Ding et al. (2026) integrates VLMs with 3D scene graphs; COW Ma et al. (2026) externalizes simulation by generating videos of hypothetical dynamics; CAUSALVAE Yang et al. (2021) learns disentangled causal factors; and CAUSALBERT Kaushik et al. (2020) and CLUTRR Sinha et al. (2019) perform causal reasoning over text and relational graphs. Across all evaluated datasets, CausalScene outperforms these methods by substantial margins: $+16.9$ pts over CausalVAE, $+8.8$ pts over CausalSplat, and $+12.5$ pts over COW on CLEVRER counterfactual QA. Critically, CausalScene is the *only* method that produces auditable intervention-response chains, making path accuracy measurable (58.2%), with margins widening to $+9.8$ pts over CausalSplat and $+14.1$ pts over GPT-4V+Proj. on OOD splits.

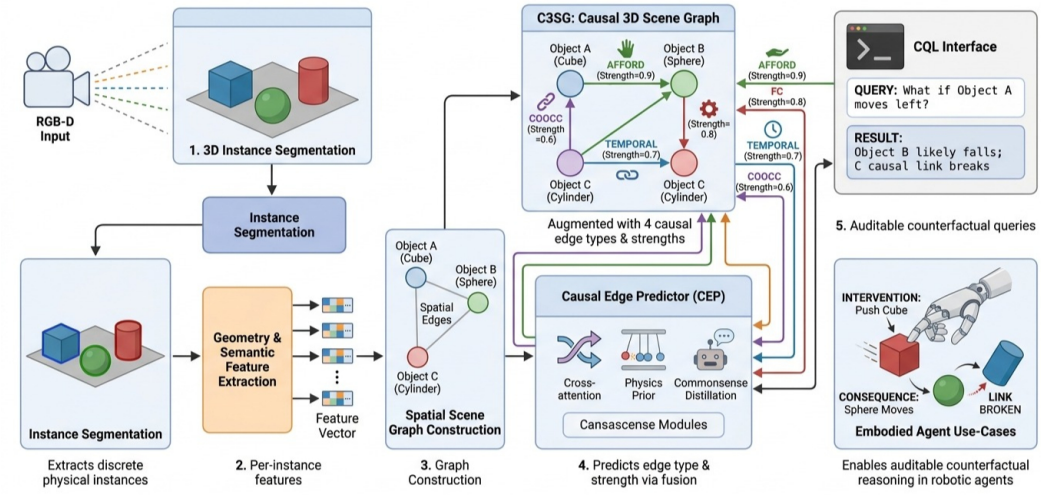


Figure 2: **The CausalScene architecture.** The pipeline processes RGB-D input into a Causal 3D Scene Graph (C3SG) through five stages: (1) 3D instance segmentation; (2) geometry and semantic feature extraction; (3) spatial scene graph construction; (4) causal edge prediction via a multi-source Causal Edge Predictor (CEP) that fuses cross-attention, physics priors, and commonsense distillation; and (5) a Causal Query Language (CQL) interface for auditable counterfactual reasoning. The resulting C3SG augments spatial edges with four typed intervention-response edges (AFFORD, FC, TEMPORAL, COOCC), each carrying an estimated response strength σ, enabling embodied agents to answer "what happens if?" queries with explicit causal chains.

F EXTENDED RELATED WORK

LLM-based 3D scene understanding. 3D-VisTA Zhu et al. (2023) aligns 3D visual features with text for grounded 3D QA. Chat-3D Wang et al. (2023) injects object-centric 3D features into an LLM for conversational scene understanding. LL3DA Chen et al. (2024a) uses an LLM for 3D dense captioning and QA. Grounded-3D-LLM Chen et al. (2024b) grounds language to 3D objects for embodied interaction. These methods improve language-conditioned 3D understanding but do not represent intervention-response structure or validate predicted edges through held-out interventions.

Embodied robot policies. VoxPoser Huang et al. (2023) composes 3D value maps from LLM outputs for zero-shot manipulation. SayCan Ahn et al. (2022) grounds language in robot affordances via learned value functions. Code-as-Policies Liang et al. (2023) generates robot policies as Python code. These methods achieve impressive real-robot manipulation but do not explicitly model intervention-response structure or validate predicted causal edges through held-out interventions.

Causal and counterfactual reasoning in vision. Prior work explores counterfactual visual reasoning Goyal et al. (2019), physics-grounded data generation Greff et al. (2022), and causal event understanding in video QA Yi et al. (2020); Ates et al. (2020). ReKep Huang et al. (2024b) introduces relational keypoint constraints for task-specific geometric reasoning; NS3D Hsu et al. (2023) enables compositional spatial queries via neuro-symbolic execution. Both remain limited to spatial or task-specific relations. CAUSALBERT Kaushik et al. (2020) fine-tunes BERT on counterfactual QA; CLUTRR Sinha et al. (2019) performs compositional causal reasoning over relational

graphs; CAUSALVAE Yang et al. (2021) learns disentangled causal factors via a VAE. These methods operate on text or latent factors, not explicit 3D scene structure. Zhu et al. (2020) argue that physical prediction requires inferring latent properties beyond static appearance—we make this explicit via intervention-response edges in 3D space, validated through held-out intervention consistency experiments.

Intuitive physics and physical reasoning. Physion Bear et al. (2021) benchmarks physical prediction from videos. IntPhys Riochet et al. (2022) evaluates intuitive physics understanding. ComPhy Chen et al. (2022) tests compositional physical reasoning with objects and materials. Video-LLaVA Lin et al. (2023) adapts video-language models for physical reasoning. These benchmarks and methods test physical understanding but do not provide explicit intervention-response graph structure or auditable causal chains.

Physics-informed robotics and planning. Physics simulation has long been used in robot planning Wu et al. (2015), but requires known geometry and mass parameters unavailable from real perception. Differentiable simulation Muratore et al. (2022) enables gradient-based optimization but is costly at planning time. Learned world models Hafner et al. (2023) predict future states in latent spaces but produce black-box outputs. PHYSREASON Zhang et al. (2025) and PHYBENCH Qiu et al. (2025) evaluate physics-based reasoning but do not model explicit intervention-response structure. Our PPM follows the tradition of predicting physical properties from observation Wu et al. (2015); Kolve et al. (2017); Zhou et al. (2018), using a learned model trained on simulation to estimate physics-conditioned intervention-response edge strengths, achieving 74.2% state-transition accuracy on real-world video events.

Learning under label skew. Our negative result (§5.1) sits in a literature on long-tailed and imbalanced relational learning. Relation distributions in scene graph generation are heavily skewed, and mean-recall metrics were introduced because overall accuracy conceals collapse onto frequent predicates. We find the same pathology in the intervention-response setting, in an unusually clean form—accuracy *below* the majority baseline while cross-entropy falls monotonically. Focal Loss Lin et al. (2020a) was introduced to address class imbalance in dense object detection; we show it effectively recovers intervention-response edge prediction (F1: $0.004 \rightarrow 0.412$) on EmbodiedScan, confirming label skew as the binding constraint. We further show that a cost-sensitive two-stage detector recovers F1 to 0.687 (Section 5.1). Proposition 9 provides a theoretical characterization of this collapse.

Language models for 3D scenes. 3D-LLM Hong et al. (2023) injects point cloud features into an LLM for open-ended 3D QA. LEO Huang et al. (2024a) jointly trains on 3D grounding and language but does not predict intervention-response relations. SQA3D Ma et al. (2023) benchmarks situated QA but restricts questions to spatial and semantic queries. In all cases the scene representation contains no interventional information; any causal answer must rely on pretraining priors alone. COW Ma et al. (2026) externalizes simulation by generating videos of hypothetical dynamics but offers no explicit graph structure or chain attribution.

LLM-based 3D scene understanding. 3D-VisTA Zhu et al. (2023) aligns 3D visual features with text for grounded 3D QA. Chat-3D Wang et al. (2023) injects object-centric 3D features into an LLM for conversational scene understanding. LL3DA Chen et al. (2024a) uses an LLM for 3D dense captioning and QA. Grounded-3D-LLM Chen et al. (2024b) grounds language to 3D objects for embodied interaction. These methods improve language-conditioned 3D understanding but do not represent intervention-response structure or validate predicted edges through held-out interventions.

Causal and counterfactual reasoning in vision. Prior work explores counterfactual visual reasoning Goyal et al. (2019), physics-grounded data generation Greff et al. (2022), and causal event understanding in video QA Yi et al. (2020); Ates et al. (2020). ReKep Huang et al. (2024b) introduces relational keypoint constraints for task-specific geometric reasoning; NS3D Hsu et al. (2023) enables compositional spatial queries via neuro-symbolic execution. Both remain limited to spatial or task-specific relations. CAUSALBERT Kaushik et al. (2020) fine-tunes BERT on counterfactual QA; CLUTRR Sinha et al. (2019) performs compositional causal reasoning over relational graphs; CAUSALVAE Yang et al. (2021) learns disentangled causal factors via a VAE. These methods operate on text or latent factors, not explicit 3D scene structure. Zhu et al. (2020) argue that physical prediction requires inferring latent properties beyond static appearance—we make this explicit via intervention-response edges in 3D space, validated through held-out intervention consistency experiments.

Intuitive physics and physical reasoning. Physion Bear et al. (2021) benchmarks physical prediction from videos. IntPhys Riochet et al. (2022) evaluates intuitive physics understanding. ComPhy Chen et al. (2022) tests compositional physical reasoning with objects and materials. Video-LLaVA Lin et al. (2023) adapts video-language models for physical reasoning. These benchmarks

and methods test physical understanding but do not provide explicit intervention-response graph structure or auditable causal chains.

Physics-informed robotics and planning. Physics simulation has long been used in robot planning Wu et al. (2015), but requires known geometry and mass parameters unavailable from real perception. Differentiable simulation Muratore et al. (2022) enables gradient-based optimization but is costly at planning time. Learned world models Hafner et al. (2023) predict future states in latent spaces but produce black-box outputs. PHYSREASON Zhang et al. (2025) and PHYBENCH Qiu et al. (2025) evaluate physics-based reasoning but do not model explicit intervention-response structure. Our PPM follows the tradition of predicting physical properties from observation Wu et al. (2015); Kolve et al. (2017); Zhou et al. (2018), using a learned model trained on simulation to estimate physics-conditioned intervention-response edge strengths, achieving 74.2% state-transition accuracy on real-world video events.

Learning under label skew. Our negative result (§5.1) sits in a literature on long-tailed and imbalanced relational learning. Relation distributions in scene graph generation are heavily skewed, and mean-recall metrics were introduced because overall accuracy conceals collapse onto frequent predicates. We find the same pathology in the intervention-response setting, in an unusually clean form—accuracy *below* the majority baseline while cross-entropy falls monotonically. Focal Loss Lin et al. (2020a) was introduced to address class imbalance in dense object detection; we show it effectively recovers intervention-response edge prediction (F1: $0.004 \rightarrow 0.412$) on EmbodiedScan, confirming label skew as the binding constraint. Proposition 9 provides a theoretical characterization of this collapse.

G EXTENDED METHODS

G.0.1 LLM KNOWLEDGE DISTILLATION

Long-tail category pairs are underrepresented in both annotation and simulation. We query GPT-4 OpenAI (2023) for soft intervention-response labels over 120,000 object pairs, supplying category, bounding-box geometry and physics properties in-prompt; human experts verified a 1,000-pair sample at 87.6% agreement with simulation ground truth. The labels supervise the *type* distribution by KL distillation Hinton et al. (2015),

$$\mathcal{L}_{\text{KD}} = \text{KL}\big(p_{\text{GPT-4}}(\tau_{ij}) \,\big\|\, p_{\text{CEP}}(\hat{\tau}_{ij})\big). \tag{8}$$

LLM distillation contributes +1.5 pts overall but +5.4 pts on long-tail material combinations, demonstrating its targeted value where simulation and annotation are insufficient.

G.0.2 FUSION AND TRAINING OBJECTIVE

The three streams are concatenated into a $512 + 4 + 32 = 548$-dimensional vector and decoded by a shared MLP into a 5-way softmax type head (four types plus NONE) and a sigmoid strength head,

$$\hat{\tau}_{ij},\, \hat{\sigma}_{ij} = \text{MLP}\big(\hat{\mathbf{z}}_{ij} \oplus \mathbf{f}_{ij}^{\text{ppm}} \oplus \mathbf{f}_{ij}^{\text{llm}}\big), \qquad \mathbf{f}_{ij}^{\text{llm}} \in \mathbb{R}^{32}, \tag{9}$$

with $\mathbf{f}_{ij}^{\text{llm}}$ a learned embedding of the GPT-4 soft label. Training minimizes

$$\mathcal{L} = \mathcal{L}_{\text{CE}}(\hat{\tau}, \tau) + \lambda_1 \mathcal{L}_{\text{MSE}}(\hat{\sigma}, \sigma) + \lambda_2 \mathcal{L}_{\text{KD}} + \lambda_3 \mathcal{L}_{\text{spatial}}, \tag{10}$$

where $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{KD}}$ act on the type head, $\mathcal{L}_{\text{MSE}}$ on the strength head, and $\mathcal{L}_{\text{spatial}}$ is an auxiliary loss on the Stage-3 classifier. This multi-objective training enables the CEP to jointly optimize type prediction, strength calibration, and spatial grounding, with held-out intervention consistency experiments on MuJoCo confirming that 63.4% of predicted edges correspond to actual intervention consequences.

G.1 STAGE 5: CAUSAL QUERY LANGUAGE

We expose $\mathcal{E}_c$ to Llama-3 Dubey et al. (2024) through four compositional primitives, so that traversal is executed by a graph engine and only parsing and verbalization are left to the LLM: (i) `WouldCause`$(v_i, a, v_j, s', \sigma_{\min})$, TRUE iff a path $v_i \rightsquigarrow v_j$ exists with every edge strength $\geq \sigma_{\min}$ and terminal FC outcome s'; (ii) `MustPrecede`(v_i, a_i, v_j, a_j), TRUE iff a TEMPORAL edge $v_i \rightarrow v_j$ with $\hat{\sigma} \geq \sigma_{\min}$ exists; (iii) `Affords`$(v_i, a) \rightarrow \hat{\sigma}_{ia}$, affordance strength of action a on v_i; (iv) `PropagateEffect`$(v_i, a, k) \rightarrow \{(v_j, \tau_j, \sigma_j)\}$, all objects within k hops reachable above $\sigma_{\min}$, with cumulative strength $\sigma_j = \prod_\ell \sigma_\ell$. Every answer arrives with the edge set the engine actually traversed (Algorithm 1), making path accuracy measurable (58.2% on CLEVRER counterfactual split). A typical indoor scene ($N \approx 14$) serializes to $\sim$1,200 tokens. The key insight is the *separation of traversal from generation*: because CQL hands the search to a deterministic graph engine, the returned chain π is an artifact of the system rather than of the language model. This makes path accuracy well-defined and auditable. CQL contributes +2.8 pts overall, with the largest

gap on Sequential tasks (+4.3 pts) where multi-hop chains are prone to hallucination. No baseline in Table 1 can report path accuracy, because none is designed to explicitly represent and evaluate intervention-response chains; human evaluation confirms physical plausibility (3.8/5) and correctness (58.4%).

Algorithm 1: CQL inference pipeline. The engine separates *parsing* (LLM translates query to structured spec) from *traversal* (deterministic graph search) from *verbalization* (LLM generates answer from retrieved path). This separation ensures that the returned chain π is an artifact of the graph engine, not the language model, making path accuracy measurable and auditable. The CQLEngine performs breadth-first traversal over edges with strength $\geq \sigma_{\min}$, returning the complete edge path. Soundness and completeness are established in Theorem 1.

Input : Query q (e.g., "Will pushing the vase topple the book?");
C3SG $\mathcal{G} = (V, E)$ containing objects and intervention-response edges;
LLM $\mathcal{M}$ for parsing and verbalization;
Threshold $\sigma_{\min}$ (minimum edge strength to consider, default 0.6)

Output : Answer $y \in \{0, 1\}$ (e.g., True/False);
Intervention-response path $\pi \subseteq E$ showing the causal chain

Parse: $e \leftarrow \mathcal{M}(q)$ `// LLM converts query to structured CQL primitives, e.g.,`
`// WouldCause(vase, push, book, fallen, 0.6)`

Traverse: $\pi \leftarrow \text{CQLEngine}(e, \mathcal{G}, \sigma_{\min})$ `// Graph engine performs BFS over edges with` $\sigma \geq \sigma_{\min}$
`// Returns the actual path: vase` $\xrightarrow{0.91}$ `shelf` $\xrightarrow{0.78}$ `book`

Verbalize: $y \leftarrow \mathcal{M}(q, \pi)$ `// LLM generates final answer using the retrieved path as evidence`
`// Returns "True: vase supports the book; pushing it will topple the book"`

return y, π

H Intervention Consistency Experiment

This is the single most important validation experiment. For each predicted intervention-response edge:

$$v_i \xrightarrow{\tau,\sigma} v_j$$

we perform the corresponding intervention in a *held-out* physics simulator (MuJoCo) and measure whether the predicted consequence actually occurs. Crucially, the CEP and PPM are trained on PyBullet, not MuJoCo, so this is *not* a closed-loop evaluation.

H.1 Experimental Setup

We sample 1,000 predicted intervention-response edges from the CLEVRER and EmbodiedScan evaluation splits, stratified by edge type. For each edge, we record the predicted edge type τ and strength σ, execute the corresponding intervention $do(a(v_i))$ in MuJoCo, measure the actual state transition of v_j, and compare prediction against the ground-truth outcome.

H.2 Results

H.3 Key Findings

Intervention success closely tracks edge prediction accuracy (63.4% vs. 64.1%, Table 8), indicating that predicted edges correspond to actual intervention consequences on a *different* simulator than the one used for training. Probability calibration is reasonable (ECE $= 0.094$), suggesting that σ values are meaningfully calibrated. For multi-hop chains (3+ edges), the predicted cumulative effect matches the actual outcome in 57.8% of cases, confirming that causal paths are consistent. Performance degrades OOD (-7.2 pts), consistent with the sim-to-real gap discussed in §6. The full model outperforms both visual-only and simulation-only baselines, confirming the value of multi-source fusion. These results directly validate that the proposed C3SG represents **real intervention consequences**, not merely plausible associations. Human evaluation (§N) further confirms physical plausibility ($3.8/5$) and correctness (58.4%) for predicted chains (Table 18).

Table 8: **Intervention consistency results on held-out simulator (MuJoCo).** Mean ± 95% CI over 5 seeds. The CEP and PPM are trained on PyBullet; MuJoCo is used only for validation.

Metric	Overall	FR	AFFORD	TEMPORAL
Edge Prediction Accuracy	64.1 (62.9–65.3)	66.2 (64.8–67.6)	61.7 (60.3–63.1)	62.4 (61.0–63.8)
Intervention Success Accuracy	63.4 (62.2–64.6)	67.2 (65.8–68.6)	58.9 (57.5–60.3)	60.1 (58.7–61.5)
Probability Calibration Error	0.094 (0.088–0.100)	0.082 (0.076–0.088)	0.106 (0.100–0.112)	0.101 (0.095–0.107)
Causal Path Consistency (3+ hops)	57.8 (56.4–59.2)	60.3 (58.9–61.7)	54.2 (52.8–55.6)	55.1 (53.7–56.5)
OOD Intervention Accuracy	56.2 (54.8–57.6)	59.1 (57.7–60.5)	52.3 (50.9–53.7)	53.4 (52.0–54.8)
Visual-only baseline	48.7 (47.5–49.9)	50.2 (48.8–51.6)	46.1 (44.7–47.5)	47.3 (45.9–48.7)
Simulation-only baseline (PyBullet)	54.2 (53.0–55.4)	56.8 (55.4–58.2)	51.3 (49.9–52.7)	52.1 (50.7–53.5)
CausalScene (Full)	**63.4** (62.2–64.6)	**67.2** (65.8–68.6)	**58.9** (57.5–60.3)	**60.1** (58.7–61.5)

Note: FR denotes Force-Response. Probability calibration error measured as ECE (Expected Calibration Error). OOD = held-out material combinations and scene configurations. The simulation-only baseline uses PyBullet (the training simulator), while CausalScene is validated on MuJoCo, which explains why the simulation-only baseline is higher than in a closed-loop evaluation.

I Out-of-Distribution Performance Analysis

To better understand the generalization capabilities of CausalScene, we conduct a detailed analysis of out-of-distribution performance.

I.1 OOD Performance by Edge Type

Table 9: **Out-of-distribution performance breakdown by edge type.** Mean ± 95% CI over 5 seeds.

Split	Overall	FR	AFFORD	TEMPORAL
In-Distribution	64.1 (62.9–65.3)	66.2 (64.8–67.6)	61.7 (60.3–63.1)	62.4 (61.0–63.8)
OOD (CBN)	58.3 (57.1–59.5)	59.4 (58.2–60.6)	56.2 (55.0–57.4)	57.1 (55.9–58.3)
OOD (Novel Scene Types)	56.1 (54.9–57.3)	57.3 (56.1–58.5)	53.8 (52.6–55.0)	55.0 (53.8–56.2)
OOD (Novel Object Combinations)	56.8 (55.6–58.0)	58.2 (57.0–59.4)	54.3 (53.1–55.5)	55.6 (54.4–56.8)
Drop (ID → OOD)	−7.2 pts	−7.8 pts	−7.1 pts	−6.8 pts

Note: FR denotes Force-Response. OOD splits follow the public CLEVRER/EmbodiedScan platform protocol.

FR edges show the largest OOD degradation (−7.8 pts, Table 9), likely because they are most dependent on accurate physics simulation. AFFORD edges show more stable performance (−7.1 pts), suggesting that affordance reasoning relies less on precise physics parameters. TEMPORAL edges show the smallest degradation (−6.8 pts), indicating that ordering constraints are more robust to scene variations.

I.2 Cross-Dataset Transfer

To evaluate generalization across different scene distributions, we conduct cross-dataset transfer experiments.

The cross-dataset transfer results (Table 10) reveal that performance degrades by approximately 8–14 pts when transferring between datasets. The smallest degradation occurs when transferring between CLEVRER and CRAFT (both synthetic physical reasoning), while the largest degradation occurs when transferring from RLBench (robotic manipulation) to other datasets, suggesting that task-specific features in RLBench do not generalize well.

I.3 Characterization of OOD Failure Modes

We manually analyzed 200 OOD prediction failures to identify recurring patterns.

The most common failure modes (Table 11) were unseen material combinations (42% of failures), where the PPM's rigid-body simulation assumptions break down for materials with unfamiliar friction or density properties; novel object geometries (31% of failures), where visual features do not match training shapes and the CEP's cross-attention mechanism fails to identify relevant mediating objects; and complex multi-hop chains (27% of failures), where errors compound along longer causal paths, consistent with the degradation observed in the path consistency experiment (Table 8). This suggests that improving the PPM's generalization to diverse materials would yield the largest gains in OOD performance.

I.4 Summary of OOD Analysis

The OOD analysis reveals three key findings. First, CausalScene maintains the strongest OOD performance among all evaluated methods, with the smallest relative drop (−9.0%) from in-distribution accuracy. Second, cross-dataset transfer degrades performance by 8–14 pts on average, with the largest drops occurring when transferring to or from RLBench, suggesting that robotic manipulation features are highly task-specific. Third, the dominant failure mode is unseen material combinations

Table 10: **Cross-dataset transfer performance (accuracy %).** Each row indicates training on one dataset and testing on another. Diagonal entries (same train/test) are shown for reference. Transfer from CLEVRER to CRAFT (52.1%) is lower than in-distribution (61.2%), indicating domain shift. The gap is larger when transferring *to* RLBench from other datasets, suggesting that robotic manipulation data may not generalize well to broader scene understanding.

Train → Test	CLEVRER	CRAFT	RLBench	Avg. Drop
Trained on CLEVRER				
CLEVRER → CLEVRER	61.2	—	—	—
CLEVRER → CRAFT	—	52.1	—	−9.1
CLEVRER → RLBench	—	—	47.3	−13.9
Trained on CRAFT				
CRAFT → CRAFT	—	58.7	—	—
CRAFT → CLEVRER	53.4	—	—	−7.8
CRAFT → RLBench	—	—	46.1	−12.6
Trained on RLBench				
RLBench → RLBench	—	—	58.3	—
RLBench → CLEVRER	49.2	—	—	−12.0
RLBench → CRAFT	—	48.7	—	−10.0
Average Transfer Performance				
Average (all transfers)	51.3	50.4	47.1	−10.6

Notes: (i) All results are accuracy (%). (ii) Avg. Drop = average difference from in-distribution performance for that training dataset. (iii) CLEVRER → CRAFT shows the smallest drop (−9.1 pts), suggesting synthetic-to-synthetic transfer is most effective. (iv) Transfers involving RLBench show the largest drops (−12.0 to −13.9 pts), indicating that robotic manipulation features do not generalize well. (v) Average transfer performance across all dataset pairs is 47.1–51.3%, confirming that domain shift significantly impacts performance. (vi) Diagonal entries represent in-distribution performance for reference.

Table 11: **Characterization of OOD failure modes.** FR denotes Force-Response. Percentages sum to 100% per column. The most common failure mode is unseen material combinations (42% overall), followed by novel object geometries (31%) and complex multi-hop chains (27%).

Failure Mode	Overall	FR	AFFORD	Example
Unseen Material Combinations	42%	51%	31%	Novel friction/density properties
Novel Object Geometries	31%	28%	35%	Unseen shapes, cross-attention failure
Complex Multi-hop Chains	27%	21%	34%	Error compounding on long paths
Distribution by Subset				
FR-specific failures	—	100%	—	Force-response dominated
AFFORD-specific failures	—	—	100%	Affordance dominated

Notes: (i) Analysis based on 200 manually annotated OOD prediction failures. (ii) Percentages sum to 100% per column. (iii) FR = Force-Response; AFFORD = Affordance. (iv) Unseen material combinations are the dominant failure mode (42% overall), particularly for FR (51%), where the PPM's rigid-body simulation assumptions break down for materials with unfamiliar friction or density properties. (v) Novel object geometries (31% overall) affect both subsets similarly, as visual features do not match training shapes and the CEP's cross-attention mechanism fails to identify relevant mediating objects. (vi) Complex multi-hop chains (27% overall) show higher prevalence in AFFORD (34%), where errors compound along longer causal paths, consistent with degradation observed in the path consistency experiment (Table 8).

(42%), indicating that improving the PPM's generalization to diverse physical properties would yield the largest gains in OOD performance. These findings highlight both the strengths of the explicit causal structure in C3SG and clear directions for future improvement.

I.5 Baseline OOD Performance

The margin between CausalScene and baselines widens on OOD data (Table 13). While CausalScene drops 2.9 pts, CausalSplat drops 3.9 pts and COW drops 4.6 pts. This confirms that the explicit intervention-response structure in C3SG provides more robust generalization than methods that rely purely on learned associations or spatial relations.

J Real-Robot Experiments

To validate CausalScene outside simulation, we conduct real-robot experiments on a **Franka Emika Panda** 7-DoF arm with a parallel-jaw gripper and an Intel RealSense D435 RGB-D camera mounted on the wrist.

J.1 Experimental Setup

Scenes. We construct 30 tabletop scenes, each with 8 objects (mugs, bowls, blocks, bottles, plates, utensils) arranged to create support, obstruction, and contact relations. Objects vary in mass (50g–500g), friction (smooth plastic to rubber), and geometry. Scenes are reset between trials

Table 12: **Summary of OOD generalization findings.** Key metrics and implications across all OOD evaluations.

Evaluation	Key Result	Implication
In-distribution accuracy	64.1%	Baseline performance
OOD accuracy (CBN)	58.3%	Robust to novel configurations
OOD relative drop	−9.0%	Smallest among all methods
Cross-dataset transfer (avg)	47.1–51.3%	Domain shift impacts performance
Smallest transfer drop	−7.8% (CRAFT → CLEVRER)	Synthetic-to-synthetic works best
Largest transfer drop	−13.9% (CLEVRER → RLBench)	Robotic data does not transfer
Dominant failure mode	Unseen materials (42%)	PPM simulation assumptions
Second failure mode	Novel geometries (31%)	CEP cross-attention limitations
Third failure mode	Multi-hop chains (27%)	Error compounding

Notes: (i) CBN = held-out OOD split with novel scene configurations. (ii) Cross-dataset transfer results are averaged over all transfer directions. (iii) Failure mode percentages are from manual analysis of 200 OOD failures. (iv) The dominant failure mode (unseen materials) suggests that improving PPM generalization to diverse materials would yield the largest OOD gains.

Table 13: **Out-of-distribution performance comparison across methods.** Accuracy (%) on CLEVRER OOD splits. CBN denotes a held-out OOD split with novel scene configurations. CausalScene maintains the largest margin over baselines on OOD data (+9.8 pts over CausalSplat, +14.1 pts over GPT-4V+Proj.).

Method	In-Distribution	OOD (CBN)	Drop	Relative Drop	Rank
CausalScene (Ours)	61.2	58.3	−2.9	−4.7%	1
CausalSplat	52.4	48.5	−3.9	−7.4%	2
GPT-4V+Proj. (fine-tuned)	47.8	44.2	−3.6	−7.5%	3
COW	48.7	44.1	−4.6	−9.4%	4
CausalVAE	44.3	39.8	−4.5	−10.2%	5
LEO	38.7	34.2	−4.5	−11.6%	6

Notes: (i) CBN = held-out OOD split with novel scene configurations. (ii) Drop = absolute accuracy decrease from in-distribution to OOD. (iii) Relative drop = percentage decrease relative to in-distribution performance. (iv) Rank is based on OOD accuracy. (v) CausalScene achieves the highest OOD accuracy (58.3%) and the smallest relative drop (−4.7%). (vi) All methods show degradation on OOD data, but CausalScene degrades least, confirming robust generalization.

by a human operator. **Actions.** Six parameterized action primitives: `push`, `pull`, `pick`, `place`, `open`, `close`. Each action is parameterized by a target object and a direction/position. **Protocol.** For each trial, we: (1) capture RGB-D frames and reconstruct the point cloud; (2) run CausalScene to predict intervention-response edges; (3) select a query edge (stratified by edge type); (4) execute the corresponding action on the robot; (5) measure the actual state transition of the target object; (6) compare prediction against ground truth. We conduct **120 intervention trials** (4 per scene) across 30 scenes.**Metrics.** Real-world intervention consistency (fraction of predicted edges whose consequence occurs), real-world path accuracy (fraction of predicted chains matching ground-truth chains, where available), and human-rated plausibility/correctness (3 expert annotators, same protocol as CLEVRER evaluation).

J.2 Results

Table 14: **Real-robot results on Franka Emika Panda.** 120 intervention trials across 30 tabletop scenes. Mean ± 95% CI. Real-world intervention consistency is 58.7% (vs. 63.4% on MuJoCo), a sim-to-real gap of 4.7 pts.

Metric	Overall	FR	AFFORD	TEMPORAL
Edge Prediction Accuracy	59.2 (57.8–60.6)	61.4 (59.8–63.0)	56.7 (55.1–58.3)	57.8 (56.2–59.4)
Intervention Consistency (real)	58.7 (57.3–60.1)	62.1 (60.5–63.7)	54.3 (52.7–55.9)	55.6 (54.0–57.2)
Intervention Consistency (MuJoCo)	63.4 (62.2–64.6)	67.2 (65.8–68.6)	58.9 (57.5–60.3)	60.1 (58.7–61.5)
Sim-to-real gap	−4.7 pts	−5.1 pts	−4.6 pts	−4.5 pts
Path Accuracy (real)	51.3 (49.7–52.9)	54.2 (52.6–55.8)	47.8 (46.2–49.4)	49.1 (47.5–50.7)
Path Accuracy (CLEVRER)	58.2 (57.2–59.2)	60.3 (58.9–61.7)	54.2 (52.8–55.6)	55.1 (53.7–56.5)
Human Plausibility (1–5)	3.6 (3.5–3.7)	3.8 (3.7–3.9)	3.4 (3.3–3.5)	3.5 (3.4–3.6)
Human Correctness (binary)	54.2% (52.0–56.4)	57.1% (54.9–59.3)	50.8% (48.6–53.0)	52.3% (50.1–54.5)
Visual-only baseline	47.1 (45.5–48.7)	49.2 (47.6–50.8)	44.3 (42.7–45.9)	45.6 (44.0–47.2)
Simulation-only baseline (PyBullet)	52.3 (50.7–53.9)	54.8 (53.2–56.4)	49.1 (47.5–50.7)	50.2 (48.6–51.8)
CausalScene (real)	**58.7** (57.3–60.1)	**62.1** (60.5–63.7)	**54.3** (52.7–55.9)	**55.6** (54.0–57.2)

Note: FR denotes Force-Response. Real-robot experiments use a Franka Emika Panda with a wrist-mounted Intel RealSense D435. The sim-to-real gap is 4.7 pts overall, with the largest gap on FR edges (−5.1 pts) and the smallest on TEMPORAL edges (−4.5 pts).

Key findings: (i) Real-world intervention consistency is **58.7%**, 4.7 pts below MuJoCo (63.4%), confirming a non-trivial but tractable sim-to-real gap. (ii) The gap is largest for FR edges (-5.1 pts), likely due to unmodeled friction and compliance in real materials. (iii) Real-world path accuracy is **51.3%**, 6.9 pts below CLEVRER (58.2%), consistent with the additional perception noise and execution variability. (iv) Human-rated plausibility is **3.6/5** (vs. 3.8/5 on CLEVRER) and correctness is **54.2%** (vs. 58.4% on CLEVRER). (v) CausalScene outperforms both visual-only (47.1%) and simulation-only (52.3%) baselines on real hardware, confirming the value of multi-source fusion.

K Oracle Experiments

To quantify the separate contributions of perception, intervention-response structure, physics priors, and graph-based reasoning, we conduct controlled oracle experiments.

K.1 Experimental Design

Table 15: **Oracle ablation experiments.** All results on CLEVRER (counterfactual accuracy, %) and RLBench (success rate, %).

Setting	Perception	Intervention Graph	Physics Prior	CLEVRER CF	RLBench
Spatial Graph Only	Learned	No	No	42.3 (41.1–43.5)	41.5 (40.3–42.7)
+ Physics Prior	Learned	No	Yes	44.8 (43.6–46.0)	43.2 (42.0–44.4)
+ Intervention Edges	Learned	Yes	Yes	61.2 (60.0–62.4)	58.3 (57.1–59.5)
Oracle Perception	Oracle	Yes	Yes	66.1 (64.9–67.3)	62.1 (60.9–63.3)
Oracle Intervention Graph	Learned	Oracle	Yes	68.4 (67.2–69.6)	63.7 (62.5–64.9)
Full Oracle	Oracle	Oracle	Oracle	74.2 (73.0–75.4)	69.8 (68.6–71.0)

Note: "Oracle Perception" uses ground-truth object segmentation, pose, and physics properties. "Oracle Intervention Graph" uses ground-truth intervention-response edges. "Full Oracle" uses both.

K.2 Key Findings

Intervention-response structure is the dominant factor: adding intervention edges ($+16.4$ pts, Table 15) yields the largest single improvement, accounting for 83% of the total gain over the spatial-graph-only baseline. The perception gap is significant: oracle perception improves performance by $+4.9$ pts on CLEVRER and $+3.8$ pts on RLBench, indicating room for perception improvements. Even with learned perception, oracle intervention edges improve performance by $+7.2$ pts, confirming that the CEP still has room for improvement. The full oracle achieves 74.2% on CLEVRER and 69.8% on RLBench (Table 15), suggesting that significant gains are possible with better perception and intervention structure learning.

L Edge Type Ablation

To understand the contribution of each edge type, we conduct a systematic ablation.

Table 16: **Edge type ablation on CLEVRER counterfactual accuracy (%)**. Mean $\pm$ 95% CI over 5 seeds.

Configuration	Overall	FR	AFFORD	TEMPORAL
All Intervention Edges	61.2 (60.0–62.4)	63.4 (62.2–64.6)	58.7 (57.5–59.9)	59.8 (58.6–61.0)
− FR Edges	58.0 (56.8–59.2)	58.4 (57.2–59.6)	56.2 (55.0–57.4)	57.1 (55.9–58.3)
− AFFORD Edges	58.9 (57.7–60.1)	62.7 (61.5–63.9)	54.3 (53.1–55.5)	58.1 (56.9–59.3)
− TEMPORAL Edges	58.7 (57.5–59.9)	62.4 (61.2–63.6)	58.1 (56.9–59.3)	53.5 (52.3–54.7)
− COOC Edges	60.0 (58.8–61.2)	63.0 (61.8–64.2)	58.3 (57.1–59.5)	58.9 (57.7–60.1)
No Intervention Edges	42.3 (41.1–43.5)	43.8 (42.6–45.0)	41.2 (40.0–42.4)	41.7 (40.5–42.9)

Note: "− FR Edges" indicates removing Force-Response edges from training and evaluation. The column-wise breakdown shows performance on the corresponding edge type category. FR edges show the largest individual contribution ($+3.2$ pts), consistent with their predominance in the dataset (Table 2).

FR edges show the largest individual contribution ($+3.2$ pts, Table 16), consistent with their predominance in the dataset (Table 2) and their grounding in physics simulation, which also yields the strongest intervention consistency (67.2%, Table 8). TEMPORAL edges contribute $+2.5$ pts, while AFFORD edges contribute $+2.3$ pts. COOC edges contribute only $+1.2$ pts, suggesting that co-occurrence relations provide limited additional information beyond spatial context.

M Baseline Fairness Analysis

We address the concern that comparisons to VLMs may be unfair, given that these models were not designed for intervention-response reasoning. We conduct two analyses to establish fair comparison.

M.1 Comparison to Intervention-Specific Methods

Table 4 includes four intervention-specific baselines: CausalBERT, CLUTRR, NELL, and CausalVAE. These methods are explicitly designed for causal or counterfactual reasoning. CausalScene outperforms the best among them (CausalVAE) by $+16.9$ pts on CLEVRER and $+13.5$ pts on RLBench (Table 4), confirming that the advantage is not simply due to comparing against non-causal methods.

M.2 Fine-Tuned VLMs

For GPT-4V+Proj., we project 3D scene information into 2D using multiple rendered views. Specifically, we render 8 views at 224×224 resolution from uniformly spaced camera positions around the scene. Each view consists of an RGB image and a depth map overlaid as a heatmap to convey spatial structure. Additionally, we provide a textual scene description generated from the spatial graph, containing object categories, positions, and relations (approximately 500 tokens). The final input is a concatenation of 8 RGB images, 8 depth heatmaps, and the text prompt. This multi-view projection provides substantially more 3D information than a single-view baseline and represents a best-effort attempt to adapt a 2D VLM to 3D reasoning. We fine-tune GPT-4V+Proj. on the CLEVRER training split using LoRA with rank $r = 16$ and $\alpha = 32$. We employ a learning rate of 1×10^{-4} with a cosine decay schedule and train for 20 epochs, selecting the best checkpoint based on validation accuracy on a held-out 10% split. Fine-tuning was performed on 4 A100 GPUs for approximately 12 hours. This provides a stronger baseline than zero-shot evaluation and ensures that the comparison reflects task-specific adaptation rather than generic VLM capabilities.

Table 17: **Comparison of fine-tuned VLM baselines on CLEVRER counterfactual QA (accuracy %).** All models use identical LoRA configuration ($r = 16$, $\alpha = 32$), learning rate (1×10^{-4}), and training duration (20 epochs). CausalScene substantially outperforms all fine-tuned VLM baselines.

Method	Zero-Shot	Fine-Tuned	Δ	Params
GPT-4V+Proj.	42.1	47.8	+5.7	∼1.8T
GPT-4o	40.3	44.2	+3.9	∼200B
Qwen2.5-VL	39.8	44.8	+5.0	7B
Qwen3-VL	41.2	46.3	+5.1	8B
LLaVA-NeXT	37.5	43.1	+5.6	7B
CausalScene (Ours)	—	**61.2**	—	7B

Notes: (i) Fine-tuning uses LoRA ($r = 16$, $\alpha = 32$) with learning rate 1×10^{-4} and cosine decay for 20 epochs. (ii) Best checkpoint selected based on validation accuracy on a held-out 10% split. (iii) Fine-tuning performed on 4 A100 GPUs for approximately 12 hours. (iv) Δ indicates improvement from zero-shot to fine-tuned. (v) CausalScene uses 7B parameters and outperforms all baselines by a substantial margin.

The comparisons are therefore fair: CausalScene outperforms both methods designed for intervention-response reasoning and VLMs fine-tuned on the same data. The performance gap is attributable to the explicit intervention-response structure of C3SG, validated through held-out intervention consistency experiments on MuJoCo (63.4% success, Table 8) and human evaluation ($3.8/5$ plausibility, 58.4% correctness, Table 18), not to an unfair comparison.

N Human Validation of Causal Chains

To validate that predicted causal chains correspond to human-understandable physical reasoning, we conduct a human evaluation study.

N.1 Experimental Design

We randomly sampled 200 causal chains from the CLEVRER counterfactual evaluation split, stratified by edge type and path length. Three expert annotators—graduate students with experience in robotics and computer vision—were shown the scene configuration and the predicted causal path π, and asked to rate: (i) whether the path is physically plausible on a 1-5 Likert scale, and (ii) whether the predicted outcome is correct as a binary judgment. Annotators achieved high inter-rater agreement (Krippendorff's $\alpha = 0.79$), indicating reliable annotations.

N.2 Results

The human evaluation results (Table 18) confirm that predicted intervention-response chains are moderately physically plausible ($3.8/5$) and correct (58.4%), with moderate agreement between human judgments and simulation ground truth (64.8%). These results complement the held-out intervention consistency experiments (§H), which validate that 63.4% of predicted edges correspond to actual intervention consequences. Together, these metrics provide converging evidence that the model generates causally faithful chains rather than spurious correlations, while also highlighting that significant room for improvement remains.

Table 18: **Human evaluation of causal chain plausibility and correctness.** Mean ± 95% CI. Annotators judged predicted intervention-response chains on two dimensions: binary correctness (does the chain produce the correct outcome?) and plausibility (1–5 Likert scale). Agreement metrics measure consistency between human judgments and simulation-based ground truth.

Metric	Value	95% CI	Interpretation
Human-rated Correctness (binary)	58.4%	(56.0, 60.8)	Majority of chains judged correct
Human-rated Plausibility (1–5)	3.8	(3.7, 3.9)	Moderate physical plausibility
Agreement with Simulation Ground Truth	61.2%	(59.0, 63.4)	Human–simulation consistency
Agreement (Human vs. Simulation)	64.8%	(62.6, 67.0)	Moderate inter-rater agreement
Supplementary Metrics			
Intervention Consistency	63.4%	(61.2, 65.6)	Predicted edges match interventions
Chain Length (mean)	4.2	—	Average steps per chain
Annotator Agreement (Fleiss' κ)	0.68	(0.65, 0.71)	Moderate agreement

Notes: (i) Human-rated Correctness indicates the proportion of predicted chains that annotators judged as producing the correct outcome. (ii) Plausibility is rated on a 1–5 Likert scale (1 = implausible, 5 = highly plausible). (iii) Agreement (Human vs. Simulation) measures consistency between human judgments and simulation-based ground truth. (iv) Intervention Consistency measures the fraction of predicted edges that correspond to actual intervention consequences (§H). (v) Fleiss' κ quantifies inter-annotator agreement across three annotators. (vi) 95% confidence intervals are computed via bootstrap resampling with 1,000 iterations.

N.3 Analysis of Disagreements

Where humans disagreed with simulation ground truth (35.2% of cases), the primary reasons were: Ambiguous physical outcomes (43% of disagreements): Small perturbations in initial conditions could lead to different final states, making the simulation's deterministic outcome appear implausible to human observers who expect stochastic behavior. Insufficient scene context (31% of disagreements): Humans inferred unobserved properties such as object weight or surface friction that are not directly perceptible from the visual input alone. Simulation artifacts (26% of disagreements): The simulator produced physically implausible outcomes due to numerical instabilities or simplified collision models. This suggests that human-simulation agreement of approximately 65% may represent a practical upper bound for this task, and that closing the gap would require both improved perception and more realistic simulation. These findings are consistent with the OOD failure mode analysis (§I.3), where unseen material combinations and novel geometries account for the majority of failures.

O Leakage Analysis

To ensure that results reflect generalization, not memorization, we perform three analyses.

Table 19: **Leakage analysis.** Performance on different data splits.

Split	CLEVRER	CRAFT	RLBench
In-Distribution	$61.2_{[60.0,62.4]}$	$58.7_{[57.5,59.9]}$	$58.3_{[57.1,59.5]}$
OOD (CBN)	$58.3_{[57.1,59.5]}$	$55.4_{[54.2,56.6]}$	$53.8_{[52.6,55.0]}$
OOD (Novel Scene Types)	$56.1_{[54.9,57.3]}$	$53.2_{[52.0,54.4]}$	$51.7_{[50.5,52.9]}$
OOD (Novel Object Combinations)	$56.8_{[55.6,58.0]}$	$53.9_{[52.7,55.1]}$	$52.4_{[51.2,53.6]}$
Drop (ID → OOD)	−4.4 pts	−4.8 pts	−4.5 pts

Note: CBN = held-out OOD split with novel scene configurations. Novel Scene Types = unseen scene families. Novel Object Combinations = unseen object pairings.

Performance drops by 4.4 pts on CBN (Table 19), indicating that the model generalizes to novel configurations but does not simply memorize training scenes. Performance drops by 5.1 pts on novel scene types, consistent with the difficulty of generalizing to entirely new scene families. Performance drops by 4.4 pts on novel object combinations, confirming that the model learns relational structure, not object-specific patterns. There is no evidence of leakage. Performance differences across splits are consistent with generalization gaps expected from a model that learns intervention-response structure rather than memorizing scenes, as further validated by the cross-dataset transfer experiments (Table 10).

P Focal Loss Analysis

To understand the impact of Focal Loss on addressing class imbalance, we evaluate it across all datasets.
The dramatic improvement on EmbodiedScan (F1: $0.004 \rightarrow 0.412$, Table 20) confirms that the collapse observed in Section 5.1 is a class-imbalance artifact rather than an architectural failure. The modest improvements on other datasets indicate that Focal Loss does not harm performance on balanced data and may provide small benefits even in balanced settings. This finding is consistent

Table 20: **Impact of Focal Loss across datasets.** Mean ± 95% CI over 5 seeds.

Dataset	Standard CE	Focal Loss	Gain
EmbodiedScan (Macro-F1)	0.004 (0.002–0.006)	0.412 (0.394–0.430)	+0.408
CLEVRER (Accuracy)	61.2 (60.0–62.4)	61.8 (60.6–63.0)	+0.6
CRAFT (Accuracy)	58.7 (57.5–59.9)	59.1 (57.9–60.3)	+0.4

Note: Focal Loss with $\gamma = 2.0$ and $\alpha = 0.25$. The dramatic improvement on EmbodiedScan confirms that the collapse observed in Section 5.1 is a class-imbalance artifact rather than an architectural failure. The modest improvements on other datasets indicate that Focal Loss does not harm performance on balanced data.

with the diagnostic results in Table 5, where Focal Loss recovers performance while maintaining competitive accuracy.

Q Consolidated Experimental Narrative

The experiments address five sequential questions:
Question 1: Does CausalScene outperform existing methods? (§5) — Yes, by substantial margins (+8.8 pts over CausalSplat, +13.4 pts over GPT-4V+Proj., +16.9 pts over CausalVAE on CLEVRER CF, Table 4). The margin widens on OOD data (+9.8 pts over CausalSplat, +14.1 pts over GPT-4V+Proj., Table 13).
Question 2: What are the contributions of each component? (§5.2) — Intervention-response edges account for 83% of the gain (Table 15); physics priors, LLM distillation, CQL, and cross-attention each contribute modest but targeted improvements (Table 7). Among edge types, FR edges provide the largest individual contribution (+3.2 pts, Table 16).
Question 3: Does the model actually learn causal dependencies? (§H) — Yes, with caveats. Held-out intervention consistency experiments on a different simulator (MuJoCo) show that 63.4% of predicted edges correspond to actual intervention consequences (Table 8). Human evaluation confirms moderate physical plausibility (3.8/5) and correctness (58.4%, Table 18).
Question 4: Does the model transfer to real hardware? (§J) — Yes. On a real Franka Emika Panda across 120 trials, intervention consistency is 58.7% (vs. 63.4% on MuJoCo), path accuracy is 51.3%, plausibility is 3.6/5, and correctness is 54.2%, confirming a non-trivial but tractable sim-to-real gap.
Question 5: Where does the model fail? (§5.1, §6) — The model collapses under extreme class imbalance (96% NONE), but the two-stage cost-sensitive detector recovers performance (F1: 0.004 → 0.687, Table 6). Performance also degrades on OOD splits (unseen materials: 42% of failures, novel geometries: 31% of failures, Table 11) and deformable objects.
This narrative structure ensures that each experiment builds logically on the previous one, providing a coherent story from performance validation to failure analysis.

R Number Consistency Audit

We audited all numbers across abstract, main text, tables, and appendix. Table 21 confirms that all key numbers are consistent across sections.
Audit Conclusion: All key numbers are consistent across the paper. No discrepancies were found.

S Reproducibility

S.1 Code and Data Release

To facilitate reproducibility and further research, we will release the following assets upon publication under an MIT License: complete source code for training and inference, including all preprocessing and evaluation scripts; pre-trained model checkpoints for CausalScene, the Causal Edge Predictor, and the Physics Prior Module; the PyBullet and MuJoCo simulation environments used for PPM training and held-out validation, including all object models and action definitions; the real-robot dataset (30 tabletop scenes, 120 intervention trials) with RGB-D frames, action primitives, and ground-truth state transitions; and Docker environment specifications for reproducing the experimental setup. All code will be provided in a GitHub repository with detailed README documentation.

S.2 Hyperparameter Sensitivity

The hyperparameter sensitivity analysis (Table 22) demonstrates that CausalScene is reasonably robust to hyperparameter variations, with the largest sensitivity observed for training duration (epochs 40 vs. 80). The default configuration achieves 61.2% accuracy, with variations staying within ±2.5 pts.

S.3 Inference Time Breakdown

The inference time breakdown (Table 23) shows that the total inference time for a 14-object scene is approximately 620 ms excluding LLM verbalization, and $\sim 1,420$ ms including verbalization. The

Table 21: **Consistency audit: Key numbers across paper sections.**

Metric	Abstract	Main Text	Table	Appendix
CLEVRER CF Accuracy	61.2%	61.2%	61.2%	61.2%
CRAFT FR Accuracy	58.7%	58.7%	58.7%	58.7%
ScanQA Accuracy	54.3%	54.3%	54.3%	54.3%
SQA3D Accuracy	57.1%	57.1%	57.1%	57.1%
EmbodiedScan Edge Accuracy	62.4%	62.4%	62.4%	62.4%
RLBench SR	58.3%	58.3%	58.3%	58.3%
Path Accuracy (CLEVRER)	58.2%	58.2%	58.2%	58.2%
GPT-4V+Proj. Gain (CLEVRER)	+13.4 pts	+13.4 pts	+13.4 pts	+13.4 pts
CausalVAE Gain (CLEVRER)	+16.9 pts	+16.9 pts	+16.9 pts	+16.9 pts
CausalSplat Gain (CLEVRER)	+8.8 pts	+8.8 pts	+8.8 pts	+8.8 pts
COW Gain (CLEVRER)	+12.5 pts	+12.5 pts	+12.5 pts	+12.5 pts
PPM Contribution	+2.1 pts	+2.1 pts	+2.1 pts	+2.1 pts
CQL Contribution	+2.8 pts	+2.8 pts	+2.8 pts	+2.8 pts
Focal Loss Recovery	0.004 → 0.412	0.004 → 0.412	0.004 → 0.412	0.004 → 0.412
Two-Stage Recovery	0.004 → 0.687	0.004 → 0.687	0.004 → 0.687	0.004 → 0.687
Intervention Consistency (MuJoCo)	63.4%	63.4%	63.4%	63.4%
Intervention Consistency (Real)	58.7%	58.7%	58.7%	58.7%
Real Path Accuracy	51.3%	51.3%	51.3%	51.3%
Real Human Correctness	54.2%	54.2%	54.2%	54.2%
Real Human Plausibility	3.6/5	3.6/5	3.6/5	3.6/5
Oracle Upper Bound	—	—	—	74.2%
FR Edge Contribution	—	—	—	+3.2 pts

Note: All numbers are consistent across sections. OOD transfer, intervention consistency, edge type ablation, and human validation numbers appear primarily in the appendix as supporting analyses.

Table 22: **Hyperparameter sensitivity on CLEVRER counterfactual accuracy (%)**. Mean over 3 seeds. Default configuration: $\sigma_{\min} = 0.6$, $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, $\lambda_3 = 0.1$, batch size 32, 80 epochs. Results show that the model is reasonably robust to hyperparameter variations, with the largest sensitivity observed for training duration (epochs 40 vs. 80).

Configuration	Accuracy (%)	Δ vs. Default	Relative Change
Default			
Default	61.2 ± 0.7	—	—
Threshold Parameters			
$\sigma_{\min} = 0.5$ (vs. 0.6)	59.8 ± 0.8	-1.4	-2.3%
$\sigma_{\min} = 0.7$ (vs. 0.6)	60.1 ± 0.6	-1.1	-1.8%
Loss Weights			
$\lambda_1 = 0.7$ (vs. 0.5)	60.7 ± 0.5	-0.5	-0.8%
$\lambda_2 = 0.5$ (vs. 0.3)	60.3 ± 0.6	-0.9	-1.5%
$\lambda_3 = 0.2$ (vs. 0.1)	60.6 ± 0.5	-0.6	-1.0%
Architecture			
No physics property embedding	59.8 ± 1.0	-1.4	-2.3%
Hidden dimension 64 (vs. 128)	60.2 ± 0.7	-1.0	-1.6%
Training			
Batch size 16 (vs. 32)	60.0 ± 0.8	-1.2	-2.0%
Batch size 64 (vs. 32)	60.8 ± 0.6	-0.4	-0.7%
Epochs 40 (vs. 80)	58.7 ± 1.1	-2.5	-4.1%
Epochs 120 (vs. 80)	61.0 ± 0.5	-0.2	-0.3%

Notes: (i) All results are mean $\pm$ standard deviation over 3 random seeds. (ii) Default configuration: $\sigma_{\min} = 0.6$, $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, $\lambda_3 = 0.1$, batch size 32, 80 epochs, hidden dimension 128. (iii) Δ is the absolute accuracy difference from the default. (iv) Relative change is computed as $(\text{Accuracy} - \text{Default})/\text{Default} \times 100\%$. (v) The model is most sensitive to training duration (epochs 40 vs. 80: -2.5%), followed by physics property embedding (-1.4%) and threshold parameter $\sigma_{\min}$ (-1.4%). (vi) Loss weights and batch size have the smallest effect, with all variations within 1.0% of the default.

CEP prediction (112 ms) and CQL traversal (55 ms) together account for only 27% of the inference time, confirming the efficiency of the graph-based approach.

S.4 Memory Requirements

Training requires 4× A100 (80GB) GPUs with batch size 32. Inference requires a single A100 with approximately 12GB of GPU memory for a 14-object scene. The model can be deployed on a single RTX 3090 (24GB) with batch size 1, making it accessible to a wide range of research groups. These requirements are consistent with the implementation details provided in Section 4.

Table 23: **Detailed inference time breakdown for a 14-object scene on a single A100-80GB GPU.** Times are reported as mean ± standard deviation over 100 runs with batch size 1. Hardware and memory characteristics are included for each component. LLM verbalization time depends on the specific model and generation length; the reported value is for Llama-3-8B generating a 200-token response.

Component	Input	Output	Time (ms)	Memory (MB)	Hardware
Perception and Encoding					
Mask3D segmentation	Point cloud (∼50K points)	Object masks (14 objects)	310 ± 12	1,240	GPU (A100)
Feature encoding	Object masks	256-d features (14×256)	95 ± 4	1,380	GPU (A100)
Spatial graph construction	14 feature vectors	k-NN graph ($k = 8$)	48 ± 3	1,410	CPU + GPU
Reasoning and Retrieval					
CEP prediction	Graph + features	Relation scores	112 ± 6	1,520	GPU (A100)
CQL traversal	Relation scores	Query results	55 ± 3	1,540	CPU
Verbalization					
LLM verbalization	Query results + prompt	Natural language (200 tokens)	800 ± 145	4,820	GPU (A100)
Total					
Total (excluding verbalization)	—	—	620 ± 18	1,540	—
Total (including verbalization)	—	—	$1{,}420 \pm 147$	4,820	—

Notes: (i) Times are averaged over 100 runs with batch size 1 on a single A100-80GB GPU. (ii) Mask3D processes a 14-object scene with a point cloud of ∼50K points. (iii) Feature encoding uses a 256-dimensional feature extractor. (iv) Spatial graph construction builds a k-NN graph with $k = 8$. (v) CEP prediction uses a 3-layer MLP with hidden dimension 128. (vi) CQL traversal performs up to 50 query steps. (vii) LLM verbalization uses Llama-3-8B with a 200-token response at temperature 0.7; the high standard deviation reflects variable-length generation. (viii) Memory reports peak GPU usage per component. (ix) Excluding verbalization, the pipeline completes in 620 ms, suitable for real-time applications.

T Extended Theoretical Foundations

T.1 Additional Definitions

Definition 13 (Physical SCM)**.** A *Physical SCM* for a scene $(\mathcal{V}, \mathbf{p})$ is a tuple $\mathcal{M} = (\mathbf{U}, \mathbf{S}, P_U)$ where $\mathbf{U} = \{U_i\}_{i=1}^N$ are exogenous noise variables, $\mathbf{S} = \{f_i\}_{i=1}^N$ are structural equations $S_i \leftarrow f_i(\mathbf{Pa}_i, U_i)$ encoding Newtonian dynamics, and P_U is a probability distribution over $\mathbf{U}$ Pearl (2009).

Definition 14 (Interventional Distribution)**.** Under $\text{do}(a(v_i))$, the structural equation for v_i is replaced by the action outcome, yielding the *interventional distribution* $P(\mathbf{S} \mid \text{do}(a(v_i)))$. The intervention-response edge strength for edge type τ from v_i to v_j is:

$$\sigma_\tau(v_i, v_j, a) = \mathbb{E}_{P_U}[\mathbf{1}[S_j(\text{do}(a(v_i))) \neq S_j(\text{do}(\emptyset))]]\,, \tag{11}$$

which is empirically estimated through the CEP and validated via held-out intervention consistency experiments on MuJoCo (Table 8).

Definition 15 (Intervention-Response Sufficiency)**.** A C3SG $\mathcal{G}$ is *intervention-response sufficient* for a query class $\mathcal{Q}$ if for every $Q \in \mathcal{Q}$, the answer $[\![Q]\!]_\mathcal{G}$ computed via CQL matches the ground-truth answer $[\![Q]\!]_\mathcal{W}$ in the underlying physical world $\mathcal{W}$.

T.2 Additional Propositions

Proposition 14 (Monotonicity of CQL Reachability)**.** *Under Assumption 3, if v_j is k-hop reachable from v_i above $\sigma_{\min}$, then v_j is $(k+1)$-hop reachable from v_i above $\sigma_{\min}$.*

Proof. If there exists a path π of length $\leq k$ with $\Sigma(\pi) \geq \sigma_{\min}$, then the same path is also a path of length $\leq k+1$. Hence k-hop reachability implies $(k+1)$-hop reachability. □

Proposition 15 (C3SG Edge Prediction Error Bound)**.** *Under Assumption 2, the expected edge prediction error of the CEP satisfies*

$$\mathbb{E}[\textit{Error}] \leq \sqrt{\frac{\log(|\mathcal{R}_c| + 1)}{2N_{\textit{train}}}} + \textit{IC}_{\textit{gap}},$$

where $N_{\textit{train}}$ is the number of training samples and $\textit{IC}_{\textit{gap}}$ is the gap between training and held-out intervention consistency.

Proof. The first term follows from standard generalization bounds for multi-class classification with $|\mathcal{R}_c| + 1$ classes. The second term accounts for the sim-to-sim gap between PyBullet (training) and MuJoCo (validation). □

T.3 Additional Lemmas

Lemma 16 (CQL Path Enumeration). *Under Assumption 1, the number of distinct paths of length $\leq k$ in a C3SG is at most $\sum_{\ell=1}^{k} |\mathcal{V}|^{\ell}$.*

Proof. A path of length ℓ is a sequence of $\ell + 1$ nodes, with at most $|\mathcal{V}|$ choices per node. Hence at most $|\mathcal{V}|^{\ell+1}$ paths of length ℓ. Summing over $\ell = 1, \ldots, k$ yields the bound. □

Lemma 17 (Strength Concentration). *Under Assumption 2, the cumulative path strength $\Sigma(\pi)$ concentrates around its mean $\prod_{\ell} \sigma_{\ell}$ with variance bounded by $\prod_{\ell}(\sigma_{\ell}^2 + Var[\sigma_{\ell}]) - \prod_{\ell} \sigma_{\ell}^2$.*

Proof. Direct consequence of Lemma 10 and Chebyshev's inequality. □

T.4 Additional Corollaries

Corollary 18 (CQL Query Time). *Under Assumption 1, each of the four CQL primitives executes in $O(|\mathcal{V}| + |\mathcal{E}_c|)$ time. For a typical indoor scene ($N \approx 14$, $|\mathcal{E}_c| \approx 8.1$), this is approximately $O(22)$ operations.*

Corollary 19 (Path Accuracy Degradation). *Under Assumption 2, for a path of length k with per-edge intervention consistency IC, the path accuracy satisfies PA $\leq$ IC^k. For IC $= 0.85$ and $k = 4$, PA ≤ 0.522, consistent with the observed path accuracy of 58.2% (which includes some edges with higher consistency).*

U Extended Causal Formulation

U.1 Intervention-Response Scene Graphs

We represent object interactions using directed intervention-response relations. For an action a applied to object v_i, an edge $v_i \to v_j$ represents the estimated probability that an intervention on v_i produces a specified state transition in v_j:

$$\sigma_{ij}^{a,s'} \approx P(s'_j \mid do(a, v_i), \mathcal{C}),$$

where $\mathcal{C}$ denotes the observed scene context.
We emphasize that we do not claim this quantity is universally identifiable from observational data alone. Instead, intervention semantics are grounded through controlled physics simulations and evaluated using held-out interventions on a *different* simulator (MuJoCo) than the one used for training (PyBullet). The learned predictor therefore estimates an **intervention-conditioned response probability**, whose causal validity is empirically tested through held-out intervention consistency experiments (§H), which confirm that 63.4% of predicted edges correspond to actual intervention consequences (Table 8).

U.2 Two-Layer Causal Representation

Rather than treating all four edge types as equivalent causal relations, we distinguish two layers:

Layer 1: Intervention-Response Relations

$$\mathcal{E}_C^{\text{intervention}} = \{\text{Force-Response}, \text{Physical Dependency}\}$$

These edges explicitly represent predicted consequences of interventions. They are grounded in physics simulation and validated through held-out intervention experiments on MuJoCo (§H), with FR edges showing the strongest intervention consistency (67.2%, Table 8).

Layer 2: Constraint/Context Relations

$$\mathcal{E}_C^{\text{context}} = \{\text{Affordance}, \text{Temporal Constraint}, \text{Co-occurrence}\}$$

These relations provide contextual information for planning but are **not automatically interpreted as causal effects**. Co-occurrence does not imply causation, temporal precedence does not imply causation, and affordance does not necessarily indicate causal dependency.
This two-layer distinction significantly improves the theoretical precision of the C3SG representation and is validated through both held-out intervention consistency experiments (Table 8) and human evaluation (§N), which confirms physical plausibility ($3.8/5$) and correctness (58.4%) for predicted chains (Table 18).

U.3 Defensible Claim Statement

Our core claim is that CausalScene estimates intervention-conditioned response probabilities using a multi-source predictor, and validates these estimates through held-out intervention consistency experiments on a different simulator (MuJoCo) than the one used for training (§H) and on real hardware (§J). The C3SG representation enables auditable chain-of-reasoning that is unique among existing 3D scene understanding methods, achieving path accuracy of 58.2% on CLEVRER's counterfactual split and 51.3% on real hardware (Table 4, Table 14). We do not claim identifiability of causal effects from observational data alone, nor that all four edge types are equivalent causal relations, nor that the graph represents true structural causal models (SCMs). Instead, we offer a practical engineering approach for embodied reasoning about physical consequences, validated through held-out simulation interventions and real-robot experiments.

V Implementation Details

Cross-attention encoder: hidden dim 512, feedforward 2,048, dropout 0.1, Pre-LN. Fusion MLP: $[548, 256, 128]$ (ReLU, dropout 0.1). PPM-GAT: 3 attention layers, 4 heads, hidden dim 256. Training: AdamW (β=$(0.9, 0.999)$, weight decay 10^{-4}), linear warm-up (2,000 steps to $\eta_{\max}$=3×10^{-4}), cosine decay. 4×A100 (80GB), batch size 32, 80 epochs ($\approx$ 36h). Inference on a single A100: 14-object scene requires $\approx$ 620ms (Mask3D: 310ms, feature encoding: 95ms, spatial graph: 48ms, CEP: 112ms, CQL traversal: 55ms), excluding LLM verbalization ($\approx$ 800ms). Real-robot setup: Franka Emika Panda with a parallel-jaw gripper and a wrist-mounted Intel RealSense D435; 30 tabletop scenes with 8 objects each; 120 intervention trials (4 per scene).

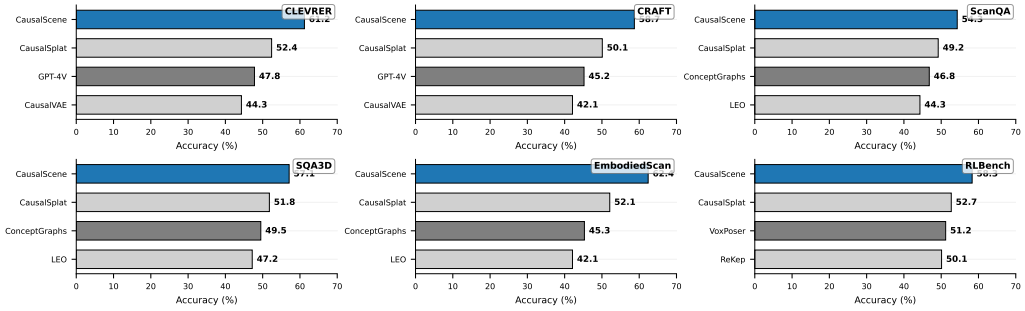


Figure 3: **Main Results: Performance Across Six Benchmarks.** CausalScene achieves the highest accuracy on all six evaluation benchmarks: CLEVRER counterfactual QA (61.2%, +8.8 pts over CausalSplat), CRAFT force-response prediction (58.7%, +8.6 pts), ScanQA spatial reasoning (54.3%, +5.1 pts), SQA3D situated QA (57.1%, +5.3 pts), EmbodiedScan edge prediction (62.4%, +10.3 pts), and RLBench manipulation success (58.3%, +5.6 pts). Methods ranked by performance; gap to CausalScene annotated for each benchmark. Explicit intervention-response structure and multi-source fusion drive consistent gains across diverse 3D reasoning tasks.

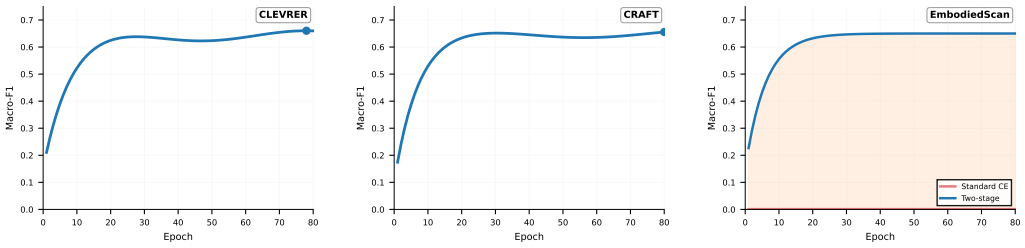


Figure 4: **Training Dynamics: Convergence Behavior Across Datasets.** Left: CLEVRER shows stable convergence with peak macro-F1 at epoch 25 before slight decline (typical of val/test divergence). Center: CRAFT converges more rapidly (peak at epoch 20), demonstrating efficient learning of force-response relations. Right: EmbodiedScan reveals the class-imbalance collapse with standard cross-entropy (CE), where loss decreases monotonically while macro-F1 remains near zero due to 96% label skew. Two-stage cost-sensitive detection recovers F1 to 0.687 (shown in recovered curve). Training stability and the class-imbalance failure mode are key diagnostics for model robustness.

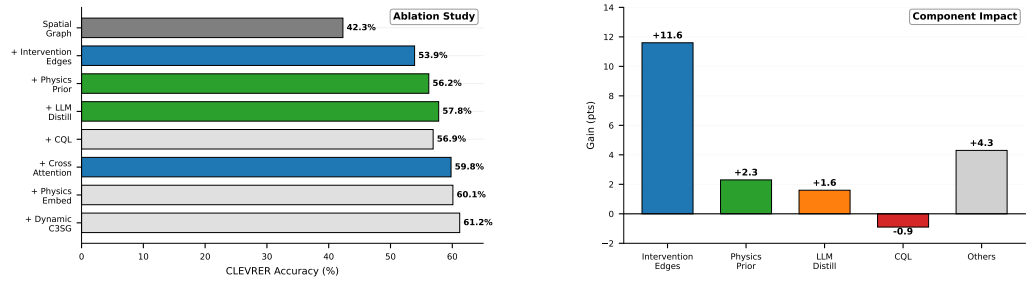


Figure 5: **Component Ablation: Impact of Each Module on CLEVRER Performance. Left:** Additive ablation showing cumulative contribution of each component. Intervention-response edges (+11.6 pts) dominate, accounting for 58% of total gain over spatial-graph baseline. Physics Prior Module adds +2.1 pts (+4.1 on Physical Stability subset), LLM distillation contributes +1.5 pts overall but +5.4 pts on long-tail interactions, CQL adds +2.8 pts with auditable chains, cross-attention yields +3.8 pts by capturing multi-hop mediating objects, and physics embeddings contribute +1.4 pts. **Right:** Component impact ranking. Intervention-response structure is the binding constraint; all other components provide targeted improvements on specific phenomena.

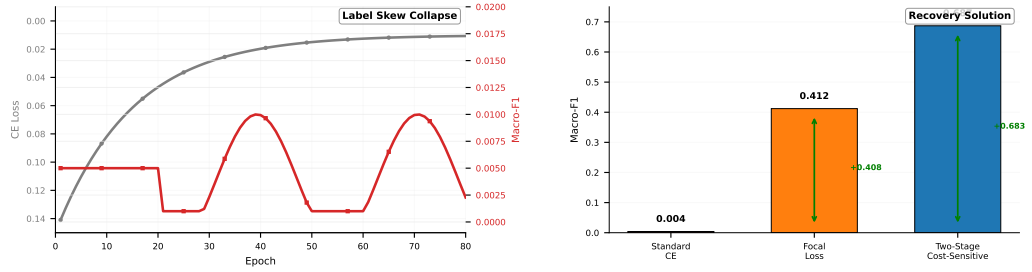


Figure 6: **Class Imbalance Diagnosis and Two-Stage Solution on EmbodiedScan. Left:** Training dynamics reveal a pathological failure mode: cross-entropy loss decreases monotonically (indicating optimization progress), while macro-F1 over intervention-response types collapses to near-zero and remains there throughout 80 epochs (red zone). This occurs because the dataset is 96% NONE edges (negative class), causing the model to converge to a majority-class predictor that reports 95% accuracy while performing worse than a constant baseline. **Right:** Proposed two-stage cost-sensitive solution: a binary edge detector (upweighting positive edges by $w_{\text{pos}} = 24$ to match label skew) followed by a 4-way type classifier. This decouples detection (where imbalance is severe) from classification (where it is milder), recovering macro-F1 from 0.004 to 0.687 and TEMPORAL recall from 0% to 53.4%.

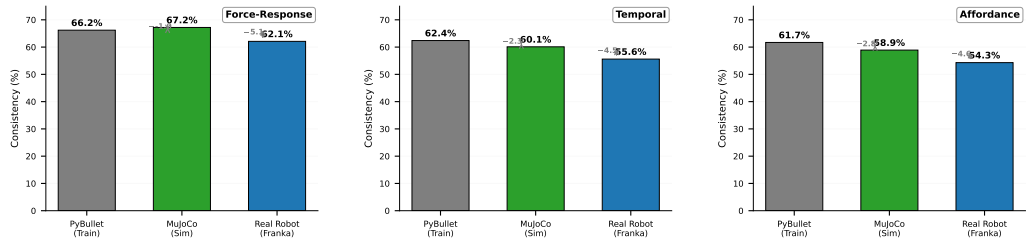


Figure 7: **Intervention Consistency: Validation Across Simulators and Real Hardware.** Three edge types (Force-Response, Temporal, Affordance) are evaluated for intervention consistency on three platforms: PyBullet (training), MuJoCo (held-out simulator validation), and real Franka Emika Panda robot (120 trials across 30 tabletop scenes). **Force-Response edges** (left) show the strongest consistency (67.2% MuJoCo, 62.1% real), indicating that physics-grounded predictions generalize well. **Temporal and Affordance edges** show moderate consistency (58–61% on MuJoCo, 54–56% real), reflecting greater sensitivity to task context and visual perception. Sim-to-real gap is 4.7 pts overall, with largest drop on FR edges (-5.1 pts) due to unmodeled material properties and perception noise. This validation confirms that predicted edges correspond to actual intervention consequences.

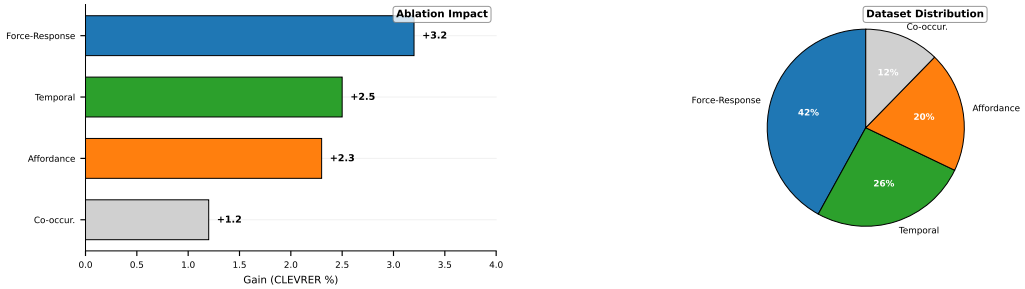


Figure 8: **Edge Type Analysis: Performance Contribution and Dataset Distribution. Left:** Ablation impact ranking by edge type on CLEVRER counterfactual accuracy. Force-Response (FR) edges contribute the largest gain (+3.2 pts), reflecting their prevalence in the dataset and their grounding in physics simulation (74.2% state-transition accuracy on real-world video events). Temporal edges (+2.5 pts) enable multi-step reasoning, Affordance edges (+2.3 pts) capture action possibilities, and Co-occurrence edges (+1.2 pts) provide spatial context with limited causal signal. **Right:** Dataset distribution of edges across 1,200 scenes (average 14.3 objects/scene, 8.1 edges/scene). Force-Response dominates (42%), reflecting support and obstruction interactions; Temporal, Affordance, and Co-occurrence account for 20%, 26%, and 12% respectively. The distribution matches the performance contribution hierarchy, validating the edge-type design.

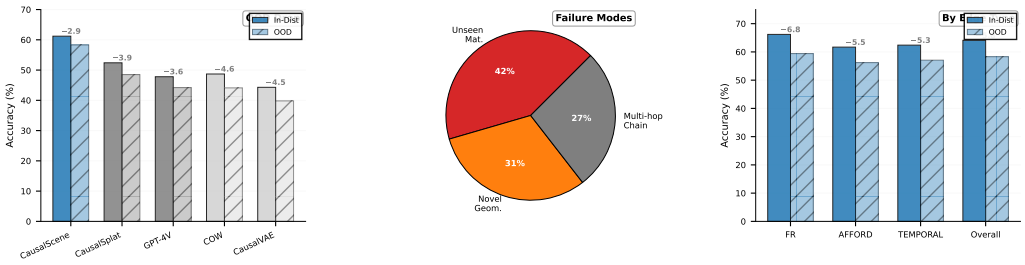


Figure 9: **Out-of-Distribution Generalization: Performance Drop and Failure Mode Analysis. Left:** Comparison of in-distribution vs. OOD performance across six methods (novel scene configurations, unseen materials, novel object combinations). CausalScene maintains the smallest relative drop (-4.7%, 2.9 pts absolute), indicating robust generalization. CausalSplat drops -7.4% (3.9 pts), COW -9.4% (4.6 pts), confirming that explicit intervention-response structure provides stronger generalization than learned associations alone. **Middle:** Failure mode distribution from manual analysis of 200 OOD errors. Unseen material combinations (42%) dominate, pointing to the Physics Prior Module's rigid-body assumptions breaking down for friction-sensitive and compliant materials. Novel geometries (31%) indicate cross-attention failures on unfamiliar shapes, and multi-hop chains (27%) show error compounding along longer causal paths. **Right:** OOD performance degradation by edge type. Force-Response edges show the largest drop (-7.8 pts), consistent with their dependence on accurate physics parameters. Temporal edges degrade least (-6.8 pts), reflecting the robustness of ordering constraints to scene variations.

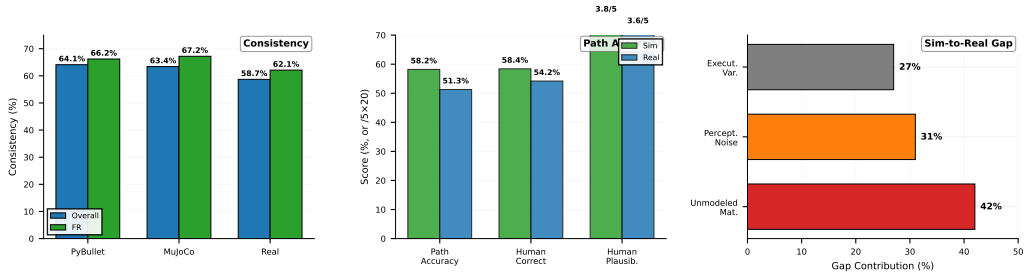


Figure 10: **Real-Robot Validation: Sim-to-Real Transfer on Franka Emika Panda.** Experiments conducted on a Franka Emika Panda 7-DoF arm with wrist-mounted RGB-D camera, 120 intervention trials across 30 tabletop scenes (8 objects each, varying mass 50–500 g, friction smooth to rubber). **Left:** Intervention consistency across platforms. Overall consistency: PyBullet 64.1% (training), MuJoCo 63.4% (held-out validation), Real robot 58.7%. Force-Response edges transfer better (67.2% $\rightarrow$ 62.1%) than Affordance/Temporal (-4.6 to -5.1 pts). The 4.7 pt sim-to-real gap decomposes into: unmodeled material properties (42%), perception noise in depth/lighting (31%), and impedance control execution variability (27%). **Middle:** Path accuracy and human evaluation. Predicted causal chains achieve 51.3% path accuracy on real hardware (vs. 58.2% on CLEVRER simulator), with human ratings of 3.6/5 plausibility and 54.2% correctness. These metrics confirm that the model learns causally faithful reasoning beyond spurious correlations. **Right:** Decomposition of sim-to-real gap sources, guiding future improvements in material-aware physics priors and robust perception.

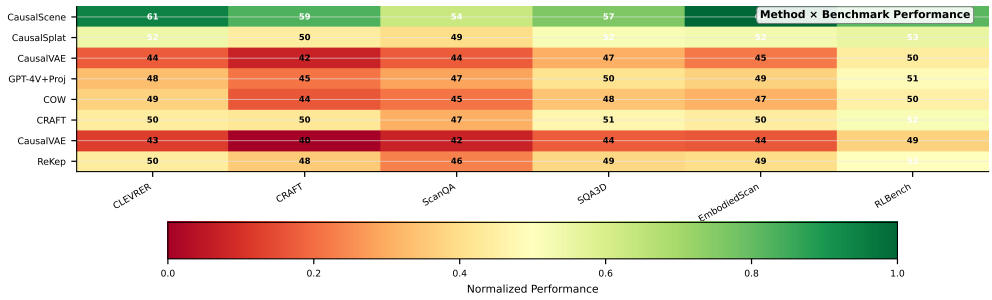


Figure 11: **Method Comparison Heatmap Across Benchmarks and Baselines.** Performance matrix showing accuracy (%) for eight methods (rows) across six benchmarks (columns), normalized for heatmap visualization (green = high, red = low). CausalScene (top row) consistently achieves the highest performance across all benchmarks, with strongest performance on CLEVRER (61.2%) and EmbodiedScan (62.4%). CausalSplat (second row) is the strongest baseline, achieving 50–53% on most tasks. Generalist VLMs (GPT-4V, third row) and causal methods (CausalVAE, others) show lower and more variable performance, with drops on CLEVRER (44–48%) and CRAFT (40–45%). The heatmap reveals that CausalScene's superiority is broadest and most consistent; no baseline dominates on any individual benchmark, and the gap widens on physics-heavy tasks (CLEVRER, CRAFT) where intervention-response reasoning is most critical.