

CauSight: Learning to Supersense for Visual Causal Discovery

Yize Zhang^{1,2,3*} Meiqi Chen^{4*} Sirui Chen^{1,5*} Bo Peng^{1,2,3*}

Yanxi Zhang^{1,4} Tianyu Li² Chaochao Lu^{1†}

¹Shanghai AI Laboratory ²Shanghai Innovation Institute

³Shanghai Jiao Tong University ⁴Peking University ⁵Tongji University

{zhangyize, luchaochao}@pjlab.org.cn

Abstract

Causal thinking enables humans to understand not just what is seen, but why it happens. To bring this ability to modern AI systems, we introduce the broader research problem of **visual causal discovery**. In this work, we operationalize it as *visual causal-dependency discovery*: inferring visually verifiable direct dependencies among entities in a single static image under an image-supported removal counterfactual. We construct the Visual Causal Graph dataset (VCG-32K), a large-scale collection of over 32,000 images annotated with entity-level causal graphs, and develop CauSight, a causally aware vision-language model for this task. Our training recipe combines data curation from VCG-32K, Tree-of-Causal-Thought (ToCT) reasoning synthesis, and reinforcement learning with a graph-structured causal reward. Experiments show that CauSight outperforms GPT-4.1, achieving over a threefold performance boost (21% absolute gain). Our code is available at <https://github.com/OpenCausaLab/CauSight>.

1 Introduction

Causal thinking is regarded as a hallmark of human intelligence (Penn and Povinelli, 2007; Udbhav and Rachna, 2017; Schulz and Bonawitz, 2007). Emulating this ability in vision-language models (VLMs) (Bordes et al., 2024; Li et al., 2025; Awadalla et al., 2023; Zhang et al., 2023) is essential for moving beyond recognition toward genuine visual reasoning (Ke et al., 2021; Krishna et al., 2016; Zellers et al., 2018; Nair et al., 2019). As shown in Figure 1, the underlying causal graph offers a mechanistic explanation for an embodied agent retrieving a laptop: the agent must first stabilize the glass, since removing the laptop directly would collapse the books and drop the glass. We

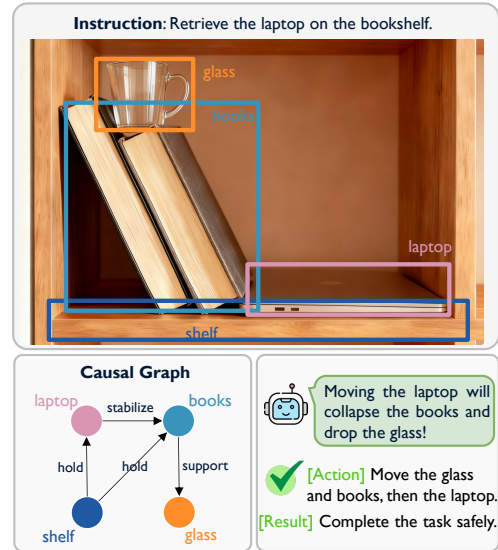


Figure 1: A causal graph captures causal mechanisms between entities. Genuine reasoning and safe deployment require VLMs to discover causal relations between entities.

view causal graphs as auditable intermediate representations: models need not always expose a graph at deployment time, but graph reconstruction provides a concrete way to measure whether visual interactions are understood mechanistically.

To discover causal graphs in real-world scenarios, primary steps involve identifying causal entities and inferring pairwise causal relations. Following the notion of “causal dispositions” in vision-language contexts (Mumford and Anjum, 2011; Lopez-Paz et al., 2016; Chen et al., 2024), we define a causal relation as one in which an entity has the inherent capacity to affect another’s state. This relation can be tested counterfactually: if the cause were absent, the effect would not maintain its current state (Frappier, 2018; Pearl, 2013). We thus position *visual causal discovery* as grounding visual concepts in language through explicit causal reasoning, and call this capability *supersensing*: perceiving not only visible entities and relations, but also their implied state-changing dependencies.

*Equal contribution.

†Corresponding author.

Our formulation is deliberately scoped to causal dependencies in a single image: a valid edge requires visible direct contact or interaction and a removal counterfactual supported by image evidence. This operational setting should not be conflated with classical causal discovery: we do not recover a structural causal model, execute interventions, or estimate interventional distributions from observational data.

However, even the most advanced VLMs currently struggle with the fundamental task of visual causal discovery (Chen et al., 2024). It requires the model not only to identify entities and their attributes, but also to infer causal relationships via counterfactual reasoning, while eliminating confounders. Our experiments show that the average recall rate of causal graphs for open-source models like Qwen2.5-VL series (Bai et al., 2025) is only **10.0%**, even proprietary models such as OpenAI o3 (openai, 2025b) and GPT-4.1 (Achiam et al., 2023) achieve only **7.3%** and **10.2%**. Such performance hinders the practical use of causal graphs in applications such as robotic and autonomous driving that require safe and reliable decision-making (Wang et al., 2023; Li et al., 2023).

To address these bottlenecks, we introduce the Visual Causal Graph dataset (VCG-32K), a large-scale collection of over 32,000 images built upon 11,428 samples from MS-COCO (Lin et al., 2014) and 20,828 from Objects365 (Shao et al., 2019). To the best of our knowledge, VCG-32K is the first benchmark at this scale to combine natural images, instance-level spatial grounding, directed removal-counterfactual dependencies, and graph-level annotations. Guided by the definition of visual causality, we employ a systematic annotation process to ensure consistency and reliability, enabling rigorous benchmarking on *visual causal discovery*.

We further develop CauSight, a VLM for discovering visually verifiable causal graphs across diverse scenarios via causally aware reasoning. Training proceeds in two stages. First, we curate data from VCG-32K and synthesize visually grounded reasoning trajectories with Tree-of-Causal-Thought (ToCT): a teacher iteratively performs *region selection*, *entity recognition*, and *causality orientation*, while Monte Carlo Tree Search (MCTS) (Browne et al., 2012) explores candidate trajectories and propagates value estimates (Zhang et al., 2025; Yao et al., 2023; Zhang et al., 2024). Rather than providing labels, the teacher proposes reasoning paths, which

are scored against human-annotated causal graphs, structurally filtered, and used as SFT demonstrations to initialize the policy model¹. Second, we refine the policy model with GRPO (Shao et al., 2024) using a graph-structured causal reward promoting accurate discovery and faithful graph generation. CauSight achieves **27.4%** gain ($\approx 8.2\times$) over its base model and **21.0%** ($\approx 3.1\times$) over GPT-4.1, and transfers from the MS-COCO-derived training set to the separately sourced Objects365-derived test set.

The main contributions of our work are:

- We formulate the task of *visual causal discovery* and introduce the VCG-32K dataset, a large-scale collection of over 32,000 images annotated with entity-level causal graphs to facilitate research on this problem.
- We introduce CauSight, which was trained using the Tree-of-Causal-Thought approach for synthesizing high-quality reasoning trajectories, and a reinforcement learning framework optimized by a graph-structured causal reward to enhance visual causal discovery.
- CauSight shows superior performance on visual causal discovery, with marked improvements over baseline models and promising cross-dataset transfer. Further analysis provides valuable insights for future research.

2 Task Formulation: *Visual Causal-Dependency Discovery*

We operationalize visual causal discovery as *visual causal-dependency discovery*: constructing an image-supported dependency graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ from a single image. The graph consists of basic entities $\mathcal{V}$ and directed dependencies $\mathcal{E}$ that satisfy the annotation criterion below. This formulation captures a restricted, visually auditable form of causal dependence rather than identifying an underlying structural causal model. In our VLM setting, the input is an image with a textual instruction, and the output is a structured graph containing entity boxes and directed causal edges. During training, ToCT exposes intermediate region-selection, entity-recognition, and causality-orientation steps; during evaluation, only the final generated graph is parsed and scored.

Entities. Each graph $\mathcal{G}$ contains N entities $\mathcal{V} =$

¹We use *policy model* to denote the model being trained; Qwen2.5-VL-7B is adopted as the base model for its favorable balance between capability and training cost.

$\{\mathbf{v}_i\}_{i=1}^N$, with $\mathbf{v}_i = (c, b)$ where c is the category label and $b \in \mathbb{R}^4$ is the bounding box.

Causal Relationships. Edges $\mathcal{E} = \{\mathbf{e}_{ij}\}_{i \neq j}$ are directed, predicate-labeled triples $\mathbf{e}_{ij} = (\mathbf{v}_i, \kappa, \mathbf{v}_j)$, where $\kappa \in \mathcal{K}$, e.g., “support”. Following causal dispositions (Mumford and Anjum, 2011; Lopez-Paz et al., 2016; Chen et al., 2024), an edge $i \rightarrow j$ exists when removing $\mathbf{v}_i$ changes the state of $\mathbf{v}_j$:

$$\begin{aligned} \mathbf{e}_{ij} \in \mathcal{E} \iff & \text{Interact}(\mathbf{v}_i, \mathbf{v}_j \mid I) = 1 \wedge \\ & \text{Maintain}(\mathbf{v}_i, \mathbf{v}_j \mid I) = 1 \wedge \\ & \text{Change}(\mathbf{v}_j; \text{remove}(\mathbf{v}_i), I) = 1. \end{aligned} \quad (1)$$

Here I is the observed image. The three terms indicate that $\mathbf{v}_i$ and $\mathbf{v}_j$ visibly interact, $\mathbf{v}_i$ maintains the observed state of $\mathbf{v}_j$, and an annotator judges from the image that hypothetically removing $\mathbf{v}_i$ would change that state. This is an image-conditioned counterfactual judgment: no intervention is executed and no interventional distribution is estimated. For example, in Figure 1, removing the laptop directly affects the books and may indirectly affect the glass via the books.

The graphs are not constrained to be directed acyclic graphs. Self-loops and duplicate edges are invalid, but reciprocal dependencies and larger cyclic structures are permitted when each directed edge independently satisfies the operational criterion. Thus, $i \rightarrow j$ and $j \rightarrow i$ are treated as two distinct edges during training and evaluation.

Operational Scope. To ensure reliable single-image annotation, VCG-32K focuses on *visually observable direct causal dependence*: an edge is included only when entities exhibit visible contact or direct interaction, and the removal counterfactual is image-supported. Thus, VCG-32K annotates direct rather than transitive effects: if $\mathbf{v}_i$ supports $\mathbf{v}_j$ and $\mathbf{v}_j$ supports $\mathbf{v}_k$, we include $i \rightarrow j$ and $j \rightarrow k$, but not $i \rightarrow k$. This scope makes the task auditable, separates causal mechanisms such as support or carrying from generic spatial predicates such as on or near, and excludes non-contact, intentional, or temporally latent relations unverifiable from a single image.

Evaluation. Predicted graphs are evaluated structurally, ignoring entity or predicate labels. Entities are matched to ground-truth via the Hungarian algorithm (Kuhn, 1955) with GIoU (Rezatofighi et al., 2019). An edge $(\mathbf{v}_i, \mathbf{v}_j)$ is correct if it exists in the ground-truth with the same causal direction $i \rightarrow j$; reciprocal edges, when present, are matched and scored independently.

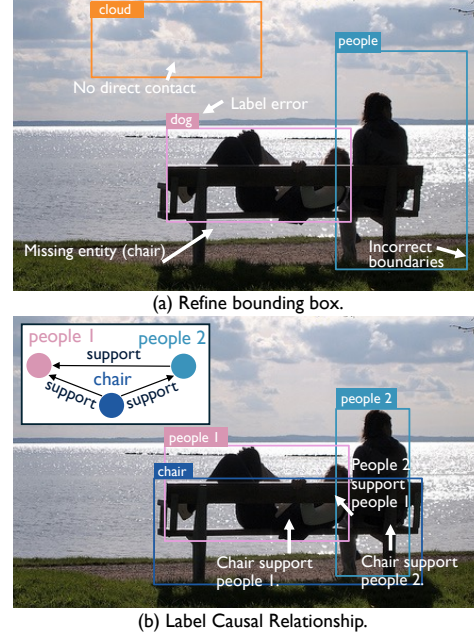


Figure 2: The two-stage annotation pipeline of VCG-32K: bounding box refinement and causal relationship labeling.

3 The VCG-32K Dataset

Existing visual understanding datasets, including scene-graph benchmarks (Chang et al., 2021; Johnson et al., 2015; Yang et al., 2022), largely ignore underlying causality. While they annotate spatial relations (e.g., “on”), they lack the causal mechanisms (e.g., “support” or “carry”) explaining how changes to one entity affect another. In addition, their bounding boxes are not optimized for causal graph reconstruction, where loose or incorrect boxes can alter contact states and thus causal labels. To address these limitations, we construct the Visual Causal Graph dataset (VCG-32K), featuring (1) refined entity boxes $\{\mathbf{v}_i\}_{i=1}^N$, (2) explicit causal relations $\mathbf{v}_i \rightarrow \mathbf{v}_j$, and (3) causal mechanism annotations κ .

VCG-32K is constructed with images from the MS-COCO (Lin et al., 2014) and Objects365 (Shao et al., 2019) datasets. We employ 50 trained annotators and 10 senior reviewers for dataset construction. Figure 2 illustrates the two-stage annotation pipeline: bounding box refinement and causal relationship labeling. In the first stage, annotators refine existing bounding boxes by correcting inaccurate boundaries or labels, removing entities without direct contact, and adding missing entities involved in causal interactions. In the second stage, annotators identify causal relationships $\mathbf{e}_{ij} = (\mathbf{v}_i, \kappa, \mathbf{v}_j)$. A causal edge $i \rightarrow j$ is defined when (1) $\mathbf{v}_i$ directly contacts $\mathbf{v}_j$, (2) $\mathbf{v}_i$ maintains the current state of

v_j , and (3) removing v_i would disrupt that state.

We performed quality control throughout both annotation stages. Annotators were first trained on a 2,000-image calibration set and required to achieve at least 95% accuracy on a 50-image qualification subset before formal annotation. After each production batch, senior reviewers audited 10% of annotation units; batches below 95% audited accuracy were corrected and used for targeted retraining. Starting from 100,000 candidate images for box review and 83,804 items for causal-relation annotation, the final VCG-32K contains 32,256 images, 299,262 entities, 2,287 entity categories, and 185,321 causal relationships, averaging 9.28 entities and 5.75 relationships per image. Additional details are provided in [Appendix E](#).

4 Method

We develop CauSight through two stages: tree-search-based trajectory synthesis for reasoning policy initialization ([subsection 4.1](#)) and reinforcement learning for further optimization ([subsection 4.2](#)). [Figure 3](#) shows part of a synthesized reasoning trajectory, while [Figure 4](#) presents the overall training pipeline.

4.1 Tree-of-Causal-Thought

Visual causal discovery seeks to ground visual concepts in language, thereby requiring VLMs to internalize a causally aware reasoning process. However, VCG-32K offers only final answers but no process supervision. Moreover, collecting expert rationales is prohibitively costly and inefficient. Therefore, we propose **Tree-of-Causal-Thought (ToCT)**, an automated approach for synthesizing candidate causally aware reasoning trajectories. We first present the key actions, then the MCTS procedure, and finally how the resulting small ToCT dataset supervises model training.

Key Actions. Constructing ToCT involves three key actions: *region selection* A_r , *entity recognition* A_e , and *causality orientation* A_c . (1) *Region selection* zooms into specific sub-regions, facilitating a more comprehensive search for causal pairs and helping to exclude confounding factors. (2) *Entity recognition* grounds entities in the selected region. (3) *Causality orientation* infers the existence and direction of causal relationships between entities. The actions are repeatedly executed in a fixed loop: $A_r \rightarrow A_e \rightarrow A_c \rightarrow A_r \rightarrow \dots$. At each step t , the reasoning state

$s_t = \{\text{regions}_{1:t-1}, \text{causality}_{1:t-1}, a_{t-1}\}$ updates all previously explored regions, discovered causal relationships and last executed action. The model selects the next action $a_t \in \{A_r, A_e, A_c\}$ following the looping order, and then generates an intermediate reasoning result $r_t \sim \pi_\theta(\cdot \mid s_t, a_t)^2$. This process iterates until a predefined step limit T is reached. The full reasoning trajectory is thus the accumulation of intermediate results: $\tau = r_1 \oplus r_2 \oplus \dots \oplus r_T$. We use Qwen2.5-VL-72B as the teacher model to propose candidate trajectories for its strong reasoning performance and low cost compared with proprietary models. The teacher is not treated as an oracle: candidate trajectories are evaluated by their final agreement with the human-annotated causal graph and by structural validity checks.

Monte Carlo Tree Search. The MCTS algorithm expands a single trajectory into a tree structure, maintaining multiple branches and propagating value estimates to search for higher-value trajectories. Each search iteration consists of four phases: *selection*, *expansion*, *simulation*, and *back-propagation*. During *selection* phase, the algorithm traverses the tree from the root to identify the most promising child node using the well-known Upper Confidence Bounds for Trees (UCT) rule ([Kocsis and Szepesvari, 2006](#)), which balances exploration and exploitation:

$$\text{UCT}(s, a) = Q(s, a) + w \sqrt{\frac{\ln N(Pa(s))}{N(s, a)}},$$

where $N(s, a)$ and $N(Pa(s))$ denote the visit counts of the current node and its parent, respectively, and $Q(s, a)$ is updated during backpropagation. In the *expansion* phase, the model executes an action a based on the current node state s to generate new child nodes $\mathcal{C}(s, a)$, thereby expanding the search tree in both width and depth. Next, in the *simulation* phase, the model performs an imagined rollout from the selected node until reaching a leaf node. This process estimates a foresighted value by completing a hypothetical reasoning trajectory that anticipates future outcomes. Finally, during *back-propagation*, the estimated value is propagated upward to update all nodes along the selected branch:

$$Q(s, a) = \frac{\sum_{c \in \mathcal{C}(s, a)} Q(c) \cdot N(c)}{\sum_{c \in \mathcal{C}(s, a)} N(c)}$$

The value signal originates from evaluating the leaf node’s final graph against the ground-truth causal

² π_θ refers to the policy of the VLM parameterized by θ .

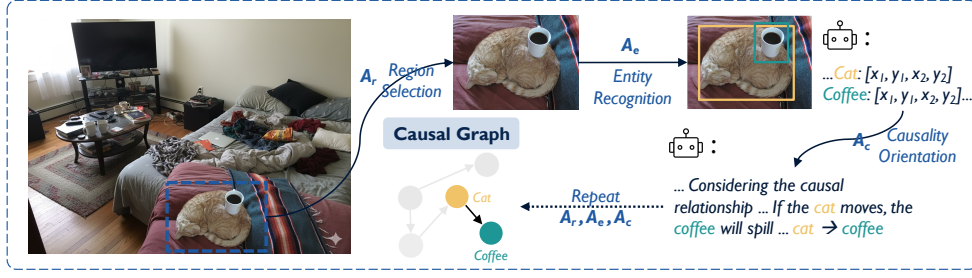


Figure 3: Illustration of a single synthesized reasoning trajectory. The teacher model can repeatedly execute three key actions to extend the reasoning trajectory.

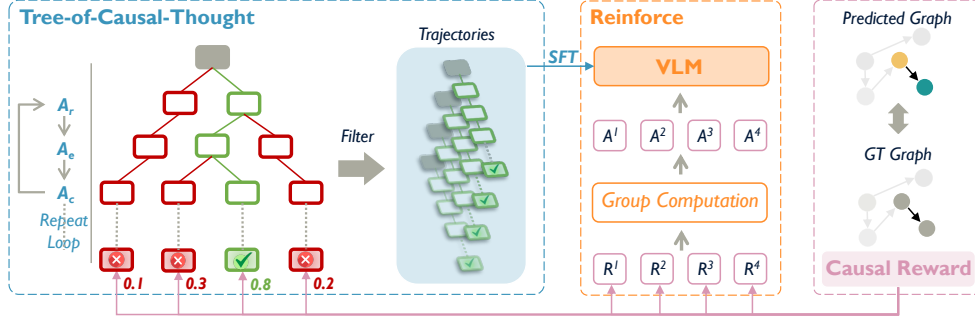


Figure 4: Overview of the training pipeline. Tree-of-Causal-Thought expands a single reasoning path into a tree structure to select high-quality trajectories, which serve as demonstrations to initialize CauSight’s reasoning policy. The subsequent RL further optimizes this policy for visual causal discovery with our designed causal reward.

graph, providing outcome supervision for value propagation. This score ranks candidate paths by final graph quality, while the step-wise process supervision comes from imitating the selected ToCT trajectory during SFT. After a fixed number of search iterations, the trajectory is obtained by greedily selecting the highest-value nodes along the reasoning tree.

Reasoning Policy Initialization. To further reduce the risk of noisy trajectories, we apply a performance-based filtering strategy: only MCTS-induced trajectories that outperform their vanilla one-step reasoning counterparts are retained for training. This biases the SFT set toward trajectories that both follow the designed reasoning pattern and lead to better final graphs, minimizing the risk of introducing suboptimal reasoning patterns. We initialize the model’s reasoning policy by conducting supervised fine-tuning on the filtered trajectories:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(q, \tau) \sim \mathcal{D}_{\text{SFT}}} [\log \pi_{\theta}(\tau \mid q)],$$

where q denotes the input, and τ is the trajectory to be learned by the policy model π_{θ} .

4.2 Reinforce with Causal Reward

Although imitation learning by SFT can instill causally aware reasoning patterns, the VLM’s capability for visual causal discovery remains limited by

the narrow coverage of synthetic trajectories. These trajectories provide useful yet constrained supervision—the model can mimic existing reasoning paths but struggles to generalize. To address this, we introduce RL to move beyond passive imitation and continually optimize the causal discovery policy through iterative interaction and feedback. We start from the SFT-trained policy model and use the GRPO algorithm. Formally, the optimization objective is given by:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{[q \sim \mathcal{D}, \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot \mid q)]} \frac{1}{G} \sum_{i=1}^G (\min(R_i A_i, \text{clip}(R_i, 1 - \epsilon, 1 + \epsilon) A_i)),$$

$$R_i = \frac{\pi_{\theta}(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}, \quad A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}.$$

Causal Reward Design. As outlined in section 2, our goal is to improve the VLM’s ability to understand and reconstruct causal graph structures under the visually verifiable causal definition of VCG-32K. Accordingly, we design a graph-structured causal reward to promote (i) accurate visual causal discovery and (ii) faithful structural generation:

$$R(\tau) = \underbrace{\lambda_r \cdot \text{Recall}(\tau) + \lambda_p \cdot \text{Precision}(\tau)}_{\text{visual causal discovery}} + \lambda_f \cdot \text{Format}(\tau).$$

The Recall and Precision are defined as:

$$\text{Recall} = \frac{|\mathcal{E}_{\text{pred}} \cap \mathcal{E}_{\text{gt}}|}{|\mathcal{E}_{\text{gt}}|}, \quad \text{Precision} = \frac{|\mathcal{E}_{\text{pred}} \cap \mathcal{E}_{\text{gt}}|}{|\mathcal{E}_{\text{pred}}|},$$

Model	VCG-32K-COCO			VCG-32K-365			Average		
	Recall	Precision	F1 Score	Recall	Precision	F1 Score	Recall	Precision	F1 Score
Gemini 2.5 Pro	5.7	2.5	3.1	3.0	1.3	1.7	4.4	1.9	2.4
OpenAI o3	8.6	7.7	7.0	6.0	6.4	5.1	7.3	7.1	6.1
GPT-4.1	16.5	15.6	15.0	3.9	3.6	3.4	10.2	9.6	9.2
GPT-5	5.1	2.9	3.4	3.0	1.5	1.7	4.1	2.2	3.6
Qwen2.5-VL-7B	4.7	7.9	5.4	2.9	6.6	3.8	3.8	7.3	4.6
Qwen2.5-VL-32B	14.0	14.1	12.3	7.7	10.3	7.7	10.9	12.2	10.0
Qwen2.5-VL-72B	19.0	26.7	20.1	11.8	20.5	13.4	15.4	23.6	16.8
	<i>in-domain</i>			<i>cross-dataset</i>					
Qwen2.5-VL-7B + SFT	11.2	14.8	12.0	12.1	19.5	13.8	11.7	17.2	12.9
CauSight (ours)	34.2	48.7	37.6	28.1	42.3	31.1	31.2	45.5	34.4

Table 1: Comparison of CauSight with proprietary, open-source, and the SFT variant. (%) We report graph-level Recall, Precision, and F1 Score on the MS-COCO and separately sourced Objects365 subsets of VCG-32K. CauSight comprehensively outperforms diverse baselines.

Model	Recall	Recall _r	Loss
Gemini 2.5 Pro	5.7	10.33	44.8
OpenAI o3	8.6	9.6	10.6
GPT-4.1	16.5	23.0	27.9
Qwen2.5-VL-7B	4.7	5.7	18.1
Qwen2.5-VL-32B	14.0	18.2	23.1
Qwen2.5-VL-72B	19.0	21.8	12.9
CauSight (ours)	34.2	37.2	8.0

Table 2: Causal Reasoning-Induced Performance Loss.(%) CauSight effectively mitigates reasoning-induced loss toward reachable upper bound.

where $\mathcal{E}_{pred}$ and $\mathcal{E}_{gt}$ denote the sets of predicted and ground-truth directed edges, respectively. The Recall term encourages the policy model to align its predicted causal structure with the ground-truth, while the Precision term is introduced to suppress false positives and balance the trade-off between completeness and correctness. The Format term evaluates the compliance of the model’s structural generation with the predefined output schema, including parseability and valid entity references. The reward is outcome-level rather than a replacement for process supervision; in our recipe, ToCT provides step-wise reasoning demonstrations and GRPO refines the final graph quality, making the two forms of supervision complementary.

5 Experiments

5.1 Setup

Baselines. We evaluate CauSight against three baseline groups: (1) proprietary frontier models, including Gemini 2.5 Pro (Google, 2025), OpenAI o3 (openai, 2025b), and the GPT series (Achiam et al., 2023; openai, 2025a); (2) open-source Qwen2.5-VL variants (Bai et al., 2025) at

different scales; and (3) an SFT baseline³ trained on ground-truth formatted labels using the same base model as CauSight. All baselines follow the same instruction and output schema. Off-the-shelf scene-graph generation systems are excluded because they target fixed spatial or semantic predicates rather than open-vocabulary entity boxes with directed removal-counterfactual mechanisms.

Training Data. We use the MS-COCO portion of the VCG-32K dataset as the training corpus, enabling cross-dataset evaluation on the Objects365 subset. The entire training set contains 11,078 samples. ToCT is performed on 6,000 of these samples, yielding 3,631 trajectories after filtering. The 3,631 retained trajectories are used for SFT initialization, and the other 7,447 training samples are used for GRPO with final graph rewards.

Benchmarks and Metrics. We evaluate models on 350 samples from the MS-COCO subset of VCG-32K for the main benchmark, and another 350 from the Objects365 subset to test cross-dataset performance. We report graph recall, precision, and F1 score computed at a GIoU threshold of 0.5.

5.2 Results on VCG-32K

Shown in Table 1, CauSight achieves a substantial improvement over all baselines on both the in-domain VCG-32K-COCO benchmark and the cross-dataset VCG-32K-365 benchmark. It surpasses the leading proprietary baseline, GPT-4.1, by an average absolute gain of **21.0%** in graph recall, and outperforms the larger open-source model (Qwen2.5-VL-72B) by **15.8%**. It also maintains

³This baseline is directly SFTed on final formatted labels, unlike the ToCT-initialized policy model used in CauSight.

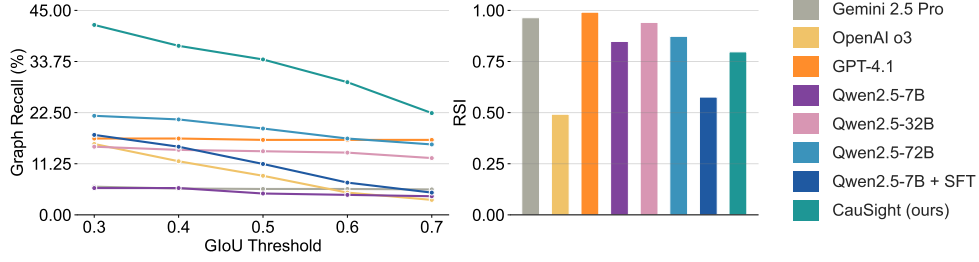


Figure 5: Detection-dependent performance stability. Graph recall is evaluated under varying GIoU thresholds, summarized by the Recall Stability Index (RSI). CauSight achieves strong causal discovery performance while maintaining high RSI, indicating a balanced integration of detection and reasoning capabilities.

strong precision, and a consistent F1 score improvement across both subsets.

Compared with the final-label SFT variant, CauSight improves substantially on both subsets: by **23.0%** recall and **25.6%** F1 on VCG-32K-COCO, and by **16.0%** recall and **17.3%** F1 on VCG-32K-365. This shows that the gain is not merely from learning the output format; the ToCT and RL stages further improve graph reconstruction on both evaluation subsets. The Objects365 result demonstrates transfer to a separately sourced dataset that is not used for training. Because MS-COCO and Objects365 are both natural-image datasets dominated by everyday objects and use the same annotation framework, we treat this result as cross-dataset transfer rather than evidence of robustness to substantially different domains.

The comparison with proprietary models further highlights that, despite their massive scale and pretraining coverage, general-purpose VLMs still struggle to establish coherent cause–effect structures from visual scenes. In contrast, CauSight explicitly integrates causal priors and decision optimization through ToCT and RL, enabling a more systematic understanding of visual interactions. These results validate the effectiveness of our recipe in bridging the gap between perception and reasoning, while the broader generalization of visually grounded causal-dependency discovery remains to be established on more diverse domains.

5.3 Reasoning and Detection

We investigate the respective impact of reasoning and detection capabilities on the task performance with graph recall results on VCG-32K-COCO.

Reasoning-Induced Performance Loss. We isolate the impact of reasoning from detection by evaluating it in a controlled setting. Specifically, for each causal graph, we compute a *reachable* $\text{Recall}_{\mathcal{R}}$, defined as the maximum achievable re-

call given the set of correctly recognized entities (i.e., assuming perfect causally aware reasoning over detected nodes). Let Recall denote the actual graph recall obtained by the model. We then quantify the reasoning-induced performance loss:

$$\mathcal{L} = (\text{Recall}_{\mathcal{R}} - \text{Recall}) / \text{Recall}_{\mathcal{R}}.$$

This metric reflects the relative gap between the model’s achievable upper bound and its actual causally aware reasoning capability. As shown in Table 2, the reasoning gap markedly affects the recall performance across models⁴. While larger models such as Qwen2.5-VL-72B and GPT-4.1 achieve solid graph recall, they still suffer substantial performance loss toward the reachable upper bounds, suggesting insufficient causal comprehension. In contrast, CauSight achieves a recall of **34.2%**, outperforming even much larger models while reducing the loss to **8.0%**, effectively mitigating reasoning-induced degradation.

Detection-Dependent Performance Stability. It is inherently challenging to evaluate a model’s detection capability alone on this task, since the detected entities are closely intertwined with their causal relations. To partially disentangle these factors, we evaluate graph recall under different GIoU thresholds (ranging from 0.3 to 0.7), which reflects its sensitivity to detection strictness, thus indicating models’ underlying detection capability. We introduce *Recall Stability Index* (RSI), denoted as:

$$\text{RSI}_{[0,1]} = \max\left(0, 1 - \frac{\text{std}(\{\text{Recall}_t\})}{\text{mean}(\{\text{Recall}_t\})}\right),$$

where Recall_t denotes graph recall at GIoU threshold t . RSI quantifies how robust graph recall is across thresholds, i.e., how strongly it depends on

⁴Since the model fine-tuned directly on ground-truth labels (the SFT variant) follows a structured output and no longer falls under the category of **reasoning models**, its relevant results are not reported.

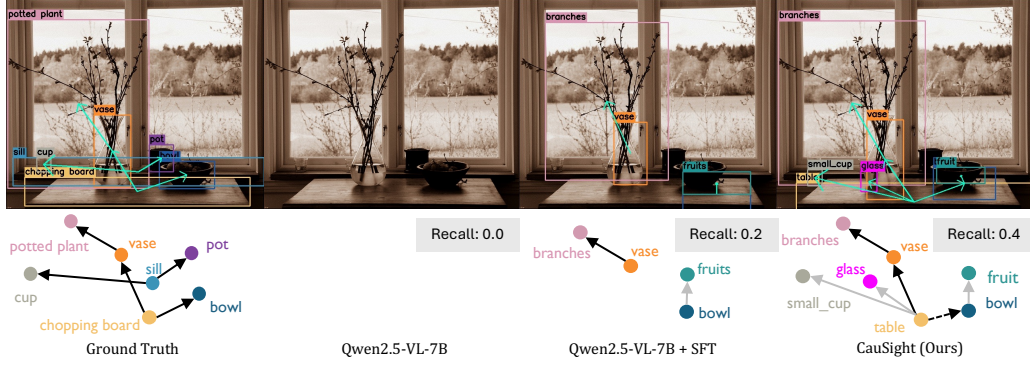


Figure 6: Qualitative comparison on the base model, the SFT variant and our CauSight.

detection capability. As Figure 5 shows, GPT-4.1 exhibits highest stability, indicating superior detection capability together with competitive recall. In contrast, directly applying SFT on the base model severely degrades detection capability, as reflected by the markedly low RSI. CauSight achieves a balanced improvement: it substantially boosts causal discovery performance with only a minor RSI drop.

Comparing reasoning-induced loss with RSI reveals a trade-off between detection and reasoning. OpenAI o3 shows low reasoning-induced loss (10.6%) but weak RSI, while Gemini 2.5 Pro achieves strong detection performance at the cost of substantially higher reasoning-induced loss (44.8%). These results highlight the difficulty of aligning visual perception with causal reasoning. In contrast, CauSight achieves leading visual causal discovery performance and balanced detection-reasoning capability.

5.4 Ablation Study on Training Recipe

We conduct an ablation study of the training recipe on both the in-domain VCG-32K-COCO benchmark and the cross-dataset VCG-32K-365 benchmark. As shown in Table 3, both ToCT and the subsequent RL stage contribute substantially to performance gains. ToCT-induced trajectories provide a supervised prior for initializing structured reasoning, while RL further refines the policy through dynamic feedback. Overall, CauSight improves graph recall by 27.4% over the base model, and by 21.1% and 6.7% over variants without RL and ToCT, respectively.

Notably, despite the limited number of ToCT-induced trajectories, SFT initialization before RL still yields a significant advantage over RL alone, improving graph recall by 10.1% on the cross-dataset benchmark. This supports the benefit of structured reasoning initialization for transfer be-

tween the two evaluated datasets. Additional ablation results are provided in the Appendix.

5.5 Qualitative Comparison

Figure 6 presents a qualitative comparison among the base model, the SFT variant, and CauSight. CauSight achieves high-quality causal discovery, producing detailed and precise outputs. Additionally, under our multi-stage training using graph reward without category-label supervision, the model can use visual attributes to adjust entity descriptions for causal discovery: e.g., it can revise the ground-truth label chopping board to table when the visual evidence supports that role. These results highlight its effectiveness in capturing complex causal patterns within visual scenes.

6 Related Work

Causal Discovery and Reasoning. Causal reasoning is typically formalized with structural causal models (SCMs) (Pearl, 2009). Classical causal-discovery methods (Shimizu et al., 2006; Peters et al., 2014) and graph-based or differentiable DAG methods (Yu et al., 2019; Lachapelle et al., 2019) aim to identify causal structure from observational or interventional data under explicit assumptions, while recent deep causal learning approaches (Dasgupta et al., 2019; Buesing et al., 2018; Nair et al., 2019) often study synthetic environments. Our setting is different: it predicts an image-supported dependency graph under a restricted operational counterfactual criterion and does not identify an SCM.

Causal Evaluation in LLMs and VLMs. LLMs have shown strong causal reasoning in text domains, including commonsense, extraction, and multi-step reasoning (Sap et al., 2019; Chen et al., 2022, 2023; Lyu et al., 2020), yet visual grounding

Model	VCG-32K-COCO			VCG-32K-365			Average		
	Recall	Precision	F1 Score	Recall	Precision	F1 Score	Recall	Precision	F1 Score
CauSight (ours)	34.2	48.7	37.6	28.1	42.3	31.1	31.2 (+27.4)	45.5 (+38.2)	34.4 (+29.8)
w/o ToCT	31.0	50.9	35.3	18.0	36.5	21.9	24.5 (+20.7)	43.7 (+36.4)	28.6 (+24.0)
w/o RL	10.7	6.8	7.3	9.4	6.3	6.6	10.1 (+6.3)	6.6 (-0.7)	7.0 (+2.4)
Qwen2.5-VL-7B	4.7	7.9	5.4	2.9	6.6	3.8	3.8	7.3	4.6

Table 3: Ablation results on the training recipe.(%) Both the ToCT approach and the subsequent RL stage play critical roles in enhancing the performance of CauSight.

remains limited. Existing visual QA and causal benchmarks (Zhu et al., 2015; Goyal et al., 2016; Wang et al., 2016; Zellers et al., 2018; Park et al., 2020) involve causal reasoning but lack explicit causal structures and fine-grained entity modeling. We address this gap by learning explicit causal graphs from images in vision-language settings.

7 Conclusions

This work introduces CauSight, a VLM for constructing image-supported causal-dependency graphs from visual scenes. We operationalize visual causal discovery as a restricted, auditable single-image task and present VCG-32K, a dataset of over 32,000 images combining instance-level grounding with directed removal-counterfactual dependencies. To develop CauSight, we propose Tree-of-Causal-Thought together with a subsequent reinforcement learning stage. Our results establish strong performance on VCG-32K and promising transfer between its MS-COCO and Objects365 subsets, while broader causal validity and domain generalization remain open problems.

Limitations

While CauSight demonstrates strong performance on visual causal discovery, we acknowledge several limitations. VCG-32K is constructed from MS-COCO and Objects365, which mainly contain indoor and everyday object scenarios; a broader range of specialized domains such as scientific diagrams or medical imaging remains to be covered in future dataset construction. VCG-32K focuses on visually observable direct-contact causality to keep annotation auditable, which excludes non-contact physical effects, intention-driven relations, path-blocking or affordance constraints, and temporal causal events that require video evidence. CauSight currently operates on single images only and therefore infers causal graphs from static visual observations. Dynamic input modalities such as video sequences are worth exploring in the future.

Acknowledgement

This work is supported in part by the Shanghai Artificial Intelligence Laboratory and in part by the National Natural Science Foundation of China under Grant 625B2131.

References

- OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, and 1 others. 2023. [Gpt-4 technical report](#).
- Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani S. Marathe, Yonatan Bitton, Samir Yitzhak Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. [Openflamingo: An open-source framework for training large autoregressive vision-language models](#). *ArXiv*, abs/2308.01390.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibong Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-vl technical report](#). *ArXiv*, abs/2502.13923.
- Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C. Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, Mark Ibrahim, Melissa Hall, Yunyang Xiong, Jonathan Lebensold, Candace Ross, Srihari Jayakumar, Chuan Guo, Diane Bouchacourt, Haider Al-Tahan, and 22 others. 2024. [An introduction to vision-language modeling](#). *ArXiv*, abs/2405.17247.
- Cameron Browne, Edward Jack Powley, Daniel Whitehouse, Simon M. M. Lucas, Peter I. Cowling, Philipp Rohlfschagen, Stephen Tavener, Diego Perez Liebana, Spyridon Samothrakis, and Simon Colton. 2012. [A survey of monte carlo tree search methods](#). *IEEE Transactions on Computational Intelligence and AI in Games*, 4:1–43.
- Lars Buesing, Théophane Weber, Yori Zwols, Sébastien Racanière, Arthur Guez, Jean-Baptiste Lespiau, and Nicolas Manfred Otto Heess. 2018. [Woulda, coulda, shoulda: Counterfactually-guided policy search](#). *ArXiv*, abs/1811.06272.

- Xiaojun Chang, Pengzhen Ren, Pengfei Xu, Zhihui Li, Xiaojiang Chen, and Alexander G. Hauptmann. 2021. [A comprehensive survey of scene graphs: Generation and application](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:1–26.
- Meiqi Chen, Yixin Cao, Kunquan Deng, Mukai Li, Kunpeng Wang, Jing Shao, and Yan Zhang. 2022. [Ergo: Event relational graph transformer for document-level event causality identification](#). *ArXiv*, abs/2204.07434.
- Meiqi Chen, Yubo Ma, Kaitao Song, Yixin Cao, Yan Zhang, and Dongsheng Li. 2023. [Learning to teach large language models logical reasoning](#). *ArXiv*, abs/2310.09158.
- Meiqi Chen, Bo Peng, Yan Zhang, and Chaochao Lu. 2024. [Cello: Causal evaluation of large vision-language models](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Ishita Dasgupta, Jane X. Wang, Silvia Chiappa, Jovana Mitrovic, Pedro A. Ortega, David Raposo, Edward Hughes, Peter W. Battaglia, Matthew M. Botvinick, and Zeb Kurth-Nelson. 2019. [Causal reasoning from meta-reinforcement learning](#). *ArXiv*, abs/1901.08162.
- Mélanie Frappier. 2018. [The book of why: The new science of cause and effect](#). *Science*, 361:855 – 855.
- Google. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *ArXiv*, abs/2507.06261.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. [Making the v in vqa matter: Elevating the role of image understanding in visual question answering](#). *International Journal of Computer Vision*, 127:398 – 414.
- Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2015. [Image retrieval using scene graphs](#). *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3678.
- Nan Rosemary Ke, Aniket Rajiv Didolkar, Sarthak Mittal, Anirudh Goyal, Guillaume Lajoie, Stefan Bauer, Danilo Jimenez Rezende, Yoshua Bengio, Michael C. Mozer, and Christopher Joseph Pal. 2021. [Systematic evaluation of causal discovery in visual model based reinforcement learning](#). *ArXiv*, abs/2107.00848.
- Levente Kocsis and Csaba Szepesvari. 2006. [Bandit based monte-carlo planning](#). In *European Conference on Machine Learning*.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2016. [Visual genome: Connecting language and vision using crowdsourced dense image annotations](#). *International Journal of Computer Vision*, 123:32 – 73.
- Harold W. Kuhn. 1955. [The hungarian method for the assignment problem](#). *Naval Research Logistics (NRL)*, 52.
- Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. 2019. [Gradient-based neural dag learning](#). *ArXiv*, abs/1906.02226.
- Tianyu Li, Li Chen, Huijie Wang, Yang Li, Jiazhi Yang, Xiangwei Geng, Shengyin Jiang, Yuting Wang, Hang Xu, Chunjing Xu, Junchi Yan, Ping Luo, and Hongyang Li. 2023. [Graph-based topology reasoning for driving scenes](#).
- Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025. [A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges](#). *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1578–1597.
- Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. [Microsoft coco: Common objects in context](#). In *European Conference on Computer Vision*.
- David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. 2016. [Discovering causal signals in images](#). *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 58–66.
- Qing Lyu, Li Zhang, and Chris Callison-Burch. 2020. [Reasoning about goals, steps, and temporal ordering with wikihow](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Stephen Mumford and Rani Lill Anjum. 2011. [Getting causes from powers](#).
- Suraj Nair, Yuke Zhu, Silvio Savarese, and Li Fei-Fei. 2019. [Causal induction from visual observations for goal directed tasks](#). *ArXiv*, abs/1910.01751.
- openai. 2025a. [Gpt-5 system card](#).
- openai. 2025b. [Openai o3 and o4-mini system card](#).
- Jae Sung Park, Chandra Bhagavatula, Roozbeh Motlaghi, Ali Farhadi, and Yejin Choi. 2020. [Visual-comet: Reasoning about the dynamic context of a still image](#). In *European Conference on Computer Vision*.
- Judea Pearl. 2009. *Causality*. Cambridge university press.
- Judea Pearl. 2013. [Structural counterfactuals: A brief introduction](#). *Cognitive science*, 37 6:977–85.
- Derek C. Penn and Daniel J. Povinelli. 2007. [Causal cognition in human and nonhuman animals: a comparative, critical review](#). *Annual review of psychology*, 58:97–118.

- Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. 2014. [Causal discovery with continuous additive noise models](#). *Journal of Machine Learning Research*, 15(58):2009–2053.
- Seyed Hamid Reza Tofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. 2019. [Generalized intersection over union: A metric and a loss for bounding box regression](#). *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 658–666.
- Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. [ATOMIC: an atlas of machine commonsense for if-then reasoning](#). In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press.
- Laura E. Schulz and Elizabeth Baraff Bonawitz. 2007. [Serious fun: preschoolers engage in more exploratory play when evidence is confounded](#). *Developmental psychology*, 43 4:1045–50.
- Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. 2019. [Objects365: A large-scale, high-quality dataset for object detection](#). *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8429–8438.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *ArXiv*, abs/2402.03300.
- Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti J. Kerminen. 2006. [A linear non-gaussian acyclic model for causal discovery](#). *J. Mach. Learn. Res.*, 7:2003–2030.
- Udbhav and Rachna. 2017. [Sapiens: A brief history of humankind](#). *Journal of the Practice of Cardiovascular Sciences*, 3:61 – 62.
- Huijie Wang, Tianyu Li, Yang Li, Li Chen, Chonghao Sima, Zhenbo Liu, Bangjun Wang, Peijin Jia, Yuting Wang, Shengyin Jiang, Feng Wen, Hang Xu, Ping Luo, Junchi Yan, Wei Zhang, and Hongyang Li. 2023. [Openlane-v2: A topology reasoning benchmark for unified 3d hd mapping](#). In *NeurIPS*.
- Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. 2016. [Fvqa: Fact-based visual question answering](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:2413–2427.
- Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. 2022. [Panoptic scene graph generation](#). *ArXiv*, abs/2207.11247.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. [Tree of thoughts: Deliberate problem solving with large language models](#). *ArXiv*, abs/2305.10601.
- Yue Yu, Jie Chen, Tian Gao, and Mo Yu. 2019. [Dag-gnn: Dag structure learning with graph neural networks](#). In *International Conference on Machine Learning*.
- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2018. [From recognition to cognition: Visual commonsense reasoning](#). *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6713–6724.
- Dan Zhang, Sining Zhou, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024. [Rest-mcts*: Llm self-training via process reward guided tree search](#). *ArXiv*, abs/2406.03816.
- Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2023. [Vision-language models for vision tasks: A survey](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:5625–5644.
- Yize Zhang, Tianshu Wang, Sirui Chen, Kun Wang, Xingyu Zeng, Hongyu Lin, Xianpei Han, Le Sun, and Chaochao Lu. 2025. [Arise: Towards knowledge-augmented reasoning via risk-adaptive search](#). *ArXiv*, abs/2504.10893.
- Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. 2015. [Visual7w: Grounded question answering in images](#). *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4995–5004.

A Synthesized Trajectories

In this section, we present a detailed analysis on the quality of ToCT-induced reasoning trajectories. As discussed in section 4.1, the ToCT approach could produce suboptimal solutions compared to their vanilla one-step reasoning counterparts on certain cases. We provide the overall quantification results in Table 4. The **ZERO** represents samples where both the ToCT and the vanilla approaches produce solutions with **0 recall**. These samples are regarded as relatively *extreme cases*, e.g., those with excessively high difficulty. We can see that after removing the extreme cases (w/o ZERO), mean recall of ToCT achieves 0.42, with a 13% absolute gain over the vanilla approach. We further analyzed the data and found that about 50% of the samples had an > 0.1 improvement (ToCT over vanilla). Figure 15 shows the specific results of 6,000 samples (scatter). This filtering is used only to select synthetic reasoning trajectories for SFT demonstrations; it does not remove hard images from the benchmark evaluation.

Our rule for filtering high-quality trajectories is that the ToCT approach is required to strictly outperform the vanilla one-step reasoning on the same image. This makes filtered trajectories more likely to follow the designed reasoning patterns and lead to superior solutions, which minimizes the risk of introducing suboptimal reasoning paths. Finally, after the filtering, we retain 3,631 from the total 6,000 trajectories. The mean recall of the filtered trajectories is 0.50, while that of their vanilla counterparts is 0.21.

Metric	w/ ZERO		w/o ZERO	
	Vanilla	ToCT	Vanilla	ToCT
Mean	0.22	0.32	0.29	0.42
Median	0.00	0.22	0.13	0.33

Table 4: Recall comparison between ToCT and vanilla approaches under different conditions. The ToCT approach is overall superior to vanilla one-step reasoning.

Cost-quality trade-off. ToCT improves trajectory quality at a substantial computational cost: under the same resources, its tree search requires approximately 10–20 times more processing time than vanilla one-step generation. This additional cost is incurred during cold-start trajectory synthesis rather than model inference. The one-step

baseline nevertheless provides weak supervision on many examples—its median recall is 0 when all 6,000 samples are included—whereas ToCT decomposes generation into structured actions and improves mean recall by 10–13 percentage points. In practice, the trade-off can be controlled through the number of synthesized examples and the depth and width of tree search.

B Ablation Study on ToCT and Reward

For the ToCT ablation, we used **sampled unfiltered trajectory subsets (1,000, 2,000, 3,631) for cold-start** from the 6,000 generated trajectory pool to compare against the filtered subset (3,631) used in the paper. Regarding reward design, we adopted a **recall + format** structure, excluding precision.⁵

Method	Recall	Precision	F1 Score
CauSight (original)	34.2	48.7	37.6
<i>cold-start</i>			
Unfiltered 3,631 subset	32.7	47.8	36.1
Unfiltered 2,000 subset	31.5	46.1	34.5
Unfiltered 1,000 subset	26.0	39.5	28.9
<i>reward design</i>			
Recall + Format	31.3	26.4	25.2

Table 5: Ablation Study on ToCT and Reward. The results demonstrate the effectiveness of our quality filtering and the causal reward design.

We attribute the significant performance drop with the 1K unfiltered subset to data scarcity, as evidenced by the substantial improvement when scaling from 1K to 2K. Further, the saturation observed from 2K to 3,631 supports the value of our quality filtering. Surprisingly, removing the precision reward not only impairs precision performance **but also harms recall performance**. This indicates that precision and recall are interdependent and mutually supportive in reasoning. This also confirms that RL with task-oriented rewards is highly effective for data-scarce causal tasks.

C Confidence Intervals on the MS-COCO Test Set

Table 6 reports 95% confidence intervals for the main MS-COCO evaluation over the 350-example test set.

⁵Evaluations are performed on MS-COCO subset.

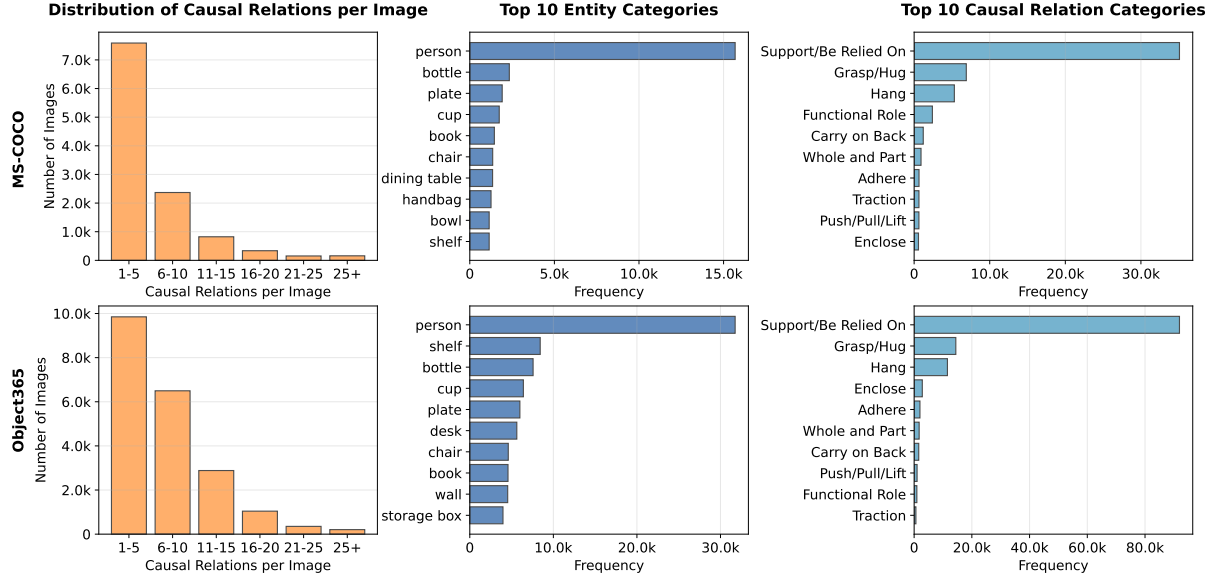


Figure 7: Dataset statistics. The figure reports the distribution of causal relations per image, frequent entity categories, and frequent causal relation categories for the MS-COCO and Objects365 subsets.

Model	Recall (95% CI)	Precision (95% CI)	F1 (95% CI)
Gemini 2.5 Pro	5.7 [3.8, 7.8]	2.5 [1.7, 3.5]	3.1 [2.1, 4.1]
OpenAI o3	8.6 [6.2, 11.1]	7.7 [5.7, 9.8]	7.0 [5.2, 8.9]
GPT-4.1	16.5 [13.3, 20.0]	15.6 [12.5, 18.8]	15.0 [12.1, 18.0]
GPT-5	5.1 [3.3, 7.2]	2.9 [1.9, 4.1]	3.4 [2.2, 4.7]
Qwen2.5-VL-7B	4.7 [2.9, 6.6]	7.9 [5.2, 10.6]	5.4 [3.5, 7.5]
Qwen2.5-VL-32B	14.0 [11.0, 17.2]	14.1 [11.2, 17.1]	12.3 [9.8, 15.0]
Qwen2.5-VL-72B	19.0 [15.8, 22.4]	26.7 [22.8, 30.8]	20.1 [16.9, 23.3]
Qwen2.5-VL-7B + SFT	11.2 [8.4, 14.2]	14.8 [11.5, 18.4]	12.0 [9.2, 15.1]
CauSight	34.2 [30.5, 38.1]	48.7 [44.3, 53.1]	37.6 [33.9, 41.4]

Table 6: Results with 95% confidence intervals on the MS-COCO test set. All values are percentages.

D Performance by Causal-Graph Density

We divide the 350-example MS-COCO test set into sparse graphs with at most five ground-truth causal edges and dense graphs with more than five edges. As shown in Table 7, CauSight performs better on sparse graphs across all metrics. Dense scenes contain more relevant entities and dependencies, increasing both structural reasoning difficulty and the challenge of accurate entity localization and matching.

Graph density	Samples	Recall	Precision	F1
Sparse (≤ 5 edges)	244	39.1	51.1	41.8
Dense (> 5 edges)	106	22.9	43.4	27.8

Table 7: CauSight performance by ground-truth causal-graph density on MS-COCO. All metric values are percentages.

Recall decreases from 39.1% to 22.9% as graph density increases, while precision decreases from

51.1% to 43.4% and F1 from 41.8% to 27.8%. These results identify dense causal-dependency graphs as a challenging direction for future improvement.

E Dataset Construction and Quality Control

This section provides the annotation protocol and quality-control details for VCG-32K. Overall dataset statistics, including the top causal relation categories, are visualized in Figure 7.

E.1 Image Sampling and Annotator Training

We sample candidate images from MS-COCO and Objects365 based on the number of visible entities and the diversity of entity labels, because images with too few entities rarely contain informative causal structures. The annotation pipeline first processed 100,000 candidate images for entity boxes, then passed 83,804 candidate items to

causal-relationship labeling. After filtering images without valid causal relations and completing reviewer adjudication, the released VCG-32K contains 32,256 images: 11,428 from MS-COCO (35.4%) and 20,828 from Objects365 (64.6%). The final annotations include 299,262 entities and 185,321 causal relationships, corresponding to 9.28 entities and 5.75 causal relationships per image on average. Data annotation was conducted by a professional labeling service, with annotators compensated through the vendor. In total, 50 junior annotators participated in labeling, and 10 senior annotators served as reviewers. Before production annotation, annotators studied the written guidelines and trained on a 2,000-image pre-annotated calibration pool covering boundary precision, entity granularity, direct-contact judgment, removal-counterfactual reasoning, and relation-category selection. Formal annotation started only after each annotator completed a 50-image qualification subset and achieved at least 95% accuracy. After each production batch, senior reviewers randomly audited 10% of completed annotation units, including image-level box refinements and causal-relation labels. Batches below the 95% reviewer-audited accuracy threshold were corrected and returned for annotator retraining on the corresponding error types. We additionally estimate inter-annotator agreement on a stratified 100-image subset independently annotated by two annotators. After matching entities with $\text{GIoU} \geq 0.5$, the agreement reaches Cohen’s $\kappa = 0.84$ for causal-edge existence and $\kappa = 0.81$ for causal-mechanism categories. The 95% figure is therefore a human-audit criterion over reviewed annotation units, rather than a model performance number.

E.2 Annotation Guidelines and Examples

Bounding box refinement. All visually grounded entities, including objects and humans, are considered candidate entities. Each entity must be annotated individually; merged boxes are corrected unless the visible parts cannot be separated. Entities smaller than 30×30 pixels are excluded. Pre-annotated boxes are corrected or deleted under three conditions: (1) boundary errors, where the box contains excessive background or truncates the entity; (2) label errors, where the category is overly broad, abstract, or factually incorrect; and (3) irrelevant isolated entities, where the entity has no direct physical contact or interaction with any other entity in the scene. New

entities are added when they participate in at least one direct causal relation but are missing from the original annotation.

Entity representation. Bounding boxes provide explicit spatial grounding, support consistent annotation, and enable standardized entity matching through overlap-based metrics. However, boxes can be imprecise for occluded or irregularly shaped entities, and their boundaries may introduce annotation subjectivity. Alternative representations such as segmentation masks or region-free grounding are promising directions for extending the benchmark.

Causal relationship labeling. Given refined entities, annotators examine candidate pairs and assign a causal edge $\mathbf{v}_i \rightarrow \mathbf{v}_j$ only if all three criteria hold: (1) $\mathbf{v}_i$ and $\mathbf{v}_j$ are in visible direct contact or interaction; (2) $\mathbf{v}_i$ maintains the current state of $\mathbf{v}_j$; and (3) removing $\mathbf{v}_i$ would cause $\mathbf{v}_j$ to lose its current state. Annotators then assign a relation category κ (e.g., “Support/Be Relied On” or “Grasp/Hug”) to explain why the edge holds. Spatial contact alone is insufficient: an edge is not annotated when two entities merely touch or co-occur without a removal-based state change. Representative decisions are shown in Table 8.

E.3 Adjudication and Scope

Disagreements are resolved by senior reviewers. For bounding boxes, the reviewer decides whether the entity should be kept, merged/split, relabeled, or redrawn. For causal edges, the reviewer first checks visible contact, then applies the removal counterfactual, and finally selects the most specific causal mechanism. Hard cases are escalated to a lead reviewer. If neither direction is visually supported, the edge is removed. If disagreements involve reciprocal or cyclic-looking local structures, reviewers re-apply the direct-contact and removal criteria to each direction independently. Reciprocal pairs and larger cyclic structures are retained when every included directed edge is visually supported; adjudication does not impose an acyclicity constraint.

The direct-contact constraint makes annotation more auditable, but it also defines the boundary of VCG-32K. It excludes at least four classes of causal relations: (1) non-contact physical effects such as light, heat, airflow, magnetism, or sound; (2) intentional or social causality such as a person looking at, commanding, or avoiding an object; (3)

Case	Visual situation	Annotation	Reason
Support	A pizza is on a plate.	plate $\rightarrow$ pizza, Support/Be Relied On	Removing the plate would change the pizza’s supported state.
Grasp/hold	A person is holding a bag.	hand/person $\rightarrow$ bag, Grasp/Hug	The hand or person maintains the bag’s position against gravity.
Spatial adjacency	A cup and a bottle are both on the same table.	table $\rightarrow$ cup and table $\rightarrow$ bottle; no cup-bottle edge	The cup and bottle are near each other, but neither maintains the other’s state.
Indirect chain	A laptop supports books, and the books support a glass.	laptop $\rightarrow$ books; books $\rightarrow$ glass	We annotate direct edges and omit the transitive laptop $\rightarrow$ glass relation unless they are also in direct contact.
Non-contact cue	A traffic light affects whether a car stops.	no edge in VCG-32K	The relation is plausible but cannot be verified as a direct dependence from single-image contact.
Ambiguous touch	A hand lightly touches an object without supporting or moving it.	no edge unless the state-change test is satisfied	Contact alone is not enough; the removal counterfactual must be visually supported.

Table 8: Representative annotation decisions. These examples distinguish causal relations from scene-graph-style spatial predicates.

path-blocking or affordance constraints where one entity prevents access to another without contact; and (4) temporal causal events that require observing before/after states, such as a ball that has just hit a glass. These exclusions are deliberate for the first version of VCG-32K: they reduce speculation from single images and make each edge verifiable. Future extensions can add video, temporal state changes, and commonsense annotations to cover these broader causal phenomena.

E.4 Dataset Format

An example of VCG-32K is shown below, where each causal relationship is denoted by entity IDs and the first entity causes the second entity.

```
{
  "dataset_id": "COCO",
  "img_id": 0,
  "entities": [
    {
      "entity_id": 0,
      "entity_name": "woman",
      "bbox": [502.6, 105.47,
              25.83, 132.38]
    },
    ...
  ],
  "causal_relationships": {
    "carry_on": [[0, 1], ...],
    "support": [[2, 3], ...]
  }
}
```

F Implementation Details.

For ToCT, we set the step limit T to 12, expand up to 10 child nodes per step, and run 20 search iterations. For GRPO, training runs for 15 epochs on 4 nodes with 8 H200 GPUs each, using a batch size of 512 and a maximum response length of 4,096 tokens. Each rollout samples 5 outputs per prompt. Additional prompt templates are shown in [Boxes](#).

G General Capability Assessment

To evaluate the generalizability of CauSight, we compare it to the base model and the SFT variant on three out-of-domain (OOD) benchmarks: Math-V and MathVista for multimodal mathematical reasoning, and BLINK for visual perception. [Figure 8](#) presents the results. Our findings indicate that CauSight largely preserves generalization capabilities comparable to the original model. Conversely, models trained directly with SFT exhibit a severe degradation. This is particularly evident on math reasoning tasks, where the performance of the SFT variant dropped by over **20%**.

H Qualitative Comparisons

We provide more qualitative comparisons on Qwen2.5-VL-7B (base model), Qwen2.5-VL-7B + SFT and our CauSight. For each, we present the ultimate generated causal graph and corresponding recall against the ground-truth. Black solid lines denote correctly constructed causal edges, black dashed lines indicate correctly inferred causal relationships but without precise bounding box

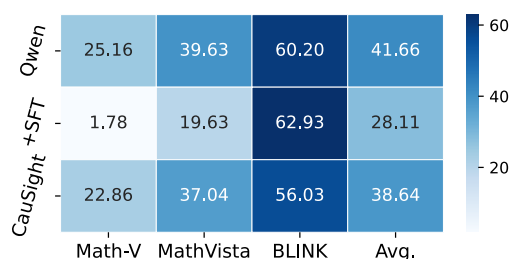
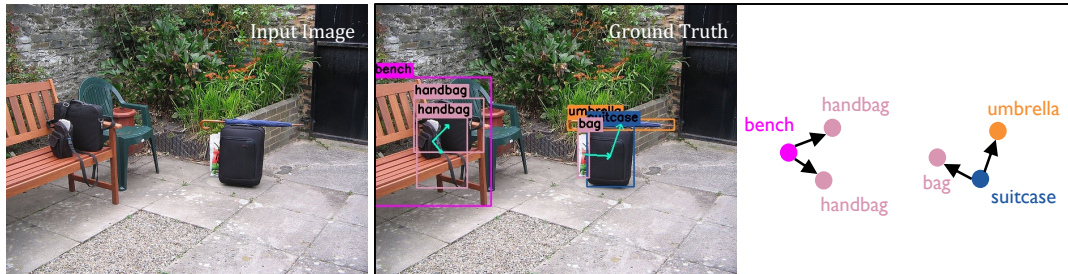


Figure 8: Model generalizability across three OOD benchmarks. Qwen refers to Qwen2.5-VL-7B, +SFT to the SFT variant in the baseline. Each cell corresponds to the model’s accuracy.

matches, and gray solid lines represent incorrectly constructed causal edges.

We observe that the base model struggles to construct a causal graph in most cases, due to its limited detection and causal discovery capabilities. In our provided cases, the base model consistently achieves a final recall of 0. The model SFTed with ground-truth labels shows slight improvement; however, it exhibits severe hallucinations in complex scenarios as Figure 10 and Figure 14 show (cases 1 and 5). For example, it incorrectly assumes that one container causally affects the food inside another container. In simpler scenarios, it tends to over-reason: although it achieves a recall of 0.5 in cases 3 and 4, its precision remains extremely low because it falsely detects many nonexistent entities and infers their nonexistent causal relationships. Our model, CauSight, after multi-stage training, demonstrates strong performance. It reliably captures the essential causal structure of a scene. In more complex cases (1, 2, and 5), although the recall does not reach 1, it correctly identifies the main causal relationships (some correct edges are counted as false due to imperfect bounding box matching, as indicated by the black dashed lines). Meanwhile, in the simpler scenarios (3 and 4), the model even achieves perfect recall, showing strong generalization even under a complex reasoning pattern.

Additionally, we provide an actual reasoning path from CauSight as shown in Figure 9. The reasoning path is still well structured to include steps like “focus on a region”, “recognize entities” and “infer causal relationships”, following the causally aware reasoning patterns we designed in subsection 4.1.



<think>
The image shows an outdoor patio area with a wooden bench, a green plastic chair...
... **Focus on a region and recognize entities**
The next logical step is to explore the area around the bench that includes the ground and surrounding plants, ... **So I need to focus on the "Ground Around Bench" region**, and the bounding box is [0,0,644,476].
By Observation, this region contains the following correlated entity pairs: [{'bench': [0, 132, 215, 358], 'bag': [78, 170, 182, 289]}, {'bench': [0, 132, 215, 358], 'backpack': [78, 204, 120, 278]}, {'green chair': [132, 129, 282, 332], ...}].
Considering the provided entity pairs and their bounding box coordinates, **let's analyze each pair for potential causal relationships based on the given criteria:** **Infer causal relationships**
1. ****{'bench': [0, 132, 215, 358], 'bag': [78, 170, 182, 289]****
- The bag is placed on the bench, indicating direct contact. The bench supports the bag, maintaining its position. Removing the bench would cause the bag to fall. Thus, the bench causes the bag's current state.
...
5. ****{'bench': [0, 132, 215, 358], 'green chair': [132, 129, 282, 332]****
- The green chair is positioned near the bench but not in direct contact. The bench does not maintain the green chair's current state. Therefore, there is no causal relationship between the bench and the green chair.
So the entity pairs with causal relationships are [{'bench': [0, 132, 215, 358], 'bag': [78, 170, 182, 289]}, {'bench': [0, 132, 215, 358], 'backpack': [78, 204, 120, 278]}, {'suitcase': [382, 215, 476, 326], 'umbrella': [358, 198, 547, 222]}].
...
</think>

Formatted output

<causal pairs>
...
</causal pairs>

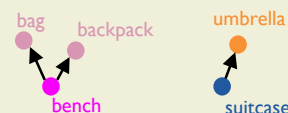
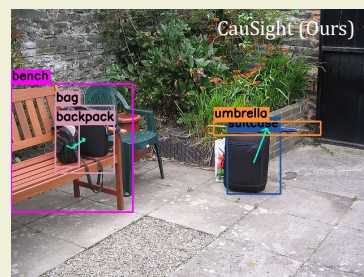


Figure 9: Illustration of actual reasoning path from CauSight.

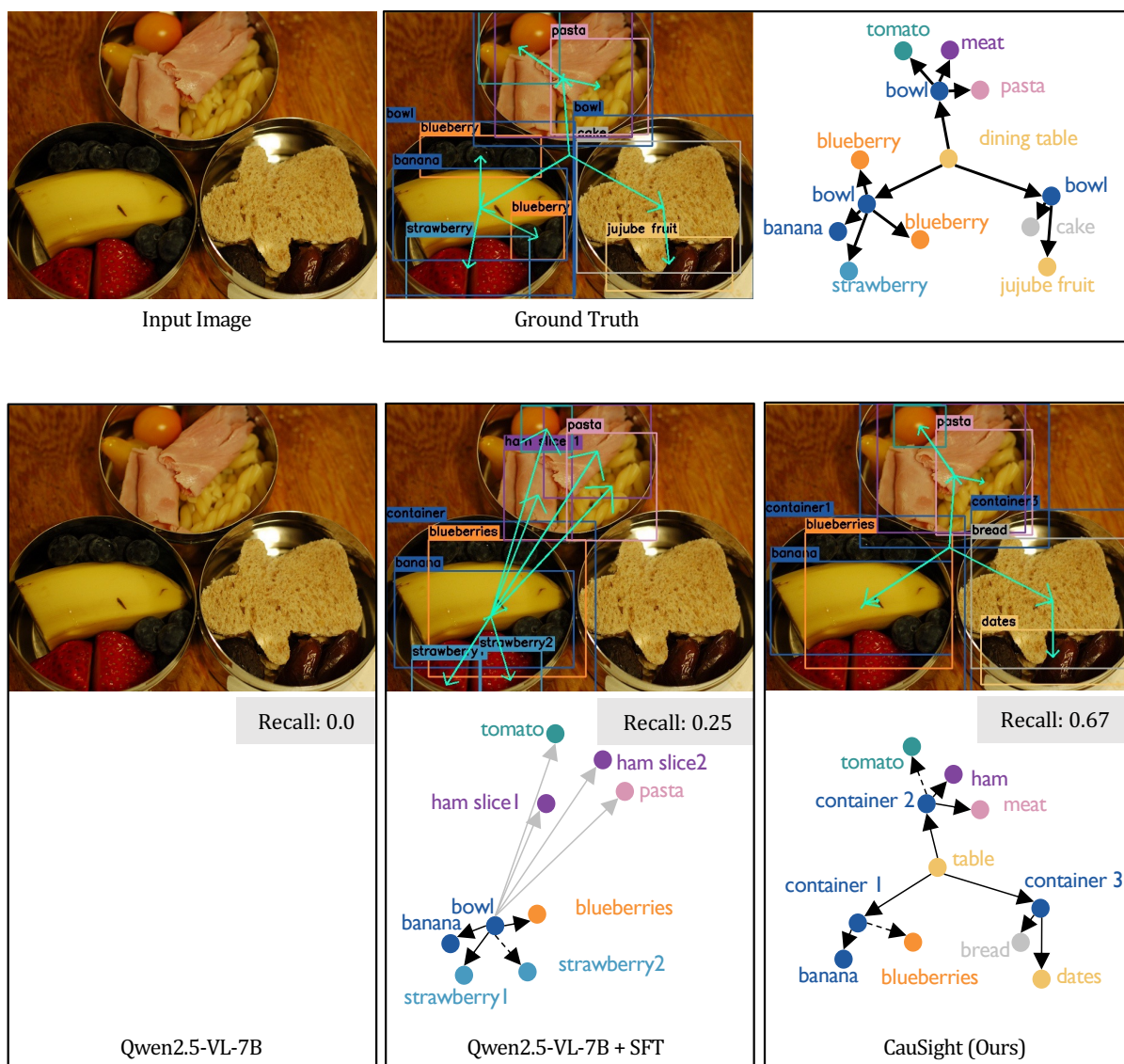


Figure 10: Qualitative comparison (1).

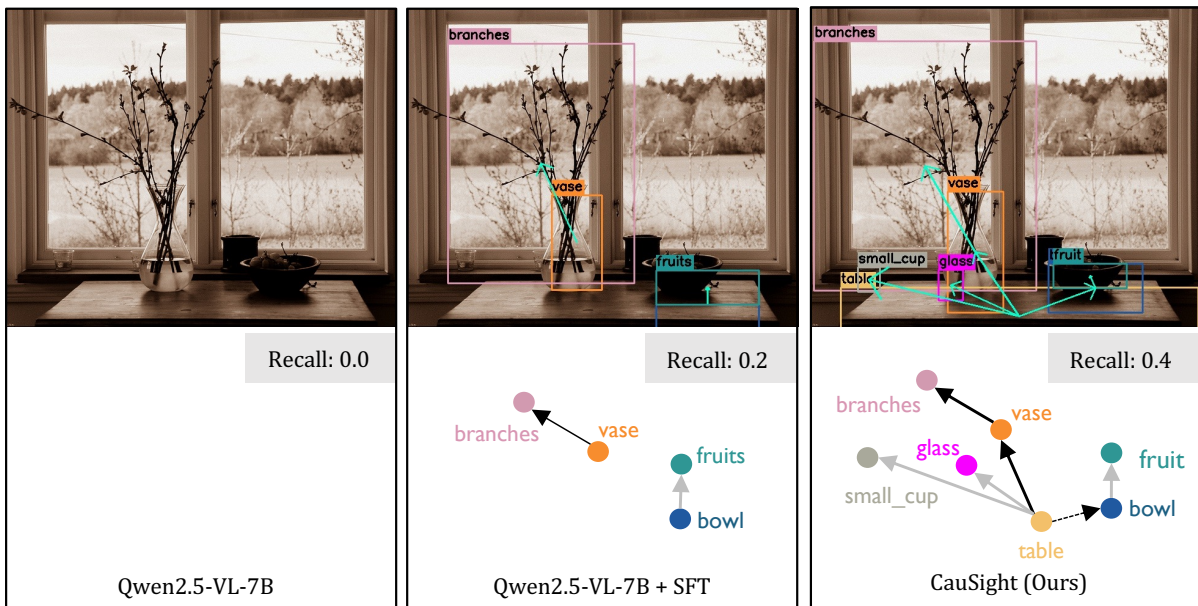
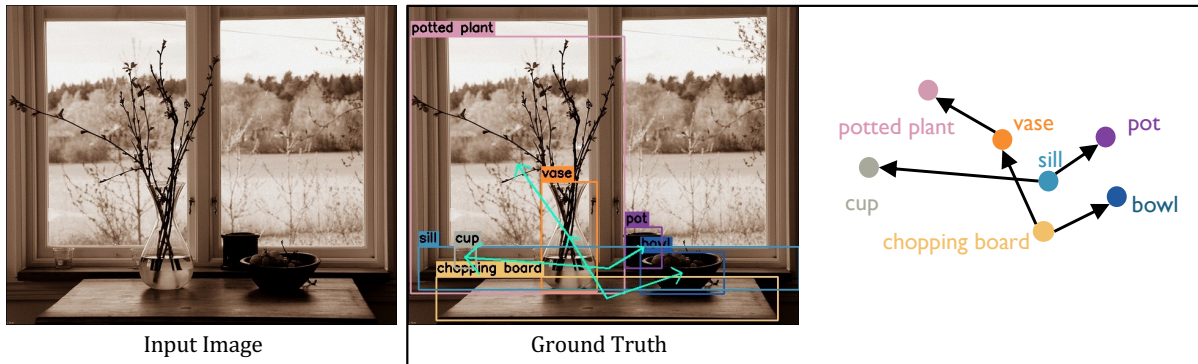


Figure 11: Qualitative comparison (2).

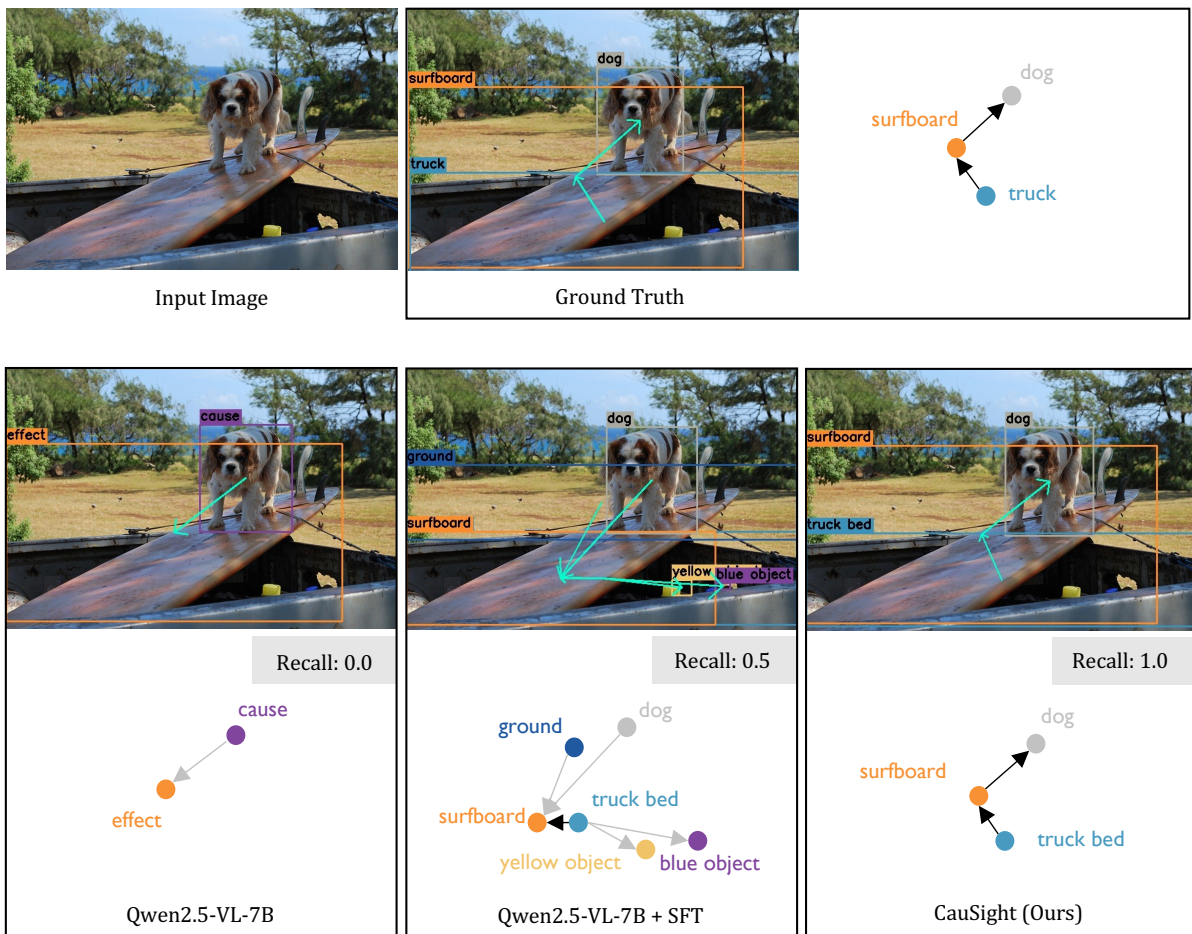


Figure 12: Qualitative comparison (3).

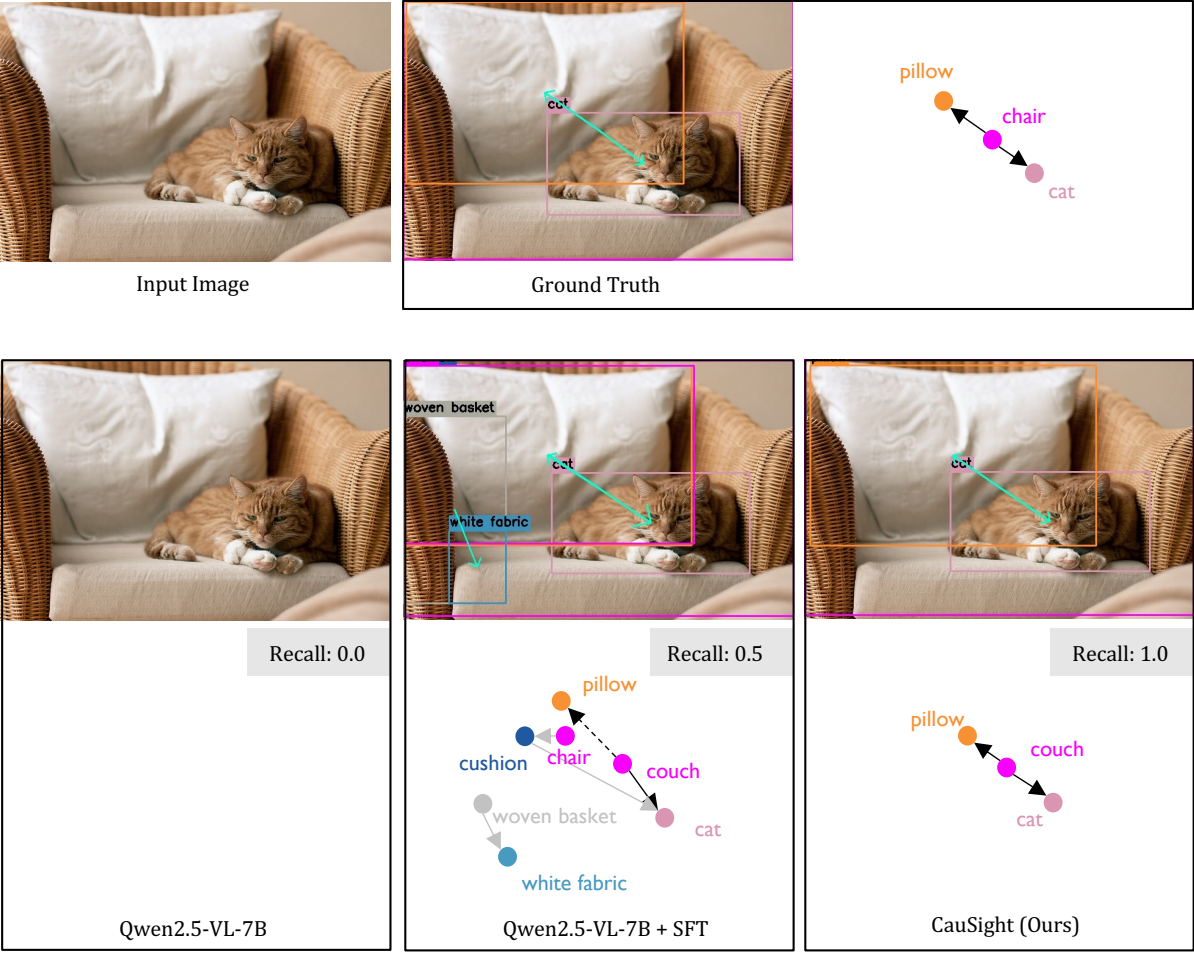


Figure 13: Qualitative comparison (4).

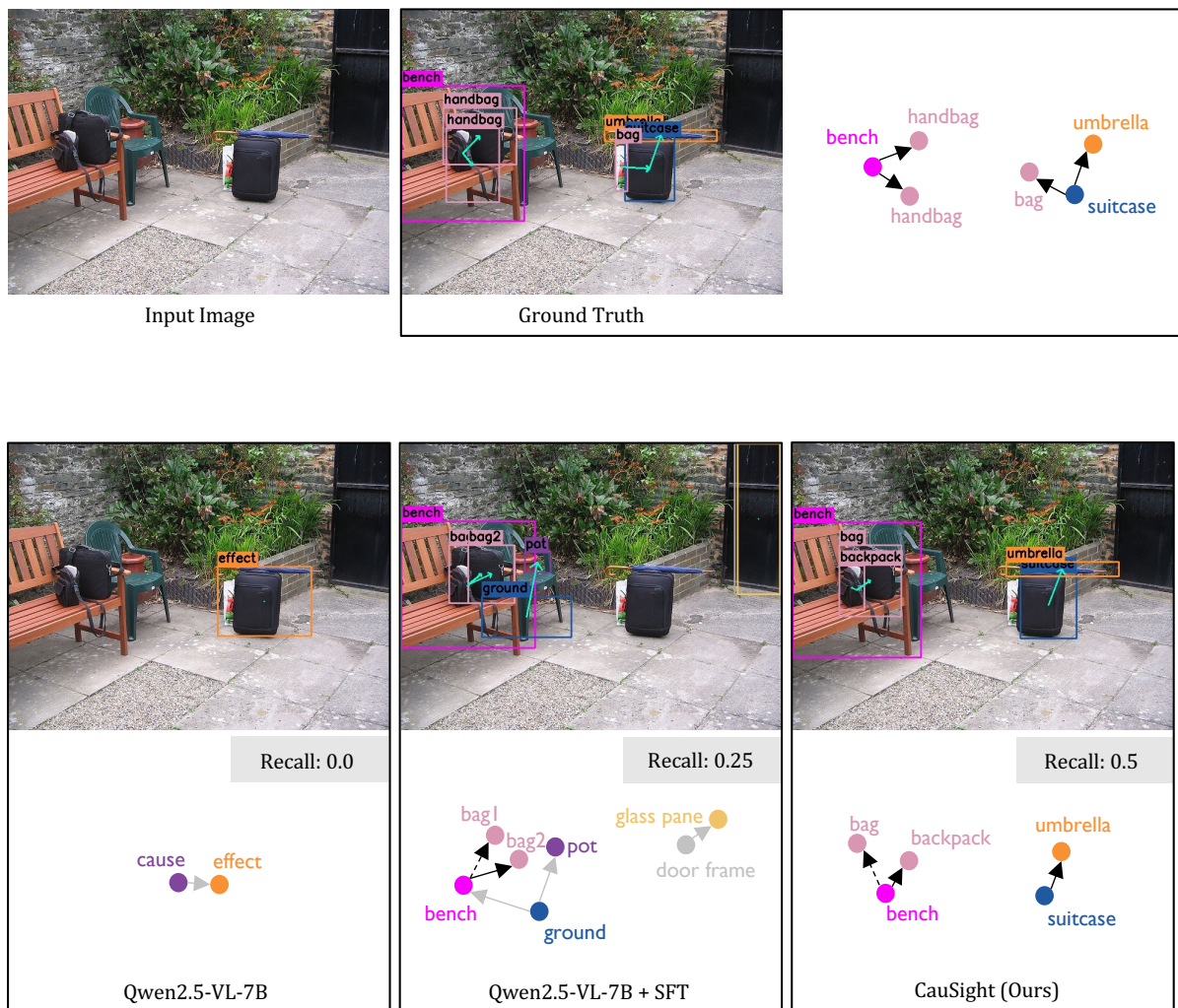


Figure 14: Qualitative comparison (5).

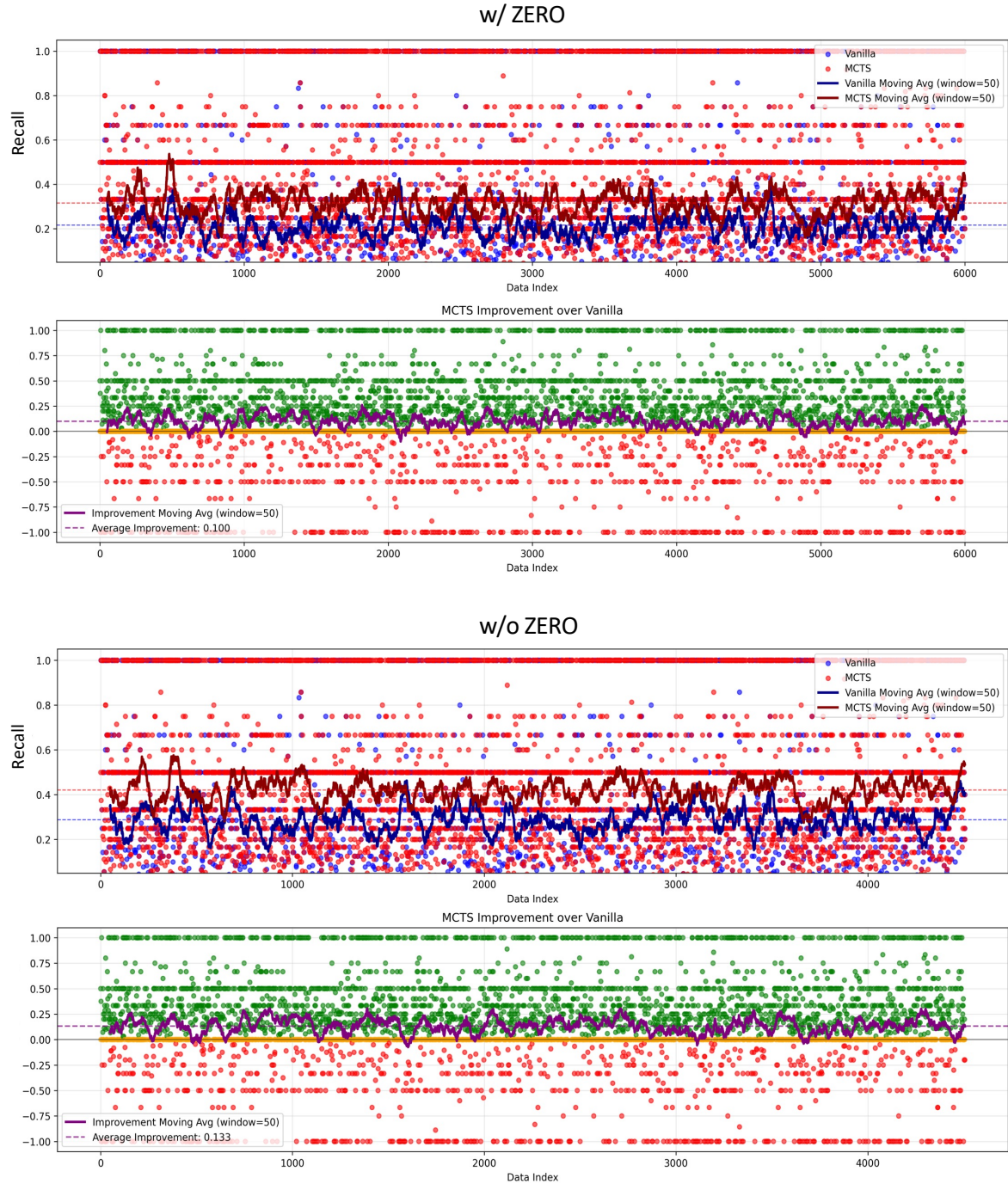


Figure 15: Specific results of 6000 samples (ToCT vs. vanilla, scatter + moving average, w/ ZERO and w/o ZERO.) ZERO represents samples where both the ToCT and the vanilla approaches produce solutions with 0 recall.

Region Selection

You are analyzing the causal relationships between entities in the image through multiple steps.

Your current reasoning trajectory is as follows:

Explored regions: {explored regions}.

Identified causal pairs: {causal pairs}.

Now we hope to look for new regions to discover more potential correlated entity pairs.

Please select the next most worthy region to focus on and explain your thinking process.

Note: the next region should be DIFFERENT from the previous explored regions.

- If you think the exploration regions and identified causal pairs are SUFFICIENTLY COMPREHENSIVE, you should DIRECTLY output "END TRACE" and nothing else.

Otherwise, your output format should be as follows:

<think>

(State the reason as concisely as possible for selecting the new focused region.)

</think>

<region name>

(Output the name of the focused region and nothing else.)

</region name>

<bounding box>

(Output the bounding box of the focused region with format [x1, y1, x2, y2] and nothing else, where (x1, y1) is the top-left coordinate and (x2, y2) is the bottom-right coordinate of the bounding box.)

</bounding box>

Entity Recognition

Your task is to identify all entity pairs that may have correlations in the image. Each pair should have obvious potential correlations such as spatial dependence, support, grasping, placement, inclusion, etc.

Think and output all these correlated entity pairs and their bounding boxes.

Your output format should be as follows:

<think>

(Provide the concise thinking process for identifying correlated entity pairs.)

</think>

<entity pairs>

(Output all the correlated entity pairs in the format of "[\"entity1\": [x1, y1, x2, y2], \"entity2\": [x1, y1, x2, y2], \"entity3\": [x1, y1, x2, y2], \"entity4\": [x1, y1, x2, y2], ...]". You should use ACTUAL ENTITY NAME to replace the placeholders "entity1", "entity2", ... in the format. (x1, y1) is the top-left coordinate and (x2, y2) is the bottom-right coordinate of the bounding box.)

</entity pairs>

Causality Orientation

Based on the image, your task is to determine whether causal relationships exist between the following entity pairs.

Entity pairs: {entity pairs}

The causality criteria are as follows:

For example, if the entity pairs are {"A": [x1, y1, x2, y2], "B": [x1, y1, x2, y2]} or {"B": [x1, y1, x2, y2], "A": [x1, y1, x2, y2]}:

- A is in direct contact with B.
- A's presence maintains B's current state.
- Removing A would cause B to lose its current state.

Then A is the cause and B is the effect.

(x1, y1) is the top-left coordinate and (x2, y2) is the bottom-right coordinate of the bounding box.

Your output format should be as follows:

<think>

(Consider entity pairs and keep the reasoning as concise as possible.)

</think>

<causal pairs>

(Output entity pairs with causal relationships only and if necessary, swap the ORDER of entities pairs to ensure the cause precedes the effect.)

</causal pairs>

E2E Visual Causal Discovery

Identify all causal relationships between entities in the image based on the following criteria:

- A is in direct contact with B.
- A's presence maintains B's current state.
- Removing A would cause B to lose its current state.

Then A is the cause and B is the effect.

Please provide your reasoning process and output all the entity pairs with causal relationships and their bounding boxes in the following format:

<think>

(Provide your reasoning process for analyzing the image.)

</think>

<causal pairs>

(Output all the entity pairs with causal relationships and their bounding boxes in the format of "[\"cause\": [x1, y1, x2, y2], \"effect\": [x1, y1, x2, y2], \"cause\": [x1, y1, x2, y2], \"effect\": [x1, y1, x2, y2], ...]". You should use ACTUAL ENTITY NAME to replace the placeholders "cause" and "effect" in the format. (x1, y1) is the top-left coordinate and (x2, y2) is the bottom-right coordinate of the bounding box.)

</causal pairs>