Beyond What Is Shown: A Benchmark for Non-Explicit Causal Reasoning in Infographics

Anonymous ACL submission

Abstract

Recent advances in Vision-Language Models (VLMs) have shown strong performance in perception and reasoning, yet their ability to perform causal inference—an essential aspect of human cognition—remains underexplored in multimodal settings. We introduce InfoCausalQA, a benchmark for evaluating causal reasoning grounded in infographics that integrate structured visual data with textual context. InfoCausalQA consists of two tasks: Task 1 evaluates quantitative causal reasoning based on inferred numerical trends, while Task 2 targets semantic causal reasoning across five relation types—cause, effect, intervention, counterfactual, and temporal. We collect 494 infographic–text pairs from four public sources and generate 1,482 multiple-choice QA pairs using GPT-4o, followed by systematic human revision to ensure that questions require genuine visual grounding rather than surface-level cues. Experimental results show that current VLMs struggle with both quantitative and semantic causal reasoning, with particularly pronounced limitations in the latter. A human evaluation on 100 Task 2 samples further reveals a substantial performance gap, with humans achieving 77% accuracy. These findings highlight the need to advance causal reasoning capabilities in multimodal AI systems.

1 Introduction

Understanding multiple causal explanations is crucial in real-world scenarios. For example, the sharp decline in stock prices could be attributed to market volatility, but it could also reflect investor panic triggered by policy uncertainty. In policy, science, and everyday decisions, grasping how events influence one another matters more than simply describing them. As a core facet of human cognition, causal reasoning enables interpretation, intention inference, and prediction (Pearl, 2009; Yi et al., 2020). To behave intelligently in complex environments—especially multimodal ones where causal links are often non-explicit—AI systems must acquire this ability (Li et al., 2025). Systematic evaluation across modalities is essential for robust, explainable decision-making.

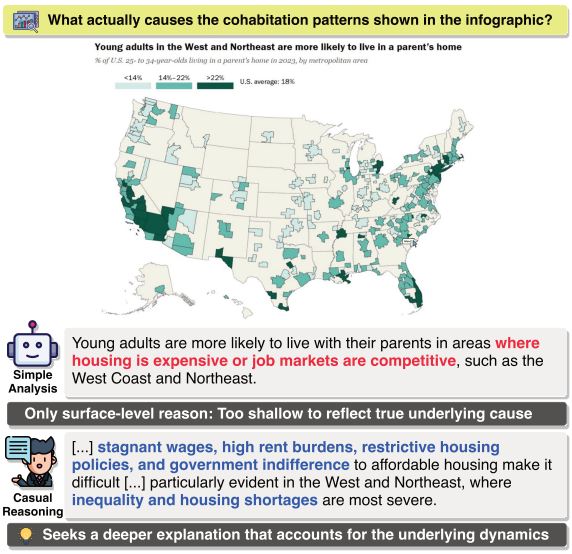

Figure 1: A motivating example for causal reasoning in infographics. While traditional infographic analysis often involves simple, superficial inference, causal reasoning requires inferring information that is not explicitly revealed.

One promising modality for evaluating such reasoning is infographics, which combines visual layouts such as charts with additional textual descriptions. For instance, as illustrated in Figure 1, the infographic shows regional differences in the rate at which young adults live with their parents. While this may appear as simple geographic data, true causal understanding involves reasoning about latent factors like wage stagnation, housing policy. Infographics thus challenge models to go beyond surface-level interpretation: they must integrate multimodal inputs and infer underlying mechanisms that are not explicitly shown.

Benchmark	# Samples	Data Types	Causal Reasoning	Main Question Types
ChartQA (Masry et al., 2022)	32,719	Bar, Line, Pie Charts	✗	Compositional question Visual question
ChartQA-Pro (Masry et al., 2025)	1,948	Bar, Line, Pie, Area, Infographic	✗	Mathematical Reasoning Visual Reasoning
ChartQA-X (Hegde et al., 2025)	30,299	Bar, Line, Pie Charts	✗	Descriptive Reasoning
InfographicVQA (Mathew et al., 2022)	30,035	Infographics	✗	Questions tagged by Evidence Questions tagged by Operation
InfoChartQA (Lin et al., 2025)	55,091	Bar, Line, Pie, Infographics	✗	Text-based questions Visual-element-based questions
InfoCausalQA (Ours)	1,482	Infographics (including various charts)	✓	Quantitative Causal Reasoning Semantic Causal Reasoning

Table 1: Comparison of InfoCausalQA with existing infographic and chart benchmarks. InfoCausalQA explicitly evaluates causal inference rather than descriptive or arithmetic reasoning.

Yet, these capabilities remain limited in current multimodal AI systems. There have been numerous cases demonstrating the evolution of visual processing tasks, from early attempts at simple image recognition to more recent efforts to assess the comprehension of infographics. (Chaudhry et al., 2020; Kahou et al., 2018; Methani et al., 2020; Kafle et al., 2018) While Vision-Language Models(VLMs) are used for analyzing infographics, most benchmarks still neglect causal inference, focusing mainly on surface-level perception and reasoning. While there have been prior attempts to evaluate visual causal reasoning in VLMs, they were limited by a narrow range of task categories (Komanduri et al., 2025) or non-reality based generality. (Wang et al., 2025) Even most current benchmarks for infographic understanding, as shown in Table 1, focus on simple numerical calculations, direct information retrieval, or basic reasoning over explicitly presented data, without addressing the crucial question (Masry et al., 2022, 2025; Hegde et al., 2025; Mathew et al., 2022; Lin et al., 2025): *Can models perform non-explicit causal reasoning based on infographic?*

To address this gap, we introduce InfoCausalQA, a dataset of 1,482 multiple-choice questions curated from 494 infographics. As shown in Figure 2, InfoCausalQA comprises two tasks: Task 1, *Quantitative Causal Reasoning*, requires models to infer causal relationships from visual trends, going beyond simple arithmetic or explicit comparisons. For example, in the rightmost Task 1 example of Figure 2, the model must predict what happens next if an income trend continues, which involves reasoning about non-explicit causal dynamics. Task 2, *Semantic Causal Reasoning*, evaluates the model's ability to reason about five core causality types—Cause, Effect, Intervention, Counterfactual, and Temporal—through both visual and textual interpretation. As shown in the rightmost Task 2 example of Figure 2, the model may be asked to identify the most likely cause of an income trend or infer the expected effect of a policy change. Such questions require contextual reasoning beyond surface-level visual understanding, relying on inference over non-explicit causal structures.

We evaluate a range of VLMs on this benchmark. Our experimental results reveal that, despite acceptable perceptual and linguistic capabilities, VLMs struggle with both causal reasoning tasks. These results suggest a fundamental gap between perception and causal inference, underscoring the need for better causal reasoning in models. We hope InfoCausalQA fosters research bridging perception and causal reasoning, toward more explainable multimodal AI. Our main contributions are as follows:

- We propose InfoCausalQA, the first benchmark specifically designed to evaluate causal reasoning over infographics by leveraging rich visual-linguistic information, including relationships not explicitly observable.
- We conduct a systematic analysis of current VLMs' performance on infographic-based causal reasoning tasks, revealing critical limitations in their inference capabilities.
- Our study offers insights into future directions for developing inference-driven and explainable AI systems, emphasizing the importance

of causal understanding in multimodal reasoning.

2 Related Works

2.1 Infographics Question Answering Benchmark

Recent efforts have explored building benchmarks for question answering on infographics and charts with VLMs. Masry et al. (2022) introduced ChartQA, a large-scale chart QA benchmark requiring compositional and visual reasoning over bar, line, and pie charts. Its follow-ups, ChartQA-Pro (Masry et al., 2025) and ChartQA-X (Hegde et al., 2025), broadened the scope by including additional chart types and question formats, yet they remained oriented towards mathematical or descriptive reasoning tasks, without evaluating causality. Mathew et al. (2022) targets rich infographic images, requiring joint text–visual understanding and categorizing questions by evidence source and operation type; however, its queries remain largely confined to retrieving information or performing basic arithmetic. Lin et al. (2025) further scales up on paired plain and infographic charts, introducing text-based and visual-element-specific queries to test design-driven comprehension, yet it similarly omits causal questions. As a result, none of these benchmarks probe a model's ability to perform causal inference over multimodal (chart+text) information. InfoCausalQA fills this gap by introducing questions explicitly focused on causal reasoning: both quantitative causal analysis of chart trends and semantic causal reasoning grounded in infographic-style data.

2.2 Causal Inference

Interest in how AI systems understand and reason about causality has surged in recent years, giving rise to a series of focused benchmarks. Early efforts such as the structured-data suite of Cai et al. (2024) and the text-centric CLadder (Jin et al., 2023), CausalBench (Wang, 2024), and CausalNet (Ashwani et al., 2024) probe large language models with hypothetical or narrative questions that hinge on non-explicit causal links. Building on this line of work, Imam et al. (2025) introduce TemporalVQA, which requires models to infer temporal–causal relations across multiple images, illustrating the need for multimodal reasoning beyond text alone. InfoCausalQA advances this agenda by asking: Can models derive causality from the richly structured, visually dense world of infographics? By combining textual, graphical, and spatial signals in a single task, InfoCausalQA challenges models to integrate heterogeneous cues when tracing cause–effect chains.

3 InfoCausalQA

InfoCausalQA is designed to systematically evaluate causal reasoning capabilities grounded in infographics. It aims to assess whether models can move beyond perceptual recognition to infer causal relationships from both visual and textual elements.

3.1 Task Definition

We define the two core tasks to systematically evaluate the ability of Vision-Language Models (VLMs) to perform complex causal reasoning in the context of infographics.

3.1.1 Task 1: Quantitative Causal Reasoning

This task evaluates a model's ability to perform causal inference grounded in numerical data presented within infographics. Unlike traditional chart-based QA tasks that focus on directly extracting, comparing, or computing visible values, it requires models to answer causal questions based on hypothetical changes or interventions. Solving these problems involves recognizing key visual elements, understanding the question's context, and performing relevant numerical reasoning. This task assesses whether VLMs possess a foundational understanding of infographics, including both basic quantitative skills and causal inference over visualized data.

3.1.2 Task 2: Semantic Causal Reasoning

This task evaluates higher-level causal reasoning that goes beyond numerical extrapolation. Unlike Task 1, which focuses on quantitative changes implied by visualized values, Task 2 targets *semantic* causal reasoning: models must connect the infographic (and its paired textual context) to plausible explanations, consequences, and hypothetical scenarios that are typically *implied* rather than explicitly stated.

We define five causality types to cover complementary aspects of causal reasoning: *Effect, Cause, Intervention, Counterfactual, and Temporal*. Each question is assigned to exactly one type, and the definitions are summarized in Table 2.

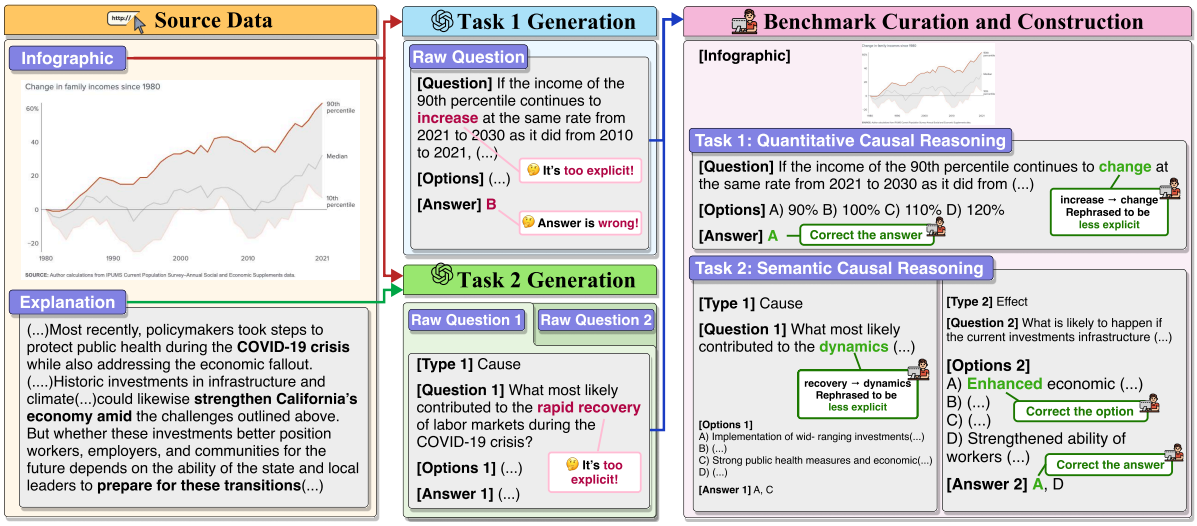

Figure 2: An overview of the benchmark generation and construction process. We manually collected infographics and their accompanying explanatory text from the source data, filtered them for quality, and generated questions for Task 1 and Task 2 using GPT-4o. We then conducted human curation to construct the benchmark.

Causality Type	Definition
Effect	Inferring about possible outcomes or situations following the situation described or depicted in the infographic.
Cause	Inferring the reasons or driving forces behind the observed trends or patterns in the infographic.
Intervention	Inferring how the infographic trend or pattern would change if a situation (action or policy) that did not occur occurred.
Counterfactual	Inferring how the infographic trend or pattern would have been different in a hypothetical scenario where a specific situation (such as a number or action) was different.
Temporal	Inferring why the trend evolves over time, including identifying plausible factors associated with turning points, accelerations, or slowdowns.

Table 2: Definitions of the five causality types used in InfoCausalQA. Each type targets a distinct form of non-explicit causal reasoning that requires integrating visual evidence with contextual semantics.

3.2 Benchmark Construction

3.2.1 Infographic Source Collection

We collect infographics from four public sources—Gallup[1] , Our World in Data (OWID)[2] , Pew Research Center (Pew)[3] , and the Public Policy Institute of California (PPIC)[4]—along with accompanying explanatory text such as captions or summaries manually. The textual context is used exclusively for Task 2 to support semantic causal reasoning, whereas Task 1 relies solely on visual information to formulate questions based on numerical trends.

To ensure quality of collected infographics, two experienced selectors independently reviewed and cross-checked all infographic–text pairs prior to question generation. Samples with low interpretability or incomplete visual elements (e.g., illegible text, missing legends or axes, or inconsistent visual–text correspondence) were removed, yielding a final set of 494 high-quality infographics.

3.2.2 Question and Answer Set Generation

Using the collected infographics and associated explanatory texts, we construct QA sets for both tasks with GPT-4o and then perform thorough human refinement. For Task 1, GPT-4o receives only the infographic image and generates a numerical causal inference question with four answer choices, where exactly one option is correct. We manually revise the generated questions to ensure clarity, numerical correctness, and alignment with non-explicit causal reasoning, while removing surface-level cues that could enable shortcut solutions. For Task 2, GPT-

[1] https://www.gallup.com
[2] https://ourworldindata.org/
[3] https://www.pewresearch.org/
[4] https://www.ppic.org/

4o takes both the infographic and its accompanying text as input and generates questions targeting two of the five causality types, where multiple options may be correct. Human annotators refine all outputs by validating question type, answer consistency, and distractor plausibility, and by ensuring that solving each question requires interpreting the visual evidence rather than relying solely on the text. For transparency and reproducibility, we provide the full prompts in the Appendix D.1.

After generation, three annotators systematically revise the questions to improve dataset quality and to better elicit non-trivial causal reasoning. In particular, they remove superficial cues that could simplify causal inference. This curation is guided by three core principles: *Non-explicitness Enforcement*, *Visual–Logical Alignment*, and *Causality-Type Fidelity*. A description of each principle can be found in Appendix F. The overall revision rate for samples that failed to meet these criteria was 60.59%. We intentionally adopt strict filtering to minimize shortcut solutions and to maintain the benchmark's focus on genuine causal reasoning.

3.3 Benchmark Statistics

3.3.1 Diversity of Infographic Types

To ensure the comprehensiveness and robustness of infographic categories, we analyze the statistical properties and diversity of our dataset. The collection of 494 infographics spans 32 distinct chart types, reflecting a wide range of visualization styles. The majority consists of line charts (43.5%) and bar charts (25.9%), which are commonly used for presenting trends and comparisons. Additional, less frequent chart types—used to assess model adaptability to more complex or uncommon formats—are detailed in Figure 4.

3.3.2 QA and Causality Type Distribution

Based on the 494 collected infographics, we construct a total of 1,482 QA instances in InfoCausalQA. Task 1 (Quantitative Causal Reasoning) includes one QA set per single infographic, yielding 494 QA sets. Task 2 (Semantic Causal Reasoning) includes two QA sets per infographic, each targeting a distinct causal reasoning type, resulting in 988 QA sets. Each Task 2 instance is labeled with one of five predefined causality types: Cause, Effect, Intervention, Counterfactual, or Temporal. Among these, *Effect* accounts for the largest proportion (43%), followed by *Cause* (31%). This distribution reflects the relative ease with which these causality types can be inferred from infographics and their accompanying textual explanations. The full breakdown is shown in Figure 5.

4 Experiments

4.1 Models

We evaluate a diverse set of widely used Vision-Language Models (VLMs), spanning both closed- and open-source models. The closed-source models include o1 (Jaech et al., 2024), GPT-4o (Achiam et al., 2023), and Claude 4 Sonnet (Anthropic, 2025), while the open-source models include InternVL3-38B-Instruct (Zhu et al., 2025), InternVL-2.5-MPO (26B, 8B) (Chen et al., 2024), Qwen2.5-VL-Instruct (72B, 32B, 7B) (Team, 2025), LLaVA-OneVision-7B (Li et al., 2024), Idefics2-8B (Laurençon et al., 2025), and Phi-3.5-Vision-Instruct (Microsoft, 2024). By covering a broad range of model families and sizes, we aim to provide a comprehensive comparison of current VLMs' causal reasoning capabilities on infographics. All experiments were conducted on eight RTX 3090 GPUs, each with 24GB VRAM.

4.2 Quantitative Causal Reasoning

4.2.1 Metric

The model selects the most appropriate one among the four given options and measures the accuracy by comparing it with the correct answer. The full prompts are provided in Appendix D.2.1 for transparency and reproducibility.

4.2.2 Results

As shown in Table 3, even the strongest models exhibited limited performance on Task 1, with Claude Sonnet 4 achieving the highest accuracy at just 57.3%. This highlights that, despite their strong language and general reasoning capabilities, current VLMs struggle with complex quantitative causal reasoning that goes beyond surface-level numerical interpretation.

Closed-source models such as Claude, GPT-4o, and o1 generally outperformed open-source counterparts like InternVL and Qwen. However, the large open-source model Qwen2.5-VL-72B-Instruct (52.0%) achieved performance comparable to GPT-4o, suggesting that open-source models are increasingly competitive.

Among open models, Qwen2.5-VL-72B-Instruct led the group, followed by InternVL3-38B-Instruct (45.7%). In contrast, smaller models such as

Model	Correct Rate (%)
Claude Sonnet 4	57.3
GPT-4o	52.0
o1	53.2
InternVL3-38B-Instruct	45.7
InternVL2.5-26B-MPO	44.1
InternVL2.5-8B-MPO	40.7
Qwen2.5-VL-72B-Instruct	52.0
Qwen2.5-VL-32B-Instruct	49.2
Qwen2.5-VL-7B-Instruct	41.7
Llava-Onevision (7B)	34.6
Idefics2-8B	32.2
Phi3.5-Vision-Instruct (4B)	33.8

Table 3: Accuracy evaluation for multiple-choice questions in the Quantitative Causal Reasoning(Task 1).

Idefics2-8B and Phi-3.5-Vision-Instruct performed substantially worse, with Idefics2-8B ranking the lowest overall. These results suggest that while model scale contributes to performance, size alone is not sufficient for mastering quantitative causal reasoning in infographics.

4.3 Semantic Causal Reasoning

4.3.1 Metric

In Task 2, models judge each of the four options independently as correct ('O') or incorrect ('X'). This design prevents models from relying on relative comparisons between options—common in fixed-number or "select all that apply" settings—and instead promotes evaluation based solely on the relationship between each option, the question, and the infographic. The full prompts are provided in Appendix D.2.2 for transparency and reproducibility.

We evaluate model performance using metrics of Select-All-That-Apply (SATA) benchmark (Xu et al., 2025), which allow for precise quantification of multi-option reasoning accuracy. A total of two categories of metrics, Performance and Count bias are used in this paper, and we use the metrics belonging to Performance as main metrics. The metrics belonging to Peformance and their meanings are as follows: *Exact Match (EM)*, *Precision*, *Recall*, *Jaccard Index (JI)*. The meaning of each metrics is shown in Appendix E.1.

4.3.2 Results

Table 4 shows the overall results for Task 2. We focus on metrics in the Performance category here, while Count Bias metrics are analyzed in Section 5.2.

Among all models, o1 achieves the highest Exact Match (EM) score (65.18%), as well as the highest Precision (90.58%) and Jaccard Index (81.90%). This indicates that even when it fails to select the exact full set of correct options, it tends to include most correct ones while avoiding incorrect answers. GPT-4o follows with an EM of 56.98%, but shows higher Recall (87.69%) than Precision (85.74%), and a lower Jaccard Index than o1. This suggests that GPT-4o tends to include more false positives, resulting in a slightly less precise prediction set.

Among open-source models, Qwen2.5-VL-72B-Instruct is the only one performing on par with GPT-4o. Most others perform considerably worse. Idefics2-8B ranks lowest in both EM and Jaccard Index (41.68%), indicating frequent failure to identify correct sets, particularly in multi-answer cases. These results highlight the current limitations of open-source models in accurately handling multi-option causal reasoning.

5 Analyses

In this section, we analyze the Count Bias results from Task 2 and the answers of models according to the causality type, and compare the performance between models and humans through human evaluation.

5.1 Qualitative Analysis

To analyze qualitative causal reasoning in Task 2, we examine model responses and justifications, with Figure 3 comparing o1, GPT-4o, and Qwen2.5-VL-72B-Instruct. In the *Cause* example (top), o1 and GPT-4o successfully integrate visual evidence with background knowledge, while Qwen2.5-VL-72B refrains from answering in the absence of explicit causal cues, indicating limited flexibility in non-explicit reasoning. In the *Counterfactual* example (bottom), only o1 reasons correctly over relative magnitudes, whereas GPT-4o partially succeeds and Qwen2.5-VL-72B exhibits counterfactual reasoning errors.

5.2 About Count Bias

To analyze selection bias in multi-option causal reasoning, we evaluate models using three count-based metrics: Count Difference (CtDif), Absolute Count Difference (CtDifAbs), and Count Accuracy (CtAcc), which measure over-/under-selection tendencies and accuracy in predicting the number of correct options. The meaning of each metrics is shown in Appendix E.2.

Overall, o1 demonstrates the most accurate and stable selection behavior, achieving the lowest Ct-

Model	Performance (%)				Count Bias		
	EM ↑	Precision ↑	Recall ↑	JI ↑	CtDif	CtDifAbs ↓	CtAcc ↑
Claude Sonnet 4	42.81	80.41	81.65	71.31	0.04	0.56	0.51
GPT-4o	56.98	85.74	87.69	79.02	0.09	0.41	0.64
o1	65.18	90.58	86.01	81.90	-0.10	0.30	0.71
InternVL3-38B-Instruct	30.67	70.52	76.11	62.28	0.17	0.57	0.50
InternVL2.5-26B-MPO	8.00	57.14	72.83	50.81	0.47	0.67	0.40
InternVL2.5-8B-MPO	7.29	53.33	62.10	44.06	0.22	0.64	0.44
Qwen2.5-VL-72B-Instruct	55.77	85.19	84.73	77.16	0	0.40	0.64
Qwen2.5-VL-32B-Instruct	15.69	65.49	76.59	57.70	0.34	0.62	0.41
Qwen2.5-VL-7B-Instruct	7.39	56.59	74.73	50.03	0.66	0.81	0.36
Llava-Onevision (7B)	24.09	74.29	76.40	60.86	0.21	0.74	0.35
Idefics2-8B	4.86	51.57	64.84	41.68	0.65	0.86	0.21
Phi3.5-Vision-Instruct (4B)	17.61	75.08	51.21	47.84	-0.76	0.84	0.32

Table 4: Performance and count bias metrics for the Semantic Causal Reasoning(Task 2).

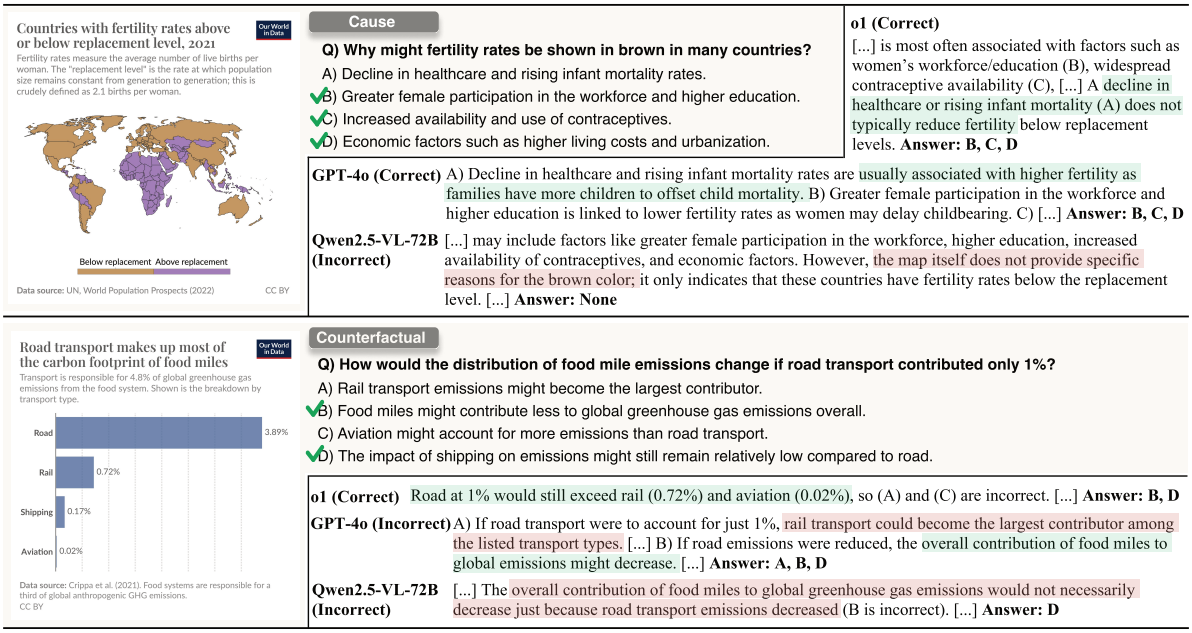

Figure 3: Qualitative results produced by o1, GPT-4o, and Qwen2.5-VL-72B. The examples illustrate common failure modes: overly conservative abstention when causality is non-explicit, partial success with spurious options, and counterfactual reasoning errors driven by incorrect magnitude/logic judgments.

DifAbs and highest CtAcc, with a slightly negative CtDif indicating a conservative strategy that avoids incorrect options. In contrast, other models exhibit pronounced biases, ranging from systematic under-selection (e.g., Phi3.5-Vision-Instruct) to over-selection (e.g., Qwen2.5-VL-7B-Instruct). These patterns suggest that, beyond identifying correct options, accurately estimating how many answers apply remains a challenge for current VLMs.

5.3 Analysis for Causality Type

We analyze model performance on Semantic Causal Reasoning by causality type, revealing consistent trends across models. The result is shown in Table 5. Most perform better on *Effect* questions than on *Cause* questions, indicating a relative strength in predicting outcomes rather than identifying underlying reasons. For example, o1 achieves 74.9% on *Effect* but drops to 57.8% on *Cause*, with Claude Sonnet 4 showing a similar pattern.

Temporal reasoning is the most difficult category for all models. Even the best-performing models, such as GPT-4o, see a steep decline in accuracy—from 64.6% on *Effect* to just 37.5% on *Temporal*. This suggests that understanding time-based causal dynamics, often non-explicit and abstract in infographics, remains a significant challenge.

While smaller open-source models (e.g., Idefics2-8B, Qwen2.5-VL-7B) perform poorly across all types, models like Llava-Onevision and

Model	E	C	I	CF	T
Claude Sonnet 4	50.7	31.2	50.5	54.2	28.4
GPT-4o	64.6	52.0	59.8	52.5	37.5
o1	74.9	57.8	71.0	67.8	35.2
InternVL3-38B-Instruct	35.7	20.8	42.1	33.9	25.0
InternVL2.5-26B-MPO	9.4	6.2	14.0	8.5	0
InternVL2.5-8B-MPO	8.0	6.2	12.2	6.8	2.3
Qwen2.5-VL-72B-Instruct	62.7	48.4	66.4	57.6	34.1
Qwen2.5-VL-32B-Instruct	17.1	12.3	14.0	18.6	20.5
Qwen2.5-VL-7B-Instruct	8.2	7.8	7.5	3.4	4.6
Llava-Onevision (7B)	24.9	25.3	30.8	22.0	9.1
Idefics2-8B	4.9	5.8	5.6	3.4	1.1
Phi3.5-Vision-Instruct (4B)	18.3	17.2	18.7	15.3	15.9

Table 5: Exact Match (EM) scores(%) by causality type in the semantic causal reasoning task: Effect (E), Cause (C), Intervention (I), Counterfactual (CF), and Temporal (T).

Model	EM (%)
Claude Sonnet 4	42.0
GPT-4o	56.0
o1	75.0
InternVL3-38B-Instruct	33.0
InternVL2.5-26B-MPO	2.0
InternVL2.5-8B-MPO	5.0
Qwen2.5-VL-72B-Instruct	54.0
Qwen2.5-VL-32B-Instruct	10.0
Qwen2.5-VL-7B-Instruct	3.0
LLaVA-OneVision (7B)	22.0
Idefics2-8B	3.0
Phi-3.5-Vision-Instruct (4B)	21.0
Human	**77.0**

Table 6: Human evaluation on Task 2 (Exact Match accuracy).

Phi3.5-Vision-Instruct demonstrate that strong causal reasoning is possible even at smaller scales. These results highlight that model size alone does not guarantee causal inference ability, underscoring the importance of training quality, architecture, and reasoning design.

5.4 Human Evaluation on Task 2

Conducting human evaluation over the full benchmark is prohibitively expensive; therefore, we perform a small-scale human evaluation by randomly sampling 100 questions from Task 2. Human annotators answer each question by selecting all correct options, following the same multi-label setting used for model evaluation in Task 2.

We report *accuracy* using Exact Match (EM), where a prediction is considered correct only if the selected option set exactly matches the gold answer set. This strict metric is consistent with the evaluation protocol of Task 2 and prevents performance inflation from partially correct selections.

Table 6 presents the comparison between human performance and VLMs on the sampled Task 2 questions. Humans achieve an EM accuracy of 77%, demonstrating that the benchmark is reliably solvable by humans even under a stringent evaluation criterion, while current VLMs exhibit a substantial performance gap in semantic causal reasoning.

This gap suggests that Task 2 requires integrating fine-grained visual evidence with non-explicit causal semantics, which remains challenging for existing models. Bridging this gap will likely require advances in multimodal grounding as well as more robust causal abstraction capabilities.

6 Conclusion

In this paper, we introduced InfoCausalQA, a novel and comprehensive benchmark designed to evaluate causal reasoning abilities of Vision-Language Models in the context of real-world infographics. By formulating two distinct but complementary tasks—Quantitative and Semantic Causal Reasoning—we move beyond existing benchmarks that focus on superficial perception or basic computations. Instead, we target the core challenge of inferring underlying causal structures from multimodal inputs, including non-explicit patterns not explicitly stated.

Through systematic evaluation across a range of state-of-the-art models—spanning both open-source and closed-source VLMs—we reveal pronounced limitations in current systems' ability to handle causal reasoning. While some closed-source models like o1 and GPT-4o show relatively higher performance, even these struggle with tasks requiring deeper integration of visual cues and non-explicit reasoning.

By surfacing these challenges, InfoCausalQA contributes a new benchmark paradigm that reflects the complexity of real-world reasoning, where understanding causality is essential for decision-making, explanation, and trust. We hope this benchmark not only drives progress in causally aware model development, but also fosters broader research at the intersection of multimodal understanding, structured reasoning, and explainable AI.

7 Limitation

InfoCausalQA is designed to evaluate non-explicit and non-explicit causal reasoning grounded in in-

fographics, rather than to provide fully formalized causal ground truth such as complete causal graphs or structural equations. Causal relations are operationalized through carefully curated multiple-choice questions that reflect plausible human interpretations of visual and contextual evidence. While this design enables scalable and reproducible evaluation, it assesses the plausibility and consistency of causal reasoning rather than definitive causal correctness in a strict causal inference sense.

Additionally, infographic-based causal reasoning inherently involves subjective interpretation, especially when causal relations are unobserved or non-explicit. Although multiple annotators systematically revised and validated generated questions to ensure answerability, visual grounding, and logical soundness, some degree of ambiguity remains unavoidable. Furthermore, the benchmark primarily draws from high-quality public infographics and adopts a multiple-choice format, which may limit generalization to more diverse infographic styles or open-ended causal reasoning scenarios.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Anthropic. 2025. Introducing claude 4. https://www.anthropic.com/news/claude-4. Accessed: 2025-05-23.

Swagata Ashwani, Kshiteesh Hegde, Nishith Reddy Mannuru, Mayank Jindal, Dushyant Singh Sengar, Krishna Chaitanya Rao Kathala, Dishant Banga, Vinija Jain, and Aman Chadha. 2024. Cause and effect: Can large language models truly understand causality? *Preprint*, arXiv:2402.18139.

Hengrui Cai, Shengjie Liu, and Rui Song. 2024. Is knowledge all large language models needed for causal reasoning? *Preprint*, arXiv:2401.00139.

Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. 2020. Leaf-qa: Locate, encode & attend for figure question answering. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3501–3510.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Shamanthak Hegde, Pooyan Fazli, and Hasti Seifi. 2025. Chartqa-x: Generating explanations for charts. *Preprint*, arXiv:2504.13275.

Mohamed Fazli Imam, Chenyang Lyu, and Alham Fikri Aji. 2025. Can multimodal llms do visual temporal understanding and reasoning? the answer is no! *Preprint*, arXiv:2501.10674.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *Preprint*, arXiv:2412.16720.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and 1 others. 2023. Cladder: Assessing causal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:31038–31065.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. 2018. Figureqa: An annotated figure dataset for visual reasoning. *Preprint*, arXiv:1710.07300.

Aneesh Komanduri, Karuna Bhaila, and Xintao Wu. 2025. Causalvlbench: Benchmarking visual causal reasoning in large vision-language models. *Preprint*, arXiv:2506.11034.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2025. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37:87874–87907.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Zhiyuan Li, Heng Wang, Dongnan Liu, Chaoyi Zhang, Ao Ma, Jieting Long, and Weidong Cai. 2025. Multimodal causal reasoning benchmark: Challenging vision large language models to discern causal links across modalities. *Preprint*, arXiv:2408.08105.

Minzhi Lin, Tianchi Xie, Mengchen Liu, Yilin Ye, Changjian Chen, and Shixia Liu. 2025. Infochartqa: A benchmark for multimodal question answering on infographic charts. *Preprint*, arXiv:2505.19028.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, and 1 others. 2025. Chartqapro: A more diverse and

challenging benchmark for chart question answering. *Preprint*, arXiv:2504.05506.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *Preprint*, arXiv:2203.10244.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Microsoft. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Qwen Team. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Zeqing Wang, Shiyuan Zhang, Chengpei Tang, and Keze Wang. 2025. Timecausality: Evaluating the causal ability in time dimension for vision language models. *Preprint*, arXiv:2505.15435.

Zeyu Wang. 2024. Causalbench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 143–151.

Weijie Xu, Shixian Cui, Xi Fang, Chi Xue, Stephanie Eckman, and Chandan K Reddy. 2025. Sata-bench: Select all that apply benchmark for multiple choice questions. *Preprint*, arXiv:2506.00643.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. 2020. CLEVRER: collision events for video representation and reasoning. In *ICLR*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

A Experimental Setup

A.1 List of Vision-Language Models Used in Paper

All the VLMs we used for this paper are as follows.

- o1 (Jaech et al., 2024)
- GPT-4o (Achiam et al., 2023)
- Claude 4 Sonnet (Anthropic, 2025)
- InternVL3-38B-Instruct (Zhu et al., 2025)
- InternVL-2.5-MPO (26B, 8B) (Chen et al., 2024)
- Qwen2.5-VL-Instruct (72B, 32B, 7B) (Team, 2025)
- LLaVA-OneVision-7B (Li et al., 2024)
- Idefics2-8B (Laurençon et al., 2025)
- Phi-3.5-Vision-Instruct (Microsoft, 2024)

A.2 Temperature Setting

For all stages of QA generation and evaluation, we use a sampling temperature of 0.2. This relatively low temperature setting is chosen to minimize randomness in model outputs and promote deterministic reasoning. Since both Task 1 and Task 2 in InfoCausalQA involve multi-step causal inference grounded in visual data, high output consistency is essential to maintain logical coherence and avoid distractive variation in question generation.

Using a lower temperature encourages the model to favor the most likely. Preliminary tests with higher temperatures (e.g., 0.7–1.0) resulted in more diverse but often logically inconsistent questions and answer choices, reinforcing our decision to fix the temperature at 0.2.

B Data Sources

To construct a benchmark grounded in authentic, real-world data, all infographics used in InfoCausalQA were collected from reputable public research organizations. These sources regularly publish high-quality visualizations on social, economic, environmental, and political topics. Below, we briefly introduce each of the four major sources used:

B.1 Gallup[5]

A global analytics and advice firm that publishes survey-based infographics on public opinion, workplace engagement, wellbeing, and world affairs. Gallup visualizations are grounded in regular polling across diverse populations and often highlight temporal trends or geographic differences.

[5]https://www.gallup.com

B.2 Our World in Data (Owid)[6]

An open-access platform that combines academic research with interactive visualization. OWID covers global issues including climate change, poverty, health, energy, and food security. The site is widely used in both educational and policy contexts for its clean and information-dense visual representations.

B.3 Pew Research Center (Pew)[7]

A nonpartisan fact tank based in the U.S. that conducts public opinion polling, demographic research, content analysis, and other data-driven social science research. Pew regularly releases infographics on topics such as technology use, political attitudes, religion, and generational trends.

B.4 Public Policy Institute of California (PPIC)[8]

A nonprofit, nonpartisan think tank focused on issues affecting California, such as education, healthcare, environment, and housing. Their visual reports often provide insight into regional policy debates using clear and data-rich infographics.

These sources ensure that InfoCausalQA reflects a broad spectrum of data types and topics while maintaining factual reliability and visual authenticity. The use of real-world infographics helps to evaluate models under conditions that resemble practical applications in media analysis, policy review, and public communication.

C Full Distribution of Infographic Types

To analyze the structural diversity of InfoCausalQA, we manually categorized all 494 infographics into 32 distinct chart types. Table 7 reports the full distribution, including both common and rare formats.

While the majority of examples are composed of Line Charts (215, 43.5%) and Bar Charts (128, 25.9%), the dataset also includes a variety of less frequent but semantically rich formats such as Choropleth Maps, Sankey Diagrams, Treemaps, and Venn Diagrams. These diverse formats ensure that the benchmark tests model robustness across a wide range of real-world infographic representations.

[6]https://ourworldindata.org/
[7]https://www.pewresearch.org/
[8]https://www.ppic.org/

D Prompts for Problem Generation and Answering

To support the construction and evaluation of the InfoCausalQA benchmark, we designed specialized prompting strategies for both the generation and answering of causal reasoning questions grounded in infographics. The prompts are tailored to each of the two benchmark tasks—Task 1: Quantitative Causal Reasoning and Task 2: Semantic Causal Reasoning—and are further adapted to reflect the unique reasoning formats required in each. This appendix documents the complete set of prompts used during dataset creation and experiments.

D.1 Problem Generation Prompts

D.1.1 For Task 1

The goal of Task 1 is to generate questions that test a model's ability to reason causally using numerical information in infographics. Figure 6 shows the prompt used to create the initial problem for Task 1. The prompt instructs the model to:

First, formulate a multiple-choice question (MCQ) based on quantitative data such as trends, percentages, or numerical values present in the infographic. Second, embed causal or temporal inference into the question by including hypothetical scenarios (e.g., "if a trend continues", "if X increases"). Third, generate four answer options, only one of which is correct, while the others are plausible but incorrect distractors. It also includes the basis for the correct answer so that people can refer to it when making corrections. This prompt design ensures that questions require causal inference rather than mere data lookup or extraction.

D.1.2 For Task 2

For Task 2, the model is asked to generate questions that evaluate higher-level semantic causal reasoning. Figure 7 shows the prompt used to create the initial problem for Task 2.The prompt is provided both an infographic, and an accompanying textual description.

Based on this information, the model must: First, select two causal reasoning types from a predefined set of five: Effect, Cause, Intervention, Counterfactual, and Temporal. Second, for each type, generate one question that adheres precisely to its definition. Third, avoid directly quoting or referencing observable trends or specific values from the infographic; instead, questions must rely on reasoning grounded in the visual-textual context. Next, provide four an-

swer choices per question, with at least two correct answers and the rest being plausible but incorrect. And last, clearly indicate which options are correct. This prompt guides the model to construct complex reasoning questions that simulate analytical interpretation of infographic content.

D.2 Multiple-Choice Answering Prompts

To evaluate model performance on the benchmark, we designed two dedicated answering prompts—one for each task—each enforcing strict output formats to facilitate automated scoring.

D.2.1 For Task 1

Figure 8 shows the prompt used to solve the problem for Task 1. The Task 1 answering prompt presents a single multiple-choice question and instructs the model to: First, examine the infographic and select the single most appropriate answer (A–D). Then, output only the chosen option letter, without any additional text or explanation. This format is compatible with exact-match accuracy metrics.

D.2.2 For Task 2

Figure 9 shows the prompt used to solve the problem for Task 2. Task 2 requires a multi-answer format. This design prevents models from relying on relative comparisons between options—common in fixed-number or "select all that apply" settings—and instead promotes evaluation based solely on the relationship between each option, the question, and the infographics.

The answering prompt provides: First, a semantic causal reasoning question with four answer options. Second, instructions to evaluate each option independently and mark it as either: O (correct), or X (incorrect). The model must output the results in a strict comma-separated format (e.g., A) O, B) X, C) O, D) X). No additional text or justification is allowed. This format allows for partial-credit scoring using multi-label evaluation metrics such as Precision, Recall, and Jaccard Index.

E Metrics Definition for Task 2

We use metrics from SATA-Bench (Xu et al., 2025) to evaluate the correct answer to Task 2. This paper uses a total of seven metrics belonging to the Performance and Count Bias categories. The definitions of each are detailed below.

E.1 Performance Metrics

- EM(Exact Match): EM measures the proportion of questions for which the model correctly selected all correct options. This is the most rigorous method of measuring multi-answer accuracy.

- Precision: Precision measures the proportion of selected options that are actually correct. It captures how "careful" the model is in selecting options.

$$Precision = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{Y}_i \cap Y_i|}{|\hat{Y}_i|} \quad (1)$$

- Recall: Recall measures the proportion of ground-truth correct options that are successfully selected by the model. It reflects the model's ability to cover all relevant answers.

$$Recall = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{Y}_i \cap Y_i|}{|Y_i|} \quad (2)$$

- JI(Jaccard Index): The Jaccard Index quantifies the similarity between the predicted and ground-truth answer sets by computing the size of their intersection over the size of their union.

$$JI = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{Y}_i \cap Y_i|}{|\hat{Y}_i \cup Y_i|} \quad (3)$$

E.2 Count Bias Metrics

These metrics measure the model's ability to estimate the number of correct answers per question, regardless of which specific options are selected.

- Mean Count Difference (CtDif):The average signed difference between the number of predicted and actual correct answers. A positive value indicates over-selection; a negative value indicates under-selection.

$$CtDif = \frac{1}{N}\sum_{i=1}^{N}\left(|\hat{Y}_i| - |Y_i|\right) \quad (4)$$

- Mean Absolute Count Difference (CtDifAbs): The average absolute error between the predicted and true number of correct answers.

$$CtDifAbs = \frac{1}{N}\sum_{i=1}^{N}\Big|\,|\hat{Y}_i| - |Y_i|\,\Big| \quad (5)$$

- Count Accuracy (CtAcc): The proportion of questions for which the model predicted the exact number of correct answers, regardless of which specific options were selected.

$$CtAcc = \frac{1}{N}\sum_{i=1}^{N} 1\left[|\hat{Y}_i| = |Y_i|\right] \quad (6)$$

F Human Refinement Procedure for Question-Answer Problem set Construction

After the initial generation of QA sets using GPT-4o, all questions and answer options undergo a comprehensive human refinement process to ensure quality and faithful grounding in the visual data. This refinement is guided by three core principles:

- **Non-explicitness Enforcement**: Any numerical values or visual trends explicitly stated in the question or answer choices are rewritten into non-explicit forms. This prevents shortcut solutions based solely on textual cues and ensures that solving the question requires visual interpretation of the infographic.
- **Visual–Logical Alignment**: Questions that are factually inconsistent with the infographic or logically ill-formed are manually revised or regenerated, including both the question and its answer choices, to ensure coherent alignment with the visual evidence.
- **Causality-Type Fidelity**: For Task 2, each question is carefully checked to ensure correct alignment with its designated causality type (e.g., Cause vs. Effect). Misclassified or ambiguous cases are corrected to preserve the intended causal reasoning structure.

Through this refinement process, we eliminate surface-level textual cues and enforce logical and semantic coherence, resulting in questions that require genuine visual-grounded causal reasoning.

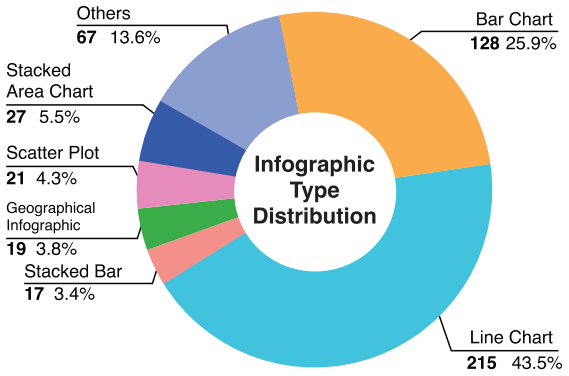

Figure 4: Infographics types distribution of InfoCausalQA.

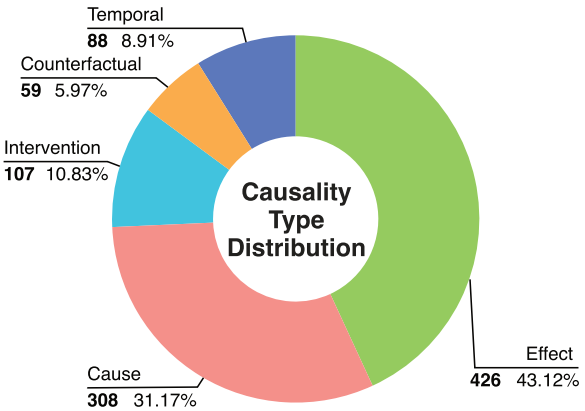

Figure 5: Causality type distribution in InfoCausalQA's Task 2.

Distribution of Infographics							
Anatomical Illustration	1	Narrative Infographic	1	Scatter Plot	21	Bar Chart	128
Mixed Bar and Pie Chart	1	Choropleth Map	16	Statistical Chart	3	Donut Chart	2
Partial Geographical Infographic	10	Simple Infographic	5	Sankey Diagram	3	Heatmap	1
World-wide Geographical Infographic	9	Simple Infographic	5	Statistical Graph	2	Area Chart	1
Horizontal Bar Chart	2	Comparative Chart	2	Bubble Chart	1	Line Chart	215
Mixed Bar and Line Chart	1	Stacked Area Chart	27	Stacked Bar Chart	17	Pie Chart	5
Informational Infographic	2	Statistical Infographic	9	Thematic Map	3	Table	1
Timeline Infographic	1	Vertical Bar Chart	1	Venn Diagram	1	Treemap	1

Table 7: Distribution of Infographics in InfoCausalQA

Prompt for Task 1(Quantitative Causal Reasoning)'s Problem Generation

You are given an infographic that includes numerical values, trends, or proportions about a specific topic. Your task is to generate a quantitative multiple-choice question that tests causality-related reasoning, using the data from the infographic.

[INSTRUCTION]
- Question type: The question should ask for a logical numerical estimate or projection (e.g., what would happen if a trend continues, if one factor changes, or if two values are combined).
- Use data from the image: Base the question on actual values, percentages, or trends shown in the infographic. Avoid adding external knowledge.
- Involve causal or temporal reasoning: The question should implicitly or explicitly involve a causal or time-based assumption (e.g., "If X continues," "If Y increases by 10%," etc.).
- If the information provided in the infographic is insufficient, you can suggest some figures through assumptions.

Provide four answer options (A–D), with only one correct answer. The distractors should be numerically plausible but incorrect.

[Output Format]
Question:
Options:
Answer:
Reason:

Figure 6: Prompt used to generate problems for Task 1(Quantitative Causal Reasoning)

Prompt for Task 2(Semantic Causal Reasoning)'s Problem Generation

You are an assistant designed to evaluate causal reasoning abilities using visual data.

You will be provided with:
- An infographic image (containing visualized data)
- A brief description of the infographic's content

Your task is to create **two questions** based on this infographic and description.
Each question should correspond to a **different causal reasoning type** chosen from the list below.
Make sure each question fits the type precisely.

[Causal Question Types (5 total)]
1. Effect
→ Ask about the likely consequence or result of a situation shown in the infographic.
e.g., "What is likely to happen if this trend continues?"

2. Cause
→ Ask about the reason or cause behind an observed pattern or trend.
e.g., "What most likely caused the drop in birth rate after 2010?"

3. Intervention
→ Ask what would happen if a certain intervention or policy had been introduced.
e.g., "What might have changed if taxes were raised earlier?"

4. Counterfactual
→ Ask about what could have happened in an alternative (non-actual) scenario.
e.g., "If country A had not implemented the law, how would the trend look?"

5. Temporal
→ Ask about changes over time or when a shift or reversal happened.
e.g., "When did the turning point occur in this trend?"

[INSTRUCTION]
- First, select two of the most relevant types based on the infographic and its description.
- Then, generate **one question for each type**.
- Question must be simple and clear. DO NOT DIRECTLY MENTION trends or information that can be seen in the infographic in the question.
- The questions should require reasoning grounded in the information from the infographic.
- Second, the options for each questions consists of four options, of which at least two are correct and the rest are very plausible but incorrect.
- Last, print out which are the correct answers.

[Output Format]
Type:
Question:
Options:
Multi-Answers: (Output only the answers (e.g. A, C))

Now, generate two questions based on the following infographic and its description:

[Information of Infographics]
{description}

Figure 7: Prompt used to generate problems for Task 2(Semantic Causal Reasoning)

Prompt for Task 1's Multiple-Choice Question Answering(Choose only one)

Please look at the following infographic and choose the most appropriate option for the question.
- Print out only answers

Question: {query}
Options:
{options}

Your answer: (Only output the answer (e.g. C))

Figure 8: Prompt used to solve Task 1(Quantitative Cuasal Reasoning)

Prompt for Task 2's Multiple-Choice Question Answering(Multiple Correct)

You will be given a question and four answer options.
Your task is to carefully read the question and options, then for each option, mark 'O' if it is correct or 'X' if it is incorrect.

[Example]
Question: What most likely explains why best-practice organizations have higher employee engagement levels compared to others?
Options:
A) They have more flexible work hours.
B) They focus on manager training and development.
C) They hire more employees than needed.
D) They pay higher salaries to their employees.
Your answer: A) O, B) O, C) X, D) X

Answer the questions with reference to the examples above.

[Important rules]
- You must provide marks(O or X) for each options.
- Answers must be in the following STRICT FORMAT: A) O, B) X, C) O, D) X (comma-separated, no extra text).
- Do not explain your reasoning. Just print the final answer only.

[You]
Question: {query}
Options:
{options}
Your answer: (Only output the answers (e.g. A) O, B) O, C) X, D) O))

Figure 9: Prompt used to solve Task 2(Semantic Causal Reasoning)