CausalVLBench: Benchmarking Visual Causal Reasoning in Large Vision-Language Models

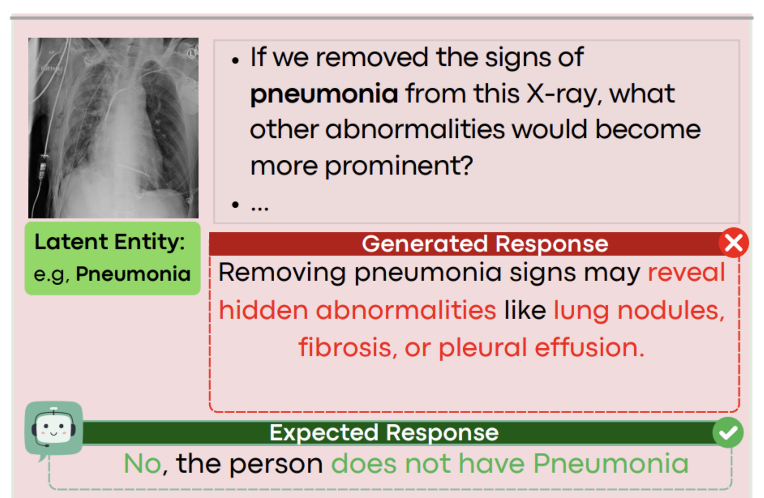
UNIVERSITY OF ARKANSAS

Aneesh Komanduri, Karuna Bhaila, and Xintao Wu
Department of Electrical Engineering and Computer Science
University of Arkansas

EMNLP 2025 Suzhou, China | 中国苏州

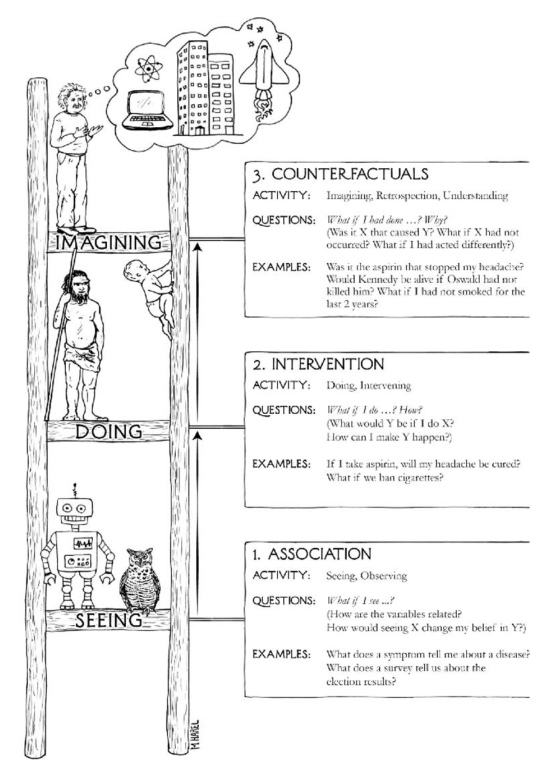
Motivation

- Large vision-language models (LVLMs) have shown remarkable ability in various tasks including object detection and visual question answering
- There have been several studies showing that LLMs and VLMs struggle in complex reasoning tasks and tend to "hallucinate"
- However, there has been little work exploring VLM reasoning from the lens of formal **causality**
- **Pearl's Ladder of Causation:** Observation (seeing), Intervention (doing), Counterfactual (imagining)



Seth et al. Towards a Systematic Evaluation of Hallucinations in Large-Vision Language Models. arXiv 2024.

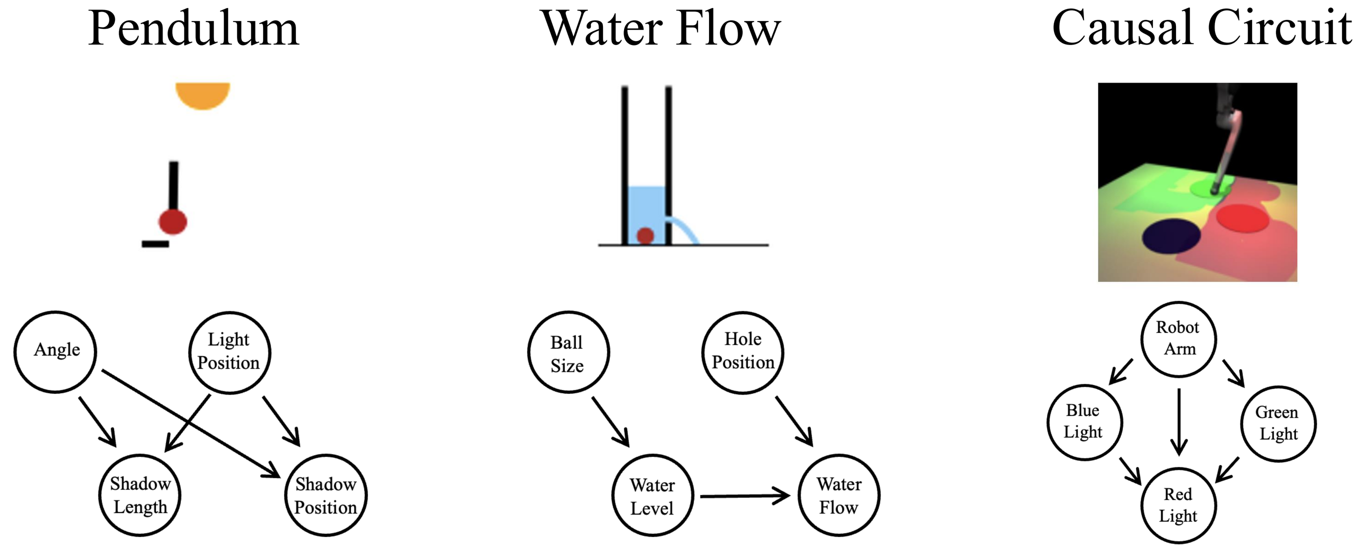
Lack of counterfactual reasoning

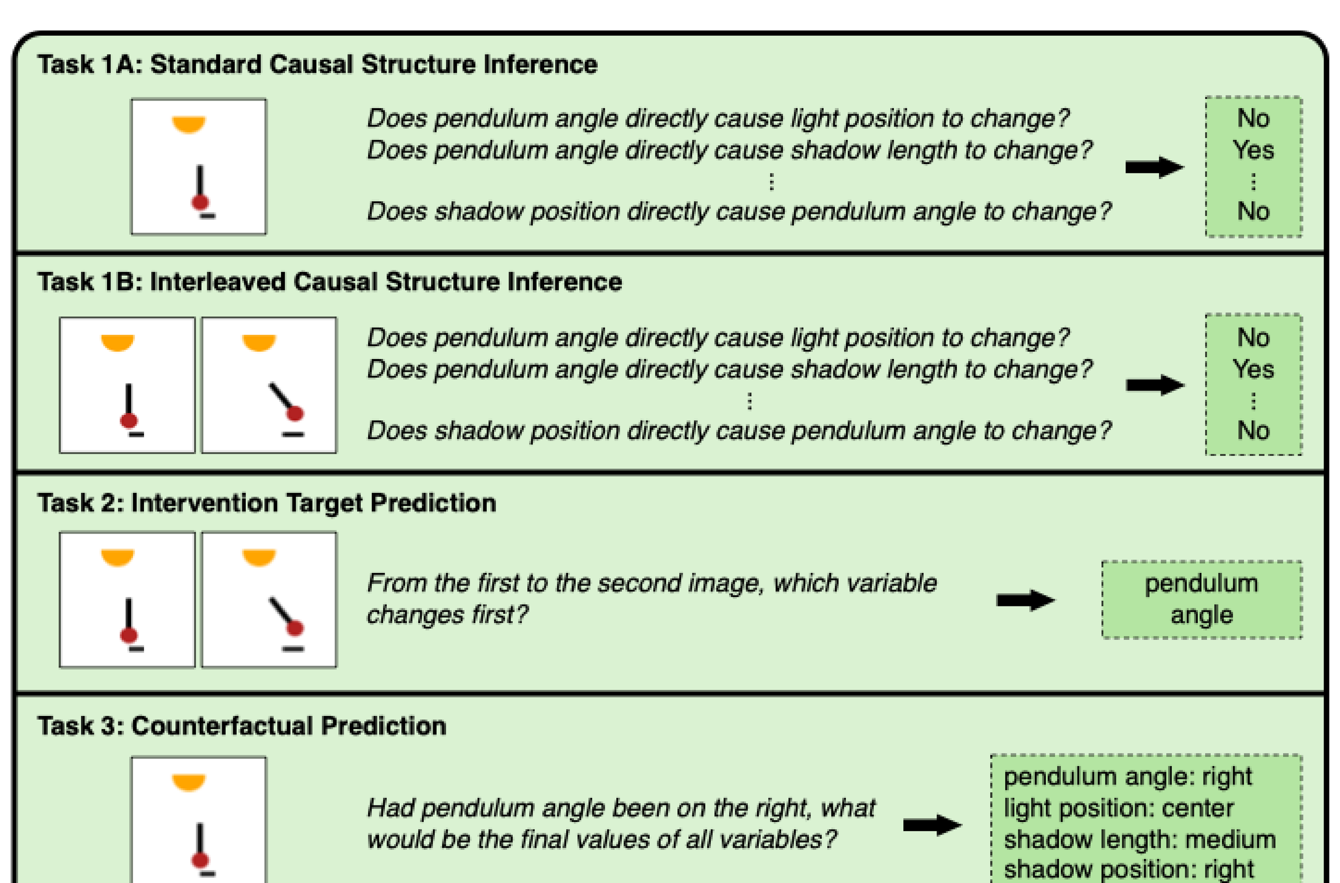


Pearl and Mackenzie. The Book of Why: The New Science of Cause and Effect. Basic Books. 2018.

Can large vision-language models perform formal visual causal reasoning as defined by Pearl's ladder of causation?

Datasets & Task Overview





CausalVLBench Framework

Task 1: Causal Structure Inference

- **Given**: an input image V or a pair of images $V = \{V_{\text{before}}, V_{\text{after}}\}$, a set of causal variables Z, and LVLM M_θ
- **Goal**: Infer the causal graph $G = (Z, E)$ where E is the set of directed edges such that $(Z_i \to Z_j) \in E$ indicates that Z_i is a direct cause of Z_j
  - Construct adjacency matrix $\hat{A}_{ij} = \mathbb{1}[\hat{Y}_{ij}] \in \{0,1\}$, Edge set: $E = \{(Z_i, Z_j) \in Z \times Z \mid \hat{A}_{ij} = 1\}$
- **Prompting**: For each pair (Z_i, Z_j) where $i \neq j$, construct input prompt as:

  Prompt: <V><*task description*><*description of causal variables*> Does Z_i directly cause Z_j to change?

  VLM output: Yes (No)
- **Evaluation**: Average Structural Hamming Distance (SHD) and accuracy for single and paired image settings

Task 2: Intervention Target Prediction

- **Given**: a pair of images $V = \{V_{\text{before}}, V_{\text{after}}\}$, a set of causal variables Z, causal graph G, and LVLM M_θ
- **Goal**: Given a pair of images, infer the *source intervention* that caused the change between the two images
- **Prompting:**

  Prompt: <$V_{\text{before}}, V_{\text{after}}$><*task description*><*description of causal variables Z and their relationships G*> From the first to the second image, which variable changes first?"

  VLM output: <*predicted source intervention variable*>
- **Evaluation**: Average accuracy of predicted source intervention for zero and few shot settings

Task 3: Counterfactual Prediction

- **Given**: an input image V, a set of causal variables Z, causal graph G, initial variable assignments, and LVLM M_θ
- **Goal**: Given an input image and initial variable assignments, infer the values of all variables had an intervention occurred
- **Prompting:**

  Prompt: <V><*task description*><*description of causal variables Z and their relationships G*><*factual state of all variables* $Z_1, Z_2, \ldots, Z_n$><*targeted intervention* $\mathbf{do}(Z_i = z_i^*)$>

  VLM output: <*counterfactual values of all variables*>
- **Evaluation**: Average accuracy of predicted counterfactual values for zero and few shot settings

Experimental Evaluation

Model	Pendulum (Naturally occurring)				Water Flow (Naturally occurring)				Causal Circuit (Induced)			
	Standard		Interleaved		Standard		Interleaved		Standard		Interleaved	
	SHD	Acc	SHD	Acc	SHD	Acc	SHD	Acc	SHD	Acc	SHD	Acc
LLaVA-OneVision-7B	$1.2_{0.01}$	$89.9_{0.06}$	$1.7_{0.02}$	$85.2_{0.14}$	$2.8_{0.01}$	$76.3_{0.09}$	$3.0_{0.00}$	$75.0_{0.00}$	$4.4_{0.03}$	$62.4_{0.24}$	$3.2_{0.01}$	$73.4_{0.10}$
Qwen-VL-Chat-9B	$1.0_{0.00}$	$83.1_{0.02}$	$0.9_{0.01}$	$87.9_{0.16}$	$2.0_{0.01}$	$74.7_{0.12}$	$2.9_{0.01}$	$68.1_{0.03}$	$3.0_{0.01}$	$74.5_{0.12}$	$2.9_{0.02}$	$75.7_{0.20}$
IDEFICS2-8B	$0.8_{0.01}$	$93.0_{0.07}$	$0.2_{0.00}$	$98.1_{0.04}$	$\mathbf{1.0_{0.00}}$	$\mathbf{91.5_{0.02}}$	$3.0_{0.00}$	$75.0_{0.00}$	$5.0_{0.00}$	$57.7_{0.08}$	$5.0_{0.00}$	$58.7_{0.02}$
Deepseek-VL2-Small	$4.0_{0.00}$	$66.6_{0.00}$	$3.7_{0.01}$	$69.1_{0.12}$	$3.0_{0.00}$	$75.0_{0.00}$	$3.0_{0.00}$	$75.0_{0.00}$	$5.0_{0.00}$	$58.3_{0.00}$	$4.9_{0.00}$	$58.8_{0.00}$
OpenFlamingo-9B	$4.0_{0.00}$	$66.6_{0.00}$	$4.0_{0.00}$	$67.6_{0.00}$	$3.0_{0.00}$	$75.0_{0.00}$	$3.0_{0.00}$	$75.0_{0.00}$	$5.0_{0.00}$	$58.3_{0.00}$	$5.0_{0.00}$	$58.3_{0.00}$
Otter-9B	$5.0_{0.00}$	$50.0_{0.00}$	$4.9_{0.01}$	$49.6_{0.12}$	$4.0_{0.00}$	$50.2_{0.00}$	$5.0_{0.00}$	$50.0_{0.03}$	$5.2_{0.02}$	$51.4_{0.18}$	$3.7_{0.00}$	$62.4_{0.20}$
Deepseek-VL2-27B	$4.0_{0.0}$	$66.7_{0.0}$	$4.0_{0.00}$	$66.7_{0.00}$	$3.0_{0.00}$	$75.0_{0.00}$	$3.0_{0.00}$	$75.0_{0.00}$	$5.0_{0.00}$	$58.3_{0.00}$	$5.0_{0.00}$	$58.3_{0.00}$
Qwen2.5-VL-Instruct-32B	$\mathbf{0.0_{0.00}}$	$\mathbf{100.0_{0.00}}$	$\mathbf{0.0_{0.0}}$	$\mathbf{100.0_{0.0}}$	$2.9_{0.01}$	$75.1_{0.04}$	$\mathbf{2.3_{0.0}}$	$\mathbf{80.1_{0.1}}$	$\mathbf{2.9_{0.03}}$	$\mathbf{75.5_{0.28}}$	$4.6_{0.0}$	$62.1_{0.1}$
Gemini-2.0-Flash	$\mathbf{0.0_{0.0}}$	$\mathbf{100.0_{0.0}}$	$0.7_{0.0}$	$94.4_{0.0}$	$\mathbf{1.0_{0.0}}$	$\mathbf{91.6_{0.0}}$	$\mathbf{2.3_{0.0}}$	$\mathbf{80.7_{0.0}}$	$3.2_{0.0}$	$73.2_{0.0}$	$\mathbf{2.8_{0.0}}$	$\mathbf{76.9_{0.0}}$

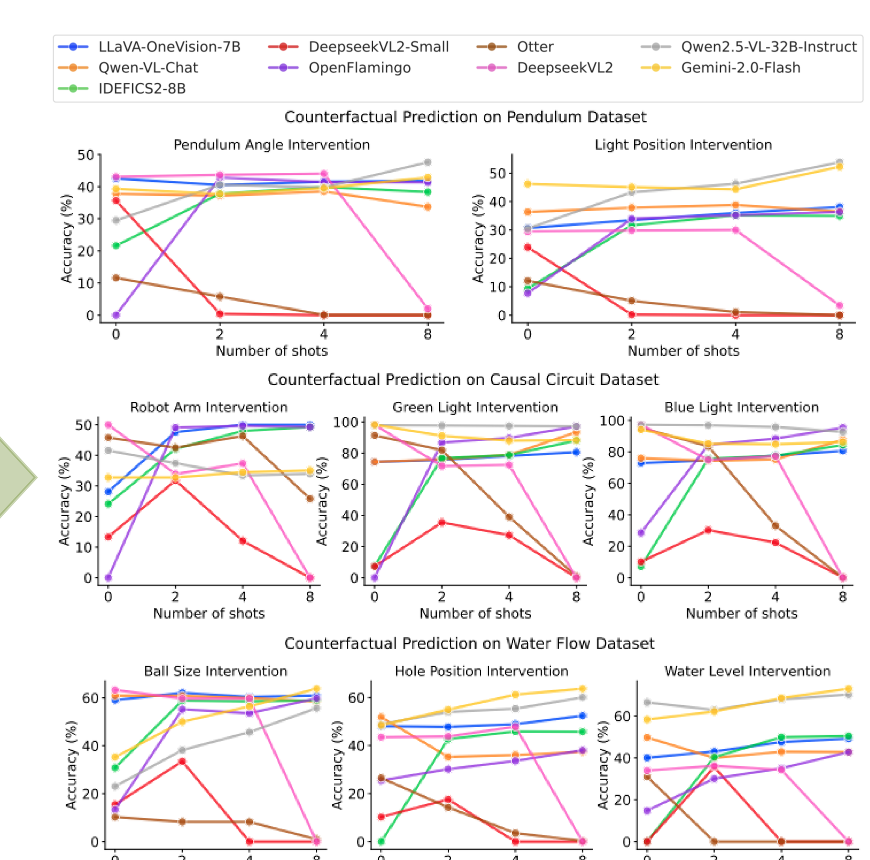
Key Observations

- Models predict more accurately for datasets based on naturally occurring causal systems
- Paired image setting tends to degrade model's ability to infer the causal structure
- Qwen2.5VL and Gemini-2.0-Flash perform best across all datasets
- IDEFICS and QwenVL-Chat show promising performance for Pendulum and Flow

Model	Pendulum				Water Flow				Causal Circuit			
	ZS	FS			ZS	FS			ZS	FS		
	0	2	4	8	0	2	4	8	0	2	4	8
LLaVA-OneVision-7B	$26.2_{1.5}$	$27.5_{1.9}$	$26.3_{1.0}$	$27.1_{0.9}$	$43.1_{0.8}$	$34.1_{2.3}$	$34.1_{1.2}$	$32.7_{1.2}$	$39.4_{0.5}$	$35.0_{0.4}$	$36.1_{0.5}$	$35.9_{0.4}$
Qwen-VL-Chat-9B	$24.9_{0.5}$	$24.8_{1.0}$	$24.3_{1.4}$	$24.7_{1.6}$	$37.8_{0.6}$	$33.1_{1.2}$	$32.9_{0.8}$	$32.1_{0.8}$	$10.4_{0.9}$	$31.0_{0.4}$	$31.8_{1.6}$	$33.0_{2.3}$
IDEFICS2-8B	$29.0_{0.4}$	$24.2_{1.9}$	$24.8_{0.9}$	$24.3_{1.1}$	$34.8_{2.1}$	$35.4_{1.8}$	$33.3_{0.3}$	$33.5_{0.8}$	$10.2_{0.4}$	$30.3_{1.2}$	$31.4_{0.9}$	$29.7_{0.5}$
Deepseek-VL2-Small-2.8B	$25.5_{1.1}$	$24.4_{0.4}$	$24.0_{0.3}$	$0.0_{0.0}$	$35.8_{0.6}$	$34.4_{0.2}$	$34.3_{0.7}$	$0.0_{0.0}$	$\mathbf{72.9_{1.1}}$	$28.1_{1.5}$	$0.2_{0.1}$	$0.0_{0.0}$
OpenFlamingo-9B	$24.8_{0.5}$	$24.7_{0.7}$	$23.7_{1.1}$	$25.2_{0.6}$	$34.2_{1.7}$	$34.5_{1.4}$	$33.0_{1.1}$	$33.1_{0.8}$	$9.8_{0.6}$	$31.6_{1.5}$	$31.9_{2.3}$	$32.3_{1.1}$
Otter-9B	$26.6_{1.9}$	$25.3_{0.3}$	$26.9_{0.4}$	$23.0_{1.2}$	$32.8_{1.1}$	$34.1_{1.0}$	$30.0_{0.9}$	$31.9_{0.9}$	$9.1_{0.7}$	$25.2_{1.4}$	$23.4_{1.4}$	$24.3_{1.4}$
Deepseek-VL2-27B	$31.9_{0.0}$	$30.4_{0.0}$	$24.1_{0.0}$	-	$44.4_{0.0}$	$36.6_{0.0}$	$31.4_{0.0}$	-	$66.1_{0.0}$	$\mathbf{43.7_{0.0}}$	$30.3_{0.0}$	-
Qwen2.5-VL-Instruct-32B	$\mathbf{44.3_{0.5}}$	$29.5_{0.3}$	$27.4_{2.0}$	$26.2_{1.2}$	$\mathbf{48.4_{0.7}}$	$37.2_{1.3}$	$37.3_{0.7}$	$36.6_{0.6}$	$32.1_{1.5}$	$32.5_{0.9}$	$32.0_{1.2}$	$34.6_{0.8}$
Gemini-2.0-Flash	$39.4_{0.0}$	$\mathbf{45.2_{0.0}}$	$\mathbf{45.3_{0.0}}$	$\mathbf{47.4_{0.0}}$	$37.6_{0.0}$	$\mathbf{46.5_{0.0}}$	$\mathbf{52.4_{0.0}}$	$\mathbf{55.7_{0.0}}$	$10.5_{0.0}$	$43.1_{0.0}$	$\mathbf{55.1_{0.0}}$	$\mathbf{66.1_{0.0}}$

Key Observations

- Overall, the best performing models are DeepseekVL2, Qwen2.5VL, and Gemini-2.0-Flash
- Gemini demonstrates the most clear upward trend as number of shots is increased
- Deepseek performs well on the more complex Causal Circuit dataset
- Most open-weight models perform quite poorly in this reasoning task



Key Observations

- Most models attain better results when interventions are on variables with no descendants, but struggle with accurately propagating causal effects to descendants
- LLaVA-OneVision-7B, Deepseek-VL2, Qwen2.5-VL, and Gemini-2.0-Flash show the best performance across datasets
- For variables with at least one descendant, we see a clear improvement as we increase the number of demonstrations

Acknowledgements: This work was supported in part by NSF 1910284, 1946391, 2119691, 2147375, and NIH P20GM139768.