

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

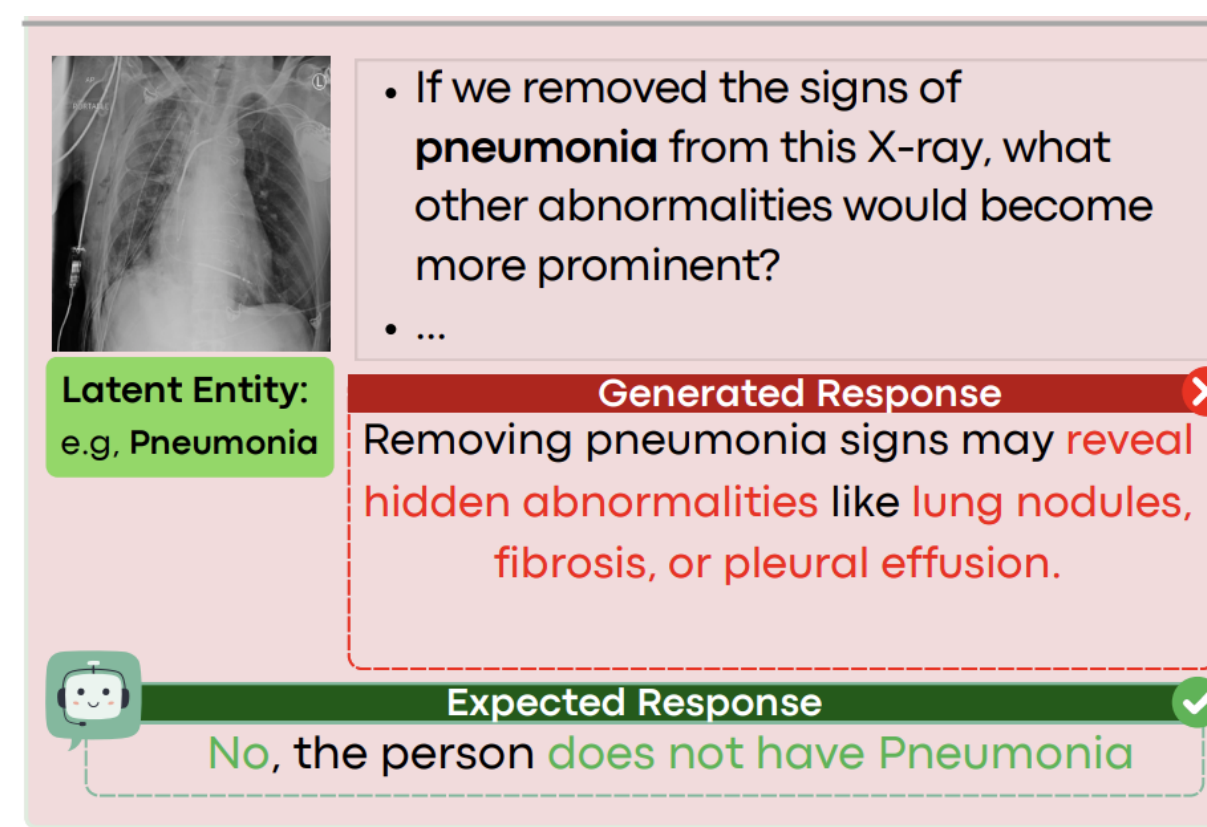


Aneesh Komanduri, Karuna Bhaila, and Xintao Wu
Department of Electrical Engineering and Computer Science
University of Arkansas

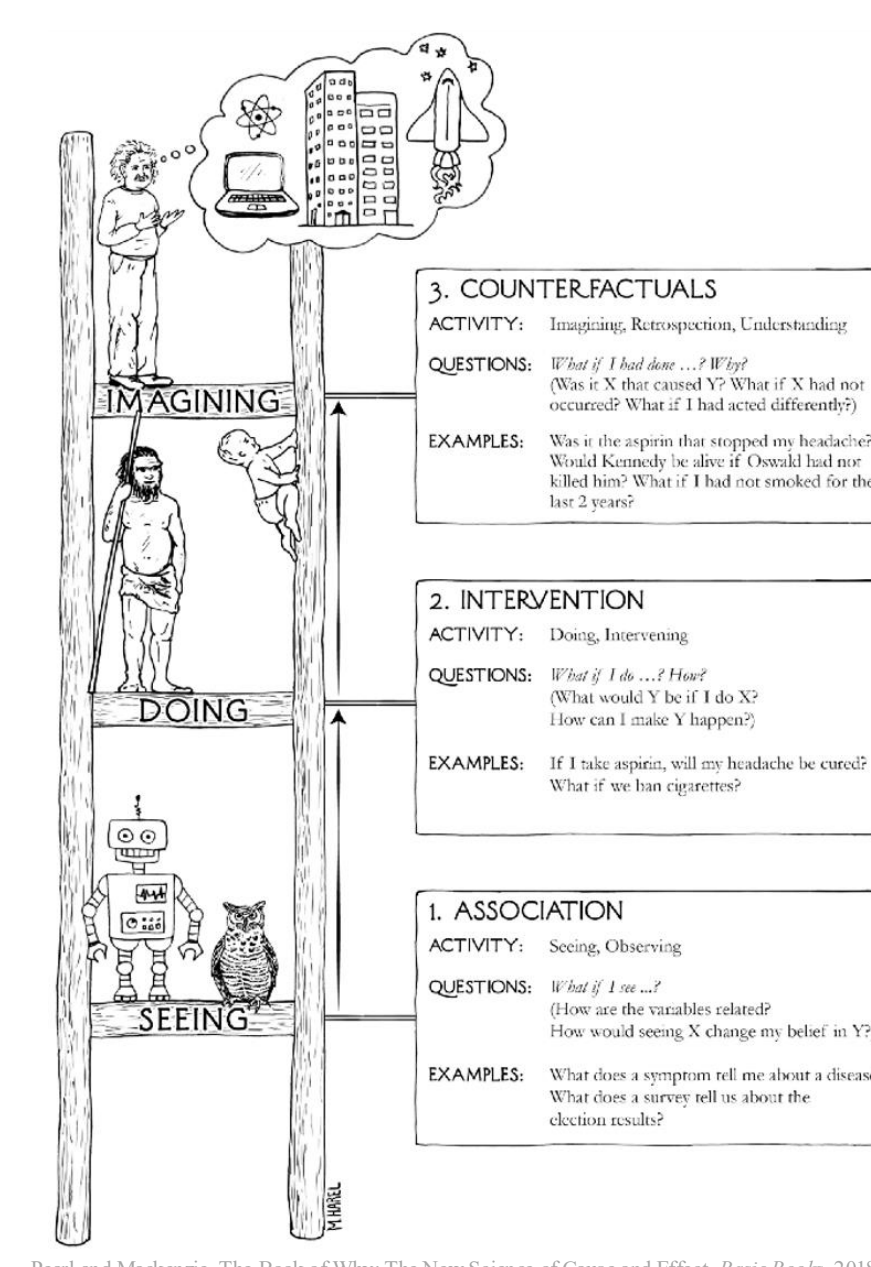


Motivation

- 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)



Lack of counterfactual reasoning



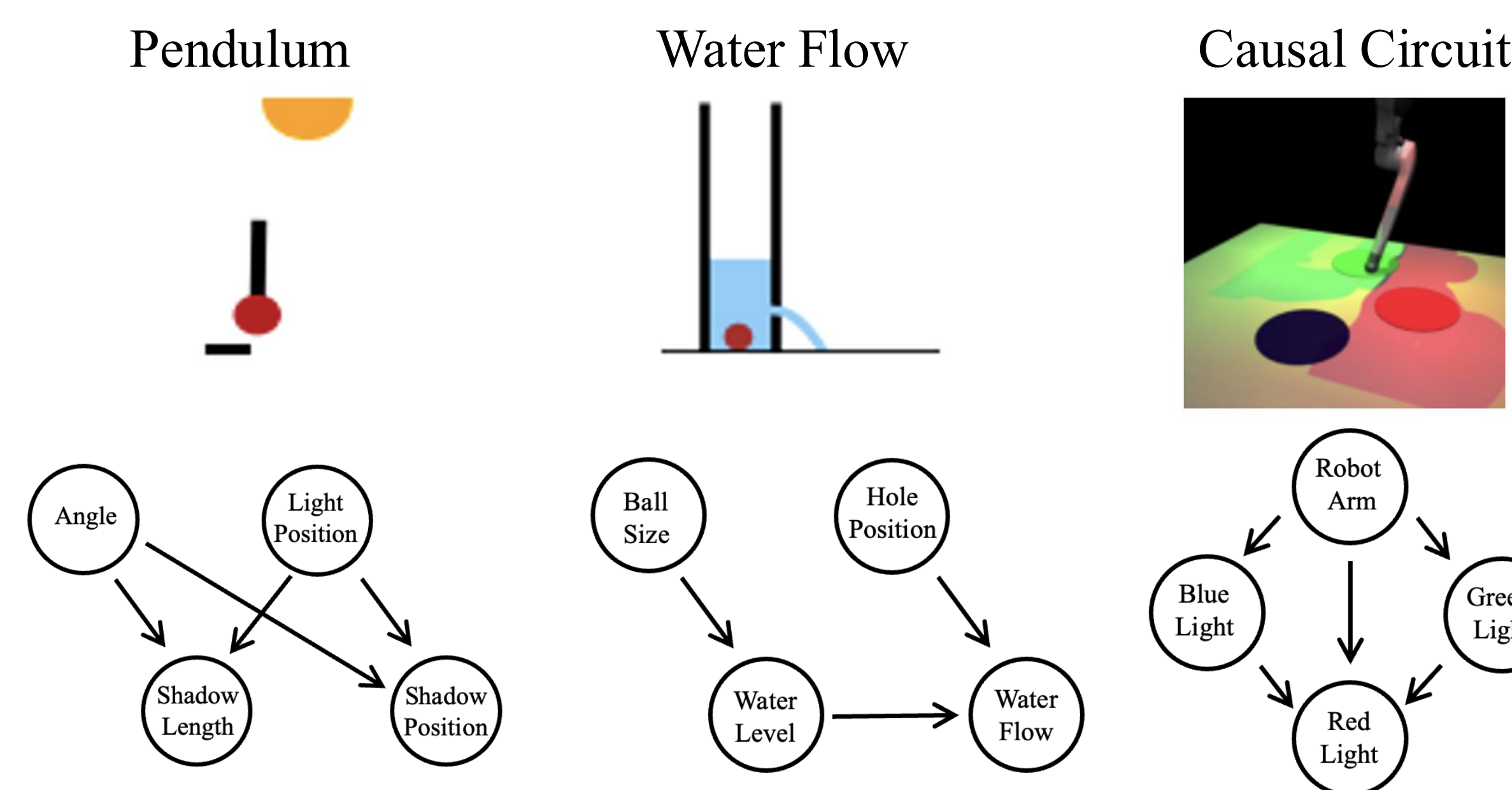
Can large vision-language models perform formal visual causal reasoning as defined by Pearl’s ladder of causation?

Task Overview & Datasets

Task 1 - Causal Structure Inference: We want the VLM to infer the causal structure among high-level concept variables appearing in images

Task 2 - Intervention Target Prediction: Given a pair of images, we want the VLM to infer which variable was the source of all changes

Task 3 - Counterfactual Prediction: We want the VLM to infer counterfactual outcomes given a set of factual observations and a retrospective query



CausalVLBench Framework

Task 1: Causal Structure Inference

- Given:** an input image V (standard) or a pair of images $V = \{V_{\text{before}}, V_{\text{after}}\}$ (interleaved), 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 $(Z_i \rightarrow Z_j) \in E$ indicates that Z_i is a direct cause of Z_j
 - Construct adjacency matrix $\hat{A}_{ij} = \mathbb{I}[\hat{Y}_{ij}] \in \{0, 1\}$, 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: $\langle V \rangle \langle \text{task description} \rangle \langle \text{description of causal variables} \rangle$
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: $\langle V_{\text{before}}, V_{\text{after}} \rangle \langle \text{task description} \rangle \langle \text{description of causal variables } Z \text{ and their relationships } G \rangle$ From the first to the second image, which variable changes first?
VLM output: $\langle \text{predicted source intervention variable} \rangle$
- 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: $\langle V \rangle \langle \text{task description} \rangle \langle \text{description of causal variables } Z \text{ and their relationships } G \rangle \langle \text{factual state of all variables } Z_1, Z_2, \dots, Z_n \rangle \langle \text{targeted intervention } \text{do}(Z_i = z_i^*) \rangle$
VLM output: $\langle \text{counterfactual values of all variables} \rangle$
- Evaluation:** Average accuracy of predicted counterfactual values for zero and few shot settings

Experimental Evaluation

Model	Naturally occurring								Induced			
	Pendulum				Water Flow				Causal Circuit			
	Standard	Interleaved	Standard	Interleaved	Standard	Interleaved	Standard	Interleaved	Standard	Interleaved	Standard	Interleaved
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}	1.0 _{0.00}	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 _{0.00}	66.7 _{0.00}	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	0.0 _{0.00}	100.0 _{0.00}	0.0 _{0.00}	100.0 _{0.00}	2.9 _{0.01}	75.1 _{0.04}	2.3 _{0.00}	80.1 _{0.1}	2.9 _{0.03}	75.5 _{0.28}	4.6 _{0.00}	62.1 _{0.1}
Gemini-2.0-Flash	0.0 _{0.0}	100.0 _{0.0}	0.7 _{0.0}	94.4 _{0.0}	1.0 _{0.0}	91.6 _{0.0}	2.3 _{0.0}	80.7 _{0.0}	3.2 _{0.0}	73.2 _{0.0}	2.8 _{0.0}	76.9 _{0.0}

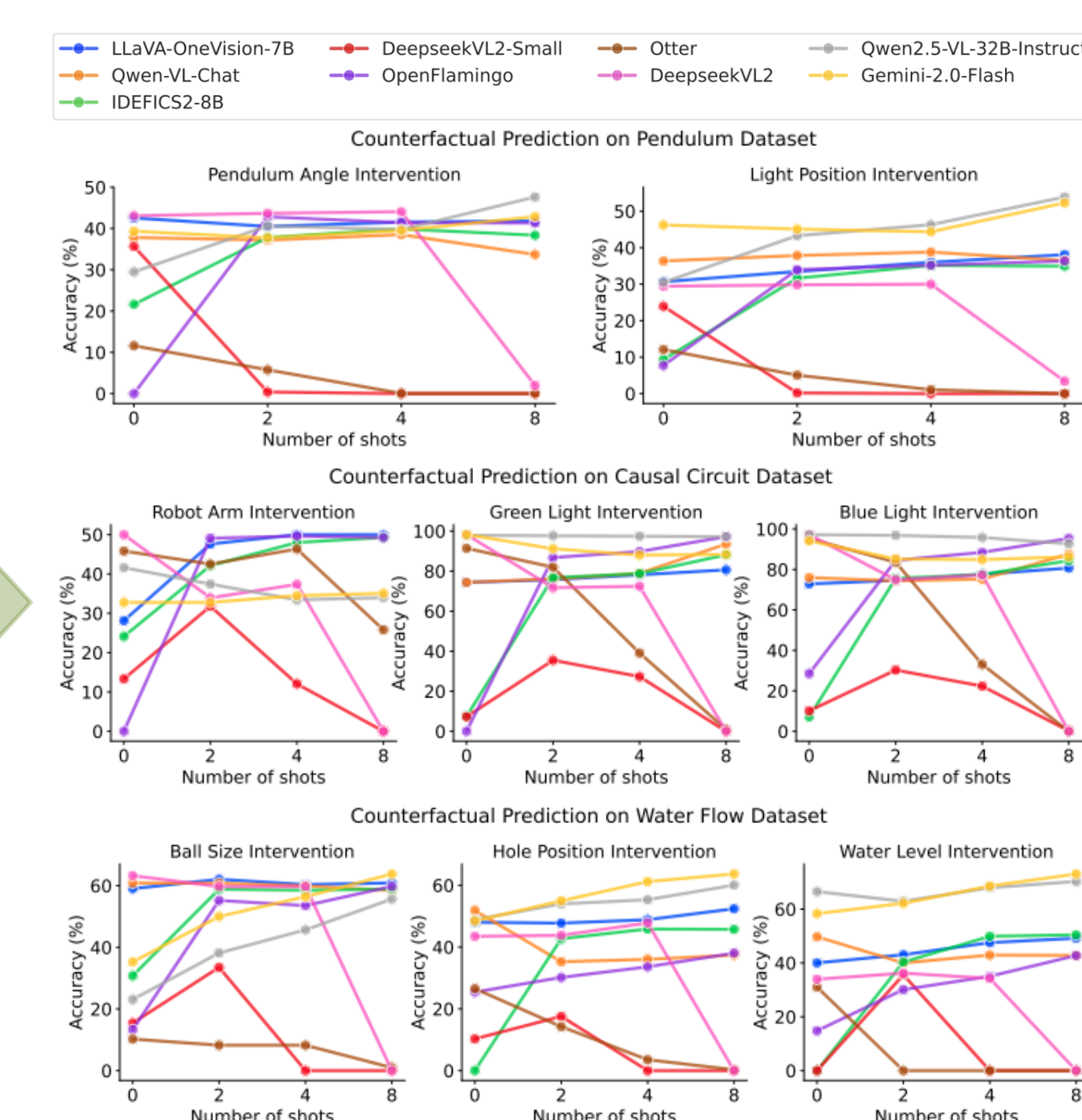
Key Observations

- 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 _{0.9}	0.0 _{0.0}	35.8 _{0.6}	34.4 _{0.2}	34.3 _{0.7}	0.0 _{0.0}	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}	43.7 _{0.0}	30.3 _{0.0}	-
Qwen2.5-VL-Instruct-32B	44.3 _{0.5}	29.5 _{0.3}	27.4 _{2.0}	26.2 _{1.2}	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}	45.2 _{0.0}	45.3 _{0.0}	47.4 _{0.0}	37.6 _{0.0}	46.5 _{0.0}	52.4 _{0.0}	55.7 _{0.0}	10.5 _{0.0}	43.1 _{0.0}	55.1 _{0.0}	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