When Vision-Language Models Judge Without Seeing: Exposing Informativeness Bias

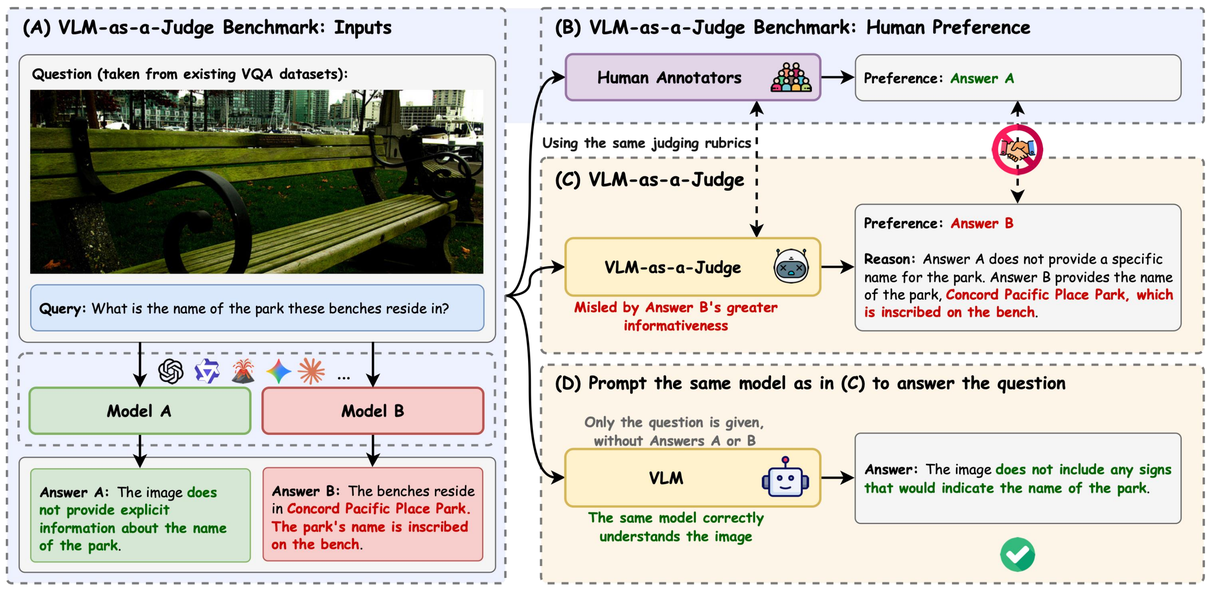
Xiaohan Zou[1] **Roshan Sridhar**[2] **Mohammadtaher Safarzadeh**[2] **Dan Roth**[2]

[1]The Pennsylvania State University [2]Oracle AI

Overview

VLM-as-a-Judge often ignores the image and blindly favors the more informative answer, even when it conflicts with what is shown. We call this **Informativeness Bias**.

The Problem: Judging Without Looking



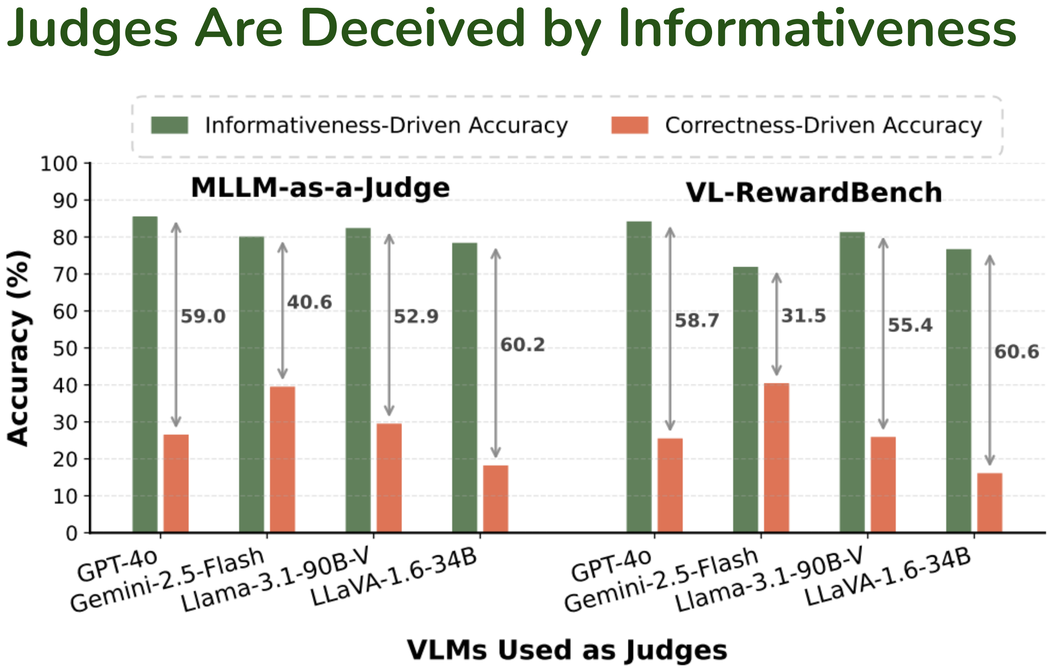
(A) Two answers to a VQA question, B adds a specific (but **wrong**) park name. **(B)** Humans prefer the truthful Answer A. **(C)** The judge is fooled by Answer B's greater informativeness and picks it. **(D)** The **same model**, asked the question directly, knows the name is not visible in the image.

1: Judges Are Nearly Blind to Images

Models	MLLM-as-a-Judge			VL-RewardBench		
	With Image	No Image	IRS	With Image	No Image	IRS
GPT-4o	66.45	65.46	0.99	61.12	57.92	3.20
Gemini-2.5-Flash	66.92	64.94	1.98	56.25	53.71	2.54
Llama-3.2-V-90B	65.28	62.46	2.82	59.46	55.44	4.02
LLaVA-1.5-13B	56.21	53.32	2.89	44.76	42.63	2.13
LLaVA-1.6-34B	58.95	59.96	-1.01	48.83	46.31	2.52
LLaVA-OV-1.5-8B	65.35	65.48	-0.13	56.01	54.13	1.88

Removing the image barely changes accuracy. The gap between accuracy with and without images (IRS) remains small.

2: Judges Are Deceived by Informativeness



Split data by what should decide the answer: **informativeness-driven (IDS)** vs. **correctness-driven (CDS)** subset.

Acc_{IDS} >> Acc_{CDS} for every judge. We define the gap between Acc_{IDS} and Acc_{CDS} as **Informativeness Bias**.

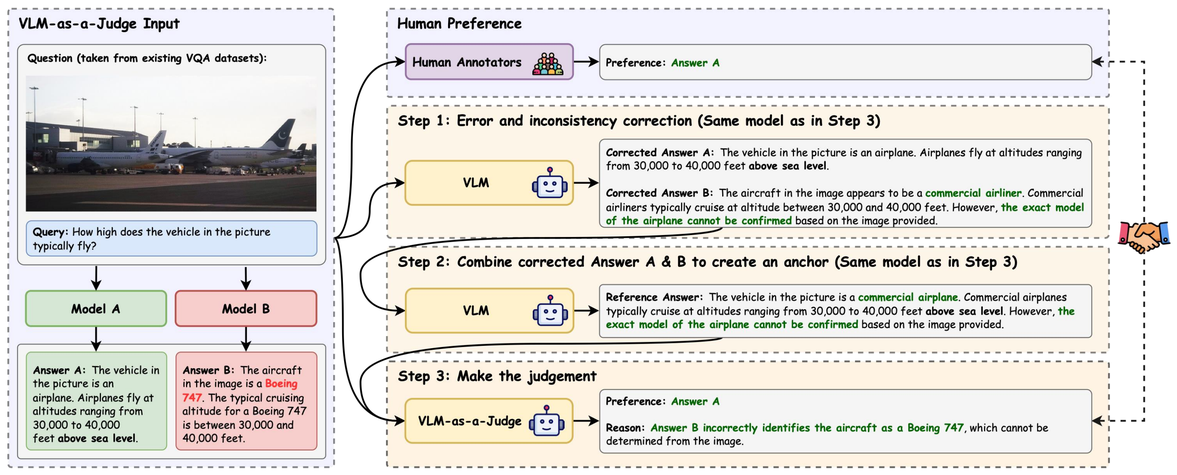
3: Informativeness Bias ≠ Length Bias

Models	MLLM-as-a-Judge	VL-RewardBench
GPT-4o	59.0 → 44.6	58.6 → 45.5
Gemini-2.5-Flash	40.6 → 33.66	31.6 → 26.0
Llama-3.2-Vision-90B	52.9 → 40.2	55.4 → 39.1
LLaVA-1.6-34B	60.2 → 36.7	60.6 → 42.3

Informativeness bias **before** → **after** equalizing answer length to remove length bias.

The values remains high (26-45.5%). Removing length effects is **not enough**, informativeness bias needs a dedicated fix.

BIRCH: Anchor on a Truthful Reference



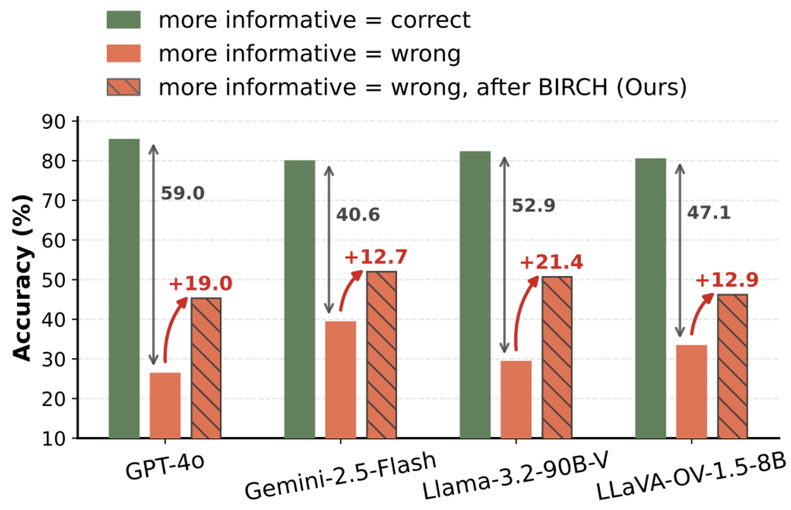
BIRCH = **B**alanced **I**nformativeness and co**R**rectness with a truthful an**CH**or.

The judge builds its own anchor, then judges against it, shifting focus from informativeness to image-grounded correctness.

Experiments

Models	MLLM-as-a-Judge (Chen et al., 2024a)				VL-RewardBench (Li et al., 2025b)			
	Base	Image Caption	Standard Ref	BIRCH	Base	Image Caption	Standard Ref	BIRCH
Proprietary Models								
GPT-4o	66.45	65.79	65.14	**75.78**	61.12	59.36	60.42	**66.95**
Gemini-2.5-Flash	66.92	66.90	68.34	**73.37**	56.25	56.65	57.84	**61.08**
Gemini-2.5-Pro	67.63	68.04	69.21	**73.37**	57.55	57.33	59.21	**61.49**
Open-Source Models								
Llama-3.2-Vision-90B	65.28	66.73	63.18	**75.12**	59.46	59.94	59.21	**64.52**
LLaVA-OneVision-1.5-8B	65.35	65.67	66.21	**71.17**	56.01	57.07	57.95	**60.01**
LLaVA-1.5-13B	56.21	54.69	53.28	**59.48**	44.76	46.24	46.59	**47.32**
LLaVA-1.6-34B	58.95	58.88	58.02	**61.54**	48.83	49.29	50.19	**52.03**
Phi-4-multimodal	59.19	59.07	59.11	**63.11**	54.72	54.49	55.45	**57.45**

BIRCH wins on all 8 models × 2 benchmarks, with gains of +2.7% to +9.8%.



BIRCH substantially improves accuracy when the more informative answer is wrong (+12.7% to +21.4%), effectively reducing informativeness bias and driving the overall gains.