Different VLMs trust different modalities, commit at different depths, and can be steered at inference.

Mechanistic Analysis and Inference-Time Control of Modality Conflict in VLMs

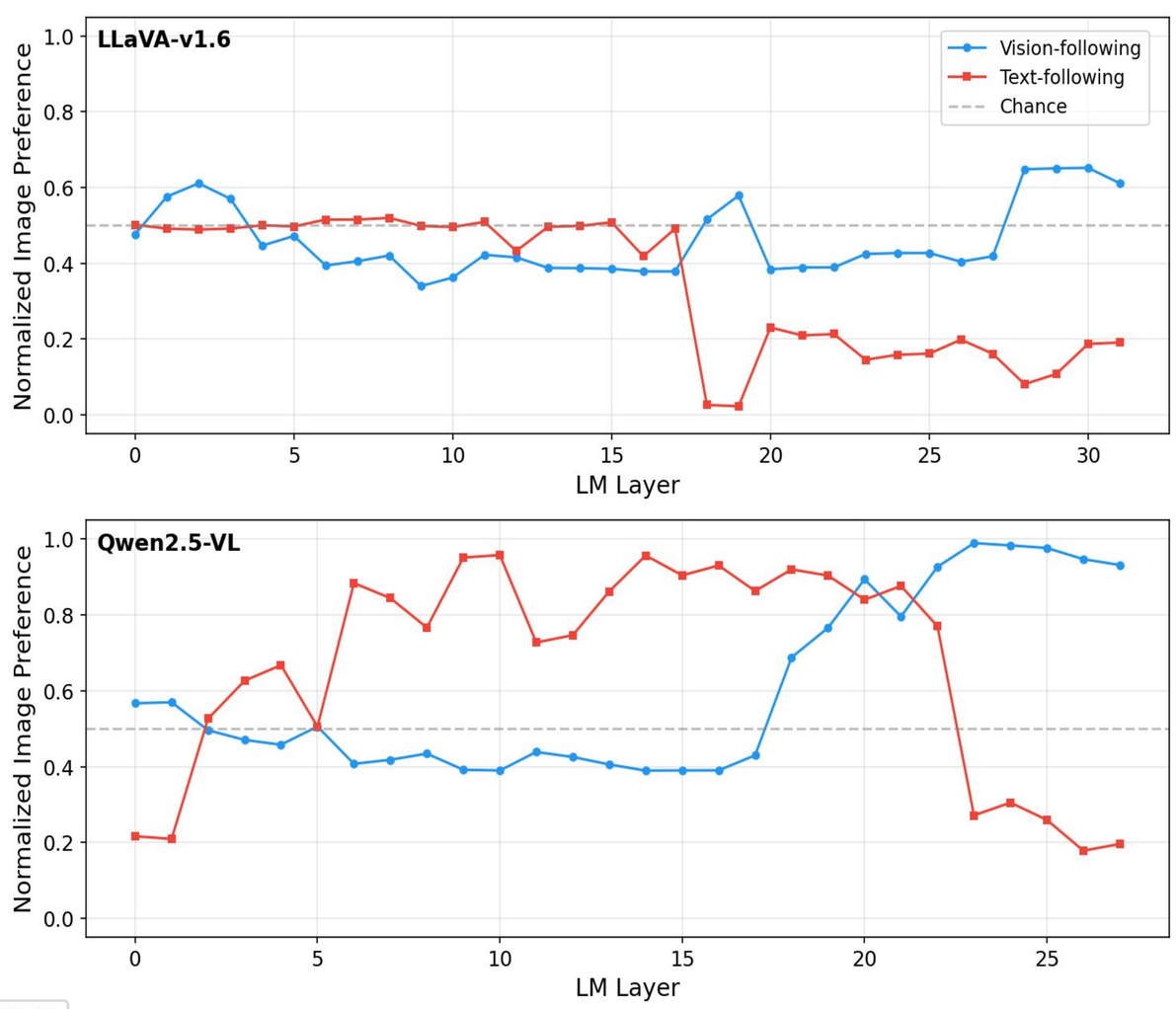
Question: When a Vision-Language Model (VLM) is given an image and text that contradict each other, which does it believe? At what point does it commit to a modality? And can we steer that choice?

We synthesized CLEVR-style dataset using blender, isolating contradictions between visual evidence and textual statements while maintaining fully verified ground truth. Each input pairs an image with a false declarative statement and a yes/no question whose correct answer is determined by the image alone.

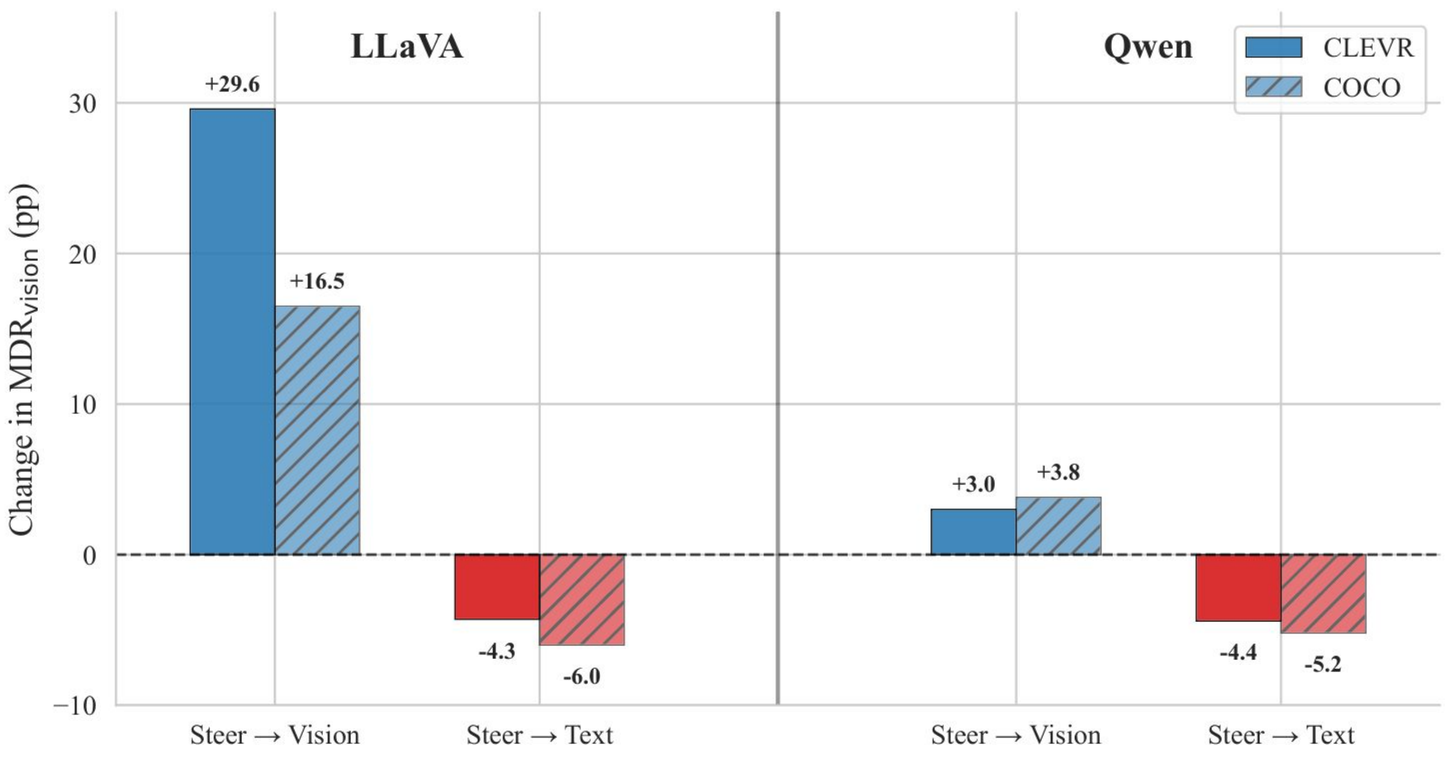
Task	LLaVA CLEVR	Qwen CLEVR
color	21.4% [14.8, 29.9]	98.3% [95.1, 99.4]
comparison	16.7% [4.7, 44.8]	87.0% [81.5, 91.0]
count	7.2% [4.1, 12.5]	92.1% [87.4, 95.2]
existence	27.5% [19.7, 36.8]	96.0% [92.3, 98.0]
identity	0.0% [0.0, 4.4]	73.0% [66.3, 78.7]
material	1.1% [0.2, 5.9]	72.4% [65.7, 78.2]
size	0.0% [0.0, 3.8]	61.3% [53.5, 68.5]
spatial	2.3% [0.6, 8.1]	59.7% [51.8, 67.2]
CLEVR Overall	**9.2% [7.3, 11.5]**	**80.7% [78.6, 82.7]**
COCO Overall	**38.4% [36.8, 40.1]**	**70.1% [68.5, 71.7]**

Table 1: Vision-following rate by task (±95% CI): LLaVA follows text, Qwen follows vision.

Result 1: LLaVA-v1.6 predominantly follows text, while Qwen2.5-VL predominantly follows vision.



Result 2: Commitment to a certain modality is determined by the specific "commitment" layers in the language model, with the mechanism being driven by MLP at that point.



Result 3: Using stateless, unit-normalized offset vectors at the commitment layers, we steer either model toward vision or text at inference, no retraining. LLaVA's vision-following jumps 9.2% → 38.8% (+29.6pp).

Limitation: The method tests two architectures, both of which utilize projection bridges. The dataset is synthesized, and contains exclusively yes/no type of questions.

PAPER

Jiahang He, Dia'a AL-Dweikat, Artur Shyutts, Abdullah AlDahoul, Oyedeji Sonuga, Kevin Zhu, Ruizhe Li, Anusha Mujumdar

Mechanistic Interpretability Workshop, ICML 2026