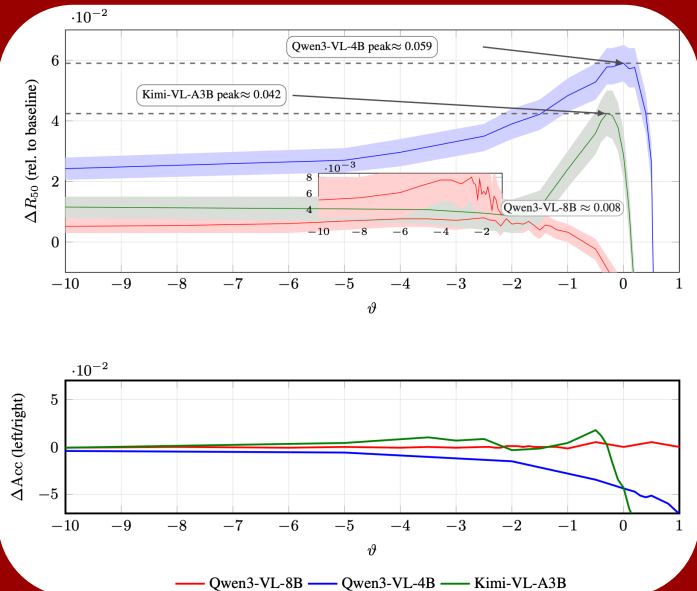
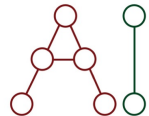


FLIP tells signal from noise in VLM interventions



FLIP: Final Layer Inference-Time Probing for Vision–Language Models

Drandreb Earl O. Juanico*, Rowel O. Atienza — University of the Philippines, Diliman



1 TL;DR

MAIN 3 TAKEAWAYS

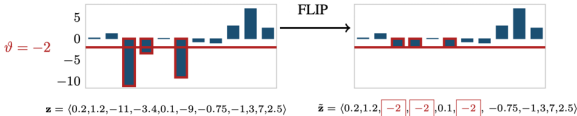
- Behavioral gains are **ambiguous** until FLIP validates task linkage.
- Only the normalized **Final site** passes all four tests.
- FLIP nominates regimes to trace; it does not steer.

2 Background / Motivating Question

- Open-weight VLMs expose logit-facing hidden states.
- Intervention gains can reflect instability, not mechanism.
- Which effects are structured and task-linked?
- Which sites merit circuit-level follow-up?

3 FLIP: Method & Protocol

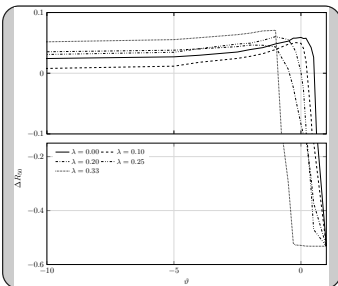
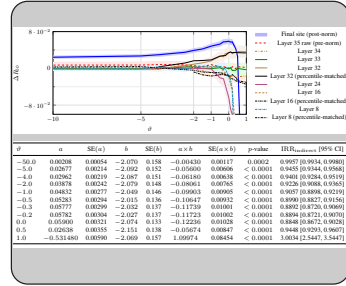
- Floor Final-site hidden activations with $\bar{z} = \max(\mathbf{z}, \vartheta)$ before logits.
- Keep model weights, prompts, decoding, and caches unchanged.
- Sweep ϑ to reveal negligible, interior, and collapse regimes.
- Validate regimes, not isolated performance gains.
- FLIP validates intervention effects through four checks: **regime structure**, **grounding alignment**, **feature coherence**, and **negative-control contrast**.



FLIP (definition used in all experiments). Given the post-normalization Final-site hidden state $\mathbf{z}$ and strength ϑ , FLIP outputs $\bar{\mathbf{z}} = \max(\mathbf{z}, \vartheta)$ element-wise before logit computation. No model weights, prompts, or decoding rules are changed.

4 Experimental Results

- Qwen3-VL-4B on 7,056 COCO pairs: ΔR_{50} peaks near 0.059; $\text{IRR}_{\text{indirect}} < 1$ supports grounding.
- Final-site sweeps recover the regime; matched raw-layer sweeps stall/collapse, ruling out threshold calibration.



- Figure 5: common-baseline permutation destroys the regime, showing feature-coherence dependence.
- Final-site flooring, clipping, and shifts show regimes; full protocol localizes specificity.

5 Limitations and Discussion

- FLIP points to targets; circuit discovery still starts afterward.
- Grounding-proxy alignment is suggestive, not causal sufficiency.
- Raw-layer failures are scoped contrasts; earlier mechanisms remain open.

- Code + data:
<https://anonymous.4open.science/r/flip-qwen3vl-EBAC>



PAPER
CODE