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

Jiahang He [1] **Dia'a AL-Dweikat** [1] **Artur Shyutts** [1 2] **Abdullah AlDahoul** [1] **Oyedeji Sonuga** [1] **Kevin Zhu** [1] **Ruizhe Li** [1 3] **Anusha Mujumdar** [1 4]

Abstract

When a vision-language model sees an image that contradicts a textual statement, which modality does it trust, and can that preference be controlled? Although prior work has answered this question behaviorally, we answer it mechanistically and causally. Behaviorally, we introduce CLEVR-CONFLICT, a controlled dataset of multimodal conflicts with verified ground truth, and show that LLaVA-v1.6 overwhelmingly follows text (9.2% vision-following), while Qwen2.5-VL predominantly follows vision (80.7%). Mechanistically, contribution patching identifies the bridge module as the causal entry point for visual information, and logit lens locates a mid-network commitment in LLaVA and a sharp late-layer flip in Qwen. Causally, MLP replacement at the commitment zone flips 73.3% of Qwen's text-following entries and 23.4% of LLaVA's, with the gap reflecting a distributed-vs-concentrated architectural distinction. We develop a stateless offset vector that achieves bidirectional inference-time control, pushing toward vision or toward text without retraining. Causal tests further reveal that the MLP is necessary but not sufficient. It enforces language priors only when conflicting text has activated them, functioning as a context-dependent amplifier rather than a standalone switch. Our findings generalize to natural images from COCO. Together, the findings offer a mechanistic account of modality conflict that moves from diagnosis to causal characterization to inference-time control.

[1]Algoverse AI Research [2]Graz University of Technology, 8010 Graz, Austria [3]University of Aberdeen, Aberdeen, Scotland, UK [4]Independent. Correspondence to: Jiahang He <jhe.primary@gmail.com>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

Vision-language models (VLMs) are increasingly deployed in domains where visual and textual evidence must be interpreted together, such as medical imaging reports, robotic perception, and document understanding (Shao et al., 2025; Jiang et al., 2025). A central reliability question is: when visual and textual information conflict, which modality does the model prioritize, and can that preference be understood and controlled? Past work has documented modality conflicts in VLMs, but these studies largely treat the phenomenon behaviorally, without identifying where in the model conflicts are resolved or why different architectures exhibit distinct behaviors (Pezeshkpour et al., 2025; Deng et al., 2025; Hua et al., 2025; Zhang et al., 2026). Prior interpretability work has begun to analyze internal components of vision–language models, including the vision encoder and the modules that connect visual tokens to the language model (Li et al., 2025c; Tartaglini et al., 2025). However, how these components behave under explicit multimodal conflict remains largely unexplored. The result is a gap between knowing that models fail under conflict and knowing why, or what to do about it.

We address this gap with three claims, each tested on LLaVA-v1.6 and Qwen2.5-VL. Our primary testbed is CLEVR-CONFLICT, a controlled dataset of multimodal conflicts we construct across eight visual reasoning tasks; we additionally generalize behavioral and mechanistic findings to natural images by constructing conflict entries from COCO (Lin et al., 2014).

Claim 1 (behavioral): *LLaVA-v1.6 and Qwen2.5-VL adopt opposing modality preferences under conflict, and the contrast reflects grounding behavior rather than perception failure.* On entries where each model answers correctly from vision alone, LLaVA overwhelmingly follows text, while Qwen overwhelmingly follows vision. The contrast holds on the shared intersection of items both models perceive correctly and reproduces on conflict entries constructed from COCO, ruling out per-model filtering and synthetic-data artifacts.

Claim 2 (mechanistic): *Conflict is resolved inside the language model at architecturally distinct commitment layers, not at the visual input stage.* Contribution patching shows

the bridge module fully transmits visual information under conflict, and null or negative effects from bridge scaling and attention reweighting rule out the visual pathway as the locus of failure. Logit lens then locates a mid-network modality commitment in LLaVA and a late-network flip in Qwen, with both layers reproducing on COCO. The commitment timing reflects each model's LM backbone rather than the input distribution.

Claim 3 (causal): *MLP computation at the commitment layers is the decisive mechanism, functioning as a context-dependent amplifier rather than a standalone switch.* Replacing commitment-zone MLP outputs with their clean-run counterparts flips most of Qwen's text-following entries and a substantial fraction of LLaVA's, with the gap reflecting a distributed-vs-concentrated architectural distinction between the two models. Injecting conflict-run MLP outputs into clean forward passes flips almost no entries, showing that the MLP is causally necessary for text-following but inert without conflicting text to amplify. A stateless offset vector applied at the commitment layers leverages this mechanism for bidirectional inference-time control.

Together these three claims move from observing that models fail under conflict, to understand why at the level of specific computations, to demonstrating that those computations can be controlled. This matters directly for the domains where VLMs are increasingly deployed: in medical imaging, a model that blindly follows a radiologist's note on what the scan shows is unreliable; in robotic perception, a model that follows verbal instructions over visual evidence may act unsafely. Bidirectional control over modality preference, the ability to dial toward vision or toward text depending on deployment context, transforms modality dominance from a fixed architectural behavior into an adjustable parameter, opening a path toward more reliable and context-aware VLM deployment. Code is released at `https://github.com/jasonhe346/modality-conflict-vlm`.

2. Related Work

Hallucination and Modality Conflict in VLMs. Prior work has shown that VLMs often generate outputs inconsistent with visual input, a phenomenon commonly described as visual hallucination (Li et al., 2023; Liu et al., 2024a; Huang et al., 2024). Several studies attribute this behavior to strong language priors that can override visual evidence, particularly when models encounter ambiguous or conflicting multimodal signals (Liu et al., 2024b; Cao et al., 2025; Wu et al., 2026; Favero et al., 2024). As a more controlled diagnostic, more recent work has begun to study modality dominance directly by constructing inputs where visual and textual evidence explicitly disagree (Deng et al., 2025; Hua et al., 2025; Zhang et al., 2025c; Nguyen et al., 2025b). However, most of these studies focus on behavioral evaluation and do not explain where within the model modality conflicts are resolved. Our work goes beyond behavioral measurement by pairing controlled conflict evaluation with mechanistic analysis, causal intervention, and inference-time control.

Mechanistic Interpretability of VLMs. A growing body of work applies interpretability techniques to analyze how VLMs integrate visual and textual information (Neo et al., 2024; Li et al., 2025a; Papadimitriou et al., 2025; Nikankin et al., 2025). Prior studies probe attention patterns, hidden representations, and cross-modal alignment to better understand multimodal reasoning and hallucination behavior (Li et al., 2025b; Shu et al., 2025; Zhang et al., 2025b). Other work has examined projection or bridge modules that connect vision encoders to language models, characterizing them as information bottlenecks between modalities (Li et al., 2025c). Despite these insights, existing analyses rarely study these components under explicit multimodal conflict conditions or localize when answers are formed during decoding. We extend this line of work by applying causal contribution patching and logit lens analysis to VLMs under controlled modality conflict.

Activation Steering. Activation steering, the injection of vectors into a model's residual stream at inference time, has demonstrated robust control over sentiment, sycophancy, and safety-relevant behaviors in LLMs (Rimsky et al., 2023; Arditi et al., 2024; Korznikov et al., 2025). Transfer to VLMs has followed, with applications spanning hallucination mitigation, safety alignment, and general behavioral control (Parekh et al., 2025; Nguyen et al., 2025a; Shi et al., 2026). Most directly related to our work, MC^2 and AMPS steer modality preference in MLLMs via representation-level intervention without additional fine-tuning (Zhang et al., 2025a; Huang et al., 2026). However, these approaches and others select injection layers empirically, without causal mechanistic justification for where or why the intervention works. Unlike prior work, we target the specific component and layers identified through causal analysis under explicit modality conflict, and demonstrate bidirectional inference-time control through MLP steering.

3. Experimental Setup and Behavioral Results

We study the following problem: given a VLM, an image I, and a text prompt containing a statement that directly contradicts I, which modality does the model trust? Each input pairs an image with a false declarative statement and a yes/no question whose correct answer is determined by the image alone.

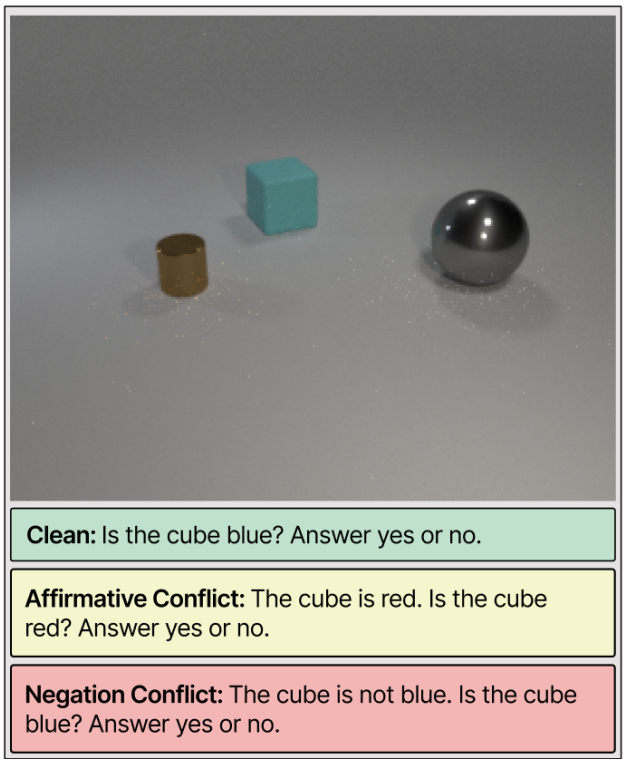

Figure 1. Each scene is paired with three prompt variants: a clean question, a false affirmation, and a false negation. Ground truth is defined by the image in all cases.

3.1. CLEVR-CONFLICT & COCO Generalization

Most existing conflict benchmarks rely on naturally occurring image-text pairs where conflict degree is hard to verify, making it difficult to distinguish modality preference from perception failure. CLEVR-CONFLICT addresses this directly: a synthetic dataset designed to isolate contradictions between visual evidence and textual statements while maintaining fully verified ground truth. We generate 200 photorealistic scenes using Blender's Cycles renderer with randomized object placement. Metadata files store all object attributes and spatial relationships, enabling programmatic question generation and exact ground-truth verification. We define eight binary visual reasoning tasks: count, color, existence, spatial, identity, material, size, and comparison, with yes/no questions generated programmatically from metadata. As illustrated in Figure 1, each conflict entry prepends a false declarative statement to a clean question, creating a direct contradiction between text and image-grounded answer. We balance two conflict directions equally: *IN-TY* (text incorrectly asserting a fact the image contradicts) and *IY-TN* (text incorrectly denying a true visual fact), yielding 1,600 conflict entries in total.

To ensure that each conflict entry tests modality interference rather than visual perception failure, we first evaluate both models on the clean questions without conflicting text, retaining only entries that each model answers correctly using vision alone. We then balance conflict directions by downsampling the larger group per task type, ensuring equal representation of both conflict directions. After filtering, LLaVA retains 738 entries across 8 task types, while Qwen retains 1,458 entries across the same 8 tasks. Evaluation subsets differ because filtering is performed per model. We intentionally condition on examples with correct visual grounding to isolate robustness to textual interference, rather than enforce a shared input distribution or compare absolute performance. To verify that this design does not drive the behavioral contrast between models, we evaluate both models on their shared intersection of correctly-answered entries (N=674, 93.5% of LLaVA's filtered set) and report consistent results in Appendix B.

To assess whether behavioral patterns generalize beyond synthetic scenes, we also generate conflict entries from COCO images and instance annotations (Lin et al., 2014). We focus on existence, count, and spatial tasks, chosen because they are the three task types supportable by COCO annotations with programmatic ground truth verification equivalent to CLEVR-CONFLICT. We do not replicate the full mechanistic pipeline on COCO; instead, we use it to verify that those core behavioral findings, commitment timing, and intervention transfer are not an artifact of synthetic data. After baseline filtering, LLaVA retains 3,314 entries and Qwen retains 3,212 entries, balanced by conflict direction.

3.2. Models and Evaluation

We evaluate two widely used 7B-scale vision–language models: LLaVA-v1.6-vicuna-7B (CLIP ViT-L/14 24-layer vision encoder, two-layer MLP projector bridge, 32-layer LLaMA-2 language model) and Qwen2.5-VL-7B-Instruct (32-layer Qwen2.5 vision encoder, linear PatchMerger bridge, 28-layer Qwen2.5 language model) (Liu et al., 2023; Yang et al., 2024). Both models use a thin bridge module rather than cross-attention to connect the vision encoder to the LM, creating a single architectural interface that is well-suited for mechanistic analysis. The models are comparable in scale but differ substantially in vision encoder depth, bridge design, and language model backbone, making them a well-suited pair for examining whether modality dominance patterns are architecture-dependent. For all experiments, we use greedy decoding with a maximum output length of 50 tokens. Each conflict response is classified as *vision* (matches image answer), *text* (matches false statement answer), or *neither*. We then report Vision Modality Dominance Rate ($\mathrm{MDR}_{\mathrm{vision}}$) with Wilson 95% confidence intervals.

Table 1. $\text{MDR}_{\text{vision}}$ with Wilson 95% confidence intervals. Per-task rows are CLEVR-CONFLICT only; COCO overall rows are from conflict entries constructed on COCO images and serve as a naturalistic generalization check. LLaVA is text-dominant throughout while Qwen is vision-dominant. Both models shift toward moderation on natural images, where conflict prompts are weaker by design than CLEVR-CONFLICT's controlled false declaratives.

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]**

3.3. Behavioral Results

LLaVA and Qwen exhibit sharply contrasting modality preferences. As shown in Table 1, LLaVA overwhelmingly follows text across all task types (overall 9.2% vision-following), with no consistent task-based gradient. Qwen predominantly follows vision (80.7%), with a clear hierarchy: perceptual tasks are most robust, while relational reasoning (spatial 59.7%) is most vulnerable. Since these results are measured only on entries that each model answered correctly from vision alone, LLaVA's text dominance reflects a failure to maintain visual grounding under linguistic interference rather than a failure of visual perception. Decomposing by conflict direction reveals a systematic negation asymmetry in Qwen: vision-following reaches 94–100% on *IN-TY* entries, but drops sharply on *IY-TN* entries for relational tasks (spatial 26%, size 25%). Since *IY-TN* corresponds to false negations (text incorrectly denying a true visual fact) and *IN-TY* to false affirmations (text incorrectly asserting a fact the image contradicts), these results indicate that false negations act as a stronger override signal than false affirmations (see Appendix Table 7).

On COCO, both models move towards a more moderate position with LLaVA's $\text{MDR}_{\text{vision}}$ rising to 38.4% and Qwen's dropping to 70.1%. This shift is consistent with CLEVR-CONFLICT's design as a maximally controlled stress test: its false declaratives are unambiguous, directly contradictory, and prepended in a fixed format that produces stronger linguistic interference than naturalistic prompts or images. The synthetic setting amplifies modality conflict by design, making commitment patterns and intervention effects cleaner to isolate. The task hierarchy for Qwen is preserved for both: existence hardest to override, followed by count and spatial, mirroring CLEVR-CONFLICT's perceptual-to-relational gradient.

4. Mechanistic Analysis

The behavioral results establish that LLaVA and Qwen adopt opposing modality strategies, but leave open *where* and *when* conflict is resolved internally. We answer both questions in turn: contribution patching localizes the causal entry point for visual information, and logit lens identifies the layer at which each model commits to a modality.

4.1. Contribution Patching

For each layer K we construct a *clean run* (image + original question) and a *corrupt run*. In the corrupt run, the false statement is prepended and vision embeddings are replaced with mean embeddings over 20 random CLEVR images to neutralize visual signal while preserving sequence structure. We inject only what layer K adds to the residual stream into an otherwise corrupt forward pass:

$$\hat{h} = h_{\text{in}}^{\text{corrupt}} + h_{\text{out},K}^{\text{clean}} - h_{\text{in},K}^{\text{clean}} \tag{1}$$

Recovery is then computed as:

$$\text{Recovery} = \frac{\delta_{\text{patched}} - \delta_{\text{corrupt}}}{\delta_{\text{clean}} - \delta_{\text{corrupt}}} \tag{2}$$

where $\delta = P(\text{image answer}) - P(\text{text answer})$. Recovery $= 1$ means full restoration of visual grounding; recovery $= 0$ means the layer has no causal effect. Bridge modules are not residual sites; we replace their full output with the clean run's bridge output instead.

Results. Contribution patching shows that the bridge module is the dominant causal entry point for visual information under conflict. As shown in Figure 2, patching the bridge fully restores visual grounding (recovery $= 1$) in both models for all tasks. This is not a trivial result: the conflicting textual statement remains present throughout the patched run, so recovery $= 1$ means that clean visual tokens alone are sufficient to restore visual grounding despite conflicting text. In contrast, vision encoder and language model layers outside the bridge have near-zero or low average recovery. Some vision encoder layers even exhibit negative recovery. This likely occurs because inserting an isolated clean contribution into an otherwise corrupt forward pass can disrupt downstream computations rather than partially recover the visual answer. Language model layers also remain low relative to the bridge with maximum recovery of $+0.410$ for LLaVA, indicating that no single LM layer acts as a dominant causal site for visual grounding.

4.2. Logit Lens

Contribution patching identifies *where* visual information enters the LM. Logit lens reveals *when* during LM process-

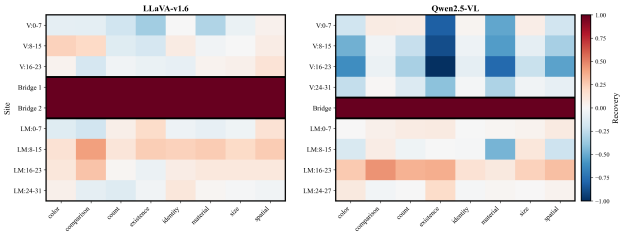

Figure 2. Contribution patching mean recovery by layer zone and task for LLaVA-v1.6 and Qwen2.5-VL. Each cell shows how much visual grounding is recovered when a single zone's clean contribution is injected into an otherwise corrupt forward pass. Rows are grouped into vision encoder layers, bridge module, and language model layers.

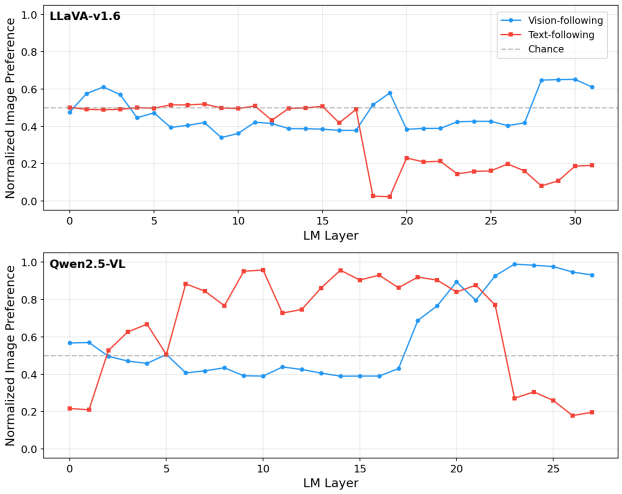

Figure 3. Normalized image preference by LM layer for LLaVA-v1.6 and Qwen2.5-VL. Each line traces the relative strength of the image-grounded answer at each layer, split by final response: blue for vision-following entries, red for text-following entries. Values near 1.0 indicate image answer dominance, near 0.0 indicate text answer dominance; dashed line marks chance (0.5).

ing the answer is formed. At each layer K, we take the hidden state of the final token, corresponding to the model's output decision, and project it through final layer normalization and language model head to obtain probabilities $P(\text{yes})$ and $P(\text{no})$. We then compute a normalized image preference (NIP) metric:

$$\text{NIP}(K) = \frac{P(\text{image_answer})}{P(\text{image_answer}) + P(\text{text_answer})} \quad (3)$$

This metric captures the relative preference for the visual answer at each layer. It is computed for all valid conflict entries, producing a layerwise trajectory of image preference per entry.

LLaVA results. As shown in Figure 3, both vision and text-following entries maintain normalized image preference near chance (≈ 0.5) from layers 0–17 in aggregate, with text-following entries diverging at layers 17–18 to stabilize around 0.2 by the final layer, while vision-following entries rise steadily to ≈ 0.65. The aggregate trajectory suggests a mid-network decision point at which language priors actively suppress the visual signal in text-following cases. As we show below, however, this aggregate description hides substantial per-entry variation.

Qwen results. Qwen reveals a strikingly different and, to our knowledge, previously unreported pattern. Through layers 5–21, text-following entries maintain *higher* normalized image preference (≈ 0.75–0.95) than vision-following entries (≈ 0.4), which remain *below* chance for most of the network. This means that entries Qwen ultimately gets wrong are not characterized by early text dominance—the visually-grounded answer is in fact the stronger signal through most of processing for text-following entries. The decisive event is a late-layer flip concentrated at layers 22–23 (quantified per entry below), where text-following entries drop sharply from ≈ 0.85 to ≈ 0.2 by the final layer, while vision-following entries rise to ≈ 0.93. This late override accounts for the 19.3% of cases where Qwen chooses to follow text.

Per-entry flip distribution. The aggregate trajectories above describe population means; per-entry trajectories may differ substantially. We define an entry's *flip layer* per text-following entry using two complementary metrics: the *max-drop* layer, which identifies the steepest local crossover and is robust to late-layer noise (used for cross-dataset comparison), and the *threshold-stable* layer (smallest L where $\text{NIP}(L) < 0.3$ and remains below through the final layer), which identifies persistent commitment. Appendix Figure 5 and Appendix Table 8 report distributions and summary statistics.

Median max-drop layers are stable across data distributions to within ± 2 layers for both models (LLaVA: $L = 16$ on CLEVR-CONFLICT, $L = 18$ on COCO; Qwen: $L = 23$ on both). This per-entry transfer matches the aggregate trajectory transfer above and provides quantitative support for commitment timing as a property of each model's LM backbone rather than of the input distribution. The threshold-stable metric on CLEVR-CONFLICT, however, reveals that this shared cross-dataset stability conceals a sharp architectural difference in *how* commitment is distributed within each model.

For Qwen, the per-entry distribution is essentially a step function at $L = 23$, with 69% of text-following entries flipping within ± 1 layer of the median. LLaVA shows a qualitatively different pattern. Flip layers span the LM with a *trimodal* structure: 31% commit early ($L = 6$–10), 35% mid-network ($L = 16$–20), and 29% late ($L = 26$–31). The aggregate "mid-network commitment" described above is therefore a population average that conceals three distinct

subpopulations: LLaVA does not commit at a specific layer but distributes commitment across three timing modes. This contrast of a tight step function for Qwen as compared to a distributed trimodal pattern for LLaVA supports the distributed-vs-concentrated architectural distinction further explored through MLP replacement results in Section 5.2.

5. Intervention Study

5.1. Ruling Out Alternative Mechanisms

Before targeting MLP computation directly, we systematically eliminate three alternative explanations for language-prior dominance.

Bridge magnitude. Contribution patching established the bridge as the main causal entry point, so we first test whether the visual signal arriving there is the problem. The bridge output delta norm between clean and conflict passes is 0.000. Thus, the bridge processes image features independently of text and its output is identical under both conditions. To test whether a stronger visual signal could overcome the text prior even though the bridge is functioning correctly, we scale bridge outputs by 1.0–5.0. $\text{MDR}_{\text{vision}}$ decreases monotonically from 9.2% to 2.6%, confirming that visual token magnitude is not the bottleneck.

Attention routing. With the bridge ruled out, we ask whether the LM is failing to attend to the visual tokens it received correctly. We measure the difference in attention weight allocated to visual token positions between clean and conflict runs across all layers, focusing on the commitment layers identified by logit lens (Section 4.2). The maximum difference is $+0.000021$ for LLaVA and $+0.000070$ for Qwen, negligible in both cases. To further test this mechanism, we add a positive bias to the pre-softmax attention logits for visual token positions, forcing the model to attend more strongly to visual tokens at the commitment layers. $\text{MDR}_{\text{vision}}$ remains unchanged from baseline. Both models attend to visual tokens equally under clean and conflict conditions; attention routing is not the mechanism.

Residual stream steering. With both the visual input and attention routing eliminated, we attempt a direct LM intervention: injecting steering vectors directly into the residual stream. We construct steering vectors using two contrast methods: outcome-group contrast (mean hidden states of vision-following vs. text-following entries) and paired clean-vs-conflict hidden states, the latter producing a $2.6\times$ stronger signal (see Appendix Figure 6). Sweeping injection layers and α values, both methods show clear layer specificity with peak effects at layers 14–17 for LLaVA and controls near baseline, independently validating the logit lens commitment point. However, paired contrast achieves only 19.6% ($+10.4$pp) at α=20 (see Appendix Figure 7). Directly modifying the residual stream produces only modest gains even at high α values, motivating a more targeted intervention at the level of individual sublayer computations.

Table 2. Unified cross-architecture comparison. MLP replacement and offset gains are relative to baseline $\text{MDR}_{\text{vision}}$. Causal necessity reports the fraction of text-following entries that flip to vision-following after MLP replacement; causal sufficiency reports the fraction of vision-following entries that flip to text-following after MLP injection.

Metric	LLaVA-v1.6	Qwen2.5-VL
Baseline $\text{MDR}_{\text{vision}}$	9.2%	80.7%
Default modality preference	Text-dominant	Vision-dominant
Commitment point (CLEVR & COCO)	L17–18	L22–23
Bridge output under conflict	Identical	Identical
Attention gap under conflict	No difference	No difference
MLP/Attn ratio at commitment	$1.7\times$	$2.8\times$
MLP replacement, commitment zone	$+21.7$pp	$+14.2$pp
MLP replacement, all layers	$+43.4$pp	$+18.4$pp
MLP offset toward vision (CLEVR)	$+29.6$pp	$+3.0$pp
MLP offset toward text (CLEVR)	-4.3pp	-4.4pp
MLP offset toward vision (COCO)	$+16.5$pp	$+3.8$pp
MLP offset toward text (COCO)	-6.0pp	-5.2pp
Causal necessity (vision)	23.4%	73.3%
Causal sufficiency (text)	0.0%	0.08%

5.2. MLP as the Decisive Mechanism

Diagnostic. We capture the difference in MLP and attention sublayer output activations between clean and conflict runs at the final token position, measuring how much each component's output changes when conflict text is introduced. As seen in Appendix Table 5, LLaVA MLP and attention deltas are comparable through layers 0–16. At the commitment point, MLP dominance emerges sharply, with the MLP/attention delta ratio reaching $1.7\times$ at layer 17, $3.9\times$ at layer 18, rising to $7.8\times$ at layer 31. For Qwen, deltas remain comparable through layers 0–21, then jump sharply at the commitment point: $2.82\times$ at layer 22, with post-commitment layers ratios of $3.22\times$. In both models, the transition co-localizes exactly with the logit lens commitment point: NIP divergence, residual steering peak, and MLP/attention ratio all converge on the same zone.

MLP replacement upper bound. To establish an upper bound on how much visual grounding MLP computation is responsible for, we replace the MLP sublayer output at the final token position with the clean-run counterpart, revealing a key architectural difference between the two models (see Appendix Table 6). For LLaVA, recovery is layer-additive with minimal regression: the commitment zone (layers 14–17) contributes $+21.7$pp in $\text{MDR}_{\text{vision}}$, post-commitment layers an additional $+21.7$pp, indicating distributed MLP reinforcement. Replacing all 32 layers raises $\text{MDR}_{\text{vision}}$ from 9.2% to 52.6%. For Qwen, the pattern is concentrated

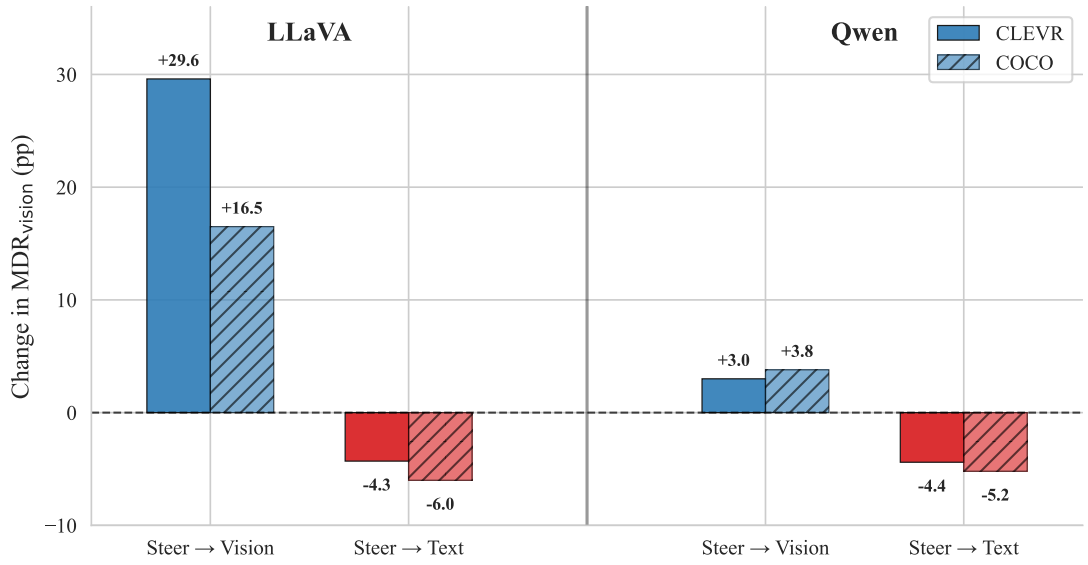

Figure 4. MLP offset toggle results for both models across both datasets. Steering toward vision or text at the commitment layers achieves bidirectional modality control without retraining. Text-direction transfer is comparable across both models and datasets, confirming the intervention is not distribution-specific.

rather than distributed: replacing just two layers (22–23) raises $\text{MDR}_{\text{vision}}$ from 80.7% to 94.9%. Replacing all 28 layers reaches 99.1% (see Table 2).

MLP offset toggle. We apply stateless unit-normalized offset vectors at the commitment layers of each model to push toward either modality at inference time without retraining. As seen in Figure 4, LLaVA (layers 14–17) steering toward vision raises $\text{MDR}_{\text{vision}}$ from 9.2% to $\mathbf{38.8}\%$ ($+29.6$pp); steering toward text lowers it to $\mathbf{4.9}\%$ (-4.3pp). For Qwen (layers 22–23), steering toward vision raises $\text{MDR}_{\text{vision}}$ from 80.7% to $\mathbf{83.7}\%$ ($+3.0$pp); steering toward text lowers it to $\mathbf{76.3}\%$ (-4.4pp). In both models, vectors derived from CLEVR-CONFLICT transfer to COCO with both directions intact: LLaVA achieves $+16.5$pp (vision) and -6.0pp (text); Qwen achieves $+3.8$pp (vision) and -5.2pp (text). Qwen's smaller vision-direction gains are consistent with its concentrated mechanism leaving less room for improvement from a high baseline, but text-direction transfer is comparable across both models.

Causal necessity test. To test whether the commitment-layer MLP is causally necessary for text-following, we take each text-following entry and replace the conflict-run MLP output at the commitment layers with the clean-run counterpart, then check whether the model flips to vision-following. Qwen flips in $\mathbf{73.3}\%$ of cases (206/281), while LLaVA flips in only $\mathbf{23.4}\%$ (157/670). The difference is directly predicted by the distributed vs. concentrated architecture distinction: Qwen's text-following behavior is almost entirely concentrated in two layers, so replacing those two layers catches most cases. LLaVA's prior accumulates gradually across 32 layers, so replacing only 4 catches a fraction of the total. This difference provides intervention-based validation that the two architectures are mechanistically distinct.

Sufficiency test. To test whether the MLP alone is sufficient to cause text-following, we take each vision-following entry and inject conflict-run MLP outputs at the commitment layers into an otherwise clean forward pass. Neither model flips: LLaVA produces $\mathbf{0/68}$ text-following responses even when all 32 layers are replaced; Qwen flips $\mathbf{1/1177}$. The MLP is therefore causally necessary but not sufficient for text-following, cross-architecture. Conflict MLP activations are inert without conflicting text present to activate the language prior. The MLP functions as a context-dependent amplifier, not a standalone switch.

6. Conclusion

We presented CLEVR-CONFLICT and a systematic mechanistic investigation of modality conflict in VLMs. Behaviorally, LLaVA-v1.6 and Qwen2.5-VL adopt opposing modality strategies. Through contribution patching and logit lens analysis, we identified the bridge module as the primary causal entry point for visual information and showed that conflict resolution is driven by commitment-layer MLP computation. Since bridge and attention are already functioning correctly under conflict, fixing modality bias requires targeting the MLP directly. The MLP offset toggle demonstrates this is achievable at inference time without retraining; learned MLP adapters represent a promising path toward

full recovery.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Arditi, A., Obeso, O., Syed, A., Paleka, D., Rimsky, N., Gurnee, W., and Nanda, N. Refusal in language models is mediated by a single direction. *ArXiv*, abs/2406.11717, 2024. URL https://api.semanticscholar.org/CorpusID:270560489.

Cao, J., Chen, Z., Wang, Z., Ma, L., Luo, W., and Qi, G.-J. When images speak louder: Mitigating language bias-induced hallucinations in vlms through cross-modal guidance. *ArXiv*, abs/2510.10466, 2025. URL https://api.semanticscholar.org/CorpusID:282058119.

Deng, A., Cao, T., Chen, Z., and Hooi, B. Words or vision: Do vision-language models have blind faith in text? *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3867–3876, 2025. URL https://api.semanticscholar.org/CorpusID:276775193.

Favero, A., Zancato, L., Trager, M., Choudhary, S., Perera, P., Achille, A., Swaminathan, A., and Soatto, S. . Multi-modal hallucination control by visual information grounding. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14312, 2024. URL https://api.semanticscholar.org/CorpusID:268553740.

Hua, T., Yun, T., and Pavlick, E. How do vision-language models process conflicting information across modalities? *ArXiv*, abs/2507.01790, 2025. URL https://api.semanticscholar.org/CorpusID:280134019.

Huang, W., Liu, H., Guo, M., and Gong, N. Z. Visual hallucinations of multi-modal large language models. *ArXiv*, abs/2402.14683, 2024. URL https://api.semanticscholar.org/CorpusID:267782634.

Huang, Z., Li, X., Surana, R., Yu, T., Wang, R., McAuley, J. J., Shang, J., and Wu, J. Amps: Adaptive modality preference steering via functional entropy. *ArXiv*, abs/2602.12533, 2026. URL https://api.semanticscholar.org/CorpusID:285606992.

Jiang, S., Wang, Y., Song, S., Zhang, Y., Meng, Z., Lei, B., Wu, J., Sun, J., and Liu, Z. Omnivmed: Scaling medical vision-language model for universal visual understanding. *ArXiv*, abs/2504.14692, 2025. URL https://api.semanticscholar.org/CorpusID:277955299.

Korznikov, A., Galichin, A. V., Dontsov, A., Rogov, O. Y., Oseledets, I. V., and Tutubalina, E. The rogue scalpel: Activation steering compromises llm safety. *ArXiv*, abs/2509.22067, 2025. URL https://api.semanticscholar.org/CorpusID:281659011.

Li, L., Qu, J., Zhou, Y., Qin, Y., Yang, T., and Zhao, Y. Treble counterfactual vlms: A causal approach to hallucination. In *Conference on Empirical Methods in Natural Language Processing*, 2025a. URL https://api.semanticscholar.org/CorpusID:276902483.

Li, Q., Ye, Z., Feng, X., Zhong, W., Ma, W., and Feng, X. Causal tracing of object representations in large vision language models: Mechanistic interpretability and hallucination mitigation. *ArXiv*, abs/2511.05923, 2025b. URL https://api.semanticscholar.org/CorpusID:282911662.

Li, W., Tang, R., Li, C., Zhang, C., Vuli'c, I., and Søgaard, A. Lost in embeddings: Information loss in vision-language models. In *Conference on Empirical Methods in Natural Language Processing*, 2025c. URL https://api.semanticscholar.org/CorpusID:281315785.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and rong Wen, J. Evaluating object hallucination in large vision-language models. In *Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://api.semanticscholar.org/CorpusID:258740697.

Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. URL https://api.semanticscholar.org/CorpusID:14113767.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *ArXiv*, abs/2304.08485, 2023. URL https://api.semanticscholar.org/CorpusID:258179774.

Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R.-Z., and Peng, W. A survey on hallucination in large vision-language models. *ArXiv*, abs/2402.00253, 2024a. URL https:

//api.semanticscholar.org/CorpusID:267365472.

Liu, S., Zheng, K., and Chen, W. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. *ArXiv*, abs/2407.21771, 2024b. URL https://api.semanticscholar.org/CorpusID:271570970.

Neo, C., Ong, L., Torr, P., Geva, M., Krueger, D., and Barez, F. Towards interpreting visual information processing in vision-language models. *ArXiv*, abs/2410.07149, 2024. URL https://api.semanticscholar.org/CorpusID:273233293.

Nguyen, D., Prasad, A., Stengel-Eskin, E., and Bansal, M. Grains: Gradient-based attribution for inference-time steering of llms and vlms. *ArXiv*, abs/2507.18043, 2025a. URL https://api.semanticscholar.org/CorpusID:280017051.

Nguyen, T., Michaels, J., Fiterau, M., and Jensen, D. D. Challenges in understanding modality conflict in vision-language models. *ArXiv*, abs/2509.02805, 2025b. URL https://api.semanticscholar.org/CorpusID:281091983.

Nikankin, Y., Arad, D., Gandelsman, Y., and Belinkov, Y. Same task, different circuits: Disentangling modality-specific mechanisms in vlms. *ArXiv*, abs/2506.09047, 2025. URL https://api.semanticscholar.org/CorpusID:279260817.

Papadimitriou, I., Su, H., Fel, T., Saphra, N., Kakade, S. M., and Gil, S. Interpreting the linear structure of vision-language model embedding spaces. *ArXiv*, abs/2504.11695, 2025. URL https://api.semanticscholar.org/CorpusID:277824176.

Parekh, J., Khayatan, P., Shukor, M., Dapogny, A., Newson, A., and Cord, M. Learning to steer: Input-dependent steering for multimodal llms. *ArXiv*, abs/2508.12815, 2025. URL https://api.semanticscholar.org/CorpusID:280676751.

Pezeshkpour, P., Aminnaseri, M., and Hruschka, E. Mixed signals: Decoding vlms' reasoning and underlying bias in vision-language conflict. In *Conference on Empirical Methods in Natural Language Processing*, 2025. URL https://api.semanticscholar.org/CorpusID:277781807.

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. M. Steering llama 2 via contrastive activation addition. *ArXiv*, abs/2312.06681, 2023. URL https://api.semanticscholar.org/CorpusID:266174252.

Shao, R., Li, W., Zhang, L., Zhang, R., Liu, Z., Chen, R., and Nie, L. Large vlm-based vision-language-action models for robotic manipulation: A survey. *ArXiv*, abs/2508.13073, 2025. URL https://api.semanticscholar.org/CorpusID:280677209.

Shi, Y., Yang, S., and Liu, D. One-shot optimized steering vector for hallucination mitigation for vlms. *ArXiv*, abs/2601.23041, 2026. URL https://api.semanticscholar.org/CorpusID:285241101.

Shu, Y., Lin, H., Liu, Y., Zhang, Y., Zeng, G., Li, Y., ZHOU, Y., Lim, S.-N., Yang, H., and Sebe, N. When semantics mislead vision: Mitigating large multimodal models hallucinations in scene text spotting and understanding. *ArXiv*, abs/2506.05551, 2025. URL https://api.semanticscholar.org/CorpusID:279244407.

Tartaglini, A. R., Grant, S., Wurgaft, D., Potts, C., and Fan, J. E. Diagnosing bottlenecks in data visualization understanding by vision-language models. *ArXiv*, abs/2510.21740, 2025. URL https://api.semanticscholar.org/CorpusID:282389282.

Wu, J., Shi, W., Shen, H., Qi, P., Tang, K., Huang, Z., Wang, B., and Yang, Z. Revis: Sparse latent steering to mitigate object hallucination in large vision-language models. 2026. URL https://api.semanticscholar.org/CorpusID:285540291.

Yang, Q. A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y.-C., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z., Quan, S., and Wang, Z. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024. URL https://api.semanticscholar.org/CorpusID:274859421.

Zhang, C., Ding, W., Liu, J., Wu, M., Wu, Q., and Mooney, R. Do images speak louder than words? investigating the effect of textual misinformation in vlms. *ArXiv*, abs/2601.19202, 2026. URL https://api.semanticscholar.org/CorpusID:285071566.

Zhang, Y., Ma, J., Hou, Y., Bai, X., Chen, K., Xiang, Y., Yu, J., and Zhang, M. Evaluating and steering modality preferences in multimodal large language model. *ArXiv*, abs/2505.20977,

2025a. URL https://api.semanticscholar.org/CorpusID:278910799.

Zhang, Z., Wang, T., Gong, X., Shi, Y., Wang, H., Wang, D., and Hu, L. When modalities conflict: How unimodal reasoning uncertainty governs preference dynamics in mllms. *ArXiv*, abs/2511.02243, 2025b. URL https://api.semanticscholar.org/CorpusID:282749400.

Zhang, Z., Zhou, W., Zhao, J., and Li, H. Robust multimodal large language models against modality conflict. *ArXiv*, abs/2507.07151, 2025c. URL https://api.semanticscholar.org/CorpusID:280150090.

A. Limitations

Several limitations bound the current findings and point to directions for future work. First, the cross-architecture claim rests on two 7B thin-bridge models; both LLaVA and Qwen use linear projection bridges rather than cross-attention. Models with multi-layer cross-attention bridges (e.g., Flamingo, BLIP-2) may exhibit fundamentally different mechanisms. Second, all experiments use binary yes/no VQA; open-ended generation under conflict may involve different commitment dynamics, and the commitment-layer story may be specific to forced-choice settings. Third, the conflict format is artificial: prepending an explicit false declarative is a controlled but unnatural form of interference. Although the mechanism transfers to COCO, the conflict format remains the same—subtler real-world conflicts, such as misleading captions or implicit contradictions, are not evaluated. We view these as natural directions for future work, and hope the mechanistic framework introduced here provides a foundation for broader investigation across architectures, tasks, and conflict types.

B. Shared Intersection Robustness Check

Motivation. The main paper evaluates LLaVA and Qwen on different filtered subsets (738 and 1,458 entries respectively), as filtering is performed per model to isolate modality interference from visual perception failure. A potential concern is that this asymmetry drives the behavioral contrast between models. We address this by identifying the shared intersection of entries both models answer correctly on the clean question and re-evaluating behavioral $\text{MDR}_{\text{vision}}$ on this common subset.

Intersection size. The shared intersection contains 674 entries, representing 93.5% of LLaVA's filtered set. This high overlap indicates LLaVA's filtered set is nearly a subset of Qwen's as almost every entry LLaVA answers correctly from vision alone, Qwen also answers correctly, reflecting Qwen's higher baseline visual accuracy across tasks.

Table 3. $\text{MDR}_{\text{vision}}$ on per-model filtered subsets (main paper) vs. shared intersection. The behavioral contrast between models is preserved on identical items.

	Per-model filtered	**Shared intersection**
LLaVA	9.2% [7.3, 11.5] (N=738)	8.3% [6.4, 10.8] (N=674)
Qwen	80.7% [78.6, 82.7] (N=1,458)	72.6% [69.1, 75.8] (N=674)
Gap	71.5pp	64.3pp

Results. LLaVA's $\text{MDR}_{\text{vision}}$ is essentially unchanged (9.2% vs. 8.3%), confirming strong text dominance on shared items. Qwen's drops modestly (80.7% to 72.6%), expected because the shared subset excludes entries that only Qwen answers correctly on the clean question, which are harder perceptual cases where Qwen's stronger visual grounding is most robust under conflict. The 64.3pp gap between models on identical items confirms the behavioral contrast is not an artifact of per-model filtering.

C. Additional Data Tables

Table 4. Cross-architecture × cross-dataset commitment timing. Commitment layers identified by logit lens are preserved in all four combinations, confirming the commitment point is an architectural property of each model's LM backbone.

	CLEVR (synthetic)	**COCO (natural)**
LLaVA	Commits L17–18	Commits L17–18 ✓
Qwen	Commits L22–23	Commits L22–23 ✓

Table 5. MLP vs. attention delta ratio at selected layers for both models. MLP dominance emerges sharply at the commitment layer in each model, co-localizing with the independently identified logit lens commitment point.

Stage	LLaVA (layer)	Qwen (layer)
Early	0.8× (L5)	0.7× (L5)
Mid	1.1× (L14)	1.2× (L14)
Pre-commit	1.0× (L16)	1.9× (L21)
Commitment	**1.7**× (L17)	**2.8**× (L22)
Post-commit	3.9× (L18)	2.0× (L23)
Late	7.8× (L31)	5.1× (L27)

Table 6. MLP replacement upper bound for both models. Replacing commitment-zone layers recovers visual grounding in both cases; Qwen's concentrated late-layer flip requires only two layers to reach 94.9%, while LLaVA's distributed prior requires all 32 layers to reach 52.6%.

	LLaVA		Qwen	
Target	**MDR**	**Gain**	**MDR**	**Gain**
1 commit layer	18.8% (L17)	+9.6pp	87.7% (L22)	+7.0pp
2 commit layers	24.1% (L16–17)	+14.9pp	94.9% (L22–23)	+14.2pp
4 commit layers	30.9% (L14–17)	+21.7pp	97.4% (L20–23)	+16.7pp
All layers	52.6% (32L)	+43.4pp	99.1% (28L)	+18.4pp

Table 7. Qwen2.5-VL $\text{MDR}_{\text{vision}}$ decomposed by conflict direction across all task types. *IY-TN* entries present a false negation (text incorrectly denying a true visual fact); *IN-TY* entries present a false affirmation (text incorrectly asserting a fact the image contradicts). Qwen maintains near-perfect vision-following on false affirmations across all tasks, but drops sharply on false negations for relational tasks, revealing a systematic negation asymmetry.

Task	IY-TN (false neg.)	IN-TY (false aff.)	Gap
color	96.6% (84/87)	100.0% (87/87)	+3.4pp
comparison	74.0% (71/96)	100.0% (96/96)	+26.0pp
count	85.3% (81/95)	98.9% (94/95)	+13.7pp
existence	92.0% (92/100)	100.0% (100/100)	+8.0pp
identity	45.9% (45/98)	100.0% (98/98)	+54.1pp
material	44.8% (43/96)	100.0% (96/96)	+55.2pp
size	25.0% (20/80)	97.5% (78/80)	+72.5pp
spatial	26.0% (20/77)	93.5% (72/77)	+67.5pp

Table 8. Per-entry flip layer statistics on text-following entries. Max-drop medians are invariant across data distributions; threshold-stable IQR captures within-CLEVR concentration. n reports text-following entries with a defined flip layer under each metric (entries that never persistently cross the threshold are excluded).

Model	Dataset	n	Max-drop median	Threshold-stable median (IQR)	% within ±1
LLaVA-v1.6	CLEVR-CONFLICT	490	18	18 ([9, 27])	35%
LLaVA-v1.6	COCO	2041	18	—	—
Qwen2.5-VL	CLEVR-CONFLICT	204	23	23 ([23, 23])	69%
Qwen2.5-VL	COCO	152	23	—	—

Table 9. Threshold-stable flip layer median is invariant to threshold choice. Per-entry medians on text-following entries from CLEVR-CONFLICT, computed at four thresholds.

Model	$\tau = 0.2$	$\tau = 0.3$	$\tau = 0.4$	$\tau = 0.5$
LLaVA-v1.6	18	18	18	18
Qwen2.5-VL	23	23	23	23

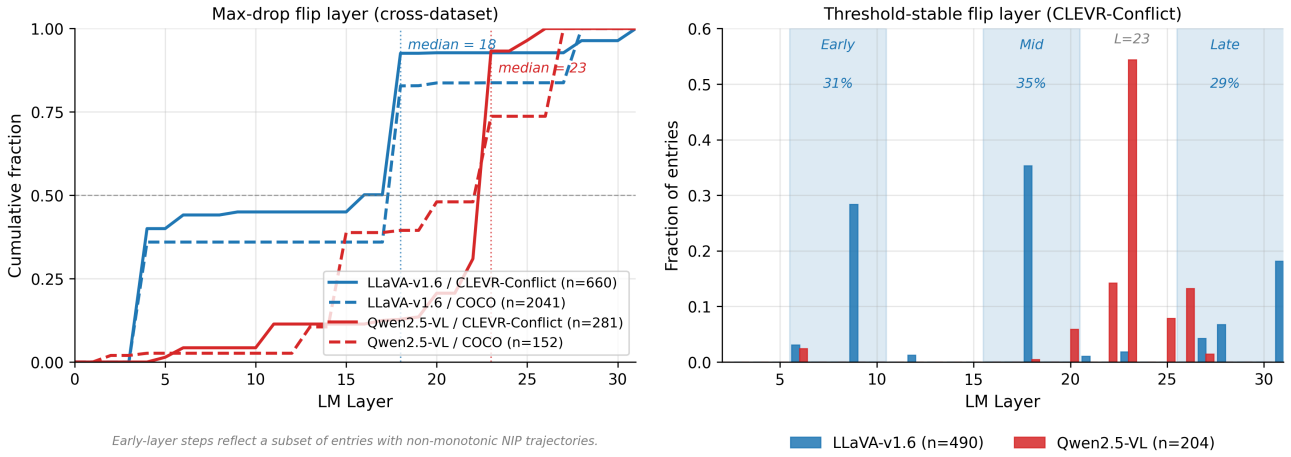

Figure 5. Per-entry flip layer distributions for text-following entries. *Left:* max-drop CDFs for LLaVA-v1.6 and Qwen2.5-VL across CLEVR-CONFLICT (solid) and COCO (dashed); medians coincide at $L = 18$ (LLaVA) and $L = 23$ (Qwen) on both datasets, supporting commitment timing as a property of each model's LM backbone rather than the input distribution. *Right:* threshold-stable flip layer histograms on CLEVR-CONFLICT showing Qwen's near-step-function commitment at $L = 23$ versus LLaVA's trimodal distribution across early ($L = 6$–10, 31%), mid ($L = 16$–20, 35%), and late ($L = 26$–31, 29%) commitment modes.

D. Residual Stream Steering

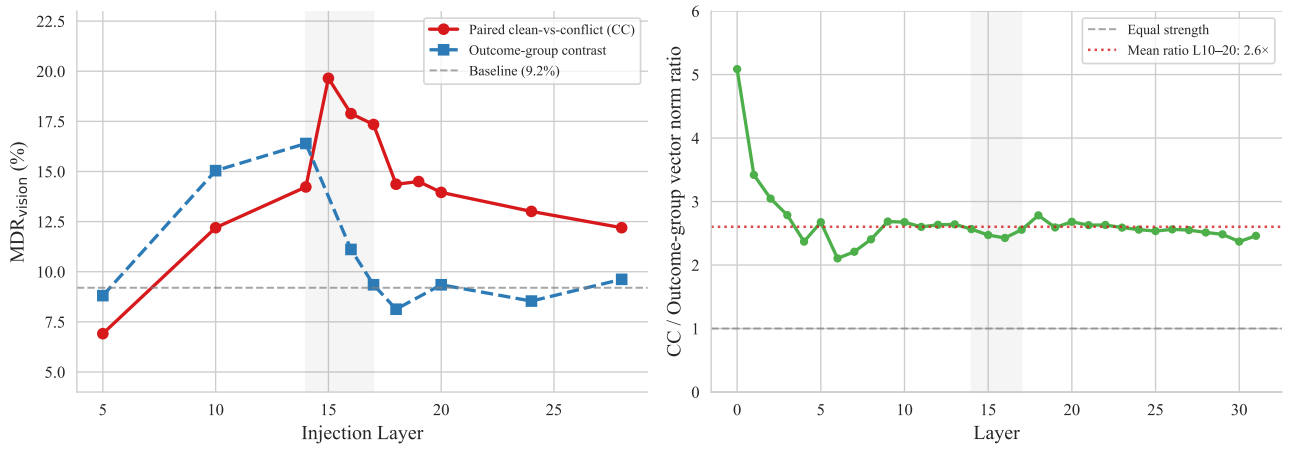

Figure 6. Comparison of two steering vector construction methods for LLaVA. *Left:* $\text{MDR}_{\text{vision}}$ as a function of injection layer for paired clean-vs-conflict (CC) vectors and outcome-group contrast vectors at α=20. CC vectors peak at the commitment zone (L14–17) while outcome-group contrast peaks earlier and falls below baseline post-commitment. *Right:* CC/outcome-group vector norm ratio across all layers, showing CC vectors are consistently $\approx$2.5$\times$ stronger in the commitment zone.

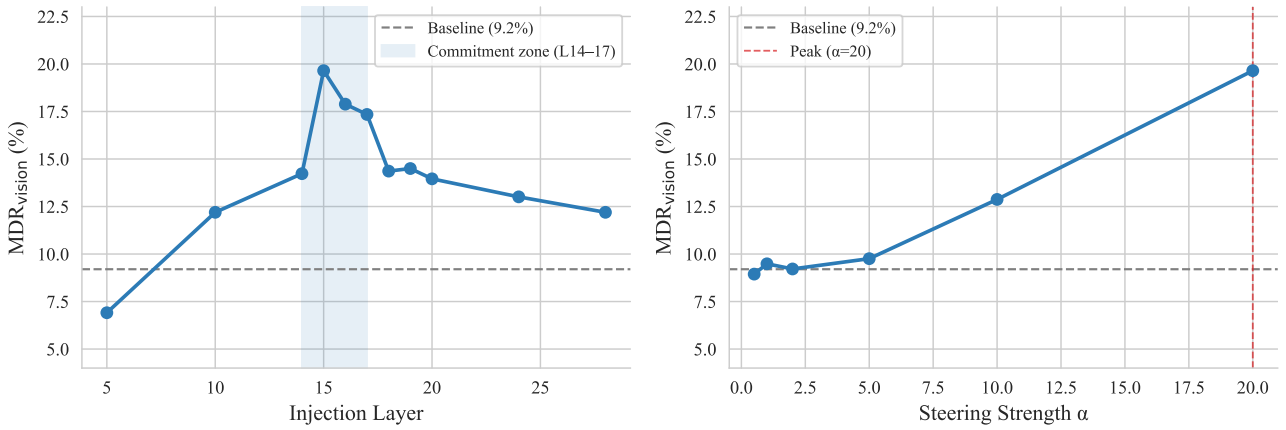

Figure 7. Residual stream steering results for LLaVA. *Left:* $\text{MDR}_{\text{vision}}$ as a function of injection layer at α=20, showing peak effect at the commitment zone (L14–17) and near-baseline performance at control layers. *Right:* $\text{MDR}_{\text{vision}}$ as a function of steering strength α at layer 15, showing monotonically increasing effect up to +10.4pp at α=20. Both results confirm the commitment zone as the effective intervention target while demonstrating that additive residual injection alone cannot overcome the language prior.