MLLMs Get It Right, Then Get It Wrong: Tracing and Correcting Late-Layer Textual Bias

Xingming Li[1], **Ao Cheng**[1], **Qiyao Sun**[1], **Xixiang He**[1], **Xuanyu Ji**[1], **Runke Huang**[2] and **Qingyong Hu**[3*]
[1]National University of Defense Technology, changsha, China
[2]Chinese University of Hong Kong, Shenzhen, China
[3]Intelligent Game and Decision Lab, Beijing, China
{lixingming, chengao18, sunqiyao18, hexixiang, jixuanyu18}@nudt.edu.cn, runkehuang@cuhk.edu.cn, huqingyong15@outlook.com

Abstract

When vision contradicts text, multimodal large language models (MLLMs) consistently favor text—even when images provide clear evidence otherwise. This bias poses risks for applications requiring visual grounding, yet its cause remains unclear. In this paper, we uncover a surprising finding: models often *get it right initially*, forming correct vision-based predictions in their intermediate layers, before *changing their minds* and favoring text in the final output. We call this "***late-layer textual override***". The visual information is encoded, it simply does not survive to the output. More intriguingly, we find that *how* predictions change reveals *whether* they're correct: 85% of failures shift toward text, while 89% of successes shift toward vision. This directional signature enables a simple but powerful intervention: when we detect a confident visual prediction being suppressed, we restore it. We propose CALRD (**C**onflict-**A**ware **L**ayer **R**eference **D**ecoding), a training-free method that recovers overridden predictions at inference time. Experiments across five MLLMs of varying architectures demonstrate up to 9.4% absolute improvements on conflict benchmarks while largely preserving standard performance, without training or external knowledge. It recovers what the model already knew but failed to preserve.

1 Introduction

Multimodal AI systems are moving into real-world use. They help diagnose diseases, moderate content, and assist drivers. This growing deployment raises a pointed question: *can we trust what they see?* Imagine a radiologist reviewing a chest X-ray with an AI assistant. The system receives both the image and the radiologist's preliminary notes. If those notes contain an error, will the AI catch it by looking at the image? Or will it simply echo the mistake?

Unfortunately, the evidence points to the second outcome. Show a model an image of a red car alongside text claiming "the car is blue," and most MLLMs will confidently answer "blue" [Liu *et al.*, 2025b; Liu *et al.*, 2025a; Wu *et al.*, 2024]. They trust the text over their own visual perception. This pattern appears across different model families [Liu *et al.*, 2023b; Dai *et al.*, 2023; Bai *et al.*, 2025b; Bai *et al.*, 2025a; Chen *et al.*, 2024] and benchmarks [Zhang *et al.*, 2025b; Jia *et al.*, 2025; Qian *et al.*, 2024]. In safety-critical applications [Bannur *et al.*, 2024; Hou *et al.*, 2025], such blind trust in text could cause serious harm.

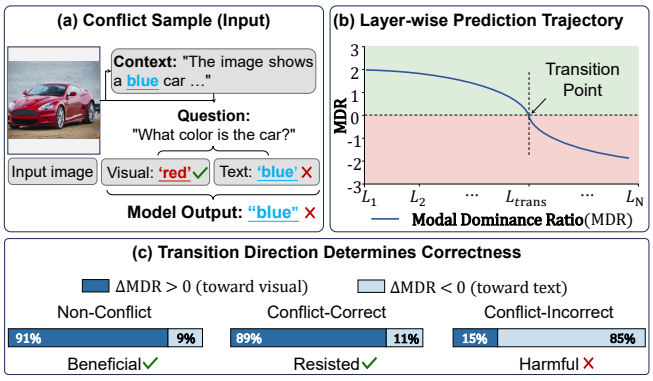

Figure 1: How predictions shift under visual-textual conflict. (a) The model sees a red car but outputs the text-suggested answer. (b) Modal Dominance Ratio (MDR) across layers: early layers favors vision, but late layers override toward text. (c) Shift direction predicts correctness—text-ward shifts correlate with failures.

Why does this happen? Prior work describes the *behavior* but not the *mechanism* [Liu *et al.*, 2025b; Zhang *et al.*, 2025b]. The failure could arise because visual information is poorly encoded, lost during cross-modal fusion, or correctly processed but later overwritten. Understanding the failure mode matters: different causes require different solutions.

To find out, we trace how a model's predictions evolve across its layers. Following [Elbayad *et al.*, 2020; Schuster *et al.*, 2022], we project hidden states at each layer to output probabilities. This lets us observe what the model "prefers" at each depth, and what we find is surprising. Surprisingly, in many conflict failures, **the model assigns higher probability to the correct visual answer in intermediate layers**. It then reverses course and outputs the text answer (Figure 1(a-b)).

*Corresponding author.

We term this phenomenon **late-layer textual override**. To measure it, we introduce Modal Dominance Ratio (MDR), which quantifies vision-versus-text preference at each layer. In failure cases, MDR starts positive (visual) and flips negative (textual) in later layers. The visual information was there; it simply didn't survive the full forward pass.

Not every late-layer shift is harmful. Such prediction shift occurs in successful cases too [Wang *et al.*, 2025b], not just failures. What distinguishes the two outcomes is the *direction* of the shift. We quantify this using the change in MDR across the transition point. As shown in Figure 1(c), in failures, 85% of transitions move toward text. In successes, 89% move toward vision. Late layers sometimes help; they sometimes hurt. The key is knowing which.

This directional asymmetry points to a practical solution. If a model holds a confident visual prediction at the transition layer but abandons it by the final layer, we suspect harmful override. We capture this with two signals derived from our analysis. *Anchor confidence* measures how certain the transition-layer prediction is. *Prediction retention* captures whether it survives to the output. High anchor confidence paired with low retention is the signature of harmful override.

Recovering what was known. These observations motivate **C**onflict-**A**ware **L**ayer **R**eference **D**ecoding (CALRD), a training-free method that restores intermediate predictions when override is detected. Our core insight is that predictions at the transition layer still carry the vision-grounded answer before it gets overwritten. By blending transition-layer logits back into the final distribution, we can recover it. CALRD first locates the layer where the output distribution shifts most sharply. It then uses anchor confidence and prediction retention to set correction strength: when both signals point to harmful override, CALRD blends in the transition-layer logits; otherwise, it leaves the output unchanged.

CALRD is straightforward. The model already has the right answer in its intermediate layers; we are helping it remember. No external knowledge or modules are involved. This points to a shift in perspective: improving MLLM reliability may be less about teaching models to see better and more about helping them retain what they have already seen.

We evaluate CALRD on five MLLMs, from InstructBLIP [Dai *et al.*, 2023] to Qwen3-VL [Bai *et al.*, 2025a]. The results suggest that late-layer override is not tied to a specific architecture. Our contributions are as follows:

- We introduce Modal Dominance Ratio to trace layer-wise modal preference and uncover late-layer textual override as a characteristic failure pattern. We show that transition direction, not mere occurrence, predicts correctness.
- We propose CALRD, a training-free method that detects harmful overrides through complementary signals and applies adaptive correction, preserving beneficial processing.
- Experiments show CALRD achieves up to 9.4% gains on conflict tasks while largely preserving standard performance. We also contribute Conflict-VQA, a diagnostic benchmark with explicit visual-textual annotations for mechanistic analysis.

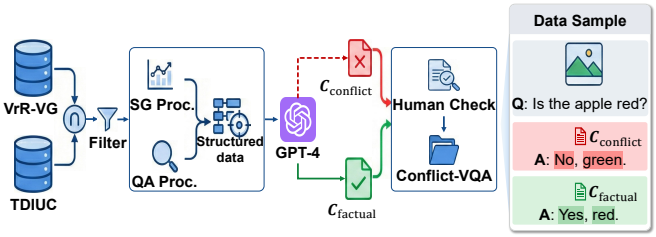

Figure 2: Conflict-VQA construction pipeline. We process images from VrR-VG and questions from TDIUC, use GPT-4o to generate C_{factual} and C_{conflict}, and verify all samples manually.

InstructBLIP	LLaVA-1.5	LLaVA-1.6	Qwen2.5-VL	Qwen3-VL
2,017	1,212	2,506	3,940	5,252

Table 1: Competent subset size per model. Total samples: 5,969. A sample is competent if the model answers correctly in all five trials under non-conflict context.

2 Related Work

Multimodal knowledge conflicts. When information sources disagree, which should a model trust? Prior work has explored this question for text-only language models [Chen *et al.*, 2022; Xu *et al.*, 2024; Xie and others, 2023]. The problem becomes more pronounced in multimodal settings, where visual perception and textual claims can point to different answers. A number of recent benchmarks document this challenge. HallusionBench [Guan *et al.*, 2024], AutoHallusion [Wu *et al.*, 2024], and PhD [Liu *et al.*, 2025a] report that MLLMs often follow textual context even when it contradicts what is visible in the image. Subsequent analyses suggest this behavior is systematic rather than accidental: models tend to rely on language priors or internal knowledge when facing conflicting evidence [Liu *et al.*, 2025b; Nguyen *et al.*, 2025; Deng *et al.*, 2025]. More targeted benchmarks explicitly construct multimodal conflicts and measure which modality dominates, further confirming this tendency [Jia *et al.*, 2025; Zhang *et al.*, 2025a; Zhu *et al.*, 2024]. While these studies clearly establish the phenomenon, they mostly describe it at the output level without explaining how conflicts are resolved inside the model. Our work addresses this gap by examining how visual and textual signals interact across layers, showing that many failures occur after visual information has already been encoded correctly.

Hallucination mitigation in MLLMs. Various inference-time methods have been proposed to reduce hallucinations. Contrastive decoding is a common strategy: VCD suppresses tokens preferred under distorted images [Liu *et al.*, 2023a], ICD contrasts standard versus perturbed instructions [Wang *et al.*, 2024c], and OPERA penalizes attention over-trust patterns [Huang and others, 2024]. Recent variants refine these ideas through explicit attention steering [Wang *et al.*, 2025c], multi-stage contrast with selective visual inputs [Park *et al.*, 2025], and probabilistic detection that avoids a second forward pass [Fieback *et al.*, 2025]. A separate line of work exploits model depth. DoLa compares predictions from shal-

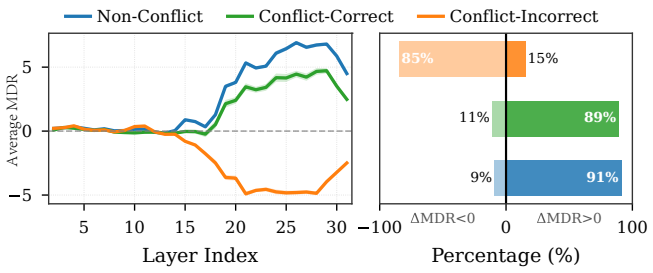

Figure 3: Layer-wise MDR analysis on LLaVA-1.6-7B. **Left:** Non-Conflict and Conflict-Correct samples maintain positive MDR, while Conflict-Incorrect samples reverse from positive to negative in late layers. **Right:** Text-ward shifts (ΔMDR<0) dominate failures (85%), vision-ward shifts (ΔMDR>0) dominate correct predictions (89–91%). Shading: ±1 std.

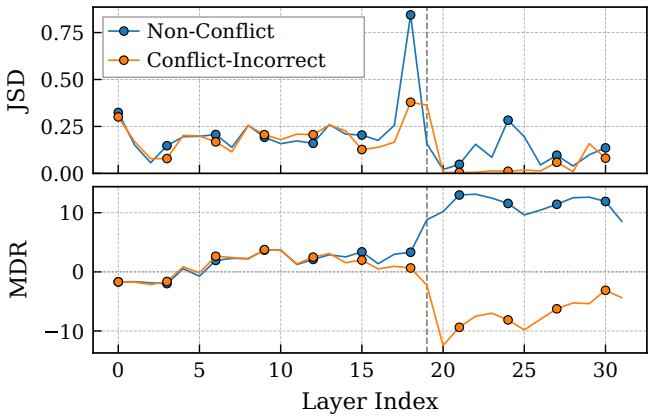

Figure 4: JSD and MDR trajectories on LLaVA-1.6-7B. JSD spikes align across categories, while MDR shifts toward vision in Non-Conflict and toward text in Conflict-Incorrect.

low and deep layers to recover factual signals [Chuang *et al.*, 2024]. DeCo, SHIFT, and related decoders rely on the observation that intermediate layers preserve stronger visual grounding [Wang *et al.*, 2025a; Wang *et al.*, 2025b; Tong *et al.*, 2025]. Training-based approaches reduce hallucinations through preference optimization or contrastive learning but require supervision [Wang *et al.*, 2024a; Yu *et al.*, 2024; Jiang *et al.*, 2024]. Most methods correct uniformly, without distinguishing helpful from harmful late-layer processing. Our finding that transition *direction* predicts correctness enables selective intervention only when needed. As we show in Section 5, uniform correction methods like DoLa can degrade performance when applied indiscriminately.

Layer-wise Analysis of Multimodal Models. Recent studies examine visual information flow in MLLMs, finding that visual signals peak in intermediate layers and weaken in deeper ones [Wang *et al.*, 2024b; Wang and others, 2025; Yin *et al.*, 2025]. Others identify failure modes such as attention drift and modality imbalance that reduce visual grounding [Jiang *et al.*, 2025; Chen *et al.*, 2025]. These findings help explain why models capture visual details early but produce text-driven answers at the end. Our work extends this line of research. We show that the *direction* of layer-wise transitions predicts final correctness, allowing us to intervene only when late-layer processing undermines visual evidence.

3 Diagnosing Late-Layer Textual Override

To understand why models favor text over vision under conflict, we need to trace how their predictions evolve during the forward pass. Does textual bias arise because visual information is never properly encoded? Or is it encoded but subsequently discarded? This section presents a layer-wise analysis that reveals a surprising answer.

3.1 Experimental Framework

Task and data. We study visual question answering where inputs consist of image I, context C, and question Q. Let w_{vis} and w_{text} denote answers supported by the image and context respectively. A *knowledge conflict* occurs when $w_{\text{vis}} \neq w_{\text{text}}$. We categorize samples into three groups based on context type and model output. **Non-Conflict**: the context is consistent with the image and the model answers correctly. **Conflict-Correct**: context contradicts image but the model follows visual evidence. **Conflict-Incorrect**: the context contradicts the image and the model follows textual claims.

Existing conflict benchmarks lack the $(w_{\text{vis}}, w_{\text{text}})$ annotations required for layer-wise analysis. To address this, we construct **Conflict-VQA**: 5,969 samples with paired consistent and conflicting contexts. We source images from VrR-VG [Liang *et al.*, 2019] and questions from TDIUC [Kafle and Kanan, 2017], covering six question types: object presence, color, attribute, counting, spatial reasoning, and activity recognition. All contexts are generated using GPT-4o and manually verified. Construction details are provided in Appendix A. For mechanistic analysis, we filter each model's competent subset, defined as samples answered correctly under consistent context across five trials. This filtering ensures that observed failures under conflict reflect conflict resolution process rather than basic visual perception errors.

Layer-wise projection. Following early-exit methods [Elbayad *et al.*, 2020; Schuster *et al.*, 2022], we project hidden states at each layer l to the output vocabulary via the language model head W_{head}: $P_t^{(l)} = \text{softmax}(W_{\text{head}} h_t^{(l)})$, where $h_t^{(l)}$ is the hidden state at the answer token position and $P_t^{(l)}(w)$ is the scalar probability assigned to vocabulary token w. This lets us observe the model's "preference" at each depth.

3.2 Visual Predictions Emerge, Then Fade

When models fail under conflict, is visual information absent from their representations, or present but ultimately ignored?

To answer this, we introduce **Modal Dominance Ratio (MDR)**, quantifying modal preference at each layer:

$$\text{MDR}^{(l)} = \log \frac{P_t^{(l)}(w_{\text{vis}})}{P_t^{(l)}(w_{\text{text}})} \tag{1}$$

Positive MDR indicates preference for the visual answer; negative indicates textual preference.

Figure 3 shows MDR trajectories averaged across sample categories. Non-Conflict and Conflict-Correct samples maintain positive MDR throughout all layers, reflecting consistent

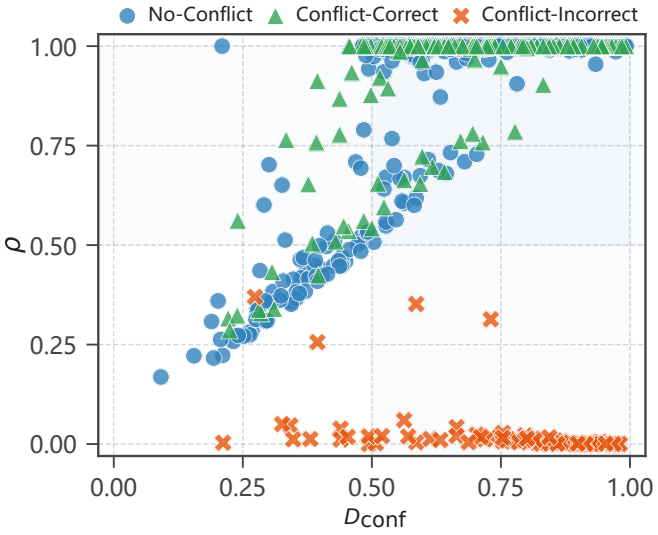

Figure 5: Detection signal distribution on LLaVA-1.6-7B. Conflict-Incorrect samples (red) exhibit high D_{conf} but low ρ. Non-Conflict (blue) and Conflict-Correct (green) maintain high ρ.

visual preference. Conflict-Incorrect samples tell a different story: **MDR is positive in early and middle layers, then reverses to negative in late layers.**

This pattern indicates that in failure cases, models assign higher probability to the visual answer in intermediate layers before shifting toward the textual answer. In other words, the visual information is present in intermediate representations but does not survive to the final output. We term this phenomenon *late-layer textual override*. This observation suggests that the failure mechanism lies in late-layer processing rather than in early-layer encoding.

3.3 Transition Direction Predicts Success

The late-layer reversal in Conflict-Incorrect samples might suggest a straightforward fix: simply use intermediate-layer predictions. But prediction transitions also occur when models answer correctly. In these cases, late-layer processing appears to refine the answer rather than override it. To intervene selectively, we need a way to tell these two situations apart.

Detecting transitions. Layer-wise predictions show high entropy in early layers and lower entropy in deeper layers as predictions converge. We define the *stability onset* L_{start} as the layer with the largest entropy drop:

$$L_{\text{start}} = \arg\max_{l} [H_t^{(l-1)} - H_t^{(l)}], \tag{2}$$

where $H_t^{(l)} = -\sum_w P_t^{(l)}(w) \log P_t^{(l)}(w)$. Across our models, L_{start} falls between 25% and 40% of model depth.

After stability onset, we detect abrupt distributional shifts using Jensen-Shannon divergence between adjacent layers:

$$\text{JSD}^{(l)} = \frac{1}{2}\left[D_{\text{KL}}(P_t^{(l)} \| M) + D_{\text{KL}}(P_t^{(l+1)} \| M)\right], \tag{3}$$

where $M = \frac{1}{2}(P_t^{(l)} + P_t^{(l+1)})$. The transition layer is: $L_{\text{trans}} = \arg\max_{l \in [L_{\text{start}}, N-1]} \text{JSD}^{(l)}$.

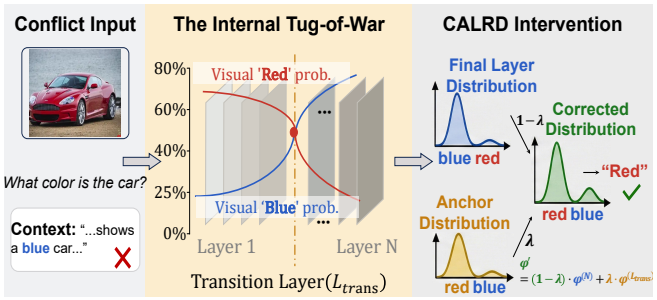

Figure 6: Overview of CALRD. Under visual-textual conflict, predictions shift from vision toward text across layers. CALRD detects this override and restores the transition-layer prediction.

Transition direction distinguishes success from failure. Figure 4 shows that both Non-Conflict and Conflict-Incorrect samples have JSD spikes at similar depths, so transitions alone do not distinguish the two groups. What does differ is the *direction* of the shift. We quantify this via MDR change across the transition:

$$\Delta\text{MDR} = \overline{\text{MDR}}_{\text{after}} - \overline{\text{MDR}}_{\text{before}} \tag{4}$$

where bars denote averages over pre/post-transition layers.

Finding. In 85% of Conflict-Incorrect cases, ΔMDR is negative, indicating a shift toward text. The pattern reverses for correct predictions: 89% of Conflict-Correct and 91% of Non-Conflict samples show positive ΔMDR, shifting toward vision or holding steady. Transitions occur in all groups, but only the direction correlates with correctness.

4 Conflict-Aware Layer Reference Decoding

The analysis in Section 3 shows that harmful overrides are characterized by a late-layer shift from visual to textual preference, and that the direction of this shift distinguishes failures from successes. However, the MDR analysis relies on ground-truth $(w_{\text{vis}}, w_{\text{text}})$ labels, unavailable at inference. Can we detect harmful overrides without labels? In this section, we develop practical detection signals and a decoding strategy that recovers overridden predictions without requiring such labels.

Specifically, we propose *Conflict-Aware Layer Reference Decoding* (CALRD), a training-free method. CALRD first identifies when a confident transition-layer prediction is being suppressed, then adjusts the output distribution to restore it. Since prediction transitions also occur in correct outputs, the intervention must be selective: we modulate its strength based on how strongly the override signature is present.

4.1 Detecting Harmful Override

The MDR analysis shows that Conflict-Incorrect samples hold confident predictions at L_{trans} that do not survive to the final layer (section 3). This observation motivates two simple signals. Let $w^* = \arg\max_w P_t^{(L_{\text{trans}})}(w)$ be the top prediction at transition. Our detection signals are:

Anchor confidence. A higher transition-layer probability for w^* indicates a stronger intermediate preference:

$$D_{\text{conf}} = P_t^{(L_{\text{trans}})}(w^*) \tag{5}$$

Prediction retention. To quantify survival, we use the final-to-transition probability ratio; $\rho \approx 1$ indicates preservation, while $\rho \approx 0$ indicates override:

$$\rho = \min\left(1, \frac{P_t^{(N)}(w^*)}{P_t^{(L_{\text{trans}})}(w^*)}\right) \tag{6}$$

Figure 5 shows how these signals distribute across sample types. Conflict-Incorrect samples cluster in the high-D_{conf}, low-ρ region: confident predictions that get overridden. Non-Conflict and Conflict-Correct samples maintain high ρ, meaning their predictions persist. This separation allows us to detect harmful overrides at test time.

4.2 Adaptive Intervention

Not all prediction transitions are harmful. As shown in Section 3.3, transitions toward vision correlate with correct outputs, while transitions toward text correlate with failures. To avoid disrupting beneficial late-layer processing, intervention strength must adapt to override severity. We define:

$$\lambda = D_{\text{conf}} \cdot (1 - \rho) \tag{7}$$

The multiplicative form ensures that λ is substantial only when *both* conditions hold: a confident transition-layer prediction exists (D_{conf} high) *and* it is being suppressed (ρ low). When D_{conf} is low, there is no preference to recover, so λ stays small regardless of retention. When ρ is high, the prediction survives to the output, so intervention is unnecessary. Only when a confident prediction gets discarded does λ become large enough to affect decoding. This design is motivated by the asymmetry in Section 3.3: harmful cases are likely to exhibit confident intermediate predictions that are later suppressed, whereas correct cases tend to preserve them.

4.3 Transition-Layer Guided Decoding

Given λ, we adjust the output logits by combining final-layer and transition-layer predictions. Let $\phi_t^{(l)} = W_{\text{head}} \cdot h_t^{(l)}$ denote the unnormalized logits at layer l. The adjusted logits are:

$$\phi_t' = (1 - \lambda) \cdot \phi_t^{(N)} + \lambda \cdot \phi_t^{(L_{\text{trans}})} \tag{8}$$

with output distribution $P_t' = \text{softmax}(\phi_t')$. When $\lambda = 0$, this reduces to standard decoding. As λ increases, the transition-layer prediction receives more weight, recovering the vision-grounded answer that would otherwise be lost. For samples without override signatures, λ remains near zero and the output is largely unchanged.

CALRD works at each token position independently. For multi-token generation, we recompute L_{trans}, D_{conf}, and ρ at each step, allowing the intervention strength to vary across the sequence. The logit-level operation makes CALRD compatible with greedy, beam search, and nucleus sampling without modification. Note that CALRD does not introduce external knowledge or require additional training. The correction comes entirely from the model's own intermediate representations. As our analysis shows, the vision-grounded answer is often already present at the transition layer; CALRD simply helps it survive to the output.

5 Experiments

We evaluate CALRD on two questions: (1) Does it improve conflict resolution? (2) Does it hurt non-conflict scenarios? Experiments cover five MLLMs with different architectures and capability levels. For each model, we report results on conflict benchmarks where context contradicts the image, as well as standard hallucination benchmarks.

5.1 Experimental Setup

Models. We evaluate five MLLMs: InstructBLIP-7B [Dai *et al.*, 2023], LLaVA-1.5-7B, LLaVA-1.6-7B [Liu *et al.*, 2023b], Qwen2.5-VL-7B [Bai *et al.*, 2025b], and Qwen3-VL-8B [Bai *et al.*, 2025a]. These models span different architectures and capability levels.

Benchmarks. We evaluate on two complementary benchmark groups. (1) For conflict resolution, we use **Conflict-VQA** (Section 3.1), which contains 5,969 samples requiring open-ended answers, and **PhD-icc** [Liu *et al.*, 2025a], the incorrect-context subset of PhD containing 16,844 yes/no question pairs. Both benchmarks include contexts that contradict visual evidence, and we report accuracy. (2) For standard hallucination without explicit conflicts, we use **POPE** [Li *et al.*, 2023] and **CHAIR** [Rohrbach *et al.*, 2018]. POPE probes object hallucination by asking "Is there a `<object>` in the image?" across random, popular, and adversarial splits. It covers 500 MSCOCO images with six questions per image for each split, and we report F1 score. CHAIR measures caption hallucination against ground-truth object annotations, reporting instance-level C_I and sentence-level C_S (lower is better). Following [Huang and others, 2024], we evaluate on 500 MSCOCO images with the same captioning prompt "Please describe this image in detail." for consistency.

Baselines. We compare against four representative methods. DoLa [Chuang *et al.*, 2024] contrasts logits from later layers against earlier layers to surface factual knowledge. VCD [Leng *et al.*, 2024] contrasts outputs from original and distorted visual inputs to reduce over-reliance on language priors. OPERA [Huang and others, 2024] penalizes over-trust patterns in self-attention and uses rollback to re-select tokens. DeCo [Wang *et al.*, 2025a] adaptively selects preceding layers where visual information is stronger and integrates their predictions into the final output. We test CALRD with greedy, beam search, and nucleus sampling. Since not all baselines are designed for every decoding strategy, we evaluate each method under the decoding regimes supported by its official implementation and compare CALRD against the same vanilla decoder in each regime. All experiments are run on NVIDIA H100 GPUs.

5.2 Main Results

Conflict resolution. Table 2 shows results on conflict benchmarks. CALRD outperforms all baselines across models and decoders. Gains are largest on PhD-icc: LLaVA-1.5 improves from 27.72% to 36.35% (+8.63%), LLaVA-1.6 from 28.23% to 36.52% (+8.29%), and InstructBLIP from 41.08% to 49.62% (+8.54%). On Conflict-VQA, improvements reach up to 7.4%.

Decoding	Method	Conflict-VQA (Acc ↑)					PhD-icc (Acc ↑)				
		InstructBLIP	LLaVA-1.5	LLaVA-1.6	Qwen2.5-VL	Qwen3-VL	InstructBLIP	LLaVA-1.5	LLaVA-1.6	Qwen2.5-VL	Qwen3-VL
Greedy	Vanilla	39.61	36.20	42.35	65.61	75.58	41.08	27.72	28.23	51.30	60.32
	DoLa	32.72	40.10	44.59	66.22	76.06	32.65	29.78	28.73	51.08	58.00
	DeCo	41.31	38.03	43.38	63.79	75.33	42.12	31.00	33.65	51.65	62.25
	CALRD (Ours)	**44.12** ↑4.5	**41.90** ↑5.7	**49.51** ↑7.2	**70.15** ↑4.5	**77.42** ↑1.8	**49.62** ↑8.5	**36.35** ↑8.6	**36.52** ↑8.3	**56.20** ↑4.9	**63.68** ↑3.4
Beam	Vanilla	39.98	38.88	43.99	65.49	75.94	40.73	28.25	28.80	51.11	62.53
	OPERA	40.12	39.00	43.74	68.49	76.43	39.78	28.93	29.95	53.70	62.62
	DeCo	40.83	38.15	43.01	63.55	75.46	42.15	31.07	34.63	51.13	61.25
	CALRD (Ours)	**44.49** ↑4.5	**44.47** ↑5.6	**50.79** ↑6.8	**70.25** ↑4.8	**79.34** ↑3.4	**49.75** ↑9.0	**36.58** ↑8.3	**37.40** ↑8.6	**56.70** ↑5.6	**64.60** ↑2.1
Nucleus	Vanilla	23.94	38.76	41.92	61.48	73.39	39.23	29.33	30.08	51.05	62.23
	VCD	23.05	42.40	46.35	64.04	75.50	40.01	31.87	31.70	53.83	63.40
	DeCo	24.91	38.45	45.44	63.55	74.79	41.83	31.25	32.90	51.62	61.50
	CALRD (Ours)	**28.07** ↑4.1	**43.40** ↑4.6	**48.72** ↑6.8	**68.91** ↑7.4	**76.55** ↑3.2	**48.62** ↑9.4	**34.95** ↑5.6	**34.12** ↑4.0	**55.83** ↑4.8	**64.52** ↑2.3

Table 2: Performance comparison on knowledge conflict benchmarks, where textual context contradicts visual evidence. We evaluate five MLLMs with three decoding strategies. Best results are in **bold**. ↑: improvement over Vanilla.

Larger gains on PhD-icc. The two benchmarks differ in question format: PhD-icc uses binary yes/no questions, while Conflict-VQA requires open-ended answers. We suspect the binary format creates conditions where textual bias is more pronounced. When context strongly suggests "yes" or "no," the model may be more easily swayed away from visual evidence. In contrast, open-ended questions require generating specific content, making it harder for context alone to override perception.

Baseline accuracy and improvement. An interesting pattern emerges when we rank models by their vanilla accuracy on PhD-icc: LLaVA-1.5 (27.72%) < LLaVA-1.6 (28.23%) < InstructBLIP (41.08%) < Qwen2.5-VL (51.30%) < Qwen3-VL (60.32%). The gains from CALRD roughly follow the inverse order, with weaker models seeing larger improvements. This pattern is expected: stronger models already resolve more conflicts correctly in their late layers, leaving fewer overridden predictions for CALRD to recover. The adaptive nature of our intervention naturally accommodates this—when predictions survive to the output (ρ remains high), λ stays small and CALRD leaves the output largely unchanged. For Qwen3-VL, the +3.36% absolute gain corresponds to an 8.5% relative error reduction, indicating that even well-calibrated models contain recoverable failures.

Comparison with other layer-based methods. DeCo and DoLa also leverage information from earlier layers, but their performance is less consistent. DoLa improves some configurations but hurts others: on InstructBLIP, Conflict-VQA accuracy drops from 39.61% to 32.72%. DeCo shows more stable behavior but still underperforms CALRD across the board. One possible explanation is that these methods apply corrections more uniformly, without distinguishing cases where late-layer processing is beneficial. CALRD's detection signals allow it to intervene selectively, reducing the risk of disrupting predictions that were already on track.

Non-conflict scenarios. Textual override is not identical to all hallucinations, so we also test CALRD on standard object-centric hallucination benchmarks without explicit conflicts. A method that improves conflict resolution but degrades standard performance would have limited practical value. Tables 3 and 4 show that CALRD largely avoids this trade-off.

Dec.	Method	POPE (F1 ↑)				
		InstructBLIP	LLaVA-1.5	LLaVA-1.6	Qwen2.5-VL	Qwen3-VL
Greedy	Vanilla	80.0	82.2	85.4	80.8	88.5
	DoLa	83.4	83.2	87.1	82.1	88.6
	DeCo	84.9	83.8	87.6	81.4	90.6
	CALRD (Ours)	**85.4** ↑5.4	**84.5** ↑2.3	**88.2** ↑2.8	**82.3** ↑1.5	**91.6** ↑3.1
Beam	Vanilla	84.4	84.9	87.1	80.7	88.6
	OPERA	84.8	85.4	87.5	80.8	89.3
	DeCo	84.9	86.7	87.9	81.4	91.2
	CALRD (Ours)	**84.7** ↑0.3	**86.9** ↑2.0	**88.3** ↑1.2	**82.6** ↑1.9	**91.9** ↑3.3
Nucleus	Vanilla	79.8	83.1	85.5	80.8	88.7
	VCD	79.9	83.1	85.6	80.9	88.9
	DeCo	81.8	84.8	87.3	81.3	91.2
	CALRD (Ours)	**83.4** ↑3.6	**85.9** ↑2.8	**87.5** ↑2.0	**81.9** ↑1.1	**91.4** ↑2.7

Table 3: Results on POPE object hallucination benchmark. CALRD results are highlighted. ↑: improvement over Vanilla.

POPE results (Table 3). CALRD maintains or improves F1 across all configurations. Under greedy decoding, InstructBLIP improves from 80.0 to 85.4, and Qwen3-VL from 88.5 to 91.6. POPE probes object hallucination without explicit visual-textual conflicts, so these results suggest that late-layer override may occur more broadly than just in conflict settings. The key is that CALRD does not intervene indiscriminately: when ρ is high, indicating the prediction is stable, λ stays near zero and the output remains unchanged.

CHAIR results (Table 4). Captioning differs from VQA in that it requires extended sequences rather than short answers. Hallucinations accumulate over multiple tokens, making it a useful stress test. CALRD reduces both C_I and C_S in most configurations. InstructBLIP shows the largest improvement: C_I drops from 23.7 to 14.6 (38% relative reduction) and C_S from 58.8 to 40.2. These results indicate that recomputing detection signals at each token position allows CALRD to adapt throughout the sequence, rather than a fixed correction.

5.3 Analysis

Ablation study. Table 5 examines the contribution of each component using LLaVA-1.6 with greedy decoding. Removing D_{conf} drops Conflict-VQA accuracy from 49.5% to 47.3%, while removing ρ causes a larger drop to 45.1%. This suggests that retention is the more informative signal for detecting harmful override, which makes sense: a prediction can be confident at the transition layer for various reasons, but a sharp drop in retention specifically indicates that something changed in late-layer processing.

Using a fixed layer at 75% depth instead of dynamic L_{trans}

Decoding	Method	InstructBLIP $C_S \downarrow$	InstructBLIP $C_I \downarrow$	LLaVA-1.5 $C_S \downarrow$	LLaVA-1.5 $C_I \downarrow$	LLaVA-1.6 $C_S \downarrow$	LLaVA-1.6 $C_I \downarrow$	Qwen2.5-VL $C_S \downarrow$	Qwen2.5-VL $C_I \downarrow$	Qwen3-VL $C_S \downarrow$	Qwen3-VL $C_I \downarrow$
Greedy	Vanilla	58.8	23.7	45.0	14.7	37.6	12.7	40.8	9.7	52.4	10.1
	DoLa	48.4	15.9	47.8	13.8	35.3	8.7	36.5	9.2	55.6	9.8
	DeCo	41.2	14.4	37.8	11.1	35.8	10.4	37.6	**9.1**	51.2	9.6
	CALRD (Ours)	**40.2** ↓18.6	**14.6** ↓9.1	**35.7** ↓9.3	**10.3** ↓4.4	**33.6** ↓4.0	**9.3** ↓3.4	**35.5** ↓5.3	**9.5** ↓0.2	**50.7** ↓1.7	**9.3** ↓0.8
Beam	Vanilla	55.6	15.8	48.8	13.9	36.0	12.1	38.9	9.4	53.6	9.7
	OPERA	46.4	14.2	44.6	12.8	35.3	12.9	39.1	10.2	54.7	11.3
	DeCo	43.8	12.7	32.0	9.7	34.4	9.0	38.2	8.5	52.0	9.3
	CALRD (Ours)	**38.4** ↓17.2	**11.4** ↓4.4	**34.7** ↓14.1	**10.9** ↓3.0	**32.8** ↓3.2	**8.7** ↓3.4	**33.4** ↓5.5	**8.2** ↓1.2	**40.6** ↓13.0	**7.9** ↓1.8
Nucleus	Vanilla	54.6	24.8	48.8	14.2	36.8	10.5	40.4	10.7	54.2	10.2
	VCD	58.0	17.0	54.0	16.0	41.2	10.2	42.9	12.3	53.4	11.0
	DeCo	43.6	12.9	42.8	13.2	36.1	10.1	40.4	9.5	54.8	9.9
	CALRD (Ours)	**42.7** ↓11.9	**12.1** ↓12.7	**37.6** ↓11.2	**12.8** ↓1.4	**34.8** ↓2.0	**11.5** ↓-1.0	**38.6** ↓1.8	**9.1** ↓1.6	**53.6** ↓0.6	**9.5** ↓0.7

Table 4: Results on CHAIR caption hallucination benchmark. Lower is better. ↓: reduction over Vanilla.

Configuration	C-VQA ↑	POPE↑
Vanilla (Greedy)	42.3	85.4
CALRD (Full)	**49.5**	**88.2**
w/o D_{conf}	47.3	87.5
w/o ρ	45.1	86.6
w/ Fixed Layer (75%)	45.9	87.1

Table 5: Ablation study on LLaVA-1.6 with greedy decoding. We evaluate the contribution of each component.

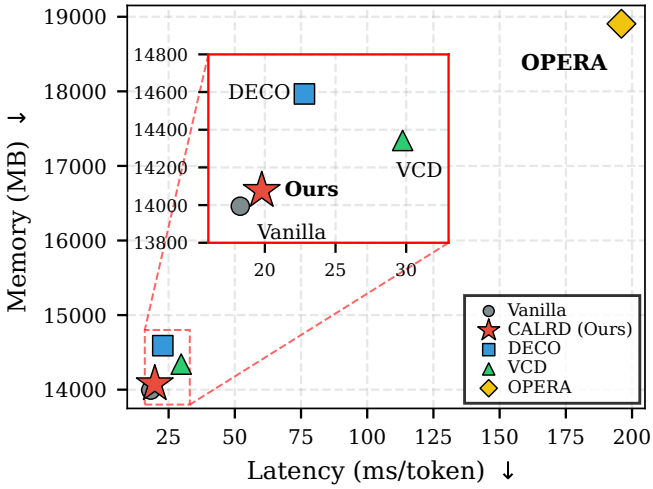

Figure 7: **Inference Efficiency on LLaVA-1.5-7B.** Latency vs. GPU memory trade-off measured on a single H100 GPU. Our CALRD achieves a superior balance, staying closest to the Vanilla baseline compared to DECO, VCD, and OPERA.

detection reduces accuracy to 45.9%. The value corresponds to the median transition layer observed in our analysis (Section 3.3), so this comparison uses the most representative fixed position rather than an arbitrary choice. The accuracy drop indicates that transition points vary substantially across samples, and a single fixed layer cannot capture this variation. We also observe that POPE follows a similar pattern, with both signals contributing to the overall improvement.

These results support the multiplicative design $\lambda = D_{\text{conf}} \cdot (1 - \rho)$: both signals provide useful information, but neither is sufficient alone. The product ensures that substantial intervention occurs only when a confident prediction exists and is being suppressed by late-layer processing.

Efficiency. Figure 7 compares latency and memory usage across methods using LLaVA-1.5 on a single H100 GPU. CALRD adds modest overhead: 19.8 ms/token compared to 18.3 ms/token for vanilla decoding, with only 80 MB additional memory. VCD nearly doubles latency due to its dual-stream requirement, and OPERA is approximately **10**× slower because of iterative rollback. DeCo falls in between. These results suggest CALRD is practical for deployment scenarios where efficiency matters.

6 Conclusion

We studied why multimodal models favor text over vision when the two conflict, and found that the problem is not poor visual encoding. In many failure cases, models assign higher probability to the correct visual answer in intermediate layers, only to override it with the textual answer in deeper layers. We call this late-layer textual override. Importantly, the direction of prediction change distinguishes failures from successes: shifts toward text correlate with incorrect outputs, while shifts toward vision correlate with correct ones. This observation motivates CALRD, a training-free decoding method that restores intermediate predictions when override signatures are detected. Experiments on five MLLMs show up to 9.4% improvements on conflict benchmarks while largely preserving performance on standard tasks. Our results suggest that improving reliability under conflict may require not better visual encoding, but better preservation of visual information through the forward pass.

Limitations and future work. Our work opens several avenues for future research. A natural next step is to examine how late-layer dynamics interact with model scale and pretraining data composition—questions that require controlled experiments with access to training corpora, beyond the scope of our mechanistic study. Additionally, exploring whether similar override patterns emerge across other modality pairs (*e.g.,* audio-text) may reveal general principles of multimodal integration. We view our mechanistic analysis as a foundation for such investigations.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62306331, 62407037, and CAAI Youth Talent Lifting Project under Grant CAAI2023-2025QNRC001.

References

[Bai *et al.*, 2025a] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, et al. Qwen3-vl technical report. *CoRR*, abs/2511.21631, 2025.

[Bai *et al.*, 2025b] Shuai Bai, Keqin Chen, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[Bannur *et al.*, 2024] Shruthi Bannur, Kenza Bouzid, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

[Chen *et al.*, 2022] Hung-Ting Chen, Michael Zhang, and Eunsol Choi. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307, 2022.

[Chen *et al.*, 2024] Zhe Chen, Jiannan Wu, Wenhai Wang, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[Chen *et al.*, 2025] Haoran Chen, Junyan Lin, Xinghao Chen, Yue Fan, Jianfeng Dong, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. Multimodal language models see better when they look shallower. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6688–6706, 2025.

[Chuang *et al.*, 2024] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *ICLR*, 2024.

[Dai *et al.*, 2023] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

[Deng *et al.*, 2025] Ailin Deng, Tri Cao, et al. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3867–3876, 2025.

[Elbayad *et al.*, 2020] Maha Elbayad, Jiatao Gu, Edouard Grave, et al. Depth-adaptive transformer. In *International Conference on Learning Representations*, 2020.

[Fieback *et al.*, 2025] Laura Fieback, Nishilkumar Balar, et al. Ecd: Efficient contrastive decoding with probabilistic hallucination detection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 21–38. Springer, 2025.

[Guan *et al.*, 2024] Tianrui Guan, Fuxiao Liu, Xiyang Wu, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.

[Hou *et al.*, 2025] Benjamin Hou, Pritam Mukherjee, Vivek Batheja, Kenneth C Wang, Ronald M Summers, and Zhiyong Lu. One year on: assessing progress of multimodal large language model performance on rsna 2024 case of the day questions. *Radiology*, 316(2):e250617, 2025.

[Huang and others, 2024] Qidong Huang et al. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 13418–13427, 2024.

[Jia *et al.*, 2025] Yifan Jia, Kailin Jiang, Yuyang Liang, Qihan Ren, Yi Xin, Rui Yang, Fenze Feng, Mingcai Chen, Hengyang Lu, Haozhe Wang, et al. Benchmarking multimodal knowledge conflict for large multimodal models. *arXiv preprint arXiv:2505.19509*, 2025.

[Jiang *et al.*, 2024] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 27036–27046, 2024.

[Jiang *et al.*, 2025] Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25004–25014, 2025.

[Kafle and Kanan, 2017] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *Proceedings of the IEEE international conference on computer vision*, pages 1965–1973, 2017.

[Leng *et al.*, 2024] Sicong Leng, Hang Zhang, et al. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024.

[Li *et al.*, 2023] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, 2023.

[Liang *et al.*, 2019] Yuanzhi Liang, Yalong Bai, Wei Zhang, Xueming Qian, Li Zhu, and Tao Mei. Vrr-vg: Refocusing visually-relevant relationships. In *Proceedings of the international conference on computer vision*, pages 10403–10412, 2019.

[Liu *et al.*, 2023a] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.

[Liu *et al.*, 2023b] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[Liu *et al.*, 2025a] Jiazhen Liu, Yuhan Fu, et al. Phd: A chatgpt-prompted visual hallucination evaluation dataset. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19857–19866, 2025.

[Liu *et al.*, 2025b] Xiaoyuan Liu, Wenxuan Wang, et al. Insight over sight: Exploring the vision-knowledge conflicts in multimodal llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17825–17846, 2025.

[Nguyen *et al.*, 2025] Trang Nguyen, Jackson Michaels, Madalina Fiterau, and David Jensen. Challenges in understanding modality conflict in vision-language models. *arXiv preprint arXiv:2509.02805*, 2025.

[Park *et al.*, 2025] Woohyeon Park, Woojin Kim, Jaeik Kim, and Jaeyoung Do. Second: Mitigating perceptual hallucination in vision-language models via selective and contrastive decoding. *arXiv preprint arXiv:2506.08391*, 2025.

[Qian *et al.*, 2024] Yusu Qian, Haotian Zhang, Yinfei Yang, and Zhe Gan. How easy is it to fool your multimodal llms? an empirical analysis on deceptive prompts. *arXiv preprint arXiv:2402.13220*, 2(7), 2024.

[Rohrbach *et al.*, 2018] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018.

[Schuster *et al.*, 2022] Tal Schuster, Adam Fisch, et al. Confident adaptive language modeling. *Advances in Neural Information Processing Systems*, 35:17456–17472, 2022.

[Tong *et al.*, 2025] Bingkui Tong, Jiaer Xia, and Kaiyang Zhou. Mitigating hallucination in multimodal llms with layer contrastive decoding. *arXiv preprint arXiv:2509.25177*, 2025.

[Wang and others, 2025] Sudong Wang et al. Towards understanding how knowledge evolves in large vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29858–29868, 2025.

[Wang *et al.*, 2024a] Fei Wang, Wenxuan Zhou, et al. mDPO: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8078–8088, Miami, Florida, USA, November 2024.

[Wang *et al.*, 2024b] Jiaqi Wang, Yifei Gao, and Jitao Sang. Valid: Mitigating the hallucination of large vision language models by visual layer fusion contrastive decoding. *arXiv preprint arXiv:2411.15839*, 2024.

[Wang *et al.*, 2024c] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics*, pages 15840–15853, 2024.

[Wang *et al.*, 2025a] Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. MLLM can see? dynamic correction decoding for hallucination mitigation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Wang *et al.*, 2025b] Sudong Wang, Yunjian Zhang, Yao Zhu, Enci Liu, Jianing Li, Yanwei Liu, and Xiangyang Ji. Shift: Smoothing hallucinations by information flow tuning for multimodal large language models. In *Proceedings of the International Conference on Computer Vision*, pages 3639–3649, 2025.

[Wang *et al.*, 2025c] Yujun Wang, Jinhe Bi, Soeren Pirk, Yunpu Ma, et al. Ascd: Attention-steerable contrastive decoding for reducing hallucination in mllm. *arXiv preprint arXiv:2506.14766*, 2025.

[Wu *et al.*, 2024] Xiyang Wu, Tianrui Guan, Dianqi Li, Shuaiyi Huang, Xiaoyu Liu, Xijun Wang, Ruiqi Xian, Abhinav Shrivastava, Furong Huang, Jordan Lee Boyd-Graber, et al. Autohallusion: Automatic generation of hallucination benchmarks for vision-language models. *arXiv preprint arXiv:2406.10900*, 2024.

[Xie and others, 2023] Jian Xie et al. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The International Conference on Learning Representations*, 2023.

[Xu *et al.*, 2024] Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, 2024.

[Yin *et al.*, 2025] Hao Yin, Guangzong Si, and Zilei Wang. Lifting the veil on visual information flow in mllms: Unlocking pathways to faster inference. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9382–9391, 2025.

[Yu *et al.*, 2024] Tianyu Yu, Yuan Yao, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024.

[Zhang *et al.*, 2025a] Yu Zhang, Jinlong Ma, et al. Evaluating and steering modality preferences in multimodal large language model. *arXiv preprint arXiv:2505.20977*, 2025.

[Zhang *et al.*, 2025b] Zongmeng Zhang, Wengang Zhou, Jie Zhao, and Houqiang Li. Robust multimodal large language models against modality conflict. *arXiv preprint arXiv:2507.07151*, 2025.

[Zhu *et al.*, 2024] Tinghui Zhu, Qin Liu, Fei Wang, Zhengzhong Tu, and Muhao Chen. Unraveling cross-modality knowledge conflicts in large vision-language models. *arXiv preprint arXiv:2410.03659*, 2024.