CACM: Conflict-Aware Cross-Modal Grounding for Mitigating Text Dominance in Vision-Language Models

Anonymous ACL submission

Abstract

Large Vision-Language Models (LVLMs) exhibit a systematic failure we term **Asymmetric Modality Reliance**: when textual context conflicts with visual evidence, models disproportionately defer to language and ignore the image. This bias persists across model scales, and existing prompt-based remedies offer only marginal, inconsistent gains. We propose the **Conflict-Aware Cross-Modal Module (CACM)**, a lightweight component that grounds textual tokens in visual evidence and softly modulates unreliable preamble tokens along a learnable conflict direction, jointly trained with an auxiliary conflict-classification objective. Trained only on a VQAv2-based dataset augmented with visually grounded conflicting preambles, CACM transfers zero-shot to four benchmarks (ScienceQA, POPE, DocVQA, MathVista) and substantially improves robustness under conflicting contexts while preserving performance under consistent inputs.

1 Introduction

Recent years have witnessed extraordinary progress in vision-language modeling (Liu et al., 2023; Dai et al., 2023; OpenAI, 2023), yet this progress has largely been measured in settings free from misleading textual priors, a convenient assumption that rarely holds in noisy, real-world deployments. Figure 1 illustrates a scenario in which a Vision-Language Model is presented with a photograph of a snow-covered mountain range alongside a conflicting textual prompt that injects a misleading semantic prior. for instance, a preamble asserting that "The mountains are completely bare" followed by the question: "Is there snow?" Remarkably, instead of relying on the visual evidence, state-of-the-art models will often succumb to the textual prior and flatly answer: "No, the mountains are completely bare." This is not an isolated curiosity. It is a systematic failure (Deng et al., 2025a; Shi et al., 2023; Jia and Liang, 2017) demonstrating how strong linguistic priors can overpower visual grounding, exposing a fundamental question: do Large Vision-Language Models (LVLM) genuinely integrate multiple modalities to answer visual questions, or do they merely default to their textual priors when faced with semantic conflict?

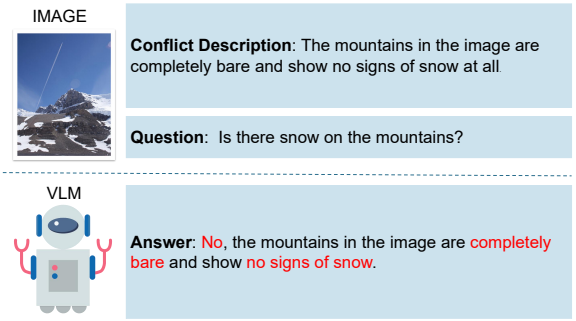


Figure 1: Illustration of asymmetric modality reliance in vision-language models under cross-modal conflict, where models tend to rely more on textual priors than visual evidence, resulting in degraded visual grounding.

The answer to what happens when modalities conflict is deeply unsettling: LVLMs systematically favor textual content over contradictory visual evidence, a phenomenon we term **Asymmetric Modality Reliance**. Rather than resolving conflict through genuine cross-modal reasoning, these models defer to language, treating vision as a disposable signal. This is corroborated by converging evidence: performance on vision-centric tasks collapses to near-chance despite intact visual representations (Fu et al., 2025a); image tokens are progressively overshadowed by textual dynamics in deeper layers (Ghosh et al., 2026a); and VLMs exhibit a pronounced tendency to rely on textual inputs, with a substantial fraction of predictions remaining consistent with misleading text even when it is demonstrably incorrect (Deng et al., 2025a). This issue resembles a well-known problem in NLP, where models rely on spurious correlations instead

of robust reasoning (Gururangan et al., 2018; McCoy et al., 2019). Although the modality setting differs, the underlying failure mode remains similar.

Yet text dominance bias remains critically underaddressed. Existing mitigation strategies largely rely on prompting heuristics or inference-time interventions, such as modality-aware prompting, attention reweighting, and representation steering (Thrush et al., 2022a; Pezeshkpour et al., 2025; Yang et al., 2025; Tang et al., 2026). However, prior work shows that these methods often produce only limited improvements, as they depend heavily on the model's latent verification capabilities and do not fundamentally reshape cross-modal representations learned during training. Recent studies further reveal that VLMs remain highly susceptible to semantically relevant but misleading text, and that larger model scales alleviate this bias only partially (Deng et al., 2025a). Moreover, hidden representations may encode cross-modal conflicts while generated outputs still fail to act on them, exposing a representation–action gap (Quang et al., 2026). Crucially, these conflicts arise from interactions between textual and visual representations inside the model, not from the surface form of the prompt; resolving them therefore demands intervening directly in the representation space rather than patching behavior at inference time.

Building on this insight, we propose the **Conflict-Aware Cross-Modal Module (CACM)**, a lightweight embedding-level component that grounds textual tokens in visual evidence and softly modulates potentially unreliable preamble tokens along a learnable conflict direction. Unlike prompt-based or inference-time interventions, which leave the underlying cross-modal representations untouched, CACM reshapes these representations directly: it performs fine-grained token-level grounding against the visual evidence and injects a learned conflict signal whose strength is adaptively controlled by a predicted per-token reliability score. This enables the model to selectively suppress misleading textual cues while preserving useful contextual information, addressing asymmetric modality reliance at its source rather than its symptoms.

To train CACM, we introduce an image-conditioned data pipeline that produces visually grounded consistent and conflicting preambles, paired with an auxiliary conflict-classification objective. Unlike prior work that synthesizes conflicts from question–answer pairs alone—often yielding preambles weakly tied to visual content—our pipeline conditions generation on the image, ensuring both consistent and conflicting preambles refer to genuinely visible entities and attributes. This yields more realistic conflicts that demand true cross-modal verification rather than exploitation of linguistic cues, while the auxiliary objective activates CACM's conflict-detection capacity. Trained only on VQAv2, CACM transfers zero-shot to four benchmarks (ScienceQA, POPE, DocVQA, MathVista) and consistently outperforms baseline VLMs, larger 13B models, and prior prompt-based methods. In summary, our main contributions are as follows:

- We propose the **Conflict-Aware Cross-Modal Module (CACM)**, a lightweight embedding-level component that grounds textual tokens in visual evidence and softly modulates unreliable preamble tokens along a *learnable conflict direction*, providing a data-driven and end-to-end trainable alternative to handcrafted heuristics and fixed steering vectors.

- We design an image-conditioned data construction pipeline that produces visually grounded consistent and conflicting preambles, paired with an auxiliary conflict-classification objective that explicitly supervises cross-modal inconsistency detection and activates the conflict-detection capacity of CACM.

- Trained only on VQAv2, CACM transfers zero-shot to four diverse benchmarks (ScienceQA, POPE, DocVQA, MathVista) and consistently outperforms baseline VLMs, larger 13B models, and prior prompt-based mitigation strategies, with comprehensive ablations validating the contribution of each design choice.

2 Related Work

Text dominance and visual underutilization. Recent studies show that VLMs often rely disproportionately on textual cues rather than visual evidence (Deng et al., 2025b; Fu et al., 2025b; Ghosh et al., 2026b; Seth et al., 2025). This failure mode has been described as a "blind faith in text" phenomenon, where models follow misleading textual context even when it contradicts the image (Deng et al., 2025b). Other work shows

that VLMs can overlook useful visual representations (Fu et al., 2025b), and that image-token representations may stabilize early across layers, suggesting limited deeper visual processing (Ghosh et al., 2026b). Related work also studies object hallucination, egocentric video hallucination, and visio-linguistic compositional failures (Li et al., 2023; Seth et al., 2025; Thrush et al., 2022b). Together, these results suggest that strong benchmark performance does not necessarily imply faithful multimodal integration.

Cross-modal conflict and robustness. Several works directly evaluate VLM behavior when visual and textual inputs conflict. MAD-Bench uses deceptive prompts to test whether multimodal models follow misleading text over visual evidence (Qian et al., 2024). Other work analyzes VLM reasoning under mixed visual and textual signals, including interventions such as mismatch detection and modality-separated reasoning (Pezeshkpour et al., 2025). Prior mitigation strategies also include image-focused instructions, token reordering, and supervised fine-tuning with text-augmented data (Deng et al., 2025b). Our work builds on this line of research but differs in focus: rather than only diagnosing text dominance or relying on prompting-based interventions, we introduce a lightweight conflict-aware cross-modal module trained to identify unreliable textual preambles and reduce asymmetric modality reliance.

Mitigation strategies for text-dominant bias. Existing mitigation strategies fall into three categories. Prompt-level methods instruct models to verify mismatches or decompose tasks across modalities before fusion (Pezeshkpour et al., 2025; Yang et al., 2025), but their effectiveness hinges on the model's intrinsic competence and often fails to elicit latent verification without explicit guidance. Training-based methods address the bias at the parameter level via text-augmented fine-tuning (Deng et al., 2025b) or LoRA adaptation with hard negative mining and curriculum learning (Zhou, 2026), yet they require extra annotation and risk overfitting to specific conflict patterns. Inference-time interventions instead steer attention toward query-relevant cues (Morini et al., 2026), suppress conflict signals via representation steering (Tang et al., 2026), or re-inject mismatch signals into the output logits (Quang et al., 2026). The latter reveals a representation–action gap: hidden states encode the conflict but outputs fail to act on it, suggesting surface-level fixes are insufficient.

3 Method

We build our model upon the LLaVA architecture (Liu et al., 2023; Cocchi et al., 2025) and introduce a lightweight **Conflict-Aware Cross-Modal Module (CACM)** that enhances robustness against conflicting textual prompts which inject misleading semantic priors. The proposed module operates at the embedding level prior to multimodal sequence encoding and is jointly trained with an auxiliary binary classification objective that explicitly supervises whether a textual prompt introduces a cross-modal conflict with the image. In parallel, the model is encouraged to preserve and exploit modality-consistent information to support downstream reasoning, while retaining the original multimodal capabilities of the backbone.

3.1 Problem Setup

Each training instance consists of an image I, a question Q, and a target answer A. An instance may additionally include a textual preamble T as a textual prompt, which is either visually consistent with the image, denoted by T^+, or in conflict with the visual content, denoted by T^-. For instances with textual preamble, we assign a binary conflict label $y \in \{0, 1\}$, where $y = 1$ indicates a conflicting textual preamble (T^-) and $y = 0$ indicates a visually consistent textual preamble (T^+). The multimodal input is constructed by concatenating the image placeholder token, the textual preamble T, and the question Q. The model is trained autoregressively to generate the target answer A.

3.2 Conflict-Aware Cross-Modal Module

Let $V \in \mathbb{R}^{B \times L_v \times D}$ denote the visual token embeddings projected from the vision encoder, where B is the batch size, L_v is the number of visual tokens, and D is the hidden dimension. We denote the textual token embeddings of the text preamble as $T \in \mathbb{R}^{B \times L_t \times D}$, and the question embeddings as $Q \in \mathbb{R}^{B \times L_q \times D}$. Our conflict-aware modulation is applied exclusively to T, while V and Q remain unchanged throughout the process.

To ground textual semantics in visual evidence, we introduce a cross-attention module in which each preamble token serves as a query and attends to all visual tokens as keys and values. Layer normalization is applied to both modalities prior to interaction to stabilize feature magnitudes. This

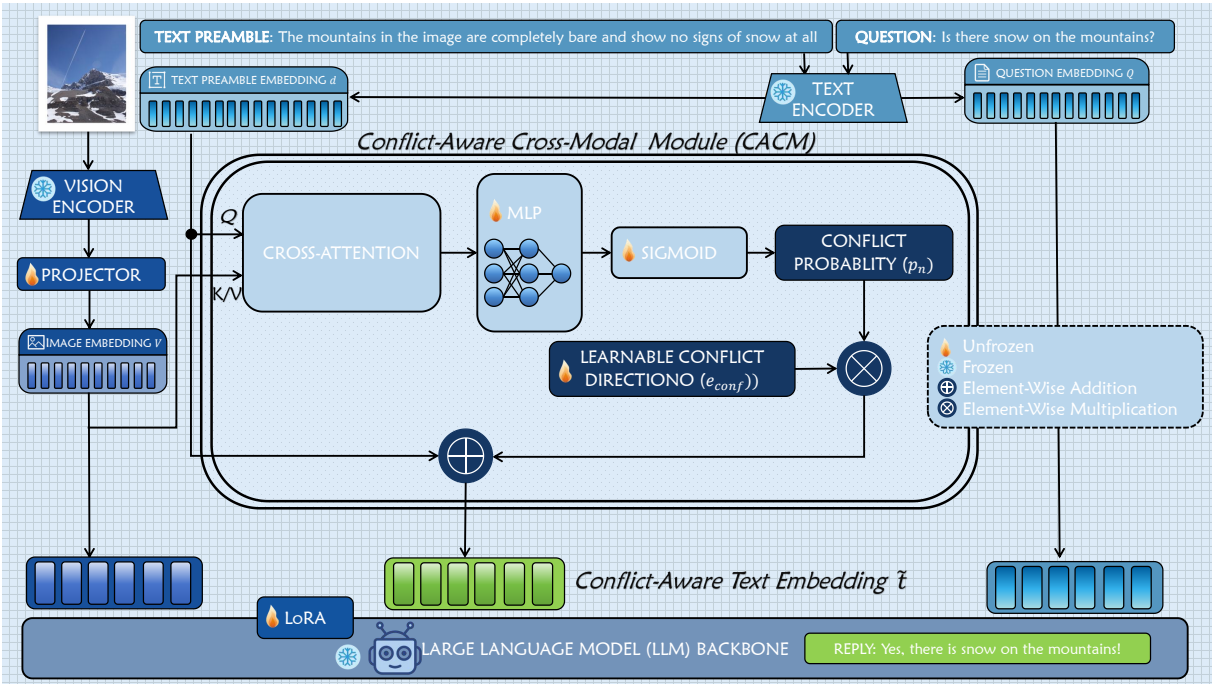


Figure 2: Overview of the proposed Conflict-Aware Cross-Modal Module (CACM) built upon the LLaVA architecture.

produces a set of visually grounded textual representations:

$$C = \mathrm{CrossAttn}(\mathrm{LN}(T), \mathrm{LN}(V), \mathrm{LN}(V)) \in \mathbb{R}^{B \times L_t \times D}$$

where the n-th token representation $c_n \in \mathbb{R}^D$ summarizes the visual evidence most relevant to the corresponding preamble token t_n. Note that C differs from T: while T encodes purely textual semantics, C encodes visual information retrieved from the image for each token. The contrast between the two thus reveals whether the textual semantics are supported by the image.

For each text token $\mathbf{t}_n$, we construct a cross-modal interaction representation inspired by prior multimodal fusion studies that explicitly model inter-modal dependencies and redundancy reduction (Zadeh et al., 2017; Fu et al., 2021).

$$\mathbf{h}_n = [\mathbf{t}_n;\ \mathbf{c}_n;\ \mathbf{t}_n - \mathbf{c}_n;\ \mathbf{t}_n \odot \mathbf{c}_n],$$

where $\odot$ denotes element-wise multiplication. A lightweight feed-forward network f_θ then predicts a scalar conflict logit and its corresponding probability:

$$z_n = f_\theta(\mathbf{h}_n), \qquad p_n = \sigma(z_n),$$

where $\sigma(\cdot)$ denotes the sigmoid function and p_n estimates the likelihood that token $\mathbf{t}_n$ is inconsistent with the visual evidence.

Rather than directly suppressing potentially conflicting tokens, we adopt a soft, embedding-level modulation that allows the model to flexibly adjust its reliance on each textual preamble token. Specifically, we introduce a learnable *conflict direction* $\mathbf{e}_{\text{conf}} \in \mathbb{R}^D$, which represents a shared subspace along which visually unreliable semantics can be encoded. For each token $\mathbf{t}_n$ within $\mathcal{S}_d$, the predicted conflict probability p_n acts as a soft gate that controls how strongly this direction is injected into the token representation:

$$\tilde{\mathbf{t}}_n = \mathbf{t}_n + p_n \cdot \mathbf{e}_{\text{conf}},$$

Intuitively, tokens deemed consistent with the visual evidence ($p_n \to 0$) remain essentially unchanged, whereas tokens identified as conflicting ($p_n \to 1$) are shifted along $\mathbf{e}_{\text{conf}}$, providing downstream layers with an explicit, learnable cue of visual unreliability without erasing the underlying semantic content.

3.3 Training Objective

Although conflict prediction is performed at the token level, supervision is provided only at the sentence (preamble) level, as the annotation indicates whether the preamble as a whole conflicts with the image. We therefore obtain a sentence-level logit by averaging token-level logits over the L_t preamble tokens $\bar{z} = \frac{1}{L_t}\sum_{n=1}^{L_t} z_n$.

For samples with valid conflict annotations, we optimize a binary cross-entropy loss (Goodfellow et al., 2016): which encourages the module to assign higher conflict scores to image-inconsistent textual preamble and lower scores to visually grounded ones.

The overall training objective combines the standard autoregressive language modeling loss with the auxiliary conflict classification loss:

$$\mathcal{L} = \mathcal{L}_{\mathrm{LM}} + \lambda_{\mathrm{BCE}}\, \mathcal{L}_{\mathrm{BCE}},$$

where λ_{BCE} controls the trade-off between generation quality and conflict supervision. For samples without auxiliary descriptions, the BCE term is omitted and optimization relies solely on the language modeling objective $\mathcal{L}_{\mathrm{LM}}$.

3.4 Training Setup

Consistent and Conflict Text Generation. We use Qwen3-VL-8B-Instruct (Bai et al., 2025) to generate both consistent and conflict textual preamble for every VQA datasets. Given an image I, a question Q, and the corresponding answer Y, we prompt the model to produce a Consistent Text and a Conflict Text following the instruction format shown in Table 1. Unlike (Deng et al., 2025b), which generates descriptions solely based on the question and answer, our method additionally conditions on the image content during generation. Although the prompt explicitly requires generating an accurate description for answering the question, generation conditioned only on the question-answer pair frequently produces descriptions that lack proper visual grounding and may contain information weakly related or entirely irrelevant to the actual image content, resulting in semantically inconsistent contextual cues.

The consistent text is grounded in visible evidence and supports the correct answer, while the conflict text refers to the same visual target but leads to a different answer by altering the relevant answer type. We extract ⟨Description 1⟩ and ⟨Description 2⟩ as T_m and T_c, respectively.

These generated texts are then used to construct training samples for improving cross-modal robustness under conflicting textual priors.

Training Data Composition. Each training instance is constructed by randomly selecting one of three settings: image-only (I), consistent text (C), or conflict text (F). Specifically, image-only samples contain only the image without additional textual preambles, consistent samples include the image paired with a visually grounded consistent description T_m, and conflict samples include the image paired with a conflict description T_c that introduces misleading semantic priors.

Table 1: Prompt for consistent and conflict text generation.

Input	Image: {image} Question: {question} Answer: {answer}
Task	Given an image, a question, and the correct answer, generate two sentences: **Description 1:** a short sentence grounded in visible evidence that supports the correct answer. **Description 2:** a short sentence about the same visual target that leads to a different answer type (object, color, count, action, location, or yes/no), while remaining plausible.
Output	Consistent Text: ⟨Description 1⟩ Conflict Text: ⟨Description 2⟩

This mixed data construction encourages the model to learn robust multimodal reasoning under both aligned and conflicting textual contexts. The effect of different sampling ratios among these three types is further studied in the ablation experiments.

To encourage CACM to learn a generalizable ability of detecting image–text conflicts rather than dataset-specific patterns, We train the module only on VQAv2 (Antol et al., 2015) using parameter-efficient fine-tuning with LoRA (Hu et al., 2021), and directly evaluate it on the other four benchmarks (ScienceQA, POPE, DocVQA, MathVista (Lu et al., 2022; Seth et al., 2025; Mathew et al., 2020; Lu et al., 2024)) without any further fine-tuning. This zero-shot transfer protocol tests whether the learned conflict-aware behavior generalizes across diverse domains.

4 Experiments

Our evaluation spans five Visual Question Answering benchmarks across three categories: general visual understanding (VQAv2), hallucination and robustness evaluation (POPE), and specialized reasoning tasks including scientific reasoning (ScienceQA), document understanding (DocVQA), and mathematical visual reasoning (MathVista).

To construct our evaluation data, we follow the setting in Table 1 and augment each original in-

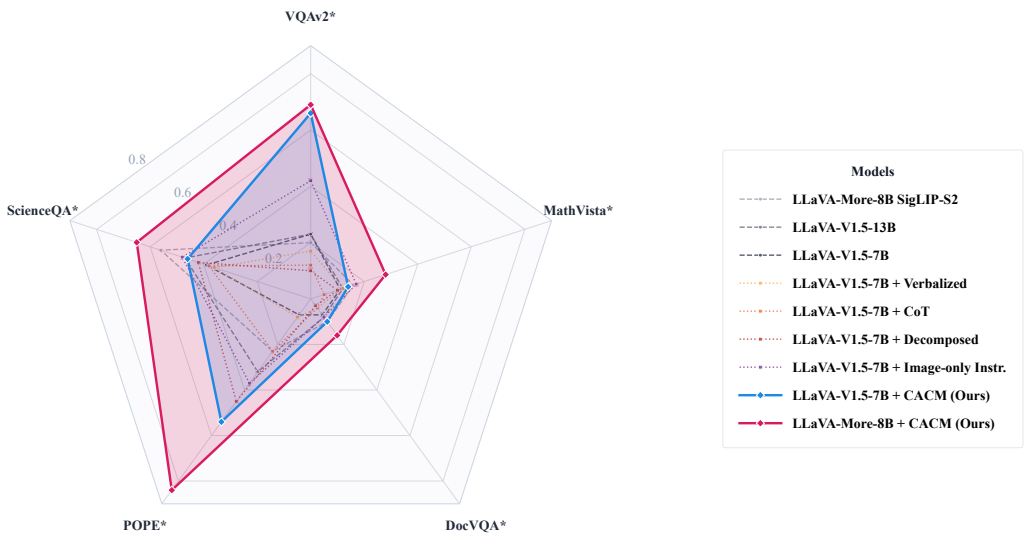


Figure 3: Visualization comparison of baseline VLMs, prior mitigation strategies, and the proposed Conflict-Aware Cross-Modal Module (CACM) on five VQA benchmarks under multimodal conflict settings. The figure highlights the robustness and performance improvements achieved by CACM across different evaluation datasets.

stance by generating conflict textual preambles. Specifically, for every sample, we condition on the image content, question, and ground-truth answer to produce a conflict preamble that introduces semantically plausible but misleading contextual information. See detailed prompts for each dataset in Appendix D.Illustrative examples of the resulting evaluation settings for each dataset are shown in Figure A1.

We compare our method with several mitigation strategies (Pezeshkpour et al., 2025) proposed for reducing modality bias in VLMs. Prompt-based approaches include Verbalized Mitigation, which explicitly instructs the model to report inconsistencies between image and text inputs; CoT Mitigation, which decomposes multimodal reasoning into multiple reasoning steps; and Decomposed Mitigation, which independently processes each modality before performing a final consistency check. Following Deng et al. (Deng et al., 2025a), We further evaluate whether stronger reasoning capabilities improve robustness by replacing the original language backbone with a larger instruction-tuned LLM. Finally, we consider modality-prioritized prompting, where the model is explicitly instructed to rely only on visual information during inference.

4.1 Main Results.

Figure 3 report the accuracy of baseline VLMs, prior mitigation strategies, and our proposed CACM across five VQA benchmarks under conflicting text-preamble settings.

Existing VLMs are highly vulnerable to conflicting textual context. All baselines suffer substantial degradation when conflicting preambles are injected: LLaVA-V1.5-7B achieves only 0.23 on VQAv2* and 0.07 on POPE*, indicating a strong tendency to over-trust text even when it contradicts the visual evidence. Crucially, neither scaling up model size (7B→13B) nor adopting a stronger visual backbone (LLaVA-More-8B with SigLIP-S2 (Cocchi et al., 2025)) yields more than marginal gains, suggesting that conflict robustness cannot be obtained through scaling or stronger pretraining alone.

Prompt-based mitigation provides limited and inconsistent gains. Among three representative prompt-level interventions, Verbalized confidence, Chain-of-Thought (CoT), and Decomposed prompting, none consistently improves robustness, and several even degrade performance (e.g., CoT drops VQAv2* from 0.23 to 0.12), implying that more elaborate reasoning can amplify rather than mitigate misleading textual cues. A simple Image-only Instruction baseline performs best among prompt-based methods (0.42 on VQAv2*, 0.37 on POPE*), confirming that redirecting attention to the visual modality helps, yet still falls short of resolving the underlying cross-modal conflict.

CACM achieves consistent, substantial, and complementary improvements. When integrated into LLaVA-V1.5-7B, CACM raises VQAv2* from

Table 2: Performance comparison under three evaluation settings: image-only (I), consistent text preamble (C), and conflicting text preamble (F), corresponding to inserting different text preambles into the image input. **Bold** indicates the best result.

Method	VQAv2			ScienceQA			POPE			DocVQA			MathVista		
	I	C	F	I	C	F	I	C	F	I	C	F	I	C	F
LLaVA-V1.5-7B	**0.69**	**0.86**	0.23	**0.59**	**0.97**	0.39	**0.89**	**1.00**	0.07	**0.12**	**0.75**	0.07	**0.24**	**0.89**	0.11
LLaVA-V1.5-7B + CACM	0.66	0.75	**0.66**	0.55	0.86	**0.46**	0.85	0.96	**0.54**	0.11	0.71	**0.10**	0.22	0.81	**0.14**
LLaVA-More-8B	**0.69**	**0.81**	0.20	**0.68**	**0.98**	0.56	**0.90**	**0.99**	0.23	**0.12**	**0.80**	0.10	**0.26**	**0.83**	0.16
LLaVA-More-8B + CACM	0.67	0.71	**0.69**	0.64	0.83	**0.65**	0.88	0.88	**0.84**	**0.12**	0.69	**0.16**	**0.26**	**0.83**	**0.28**

Table 3: Ablation studies under three evaluation setting: image-only (I), consistent text preamble (C), and conflicting text preamble (F). **Bold** indicates the best result.

Method	VQAv2			ScienceQA			POPE			DocVQA			MathVista		
	I	C	F	I	C	F	I	C	F	I	C	F	I	C	F
(a) Architecture Ablation															
LLaVA-V1.5-7B-SFT	0.65	**0.76**	0.57	0.52	0.77	0.47	0.79	**0.96**	0.45	0.10	**0.72**	0.10	0.22	0.75	0.17
LLaVA-V1.5-7B + CACM	0.66	0.75	0.66	0.55	**0.86**	0.46	0.85	**0.96**	0.54	0.11	0.71	0.10	0.22	0.81	0.14
LLaVA-More-SFT	0.62	0.65	0.59	0.58	0.82	0.60	0.63	0.82	0.69	0.10	0.67	0.16	0.25	0.77	0.24
LLaVA-More + CACM	**0.67**	0.71	**0.69**	**0.64**	0.83	**0.65**	**0.88**	0.88	**0.84**	**0.12**	0.69	**0.16**	**0.26**	**0.83**	**0.28**
(b) Loss Ablation															
$\lambda_{\text{BCE}} = 0$ (w/o $\mathcal{L}_{\text{BCE}}$)	**0.65**	**0.79**	0.47	**0.56**	**0.84**	0.42	0.83	**0.97**	0.41	0.10	**0.70**	0.07	**0.23**	0.77	0.12
$\lambda_{\text{BCE}} = 0.01$	**0.65**	0.76	0.56	**0.56**	0.74	0.64	**0.85**	0.95	**0.52**	**0.11**	0.69	0.10	0.21	**0.78**	**0.14**
$\lambda_{\text{BCE}} = 0.1$	0.64	0.71	**0.58**	0.53	0.69	**0.65**	0.83	0.94	**0.52**	0.10	0.66	**0.11**	0.21	0.74	**0.14**
(c) Data Composition Ablation (training ratio of I : C : F samples)															
I:C:F = 0.4:0.3:0.3	0.57	0.72	**0.63**	0.51	0.86	**0.46**	0.82	0.96	**0.54**	0.11	0.71	**0.10**	0.21	0.81	**0.13**
I:C:F = 0.6:0.2:0.2	0.64	0.74	**0.63**	0.54	0.88	0.44	0.86	0.97	0.52	**0.12**	0.73	**0.10**	0.22	0.84	**0.13**
I:C:F = 0.7:0.2:0.1	0.64	**0.81**	0.54	**0.57**	0.89	0.42	**0.87**	0.98	0.44	**0.12**	0.74	0.09	**0.23**	0.85	0.12
I:C:F = 0.8:0.1:0.1	**0.66**	0.77	0.53	0.55	**0.90**	0.40	**0.87**	**0.99**	0.39	**0.12**	**0.75**	0.09	**0.23**	**0.86**	0.11

0.23 to 0.66 and POPE* from 0.07 to 0.54, outperforming all prompt-based baselines and even the larger LLaVA-V1.5-13B; on the stronger LLaVA-More-8B backbone it reaches state-of-the-art levels (0.69 VQAv2*, 0.65 ScienceQA*, 0.84 POPE*, 0.16 DocVQA*, 0.28 MathVista*), more than tripling backbone accuracy on hallucination-sensitive POPE. Comparing the two backbones with CACM shows consistent additional gains across all benchmarks, indicating that stronger pretraining and the conflict-aware mechanism are mutually reinforcing. Overall, CACM offers a general architectural solution for cross-modal robustness that clearly outperforms both scaling- and prompt-based alternatives.

4.2 Ablation Studies

We conduct ablations along three complementary dimensions, all under three controlled inference settings, with the loss and data ablations performed on **LLaVA-V1.5-7B + CACM**.

Architecture ablation (Table 3) confirms that the gains stem primarily from CACM rather than training data or backbone scale. Adding CACM consistently improves conflicting-setting robustness for both backbones across nearly all benchmarks (e.g., VQAv2-F rises from 0.57 to 0.66 for LLaVA-V1.5 and from 0.59 to 0.69 for LLaVA-More), narrowing the consistent–conflicting gap on ScienceQA and POPE without sacrificing image-only performance. Notably, although LLaVA-More-SFT already outperforms LLaVA-V1.5-SFT on several benchmarks, substantial conflicting-setting gains emerge only after integrating CACM, demonstrating that the improvement originates from the architectural modification rather than stronger pretraining.

Loss ablation examines the auxiliary binary conflict classification loss $\mathcal{L}_{\text{BCE}}$ via its coefficient λ_{BCE} (Table 3(b)). Removing the supervision ($\lambda_{\text{BCE}} = 0$) markedly degrades conflicting-setting accuracy (VQAv2 $0.56 \rightarrow 0.47$, ScienceQA

$0.64 \rightarrow 0.42$, POPE $0.52 \rightarrow 0.41$), showing that the CACM module provides the structural capacity for conflict detection while the auxiliary loss is essential to activate it during training. Conversely, increasing λ_{BCE} from 0.01 to 0.1 yields only marginal conflicting-setting gains while eroding consistent-setting performance (ScienceQA-C $0.74 \rightarrow 0.69$, VQAv2-C $0.76 \rightarrow 0.71$), as overly strong supervision biases the model toward distrusting even reliable text. We therefore adopt $\lambda_{\text{BCE}} = 0.01$ as the best balance.

Data composition ablation studies how the mixing ratio of image-only (I), consistent (C), and conflicting (F) preamble samples shapes the robustness–utility trade-off (Table 3(c)). A low conflicting proportion (0.8/0.1/0.1) maximizes consistent-setting accuracy (0.90 ScienceQA-C, 0.99 POPE-C) but collapses robustness (0.40 ScienceQA-F, 0.39 POPE-F), as insufficient exposure prevents the model from learning cross-modal verification. An excessively high ratio (0.4/0.3/0.3) gives modest conflicting-setting gains (VQAv2-F 0.63) but weakens basic visual grounding (VQAv2-I drops to 0.57, POPE-I to 0.82). The 0.6/0.2/0.2 configuration strikes the best overall balance, preserving image-only and consistent-setting performance while maintaining competitive robustness, and is adopted as our default.

4.3 Training Efficiency Comparison

We further analyze the training efficiency of LLaVA + CACM compared to the standard LLaVA-SFT baseline under identical training settings. While CACM introduces additional computation due to cross-modal conflict modeling and auxiliary supervision, we aim to evaluate whether this overhead is justified by its performance gains under multimodal conflict settings. As shown in Table 4, CACM introduces a small overhead in both runtime and memory consumption. Specifically, the average per-iteration time increases from 2.01s to 2.15s, and GPU memory usage increases by 4.3%.

From a systems perspective, these results indicate that CACM achieves a favorable efficiency–performance trade-off: a minimal increase in training cost yields disproportionately large gains in robustness under multimodal conflict scenarios.

Table 4: Training efficiency comparison between LLaVA-SFT and LLaVA+CACM on a single GPU. Δ denotes relative change w.r.t. LLaVA-SFT.

Model	Time (s)$\downarrow$	ΔT	Mem (GB)	ΔM
LLaVA-SFT	2.01	–	23.1	–
LLaVA+CACM	2.15	+6.9%	24.1	+4.3%

5 Conclusion

We identified **Asymmetric Modality Reliance**, a systematic failure in which Large Vision-Language Models defer to textual priors when they conflict with visual evidence, and showed that this bias persists across scales and is only marginally addressed by prompt-based remedies. To tackle it at the source, we proposed the **Conflict-Aware Cross-Modal Module (CACM)**, a lightweight embedding-level component that grounds textual tokens in visual evidence and softly modulates unreliable preamble tokens along a learnable conflict direction, jointly trained with an auxiliary conflict-classification objective. Trained on a single source, CACM transfers zero-shot across diverse benchmarks and substantially improves robustness under conflicting inputs while preserving performance under benign ones, offering a general and deployment-friendly step toward more faithful multimodal integration.

6 Limitations

Our work has several limitations that open directions for future research. CACM mitigates reliance on misleading text but does not enhance the backbone's intrinsic reasoning ability, so gains are bounded by the base model's underlying competence. Our study also focuses on sentence-level conflicts injected via a contiguous textual preamble, whereas real-world misleading context may appear in more diverse and fine-grained forms. In addition, training relies on image-conditioned synthetic conflicts whose distribution may differ from naturally occurring or adversarial cases. Finally, CACM implicitly assumes visual reliability and modulates only the textual side; extending it to a symmetric framework that can also discount noisy visual inputs, along with broader validation across more backbones and modalities, remains important future work.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Federico Cocchi, Nicholas Moratelli, Davide Caffagni, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2025. LLaVA-MORE: A Comparative Study of LLMs and Visual Backbones for Enhanced Visual Instruction Tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. 2025a. Words or vision: Do vision-language models have blind faith in text? *Preprint*, arXiv:2503.02199.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. 2025b. Words or vision: Do vision-language models have blind faith in text? *Preprint*, arXiv:2503.02199.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. 2025a. Hidden in plain sight: Vlms overlook their visual representations. *Preprint*, arXiv:2506.08008.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. 2025b. Hidden in plain sight: VLMs overlook their visual representations. *Preprint*, arXiv:2506.08008.

Ziwang Fu, Feng Liu, Hanyang Wang, Jiayin Qi, Xiangling Fu, Aimin Zhou, and Zhibin Li. 2021. A cross-modal fusion network based on self-attention and residual structure for multimodal emotion recognition. *Preprint*, arXiv:2111.02172.

Sambit Ghosh, R. Venkatesh Babu, and Chirag Agarwal. 2026a. Do vision language models need to process image tokens? *Preprint*, arXiv:2604.09425.

Sambit Ghosh, R. Venkatesh Babu, and Chirag Agarwal. 2026b. Do vision language models need to process image tokens? *Preprint*, arXiv:2604.09425.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. *Preprint*, arXiv:1803.02324.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *Preprint*, arXiv:2305.10355.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *Preprint*, arXiv:2310.02255.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Preprint*, arXiv:2209.09513.

Minesh Mathew, Dimosthenis Karatzas, R Manmatha, and CV Jawahar. 2020. Docvqa: A dataset for vqa on document images. corr abs/2007.00398 (2020). *arXiv preprint arXiv:2007.00398*.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *Preprint*, arXiv:1902.01007.

Marco Morini, Sara Sarto, Marcella Cornia, and Lorenzo Baraldi. 2026. Look twice: Training-free evidence highlighting in multimodal large language models. *Preprint*, arXiv:2604.01280.

OpenAI. 2023. Gpt-4v(ision) system card.

Pouya Pezeshkpour, Moin Aminnaseri, and Estevam Hruschka. 2025. Mixed signals: Decoding VLMs' reasoning and underlying bias in vision-language conflict. *Preprint*, arXiv:2504.08974.

Yusu Qian, Haotian Zhang, Yinfei Yang, and Zhe Gan. 2024. How easy is it to fool your multimodal LLMs? an empirical analysis on deceptive prompts. *Preprint*, arXiv:2402.13220.

Trung Nguyen Quang, Yiming Gao, Fanyi Pu, Kaichen Zhang, Shuo Sun, and Ziwei Liu. 2026. Senses wide shut: A representation-action gap in omnimodal llms. *Preprint*, arXiv:2605.13737.

Ashish Seth, Utkarsh Tyagi, Ramaneswaran Selvakumar, Nishit Anand, Sonal Kumar, Sreyan Ghosh, Ramani Duraiswami, Chirag Agarwal, and Dinesh Manocha. 2025. EGOILLUSION: Benchmarking hallucinations in egocentric video understanding. *Preprint*, arXiv:2508.12687.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. *Preprint*, arXiv:2302.00093.

Jing Tang, Kun Wang, Haolang Lu, Hongjin Chen, Kai Chen, Zhongxiang Sun, Qiankun Li, Lingjuan Lyu, Guoshun Nan, and Zhigang Zeng. 2026. Diagnosing knowledge conflict in multimodal long-chain reasoning. *ArXiv*, abs/2602.14518.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022a. Winoground: Probing vision and language models for visio-linguistic compositionality. *Preprint*, arXiv:2204.03162.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022b. Winoground: Probing vision and language models for visio-linguistic compositionality. *Preprint*, arXiv:2204.03162.

Haiqi Yang, Jinzhe Li, Gengxu Li, Yi Chang, and Yuan Wu. 2025. Can large multimodal models actively recognize faulty inputs? a systematic evaluation framework of their input scrutiny ability. *Preprint*, arXiv:2508.04017.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *Preprint*, arXiv:1707.07250.

Lijie Zhou. 2026. Cross-modal attention analysis and optimization in vision-language models: A study on visual reliability. *Preprint*, arXiv:2604.17217.

A Additional Related Works

Large vision-language models and multimodal evaluation. Recent large vision-language models (LVLMs), such as LLaVA, InstructBLIP, and GPT-4V, have shown strong multimodal instruction-following and visual reasoning abilities (Liu et al., 2023; Dai et al., 2023; OpenAI, 2023). Their progress is commonly evaluated on benchmarks such as VQA, ScienceQA, DocVQA, POPE, MathVista, MME, and MMBench (Antol et al., 2015; Lu et al., 2022; Mathew et al., 2020; Li et al., 2023; Lu et al., 2024; Fu et al., 2023; Liu et al., 2024). These benchmarks cover visual question answering, multimodal reasoning, document understanding, hallucination, and mathematical reasoning. However, most assume that image and text inputs are semantically aligned, leaving open whether models remain grounded when text introduces misleading priors.

B Results

For completeness, we provide the full numerical results corresponding to Figure 3 in the main paper. Table A1 reports the per-benchmark accuracy of baseline VLMs, prior mitigation strategies, and our proposed CACM method on five VQA benchmarks under multimodal conflict settings. These values serve as a reference for reproducibility and to support the trends discussed in Section 4.1.

C Implementation Details

This appendix lists the fixed hyperparameters shared across all experiments (see Table A2).

C.1 Software and launcher.

We use Python 3.10 with PyTorch (CUDA), Hugging Face transformers, PEFT, and DeepSpeed. Training is launched via torchrun with 4 processes per node on the training script. We use mixed precision with BF16 and TF32 matmul, gradient checkpointing on, and external experiment tracking disabled.

C.2 DeepSpeed (ZeRO Stage-3).

All multi-GPU runs use DeepSpeed ZeRO Stage-3, which partitions optimizer states, gradients, and parameters across the four workers.

C.3 Hardware.

We train our module and fine-tune on a single node with $4\times$ NVIDIA A100-SXM4-80GB GPUs connected by NVLink. No cross-node communication is required. Training data and checkpoints reside on a shared network filesystem.

C.4 DeepSpeed (ZeRO Stage-3).

All multi-GPU runs use DeepSpeed ZeRO Stage-3, which partitions optimizer states, gradients, and parameters across the four workers.

C.5 Base models and vision encoders.

We train our method on two backbones with different language models, vision towers, and pretraining recipes. *LLaVA-1.5-7B.* The first backbone is the official LLaVA-1.5-7B checkpoint with the Vicuna-1.5 language model and the v1 conversation template. The vision tower is CLIP ViT-L/14 at 336 px, and we take patch features from layer 2. The vision-to-LLM projector is a 2-layer MLP with GELU, initialized from the official LLaVA-1.5 pretrained projector. Images are padded to square; we do not use image start/end or patch tokens. The maximum sequence length is 1024 tokens. *LLaVA-MORE LLaMA-3.1-8B (S2 + SigLIP).* The second backbone is the LLaVA-MORE stage-1 pretrained checkpoint (Cocchi et al., 2025) (LLaVA-MORE-llama_3_1-8B-S2-siglip-pretrain). The language model is Meta-LLaMA-3.1-8B-Instruct and the conversation template follows LLaMA-3. The vision tower is SigLIP-SO400M (google/siglip-so400m-patch14-384) used under the S2 (Scale-on-Scale) multi-resolution scheme at 384, 768, and 1152 px. Patch features are again taken from layer 2, and the vision-to-LLM projector is again a 2-layer MLP with GELU , initialized from the LLaVA-MORE pretrained projector. Images are padded to square; we do not use image start/end or patch tokens. The maximum sequence length is 1024 tokens.

C.6 Trainable parameters.

For both backbones we apply the same adaptation recipe: Low-Rank Adaptation (Hu et al., 2021) with rank $r = 32$, $\alpha = 64$, and dropout 0.05. LoRA is applied to all linear layers of the language model *except* those inside the MM projector, vision tower, vision resampler, and conflict module. The MM projector is tuned with a dedicated learning rate of 5×10^{-6}. The conflict module is always instantiated and trained jointly with the LoRA adapters. The backbone LLM weights (Vicuna-1.5 for LLaVA-1.5, LLaMA-3.1 for LLaVA-MORE) and the vision tower are otherwise frozen.

C.7 Conflict module and BCE auxiliary loss.

The conflict module is an image→text cross-attention block (8 heads, hidden size H) followed by a $4H \rightarrow H \rightarrow 1$ MLP scorer with GELU. Image and text streams each have their own LayerNorm. A learnable conflict direction $\mathbf{e}_{\text{conflict}} \in \mathbb{R}^H$ is initialized as $\mathcal{N}(0,\ 0.02^2)$.

C.8 Data and input construction.

Training data come from a merged, shuffled JSONL of slim auxiliary examples with images from the COCO train2014 split. We hold out 10% of records for evaluation with fixed seed 42. Each example is rendered in one of three description modes (image-only, conflict, consistent) sampled with ratios $0.6:0.2:0.2$. We enable lazy preprocessing, grouping by modality length.

C.9 Optimization and learning-rate schedule.

We optimize with AdamW. The LoRA/adapter learning rate is 2×10^{-5}; the MM projector uses a smaller 5×10^{-6}. Weight decay is 0. We use a cosine learning-rate schedule with 3% warmup. Gradient accumulation is 1, with a per-GPU train batch size of 64 (per-GPU eval batch size 16), giving an effective global batch size of $64\times4=256$.

C.10 Training duration, evaluation, and checkpointing.

We train for 3 epoch unless overridden by a fixed step budget. Evaluation runs every 100 steps; checkpoints are written every 20 steps with at most 4 kept on disk. Training metrics are logged at every step.

C.11 Efficiency micro-benchmark setup.

For the per-step training time and peak GPU-memory numbers reported in the efficiency comparison, we use a separate single-GPU configuration to isolate model-side cost from distributed-training overhead. The settings are: one A100-80GB GPU; a plain Python launcher (no DeepSpeed, no ZeRO, no NCCL); per-device batch size 8 (global batch = 8); 30 optimization steps; and a shared training subset of 4,096 samples. All remaining hyperparameters (LoRA, projector, conflict module, BCE weight, sequence length, precision, gradient checkpointing) match those above.

Table A1: Comparison of baseline VLMs, prior mitigation strategies, and our proposed CACM method on five VQA benchmarks under multimodal conflict settings. Results are reported as accuracy (%). "*" denotes evaluation with conflict text preambles added to each instance. **Bold** indicates the best result, and underlined values denote the second-best result.

Model	VQAv2*	ScienceQA*	POPE*	DocVQA*	MathVista*	Avg.
LLaVA-More-8B SigLIP-S2	0.20	0.56	0.23	0.10	0.16	0.25
LLaVA-V1.5-13B	0.23	0.47	0.32	0.08	0.12	0.24
LLaVA-V1.5-7B	0.23	0.39	0.07	0.07	0.11	0.17
LLaVA-V1.5-7B + Verbalized	0.17	0.36	0.08	0.04	0.12	0.15
LLaVA-V1.5-7B + CoT	0.12	0.38	0.23	0.04	0.05	0.16
LLaVA-V1.5-7B + Decomposed	0.10	0.42	0.45	0.03	0.10	0.22
LLaVA-V1.5-7B + Image-only Instruction	0.42	0.48	0.37	0.08	0.17	0.30
LLaVA-V1.5-7B + CACM	0.66	0.46	0.54	0.10	0.14	0.38
LLaVa-More-8B + CACM	**0.69**	**0.65**	**0.84**	**0.16**	**0.28**	**0.52**

Table A2: Summary of fixed hyperparameters for training process.

Category	Setting
Distributed	DeepSpeed ZeRO Stage-3
Precision	BF16 + TF32, gradient checkpointing
Backbone (a)	LLaVA-1.5-7B (Vicuna-1.5, CLIP ViT-L/14 @ 336 px)
Backbone (b)	LLaVA-MORE LLaMA-3.1-8B (SigLIP-SO400M, S2 $\{384, 768, 1152\}$ px)
Vision feature layer	-2 (both)
Projector	MLP2x-GELU (initialized from each backbone's stage-1 projector)
Trainable	LoRA r=32, α=64 + projector + conflict module
LR (LoRA / projector)	2×10^{-5} / 5×10^{-6}
Optimizer / scheduler	AdamW / cosine, warmup 3%
Weight decay	0
Cross-attention heads	8
Per-GPU batch	64
Global batch	256
Gradient accumulation	1
Max sequence length	1024 (LLaVA-1.5) / 1024 (LLaVA-MORE)
Description ratios	0.6 / 0.2 / 0.2 (img-only / conflict / consistent)
Eval split	0.1
Seed	42
Epochs	3
Eval / save interval	every 100 / 20 steps
Saves kept	4
Log interval	every step

D Auxiliary Text Generation Prompts

To produce the consistent and conflict text preamble for every evaluation item, we prompt Qwen3-VL-8B-Instruct with the question, the ground-truth short answer, and (where applicable) the choice list. No image pixels are provided to the language model. For each dataset we use a single joint prompt that asks the model to return one JSON object. The exact prompts used for the five datasets are shown below.

D.1 VQAv2

You are building high-quality auxiliary text for a VQA inference data. Given a VQA question and its correct short answer, generate:
- one consistent_text that clearly supports the correct answer;
- one conflict_text that clearly supports a wrong answer to the same question.

Requirements:
1. Both texts must be directly relevant to the main target of the question.
2. consistent_text must clearly make the correct answer inferable.
3. conflict_text must clearly make a different wrong answer inferable.
4. conflict_text must contradict the correct answer.
5. The two texts should be parallel whenever possible: same target, similar wording, similar length, minimal difference in answer-related content.
6. Do not add unnecessary scene details, background events, or speculative context.
7. Use exactly one sentence for each text.
8. Do not mention that the text is consistent or conflicting.
9. Do not include explanations.

Question: {question}
Correct answer: {answer}

Return exactly one valid JSON object:
{"consistent_text": "...", "conflict_text": "..."}

D.2 POPE

You are building auxiliary text for a yes/no visual presence question (POPE-style). Each line is a distinct image instance. Identifier below is only to diversify wording when the same question repeats; you must not claim you saw pixels or invent scene-specific facts tied to that id.

Given the question and the correct answer ("yes" or "no"), generate:
- consistent_text: one sentence that clearly supports the correct answer ({answer}).
- conflict_text: one sentence that clearly supports the opposite answer ({wrong_answer}) and contradicts the correct one.

Requirements:
1. Both sentences must be directly about what the question asks (presence or absence in the image).
2. consistent_text must make the correct answer ({answer}) easy to infer.
3. conflict_text must make only the wrong side ({wrong_answer}) inferable; do not hedge or imply {answer}.
4. Prefer parallel wording: same focus, similar length, minimal unrelated detail.
5. When the same question and answer repeat for different instance identifiers, vary phrasing; do not reuse the same two sentences verbatim as another instance.
6. Use exactly one sentence for each text.
7. Do not label the texts as consistent or conflicting.
8. Do not add meta commentary.
9. Return exactly one valid JSON object: {"consistent_text": "...", "conflict_text": "..."}

Instance identifier (phrasing anchor only): {instance_key}
Question: {question}
Correct answer: {answer}

D.3 DocVQA

You are building auxiliary text for a document / infographic visual question answering item. Each line is a distinct instance. Identifier below is only to diversify wording when the same question repeats; you must not claim you saw pixels or invent document-specific facts tied to that id.

Given the question and the correct short answer, generate:
- consistent_text: one sentence that clearly supports the correct answer.
- conflict_text: one sentence that clearly supports a different wrong answer to the same question and contradicts the correct one.

Requirements:
1. Both sentences must be directly about what the question asks (text or values on the document/infographic).
2. consistent_text must make the correct answer easy to infer.
3. conflict_text must make a plausible wrong answer inferable; do not hedge or restate the correct answer.
4. Prefer parallel wording: same focus, similar length, minimal unrelated detail.
5. When the same question and answer repeat for different instance identifiers, vary phrasing; do not reuse the same two sentences verbatim as another instance.
6. Use exactly one sentence for each text.
7. Do not label the texts as consistent or conflicting.
8. Do not add meta commentary.
9. Return exactly one valid JSON object: {"consistent_text": "...", "conflict_text": "..."}

Instance identifier (phrasing anchor only): {instance_key}
Question: {question}
Correct answer: {answer}

D.4 ScienceQA

ScienceQA is multiple-choice; the prompt receives the indexed choice list and the correct index.

You are building auxiliary text for a multiple-choice science or education question. Given the question, all choices, the correct index, and the list of wrong choices, generate:
- consistent_text: one sentence that clearly supports the correct choice (index {answer}: "{correct_text}").
- conflict_text: one sentence that clearly and specifically argues for exactly ONE wrong choice listed under "Wrong choices" below. A reader who trusts conflict_text alone should be led to pick that wrong index, not the correct one. Use wording tied to that wrong option (reuse its key nouns/phrases when natural). Do not endorse, name, or imply the correct option.

Requirements:

1. Both sentences must be directly about what the question asks.
2. consistent_text must make the correct choice easy to infer.
3. conflict_text must target a single wrong index among the wrong choices; it must be obviously incompatible with the correct choice and must not hedge between wrong options.
4. Prefer parallel wording: similar length, same focus, minimal unrelated detail.
5. Use exactly one sentence for each text.
6. Do not label the texts as consistent or conflicting.
7. Do not add meta commentary.
8. Return exactly one valid JSON object: {"consistent_text": "...", "conflict_text": "..."}

Question: {question}
Choices: {choices_block}
Correct answer index: {answer}
Wrong choices (conflict_text must clearly support exactly one of these indices; never {answer}): {wrong_choices_block}

D.5 MathVista (multiple-choice)

MathVista contains both multiple-choice and free-form items; we use two parallel templates selected by the question type.

You are building auxiliary text for a multiple-choice math/visual reasoning question. Instance identifier (vary wording when it changes; do not claim you saw the image): {instance_key}

Given the question, all choices, and the correct answer text, generate:
- consistent_text: one sentence that clearly supports the correct answer ("{correct_text}").
- conflict_text: one sentence that clearly argues for exactly ONE wrong choice listed below and contradicts the correct answer.

Requirements:
1. Both sentences must be directly about what the question asks.
2. conflict_text must target a single wrong option; do not hedge.
3. Use exactly one sentence for each text.
4. Return exactly one valid JSON object: {"consistent_text": "...", "conflict_text": "..."}

Question: {question}
Choices: {choices_block}
Correct answer: {correct_text}
Wrong choices (conflict_text must support exactly one of these; never "{correct_text}"): {wrong_choices_block}

D.6 MathVista (free-form)

You are building auxiliary text for a math/visual reasoning question. No image is available.

Instance identifier (vary wording when it changes; do not claim you saw the image): {instance_key}

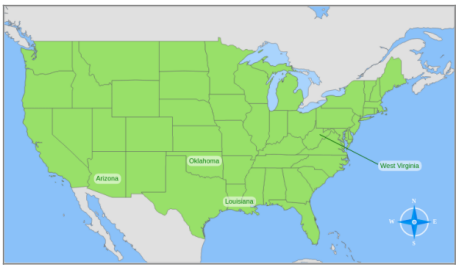
Given the question and the correct short answer, generate:
- consistent_text: one sentence that clearly supports the correct answer.
- conflict_text: one sentence that clearly supports a different wrong answer and contradicts the correct one.

Requirements:
1. Both sentences must be directly about what the question asks.
2. Use exactly one sentence for each text.
3. Return exactly one valid JSON object: {"consistent_text": "...", "conflict_text": "..."}

Question: {question}
Correct answer: {answer}

E Dataset Statistics and Visualization

For each dataset, we present one example containing both consistent and conflicting text. Each evaluation set consists of 1,000 samples per dataset, see Figure A1.



ScienceQA

Q: Which of these states is farthest north?
Choices: (A) West Virginia (B) Louisiana (C) Arizona (D) Oklahoma

Cons.: *West Virginia is located farther north than the other listed states.*

Conf.: *Louisiana is located farther north than Arizona and Oklahoma, making it a more northern state compared to those in the southwestern United States.*

GT: A. West Virginia

POPE

Q: Is there a truck in the image?

Cons.: *A large vehicle is clearly visible in the foreground.*

Conf.: *No vehicles are present in the scene.*

GT: yes

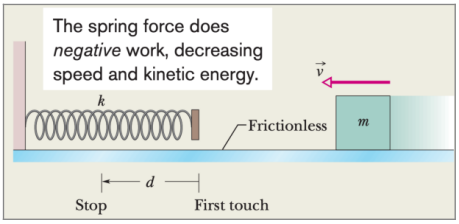
DocVQA

Q: Which social platform has heavy female audience?

Cons.: *Pinterest is known for having a large and active female user base.*

Conf.: *Instagram is popular among younger demographics and has a significant male following.*

GT: pinterest



MathVista

Q: When a spring does work on an object, we cannot find the work by simply multiplying the spring force by the object's displacement. The reason is that there is no one value for the force-it changes. However, we can split the displacement up into an infinite number of tiny parts and then approximate the force in each as…

Cons.: *The spring compression distance is calculated using the conservation of energy, equating the initial kinetic energy of the canister to the elastic potential energy stored in the spring.*

Conf.: *The spring compression distance can be found by simply multiplying the spring force by the displacement, as the force is constant throughout the compression.*

GT: 1.2

VQAv2

Q: How many children are in the image?

Cons.: *There is one child playing on the swing.*

Conf.: *There are two children playing on the swing.*

GT: 1

Figure A1: Examples from the evaluation datasets under multimodal conflict settings.