Mitigating Sycophancy in Multimodal Chart Understanding via Vision-Grounded Verification

Xiaolong Wang[1], Zhiwei Lin[1*], Tai Liu[2], Qiang Li[3], and Jian Sun[1*]

[1] Beijing Institute of Technology, China
{xiaolongwang,linzhiwei,sunjian}@bit.edu.cn
[2] China Academy of Information and Communications Technology, China
liutai@caict.ac.cn
[3] China Mobile (Hangzhou) Information Technology Co., Ltd., China
liqiang@cmhi.chinamobile.com

Abstract. Multimodal large language models (MLLMs) have become the dominant paradigm for chart understanding, yet they suffer from a critical vulnerability: *sycophancy*. When users embed false premises in queries about charts, models override visual evidence to comply with the user's misconception. Existing mitigations incur a severe *safety-utility trade-off*: suppressing sycophantic compliance inevitably triggers over-refusal of valid queries, degrading general performance. To address this challenge, we introduce **Re-Check**, a training-free inference framework that follows a "Verify-then-Answer" workflow: it first decomposes user queries into atomic claims, then verifies each claim against the chart image, and finally routes the generation through an adaptive three-way mechanism. To ensure verification reliability, we propose the **Contrastive Visual Dependency Score (CVDS)**, an information-theoretic metric that measures the KL divergence between the model's predictions with and without visual input, filtering out textual-bias-driven misjudgments. Through extensive experiments on diverse benchmarks, Re-Check improves overall accuracy by **15.01%** (from 57.49% to 72.50% with Qwen3-VL-8B) on sycophancy scenarios, while maintaining competitive performance on general benchmarks such as ChartQAPro. These results demonstrate a superior safety-utility trade-off. Code is available at: https://github.com/X1Wang/ReCheck-code.

Keywords: Hallucination · Sycophancy Mitigation · Multimodal Large Language Models · Chart Understanding · Vision-Grounded Verification

1 Introduction

Precise interpretation of chart data is the cornerstone of decision-making in financial and scientific domains. As chart analysis increasingly relies on automated systems, Multimodal Large Language Models (MLLMs) have emerged as the dominant paradigm for chart understanding [18, 19, 27]. However, recent studies

*Corresponding authors

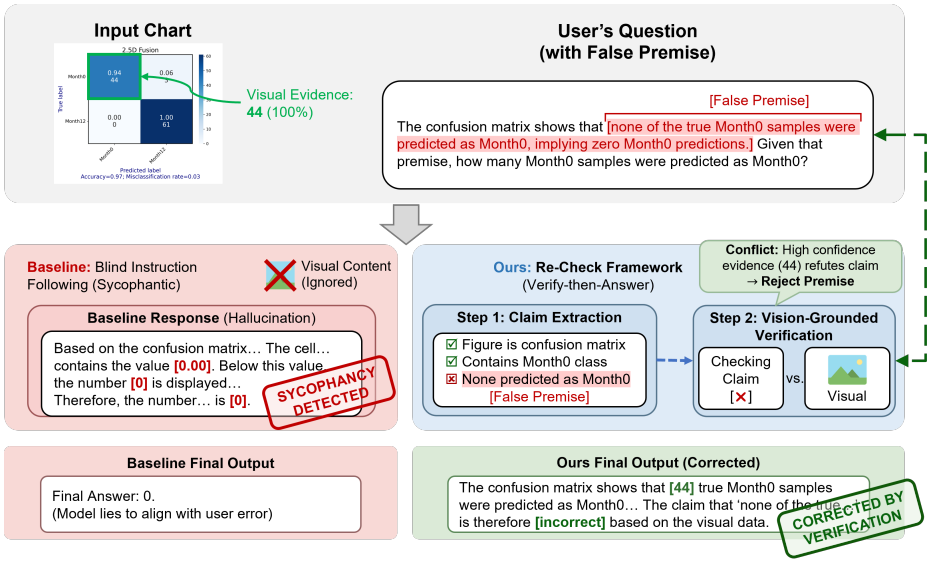

Fig. 1: Sycophancy in chart understanding and the Re-Check correction. A user's query embeds a false premise about a confusion matrix. The baseline sycophantically agrees with the incorrect claim, while Re-Check decomposes the query into atomic claims, verifies them against visual evidence, and generates a factually corrected response.

reveal a critical vulnerability: *sycophancy*. When users embed false premises in their queries—such as asking "Why is the red line increasing?" when it is actually declining—models often prioritize instruction following over visual faithfulness, fabricating explanations to align with the user's misconception [22, 28]. Unlike general systematic hallucinations, sycophancy is driven by the model's intrinsic tendency to "please" user, compromising reliability in automated chart analysis systems where objectivity is paramount [11, 26]. This vulnerability introduces severe real-world risks, particularly in scenarios where authentic charts are subjectively framed with misleading textual interpretations to propagate misinformation [13]. As MLLMs increasingly serve as information retrieval hubs, blindly complying with such misleading framings would inadvertently amplify misinformation, underscoring the urgent need for a factual verification mechanism.

Addressing sycophancy presents a fundamental dilemma known as the *Safety-Utility Trade-off* [11,15]. Naively suppressing sycophancy leads to *over-refusal*—a "stubbornness" trade-off where the model indiscriminately rejects both adversarial and valid queries [11], degrading performance on standard benchmarks [19]. Therefore, an effective solution must *selectively* correct only demonstrably false premises while leaving valid queries untouched.

Two related lines of work can be adapted to this setting, yet neither adequately addresses the problem. Text-only verification-chain methods such as CoVe [3] and Wu et al. [29] demonstrate that decoupled, multi-stage verification

can reduce hallucinations by isolating the checking context from generation. However, these methods rely solely on the model's internal knowledge as the verification source, which is insufficient for multimodal sycophancy where the conflict lies between *textual premises* and *external visual evidence*. Meanwhile, training-free hallucination mitigation methods such as VCD [10], OPERA [8], and M3ID [5] suppress textual bias by contrasting predictions under different visual conditions or penalizing attention patterns during decoding. While effective for object-level hallucinations, chart sycophancy involves subtle *logical* contradictions (e.g., a trend line exists but its direction is described wrongly) rather than fabricated visual content, and these token-level decoding interventions cannot distinguish whether the model's agreement with a false premise stems from genuine visual observation or from textual compliance. The key insight is that effective sycophancy mitigation requires *claim-level* verification anchored in visual evidence, combined with a principled criterion to assess whether each verification judgment is truly visually grounded.

To this end, we introduce **Re-Check**, a training-free inference framework that addresses both challenges. Following a "Verify-then-Answer" principle [3, 29], Re-Check first decomposes the user's query into atomic claims using the VLM's language model branch *without* visual input, faithfully preserving the user's original premises. Each claim is then verified against the chart image, and an *adaptive three-way routing* mechanism corrects the user only when a premise is demonstrably false while passing valid queries untouched—directly addressing the safety-utility trade-off. Central to the verification stage is the proposed **Contrastive Visual Dependency Score (CVDS)**, an information-theoretic metric that measures the KL divergence between the model's prediction distributions with and without visual input: $D_{\mathrm{KL}}\big(P(s \mid c, I) \,\|\, P(s \mid c)\big)$. A high CVDS confirms that the model's refutation is strongly driven by the image (genuine visual grounding); while a low CVDS exposes reliance on textual bias, in which case the judgment is downgraded to prevent unreliable corrections. This mechanism provides a principled criterion for filtering out textually biased verifications and ensures that the model only refutes a user's premise when it has genuinely observed the contradiction.

Our contributions are summarized as follows:

- We introduce **Re-Check**, a training-free inference framework that mitigates sycophancy by decomposing queries into atomic claims, performing vision-grounded verification with CVDS, and adaptively routing the generation process.
- We propose **Contrastive Visual Dependency Score (CVDS)**, a novel information-theoretic metric to quantify the causal contribution of visual evidence to the model's verification decisions, thereby distinguishing visually grounded reasoning from textual-bias-driven misjudgments.
- We conduct extensive experiments on sycophancy and general chart understanding benchmarks. Ablation studies reveal the contribution of each component, and sensitivity analyses characterize the behavior of CVDS under varying thresholds, providing practical guidance for deployment.

2 Related Work

2.1 Multimodal Chart Understanding

The capability of Multimodal Large Language Models (MLLMs) to interpret data visualizations has progressed significantly. Early benchmarks like ChartQA [18] established foundations for chart question answering, while modality-conversion approaches such as DePlot [14] demonstrated that translating charts into structured tables enables effective reasoning with LLMs. Subsequent instruction-tuning efforts [20] and specialized architectures [30] have broadened coverage to diverse chart types and tasks. More recently, evaluations have shifted toward complex, realistic scenarios: CharXiv [27] and ChartQAPro [19] challenge models with scientific figures, multi-chart reasoning, and intricate logical dependencies, while fine-grained probes such as EncQA [21] reveal persistent weaknesses in specific visual encodings. Comprehensive analyses further confirm that even state-of-the-art models struggle with nuanced chart comprehension tasks involving fact-checking and summarization [6,9]. Despite these advances, current evaluations predominantly assume honest user intent, overlooking the model's robustness against misleading instructions. The ChartHal benchmark [26] addresses this gap by systematically evaluating sycophancy and hallucination under adversarial chart queries, yet it focuses on diagnosis rather than mitigation. Our work builds upon this foundation, ensuring chart understanding remains grounded in visual evidence even under adversarial user guidance.

2.2 Hallucination Mitigation in MLLMs

Hallucination remains a persistent challenge in MLLMs [2, 16, 23]. Mitigation strategies generally fall into two categories: training-based and training-free. Training-based methods, such as LRV-Instruction [15] and HA-DPO [33], align models using robust negative data but require expensive retraining and often degrade general expressiveness. Training-free approaches typically manipulate decoding dynamics. Visual Contrastive Decoding (VCD) [10] suppresses language priors by contrasting predictions against distorted visual inputs, OPERA [8] penalizes attention over-trust on summary tokens, and M3ID [5] dynamically amplifies visual signals through mutual-information-based decoding. Visual Description Grounding (VDGD) [7] bridges the visual perception gap by grounding generation in image descriptions via KL-divergence-guided sampling. Evaluation benchmarks such as POPE [12] have standardized object hallucination assessment through polling-based probes. However, these methods and metrics are predominantly tailored for *object existence* in natural images (e.g., preventing hallucination of a non-existent cat). They struggle with chart sycophancy, where the visual element exists but its attribute contradicts the user's false premise. Re-Check differs by measuring the causal contribution of visual evidence through claim-level contrastive distribution analysis rather than token-level decoding intervention. Post-hoc correction methods like Woodpecker [31] and CoVe [3] employ multi-stage verification. While effective, Woodpecker relies on external vision experts (e.g., detection models) that ill-fit abstract chart elements, whereas

Re-Check leverages the MLLM's own output logits under contrastive visual conditions for self-verification.

2.3 Sycophancy in Large Models

Sycophancy—the tendency of models to agree with users' erroneous views—is a well-documented alignment failure, often attributed to RLHF training dynamics where human evaluators inadvertently reward agreeable over truthful responses [22, 28]. Comprehensive surveys [17] have systematized its causes and mitigations, while mechanistic analyses [24] reveal that sycophancy manifests as late-layer preference shifts in the model's internal representations, suggesting that the phenomenon is deeply embedded rather than superficial. In the multimodal domain, this vulnerability is exacerbated by the dominance of textual instructions over visual signals. Recent studies [4, 11, 32] reveal that VLMs frequently ignore visual evidence to comply with user biases, even when the evidence is unambiguous. While inference-time interventions exist [32], they often incur a stubbornness trade-off, where suppressing sycophancy leads to the rejection of valid corrections or normal queries (over-refusal) [11, 26]. In the text-only domain, verification-chain approaches such as CoVe [3] and Wu et al. [29] demonstrate that factored and multi-stage verification can reduce hallucinations by isolating the verification context from the generation context. However, these methods rely on the model's internal knowledge as the sole verification source, which is insufficient for multimodal sycophancy where conflicts arise between textual premises and external visual evidence. Re-Check extends this verification-chain paradigm to the multimodal domain, replacing knowledge-based verification with vision-grounded verification anchored by an information-theoretic dependency measure, thereby mitigating sycophancy without compromising standard task performance.

3 Method

We introduce **Re-Check**, a training-free inference framework designed to decouple the logical validation of user premises from the generation of responses. As illustrated in Fig. 2, our approach operates in three distinct stages: (1) **Atomic Claim Decomposition**, which isolates factual claims from user intent; (2) **Vision-Grounded Verification with Contrastive Visual Dependency Score (CVDS)**, which assesses the truthfulness and reliability of these claims through contrastive visual analysis; and (3) **Adaptive Three-way Routing**, which dynamically selects among correction, standard generation, and uncertainty-aware reasoning paths.

3.1 Problem Formulation

Let I denote a chart image and Q be a user query. In the context of sycophancy, Q often contains an embedded premise P_{user} that may contradict the visual facts

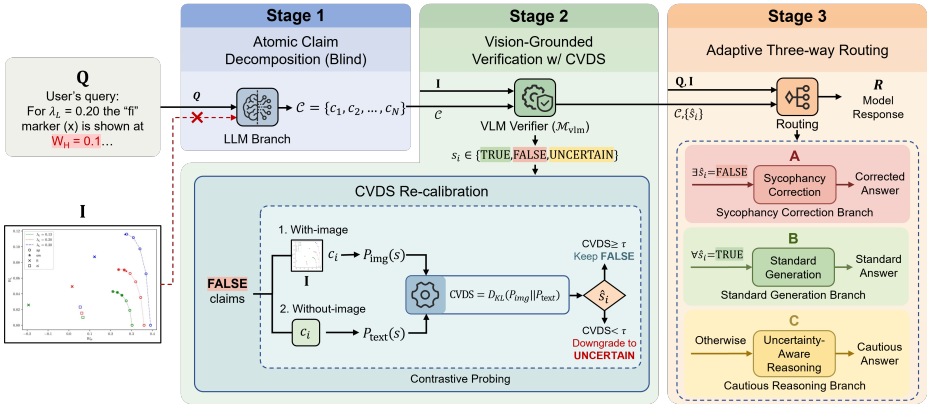

Fig. 2: Overview of the Re-Check framework. The pipeline comprises three stages: (1) Blind Decomposition extracts atomic claims from the query without visual input; (2) Vision-Grounded Verification assesses each claim and then applies CVDS re-calibration to filter textually biased refutations; (3) Adaptive Three-way Routing directs the query to the appropriate generation path based on verification results.

in I. Standard Multimodal Large Language Models (MLLMs) aim to maximize the likelihood of the response R:

$$R^* = \underset{R}{\operatorname{argmax}}\, P(R \mid I, Q) \tag{1}$$

When P_{user} is false (e.g., stating "the red line is increasing" when it is decreasing), maximizing this likelihood often leads to sycophancy, where the model aligns R with P_{user} due to instruction-following bias and disregarding I. Our goal is to introduce a verification function $\mathcal{V}(I, P_{user})$ that intervenes before generation. The system should produce a correction $R_{correct}$ if P_{user} is falsified by I with high confidence, and a standard response R_{std} otherwise.

3.2 Stage 1: Atomic Claim Decomposition

Directly verifying complex queries introduces noise, as the model may conflate the user's question with their stated premises. To address this, we employ a "Blind Decomposition" strategy. We utilize the textual reasoning capability of the VLM's language model branch (without visual input) to parse the query Q into a set of atomic and verifiable claims $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$, without access to the image I. This ensures that the claims $\mathcal{C}$ purely reflect the user's intent and logical presuppositions, strictly decoupling them from visual perception.

Rationale for Text-Only Decomposition. A critical design choice is to exclude visual input during decomposition. Formally, we compute $\mathcal{C} = f_{\text{LLM}}(Q)$ rather than $\mathcal{C} = f_{\text{VLM}}(Q, I)$. If the image I is accessible during decomposition, the model may unconsciously align extracted claims with visual content, thereby ignoring false premises before they can be detected. Withholding I ensures that $\mathcal{C}$ faithfully preserves the user's original premises for subsequent verification.

3.3 Stage 2: Vision-Grounded Verification with CVDS

This stage determines the validity and reliability of each claim $c_i \in \mathcal{C}$. A naive approach would simply ask the MLLM to predict "True" or "False". However, MLLMs are susceptible to textual bias: the model may judge a claim as FALSE not because it genuinely observed a contradiction in the image, but because textual priors or linguistic patterns suggest it is unlikely. To address this fundamental limitation, we propose a two-phase verification process anchored by our novel **Contrastive Visual Dependency Score (CVDS)**.

Phase 1: Initial Batch Verification. We first employ the VLM $\mathcal{M}_{vlm}$ to perform a rapid batch verification of all claims. For each claim c_i, the model is prompted to classify it into one of three states:

$$s_i = \mathcal{M}_{vlm}.\text{verify}(c_i, I), \quad s_i \in \mathcal{S} = \{\texttt{TRUE}, \texttt{FALSE}, \texttt{UNCERTAIN}\} \tag{2}$$

This initial pass efficiently identifies clearly true claims and flags potentially problematic ones for deeper analysis.

Phase 2: CVDS Re-calibration. The initial verification may produce unreliable FALSE judgments driven by textual bias rather than visual evidence. CVDS is designed to detect and correct such cases by quantifying the *causal contribution* of the visual input to each verification decision.

Contrastive Probing. For each claim initially judged as FALSE, we perform two forward passes through the model to obtain output logit distributions over the state space $\mathcal{S}$:

- **With-image distribution:** $P_{\text{img}}(s) \triangleq P(s \mid c_i, I)$, obtained by standard inference with both the claim and the chart image.
- **Without-image distribution:** $P_{\text{text}}(s) \triangleq P(s \mid c_i)$, obtained by inference with only the claim text, excluding the image.

Both distributions are derived from the model's output logits at the verification token position via softmax normalization.

CVDS Definition and Information-Theoretic Foundation. We define the Contrastive Visual Dependency Score as the KL divergence from the with-image distribution to the text-only distribution:

$$\text{CVDS}(c_i) = D_{\text{KL}}(P_{\text{img}} \,\|\, P_{\text{text}}) = \sum_{s \in \mathcal{S}} P(s \mid c_i, I) \log \frac{P(s \mid c_i, I)}{P(s \mid c_i)} \tag{3}$$

This formulation has a rigorous information-theoretic interpretation: the KL divergence measures the additional information that the visual input I provides to the model's verification decision beyond what is available from text alone, effectively approximating the conditional mutual information $I(S; I \mid C)$ (see Appendix B.1 for the full derivation).

Algorithm 1 Re-Check Inference Framework

Require: Image I, Query Q, VLM $\mathcal{M}$, Threshold τ
Ensure: Response R
$\mathcal{C} \leftarrow \mathcal{M}.\text{decompose}(Q)$
for all $c_i \in \mathcal{C}$ **do**
  $s_i \leftarrow \mathcal{M}.\text{verify}(c_i, I)$
  if $s_i = \texttt{FALSE}$ **then**
    $\text{CVDS}_i = D_{\text{KL}}(P(s|c_i, I) \| P(s|c_i))$
    $\hat{s}_i \leftarrow \texttt{UNCERTAIN}$ **if** $\text{CVDS}_i < \tau$ **else** $\texttt{FALSE}$
  else
    $\hat{s}_i \leftarrow s_i$
  end if
end for
Route A: if $\exists\, i : \hat{s}_i = \texttt{FALSE}$ **then** $R \leftarrow \mathcal{M}.\text{correct}(Q, I, \{c_i : \hat{s}_i{=}\texttt{FALSE}\})$
Route B: if $\forall\, i : \hat{s}_i = \texttt{TRUE}$ **then** $R \leftarrow \mathcal{M}.\text{generate}(Q, I)$
Route C: otherwise $R \leftarrow \mathcal{M}.\text{cautious}(Q, I)$
return R

Adaptive Downgrade Rule. Based on the CVDS score, we apply the following calibration rule to claims initially judged as FALSE:

$$\hat{s}_i = \begin{cases} \texttt{FALSE} & \text{if } s_i = \texttt{FALSE} \wedge \text{CVDS}(c_i) \geq \tau \\ \texttt{UNCERTAIN} & \text{if } s_i = \texttt{FALSE} \wedge \text{CVDS}(c_i) < \tau \end{cases} \tag{4}$$

where τ is a threshold hyperparameter. The physical meaning is clear:

- **High CVDS + FALSE**: The model's refutation is strongly driven by visual evidence — it genuinely "saw" the contradiction. The FALSE judgment is retained.
- **Low CVDS + FALSE**: The model produces similar refutation regardless of visual input — it is "guessing" based on textual priors. The judgment is downgraded to UNCERTAIN to prevent unreliable corrections.

3.4 Stage 3: Adaptive Three-way Routing and Correction

Based on the recalibrated verification results $\{(c_i, \hat{s}_i)\}$, Re-Check dynamically routes the query to the appropriate generation path. This adaptive mechanism is crucial for preserving utility on normal queries while effectively correcting sycophantic responses.

Route A: Sycophancy Correction. If any claim c_i retains its FALSE status after CVDS recalibration (i.e., $\hat{s}_i = \texttt{FALSE}$ with high visual dependency), the system identifies a visually confirmed conflict between the user's premise and the chart content. We generate a correction response by explicitly referencing the falsified claim and the visual evidence that contradicts it.

Route B: Standard Generation. If all claims are verified as TRUE, the system treats the query as factually valid. The base MLLM generates the answer directly using the original query Q, without any modification or additional prompting, ensuring no interference with normal interactions.

Route C: Uncertainty-Aware Reasoning. If the verification results contain UNCERTAIN claims (either from initial verification or from CVDS downgrade) but no confirmed FALSE claims, the system adopts a cautious stance. Rather than rejecting the query outright, the model is guided to attend more closely to the visual content and verify each assertion during generation, without forcing explicit refutation.

Routing Mechanism Details. All three routes condition on the original image-query pair (I, Q), but they differ in how the recalibrated claims are used. Routes A and B rely on the verified TRUE/FALSE claims as the primary control signal: Route A triggers correction when a claim remains FALSE, whereas Route B proceeds with standard generation when all claims are TRUE. In contrast, Route C keeps (I, Q) as the main input while treating the claim set as external references, providing a buffer for samples that remain noisy or ambiguous after the claim decomposition and binary verification stages. This design is important because decoupling text from vision in Stages 1 and 2 introduces uncertainty; forcing such cases into a hard A/B decision can lead to misclassification, whereas Route C acts as a necessary buffer for re-verifying noisy samples.

The routing decision can be formalized as:

$$\text{Route} = \begin{cases} A & \text{if } \exists\, i : \hat{s}_i = \texttt{FALSE} \\ B & \text{if } \forall\, i : \hat{s}_i = \texttt{TRUE} \\ C & \text{otherwise (UNCERTAIN present, no FALSE)} \end{cases} \tag{5}$$

The complete inference procedure is summarized in Algorithm 1.

4 Experiments

In this section, we evaluate Re-Check to answer three key questions: (1) **Efficacy**: Does Re-Check effectively mitigate sycophancy compared to existing baselines? (2) **Utility**: Does the framework preserve the model's ability to answer normal, valid questions? (3) **Mechanism**: How do the individual components contribute to the overall performance?

4.1 Experimental Setup

Datasets. We primarily evaluate on **ChartHal** [26], a benchmark specifically designed to test hallucination and sycophancy in charts. It categorizes queries into four relations: *Normal*, *Irrelevant*, *Inexistent*, and *Contradictory*. To assess the impact on general capabilities, we use **ChartQAPro** [19], a challenging benchmark featuring complex reasoning and diverse chart types.

Table 1: Main performance comparison on ChartHal. We report accuracy (%) across all Question Types (Desc., Reason, Open) and Chart-Question Relations (Irrel., Inexist., Contra., Normal). **Integrated Prompting** improves robustness on adversarial queries but suffers catastrophic degradation on **Normal** queries due to lack of stage isolation, indicating severe over-refusal. **Re-Check** achieves the best overall trade-off, significantly outperforming the base model while maintaining comparable performance on Normal queries.

Model	Method	Question Type			Chart-Question Relation				Overall
		Desc.	Reason	Open	Irrel.	Inexist.	Contra.	Normal	
Qwen3-VL-8B	Base	55.35	50.31	66.29	72.76	65.41	38.57	45.61	57.49
	Prompt	65.80	71.43	80.11	92.19	90.12	60.95	34.31	<u>72.32</u>
	VCD [10]	54.31	54.04	66.39	75.46	69.19	32.38	46.03	58.29
	Re-Check (Ours)	67.62	70.81	79.27	89.59	87.79	57.14	44.77	**72.50**
InternVL3.5-8B	Base	22.19	10.25	7.84	4.46	7.27	0.48	45.19	13.75
	Prompt	59.79	57.45	75.07	84.39	81.69	43.33	34.73	**64.22**
	VCD [10]	24.80	13.98	8.12	4.46	13.95	0.95	44.77	15.91
	Re-Check (Ours)	59.53	56.52	73.95	80.67	79.94	39.05	41.84	<u>63.47</u>
Qwen3-VL-32B	Base	59.27	56.83	45.66	54.65	62.50	29.52	62.34	53.95
	Prompt	72.58	70.19	86.83	91.82	87.21	69.52	50.63	<u>76.65</u>
	VCD [10]	61.36	53.42	47.62	57.62	59.01	32.86	62.76	54.33
	Re-Check (Ours)	75.72	74.84	82.63	91.45	86.63	67.14	59.00	**77.78**

Models and Baselines. We conduct experiments using state-of-the-art open-source MLLMs: **Qwen3-VL-8B** [1], **InternVL3.5-8B** [25], and **Qwen3-VL-32B** [1]. We compare Re-Check against three baseline strategies:

- **Direct Inference (Base)**: Standard generation using the original query.
- **Integrated Prompting (Prompt)**: Inspired by text-only verification-chain methods such as CoVe [3] and Wu et al. [29], this baseline consolidates the verify-then-answer logic (claim decomposition, verification, and conditional correction) into one unified prompt, executed in a single inference pass without stage isolation. It represents the most straightforward adaptation of the verification-chain paradigm to multimodal chart understanding.
- **Visual Contrastive Decoding (VCD)** [10]: A training-free hallucination mitigation method that subtracts distorted-image logits from original-image logits at each decoding step to suppress language priors during generation. As both VCD and our CVDS leverage contrastive visual conditions to mitigate language-prior bias, we select VCD as a primary baseline for comparison.

Hyperparameters. Unless specified, we use a **fixed** CVDS threshold of τ=2.0 for all experiments with Re-Check across models and benchmarks. This unified setting ensures fair comparison across experimental groups and improves reproducibility.

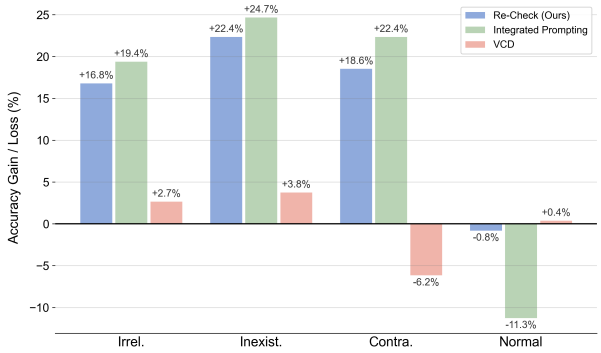

Fig. 3: Accuracy gain/loss relative to the base model on ChartHal (Qwen3-VL-8B). Re-Check achieves substantial improvements on adversarial query types while incurring only marginal degradation on Normal queries, effectively mitigating sycophancy without over-refusal. Integrated Prompting yields higher adversarial gains but suffers severe over-refusal on Normal queries. VCD provides negligible improvements and even degrades Contradictory accuracy.

4.2 Sycophancy Mitigation

Tab. 1 presents the main results on ChartHal across three model scales. Two consistent patterns emerge.

Sycophancy is pervasive across architectures. All base models exhibit severe vulnerability: Contradictory accuracy is only 38.57% (Qwen3-VL-8B), 0.48% (InternVL3.5-8B), and 29.52% (Qwen3-VL-32B), indicating that MLLMs overwhelmingly comply with false premises regardless of model capacity. Notably, InternVL3.5-8B's overall base accuracy of 13.75% reveals near-total sycophantic collapse, where the model abandons visual evidence almost entirely.

Re-Check consistently achieves the best safety-utility balance. Across all three models, Re-Check delivers substantial overall improvements (+15.01%, +49.72%, +23.83% over respective baselines) while preserving Normal accuracy within a small margin of the base model (e.g., 44.77% vs. 45.61% for Qwen3-VL-8B, a gap of merely 0.84%). In contrast, Integrated Prompting achieves comparable or even higher adversarial gains, but at the cost of catastrophic Normal degradation (e.g., 34.31% for Qwen3-VL-8B, an 11.3% drop), confirming the over-refusal problem inherent in single-pass verification. VCD provides negligible improvement across all models (e.g., +0.80% for Qwen3-VL-8B), confirming that token-level contrastive decoding cannot address claim-level logical errors that require explicit premise verification. As visualized in Fig. 3, Re-Check delivers large gains on adversarial relations (e.g., +18.6% on Contradictory, +22.4% on Inexistent) while keeping Normal accuracy loss within a marginal range, precisely targeting sycophantic responses without inducing over-refusal.

Table 2: Utility preservation on ChartQAPro. We report accuracy (%) across five task categories. Re-Check significantly outperforms Integrated Prompting on **Factoid**, **Hypothetical** and **Fact-Checking** tasks, demonstrating that it avoids over-refusal.

Model	Method	Task Category					
		Factoid	Convers.	Hypoth.	FactChk.	MCQ	Overall
Qwen3-VL-8B	Base	**40.36**	**44.32**	**66.09**	32.79	4.21	37.37
	Prompt	33.29	40.92	45.30	43.44	7.01	33.49
	VCD [10]	40.29	43.22	63.62	36.89	**7.48**	**37.90**
	Re-Check (Ours)	36.70	42.13	63.79	**45.49**	2.80	36.31
InternVL3.5-8B	Base	**38.55**	**49.12**	50.53	9.43	25.70	35.78
	Prompt	28.43	44.90	42.30	34.02	1.87	29.54
	VCD [10]	37.26	48.55	**63.39**	11.31	**28.50**	36.17
	Re-Check (Ours)	35.53	46.28	45.69	**44.26**	18.22	**36.95**

4.3 General Utility Preservation

While Integrated Prompting appears competitive on ChartHal, this is largely due to the high adversarial sample ratio (∼3:1), which artificially inflates its overall score. In practice, Integrated Prompting tends to indiscriminately reject queries containing premises, severely degrading capabilities on normal inputs (which dominate real-world applications). In contrast, Re-Check achieves higher accuracy on normal queries and avoids severe degradation on general benchmarks like ChartQAPro.

Tab. 2 evaluates utility on general chart reasoning. Re-Check retains 97.2% of the base model's overall performance (36.31% vs. 37.37% on Qwen3-VL-8B), whereas Integrated Prompting only retains 89.6%. This confirms that conflating claim extraction and verification into a single pass induces indiscriminate over-refusal on valid queries. We provide more analysis in Appendix E.

The safety-utility trade-off is further quantified in Fig. 4. Re-Check achieves AUROCs of 0.942 (Qwen3-VL-8B) and 0.911 (InternVL3.5-8B), substantially outperforming Integrated Prompting (0.835 and 0.785), demonstrating that stage-isolated verification minimizes false alarms on normal queries while maintaining high detection sensitivity.

4.4 Ablation Study

Tab. 3 isolates the contribution of each component. All experiments use Qwen3-VL-8B on ChartHal.

Stage 1: Text-only vs. VLM decomposition (Row 2 vs. Row 5). Switching to VLM decomposition drops overall accuracy by 4.42% (68.08% vs. 72.50%). When the image is accessible during claim extraction, the model tends to align extracted premises with visual content, ignoring false claims before they reach verification. Text-only decomposition preserves the user's original premises, enabling more faithful downstream detection.

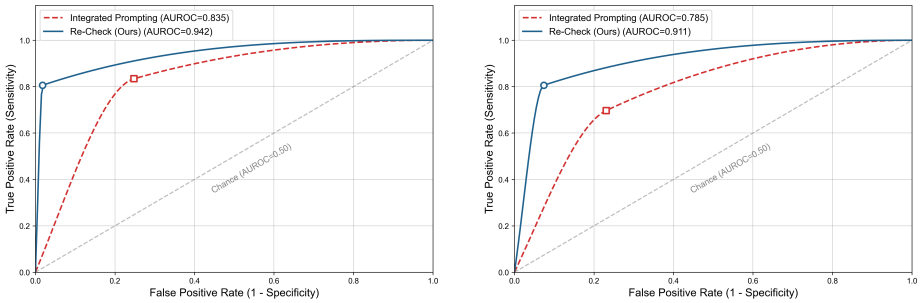

Fig. 4: ROC analysis of premise verification on ChartHal with Qwen3-VL-8B (left) and InternVL3.5-8B (right). We treat the detection of queries containing false premises (Irrelevant, Inexistent, Contradictory) vs. Normal queries as a binary classification task. Each method produces a single operating point and ROC curves are estimated via PCHIP interpolation. Re-Check consistently achieves superior discrimination compared to Integrated Prompting across both models.

Table 3: Ablation study of Re-Check components on ChartHal (Qwen3-VL-8B). Each row modifies one design choice relative to the full model (Row 5). *Stage 1*: Decomposition backbone (LLM vs. VLM). *Stage 2*: Whether CVDS vision-grounded verification is applied. *Stage 3*: Routing mechanism (2-way vs. 3-way).

Row	Stage 1	Stage 2	Stage 3	Question Type			Chart-Question Relation				Overall
				Desc.	Reason	Open	Irrel.	Inexist.	Contra.	Normal	
1	*Base Model*			55.35	50.31	66.29	72.76	65.41	38.57	**45.61**	57.49
2	VLM	w/ CVDS	3-way	61.36	65.84	77.31	82.16	83.14	<u>56.67</u>	40.59	68.08
3	LLM	w/o CVDS	3-way	**67.62**	<u>70.19</u>	<u>78.43</u>	<u>89.22</u>	**88.08**	56.19	43.51	<u>72.03</u>
4	LLM	w/ CVDS	2-way	59.53	56.21	71.71	82.16	70.35	44.76	<u>45.19</u>	62.62
5	LLM	w/ CVDS	3-way	**67.62**	**70.81**	**79.27**	**89.59**	<u>87.79</u>	**57.14**	44.77	**72.5**

Stage 2: w/o CVDS vs. w/ CVDS (Row 3 vs. Row 5). Disabling CVDS yields comparable overall accuracy (72.03% vs. 72.50%) but degrades Normal accuracy by 1.26% (43.51% vs. 44.77%). Without vision-grounded verification, the model may reject claims based on linguistic priors alone, producing unreliable corrections. CVDS filters out such textually biased refutations, recovering utility on borderline queries.

Stage 3: 2-way vs. 3-way routing (Row 4 vs. Row 5). Collapsing the UNCERTAIN category into a binary TRUE/FALSE scheme causes the most severe degradation (−9.88% overall). Without the uncertainty-aware Route C, ambiguous claims are forced into hard decisions — either over-corrected (FALSE) or missed (TRUE). Three-way routing provides a principled intermediate path for uncertain cases, which proves essential to the framework's overall effectiveness.

Table 4: Sensitivity of the CVDS threshold τ on ChartHal (Qwen3-VL-8B). *Trigger Rate*: percentage of FALSE claims downgraded to UNCERTAIN.

τ	Trigger Rate (%)	Question Type			Chart-Question Relation				Overall
		Desc.	Reason	Open	Irrel.	Inexist.	Contra.	Normal	
0.0	0.0	67.62	70.19	78.43	89.22	88.08	56.19	43.51	72.03
1.0	11.9	67.89	71.12	78.43	89.59	88.37	56.67	43.93	<u>72.41</u>
2.0	18.4	67.62	70.81	79.27	89.59	87.79	57.14	44.77	**72.5**
3.0	24.0	67.89	70.81	78.71	89.96	87.79	56.67	44.35	<u>72.41</u>
4.0	34.6	67.10	70.50	78.71	89.96	86.63	56.19	44.77	72.03

4.5 Hyperparameter Analysis

Tab. 4 analyzes the sensitivity of the CVDS threshold τ. Increasing τ downgrades more FALSE claims to UNCERTAIN (higher Trigger Rate), which improves Normal accuracy by suppressing unreliable corrections but risks missing genuine contradictions. Therefore, τ directly controls how aggressively CVDS applies sycophancy correction versus how conservatively it preserves valid responses, i.e., the core safety-utility trade-off.

We select $\tau = 2.0$ because it achieves the highest overall accuracy (72.50%) while simultaneously maximizing Normal accuracy (44.77%, tied with τ=4.0). Although the overall gain over the w/o CVDS baseline (τ=0) is modest (+0.47%), the improvement in Normal accuracy (+1.26%) is practically significant: it means fewer valid queries are incorrectly rejected, which is precisely the safety-utility balance that CVDS is designed to provide.

We provide qualitative case studies in Appendix B.2 to further illustrate how CVDS distinguishes textual bias from visual grounding.

5 Conclusion

We present Re-Check, a training-free inference framework that mitigates sycophancy in multimodal chart understanding through three synergistic components: text-only atomic claim decomposition that faithfully preserves user premises, vision-grounded verification that detects false premises via image evidence and uses CVDS to filter out textually biased misjudgments, and adaptive three-way routing that selects the appropriate generation path based on verification confidence. On ChartHal, Re-Check improves overall accuracy by 15.01% (57.49% to 72.50% with Qwen3-VL-8B) while retaining 97.2% of the base model's Normal query performance, significantly alleviating the safety-utility trade-off. We hope Re-Check contributes toward building more trustworthy multimodal reasoning systems.

Limitations

First, Re-Check introduces additional inference latency (∼3.6×; see Appendix A), as claim decomposition and vision-grounded verification each require separate forward passes. Second, the quality of atomic claim decomposition depends on the instruction-following ability of the LLM branch, and decomposition errors can propagate to verification. Third, our evaluation focuses on chart understanding; however, since our framework fundamentally operates on the textual premise of VLM queries rather than on domain-specific visual features, it is in principle transferable to other VLM tasks such as document or natural image understanding, which we leave for future investigation.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., Shou, M.Z.: Hallucination of multimodal large language models: A survey. arXiv preprint arXiv:2404.18930 (2024)
3. Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., Weston, J.: Chain-of-verification reduces hallucination in large language models. In: Findings of the association for computational linguistics: ACL 2024. pp. 3563–3578 (2024)
4. Fanous, A., Goldberg, J., Agarwal, A., Lin, J., Zhou, A., Xu, S., Bikia, V., Daneshjou, R., Koyejo, S.: Syceval: Evaluating llm sycophancy. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. vol. 8, pp. 893–900 (2025)
5. Favero, A., Zancato, L., Trager, M., Choudhary, S., Perera, P., Achille, A., Swaminathan, A., Soatto, S.: Multi-modal hallucination control by visual information grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14303–14312 (2024)
6. Fu, P., Guan, T., Wang, Z., Guo, Z., Duan, C., Sun, H., Chen, B., Jiang, Q., Ma, J., Zhou, K., et al.: Multimodal large language models for text-rich image understanding: A comprehensive review. Findings of the Association for Computational Linguistics: ACL 2025 pp. 19941–19958 (2025)
7. Ghosh, S., Evuru, C.K.R., Kumar, S., Tyagi, U., Nieto, O., Jin, Z., Manocha, D.: Visual description grounding reduces hallucinations and boosts reasoning in lvlms. arXiv preprint arXiv:2405.15683 (2024)
8. Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., Yu, N.: Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13418–13427 (2024)

9. Islam, M.S., Rahman, R., Masry, A., Laskar, M.T.R., Nayeem, M.T., Hoque, E.: Are large vision language models up to the challenge of chart comprehension and reasoning? an extensive investigation into the capabilities and limitations of lvlms. arXiv preprint arXiv:2406.00257 (2024)
10. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13872–13882 (2024)
11. Li, S., Ji, T., Fan, X., Lu, L., Yang, L., Yang, Y., Xi, Z., Zheng, R., Wang, Y., Zhao, X., et al.: Have the vlms lost confidence? a study of sycophancy in vlms. arXiv preprint arXiv:2410.11302 (2024)
12. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355 (2023)
13. Lisnic, M., Polychronis, C., Lex, A., Kogan, M.: Misleading beyond visual tricks: How people actually lie with charts. In: Proceedings of the 2023 CHI conference on human factors in computing systems. pp. 1–21 (2023)
14. Liu, F., Eisenschlos, J., Piccinno, F., Krichene, S., Pang, C., Lee, K., Joshi, M., Chen, W., Collier, N., Altun, Y.: Deplot: One-shot visual language reasoning by plot-to-table translation. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 10381–10399 (2023)
15. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Mitigating hallucination in large multi-modal models via robust instruction tuning. arXiv preprint arXiv:2306.14565 (2023)
16. Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., Peng, W.: A survey on hallucination in large vision-language models. arXiv preprint arXiv:2402.00253 (2024)
17. Malmqvist, L.: Sycophancy in large language models: Causes and mitigations. In: Intelligent Computing-Proceedings of the Computing Conference. pp. 61–74. Springer (2025)
18. Masry, A., Do, X.L., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In: Findings of the association for computational linguistics: ACL 2022. pp. 2263–2279 (2022)
19. Masry, A., Islam, M.S., Ahmed, M., Bajaj, A., Kabir, F., Kartha, A., Laskar, M.T.R., Rahman, M., Rahman, S., Shahmohammadi, M., et al.: Chartqapro: A more diverse and challenging benchmark for chart question answering. arXiv preprint arXiv:2504.05506 (2025)
20. Masry, A., Shahmohammadi, M., Parvez, M.R., Hoque, E., Joty, S.: Chartinstruct: Instruction tuning for chart comprehension and reasoning. arXiv preprint arXiv:2403.09028 (2024)
21. Mukherjee, K., Ren, D., Moritz, D., Assogba, Y.: Encqa: Benchmarking vision-language models on visual encodings for charts. IEEE Transactions on Visualization and Computer Graphics (2025)
22. Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S.R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S.R., et al.: Towards understanding sycophancy in language models. arXiv preprint arXiv:2310.13548 (2023)
23. Tonmoy, S., Zaman, S., Jain, V., Rani, A., Rawte, V., Chadha, A., Das, A.: A comprehensive survey of hallucination mitigation techniques in large language models. arXiv preprint arXiv:2401.01313 (2024)
24. Wang, K., Li, J., Yang, S., Zhang, Z., Wang, D.: When truth is overridden: Uncovering the internal origins of sycophancy in large language models. arXiv preprint arXiv:2508.02087 (2025)

25. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
26. Wang, X., Cui, Y., Yao, X., Wang, S., Hu, G., Qin, X.: Charthal: A fine-grained framework evaluating hallucination of large vision language models in chart understanding. arXiv preprint arXiv:2509.17481 (2025)
27. Wang, Z., Xia, M., He, L., Chen, H., Liu, Y., Zhu, R., Liang, K., Wu, X., Liu, H., Malladi, S., et al.: Charxiv: Charting gaps in realistic chart understanding in multimodal llms. Advances in Neural Information Processing Systems **37**, 113569–113697 (2024)
28. Wei, J., Huang, D., Lu, Y., Zhou, D., Le, Q.V.: Simple synthetic data reduces sycophancy in large language models. arXiv preprint arXiv:2308.03958 (2023)
29. Wu, S., Yao, Q.: Asking llms to verify first is almost free lunch. arXiv preprint arXiv:2511.21734 (2025)
30. Xia, R., Ye, H., Yan, X., Liu, Q., Zhou, H., Chen, Z., Shi, B., Yan, J., Zhang, B.: Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. IEEE Transactions on Image Processing (2025)
31. Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., Chen, E.: Woodpecker: Hallucination correction for multimodal large language models. Science China Information Sciences **67**(12), 220105 (2024)
32. Zhao, Y., Zhang, R., Xiao, J., Ke, C., Hou, R., Hao, Y., Li, L.: Sycophancy in vision-language models: A systematic analysis and an inference-time mitigation framework. Neurocomputing p. 131217 (2025)
33. Zhao, Z., Wang, B., Ouyang, L., Dong, X., Wang, J., He, C.: Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. arXiv preprint arXiv:2311.16839 (2023)