Taming Object Hallucinations with Verified Atomic Confidence Estimation

Jiarui Liu[1], **Weihao Xuan**[2,3], **Zhijing Jin**[4,5], **Mona Diab**[1]
[1]CMU, [2]The University of Tokyo, [3]RIKEN AIP, [4]MPI, [5]The University of Toronto
jiaruil5@andrew.cmu.edu

Abstract

Multimodal Large Language Models (MLLMs) often suffer from hallucinations, particularly errors in object existence, attributes, or relations, which undermine their reliability. We introduce TACO (Verified Atomic Confidence Estimation), a simple framework that mitigates hallucinations through self-verification and confidence calibration without relying on external vision experts. TACO decomposes responses into atomic queries, paraphrases them to reduce sensitivity to wording, and estimates confidence using self-consistency (black-box) or self-confidence (gray-box) aggregation, before refining answers with a language model. Experiments on five benchmarks (POPE, MME, HallusionBench, AMBER, and MM-Hal Bench) with two MLLMs (`LLaVA-1.5-7B` and `CogVLM2`) show that TACO consistently outperforms direct prompting and Visual Contrastive Decoding, reduces systematic biases, and improves confidence calibration, demonstrating its effectiveness in enhancing the faithfulness of MLLMs.

1 Introduction

Multimodal Large Language Models (MLLMs) have gained significant attention for their ability to bridge computer vision and natural language processing, excelling in tasks such as Visual Question Answering (VQA) (Liu et al., 2023; Yin et al., 2024; Wu and Xie, 2024). Despite this progress, MLLMs remain vulnerable to hallucination, producing responses that are unfaithful to the visual input (Yin et al., 2023; Bai et al., 2024; Huang et al., 2024; Favero et al., 2024). Unlike hallucinations in purely text-based LLMs, hallucinations in MLLMs often manifest as object hallucinations (Li et al., 2023), including errors about an object's existence, attributes, or relations. These errors reduce trustworthiness and limit the adoption of MLLMs in high-stakes applications.

A key cause of hallucination is that MLLMs are overly sensitive to textual variations (Chowdhury and Soni, 2025; Ismithdeen et al., 2025). User queries are naturally diverse and fragmentary, yet current models often struggle to align visual content with such variations. Consequently, small differences in wording can lead to inconsistent predictions, undermining reliability (Guan et al., 2023). Existing methods to mitigate hallucinations typically rely on external experts such as object detectors or image captioning models (Chen et al., 2024a; Yu et al., 2024), or on post-hoc calibration strategies (Min et al., 2023). However, these approaches introduce dependencies on auxiliary models that may not generalize well and can be computationally expensive.

In this work, we introduce TACO (Verified Atomic Confidence Estimation), a simple yet effective framework to address multimodal hallucinations through self-verification and confidence calibration, without relying on external vision models. TACO operates in four stages: (1) atomic query generation, decomposing user queries and model answers into fine-grained atomic queries that can be verified independently; (2) query reformulation, paraphrasing atomic queries into semantically equivalent variations to mitigate sensitivity to surface-level phrasing; (3) confidence estimation, aggregating responses to reformulated queries using either self-consistency (black-box) or self-confidence (gray-box with logits) estimation to identify the most reliable answer; and (4) response refinement, leveraging a language model to integrate verified atomic answers back into a coherent response. Through this design, TACO systematically detects and corrects hallucinations, making MLLM predictions more consistent and faithful to visual input.

We conduct comprehensive experiments across five benchmarks: POPE (Li et al., 2023), MME (Fu et al., 2023), HallusionBench (Guan et al., 2023),

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5430–5444
March 24-29, 2026 ©2026 Association for Computational Linguistics

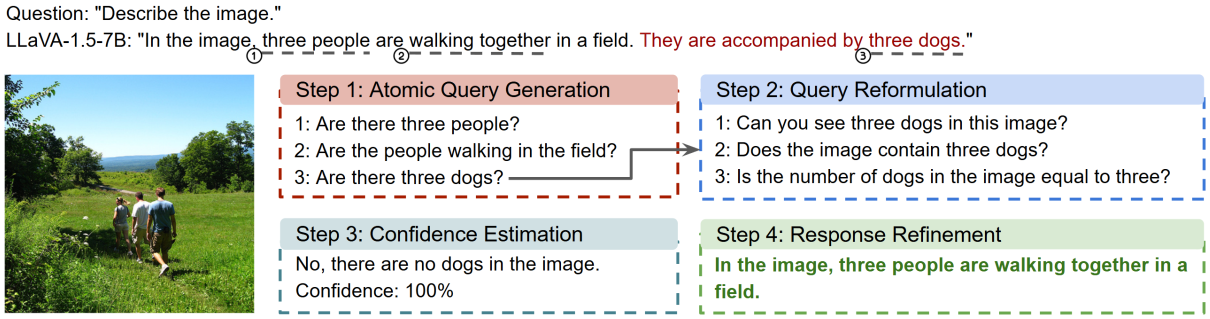

Question: "Describe the image."
LLaVA-1.5-7B: "In the image, three people are walking together in a field. They are accompanied by three dogs."
①
②
③
Step 1: Atomic Query Generation
1: Are there three people?
2: Are the people walking in the field?
3: Are there three dogs?
Step 2: Query Reformulation
1: Can you see three dogs in this image?
2: Does the image contain three dogs?
3: Is the number of dogs in the image equal to three?
Step 3: Confidence Estimation
No, there are no dogs in the image.
Confidence: 100%
Step 4: Response Refinement
In the image, three people are walking together in a field.

Figure 1: Illustration of the TACO pipeline using a generative example across four steps. First, atomic facts are extracted from the query and the original answer, and each fact is framed as a binary atomic query. Second, each atomic query is reformulated into multiple semantically equivalent variations to mitigate the over-sensitivity of MLLMs to surface text. Third, the MLLM's responses to these queries are aggregated, and confidence is estimated using either self-consistency (black-box) or self-confidence (gray-box) to select the more reliable answer. Finally, an LLM refines the MLLM's initial response by incorporating the corrected atomic answers.

AMBER (Wang et al., 2023), and MM-Hal Bench (Sun et al., 2023), and two state-of-the-art MLLMs, `LLaVA-1.5-7B` (Liu et al., 2023) and `CogVLM2` (Hong et al., 2024). Results demonstrate that TACO consistently reduces hallucinations in both discriminative and generative tasks, outperforming direct prompting and Visual Contrastive Decoding (VCD). Notably, we find that self-confidence estimation outperforms self-consistency, showing the advantage of gray-box calibration. Beyond benchmark performance, our analysis further reveals how TACO mitigates systematic biases (e.g., "yes"-answer bias) and improves reliability under query reformulations.

In summary, our contributions are threefold:

1. We propose TACO, a unified framework for mitigating hallucinations in MLLMs through verified atomic confidence estimation.

2. We demonstrate that TACO improves calibration in both black-box and gray-box settings, offering insights into the strengths of self-confidence over self-consistency.

3. We validate TACO across multiple benchmarks and models, showing consistent improvements and providing deeper analysis of its effects on error patterns and biases.

2 Related Work

Fact Verification Existing fact verification methods for text generation typically follow a multi-stage pipeline that leverages external knowledge bases or domain experts (Zhong et al., 2020; Guo et al., 2022; Durmus et al., 2020; Honovich et al., 2022). In VQA, analogous strategies employ external vision experts, such as object detection models (Rohrbach et al., 2018; Li et al., 2023; Wang et al., 2023; Chen et al., 2024b; Sahu et al., 2024; Chen et al., 2024a; Zhou et al., 2025) or image captioning models (Yin et al., 2023; Yu et al., 2024), to provide verification evidence. However, these approaches often inherit the limitations of expert outputs, reducing their robustness on out-of-distribution tasks (Manakul et al., 2023). By contrast, self-verification has been explored in the text domain as a means of validating reasoning steps without reliance on external inputs (Fabbri et al., 2022; Weng et al., 2023; Miao et al., 2023; Ling et al., 2023; Zhang et al., 2023; Li et al., 2024). Inspired by sampling-based self-check techniques in text generation (Manakul et al., 2023) and fact extraction in image-to-text tasks (Min et al., 2023; Cho et al., 2024), our work investigates the potential of self-verification to mitigate multimodal hallucinations, thereby eliminating the need for external vision experts.

Confidence Calibration Modern neural networks are widely recognized for producing poorly calibrated predictions (Guo et al., 2017; Wang et al., 2020; Minderer et al., 2021; Jiang et al., 2021; Xiong et al., 2023). Traditional calibration methods often rely on retraining or the construction of dedicated calibration datasets (Lakshminarayanan et al., 2017; Gal and Ghahramani, 2016; Lee et al., 2018; Yoo et al., 2022). With the advent of LLMs, new approaches that avoid full retraining have emerged. These methods can be broadly categorized into verbalization-based techniques (Lin et al., 2022; Zhou et al., 2023; Mielke et al., 2022; Band et al., 2024), consistency-based techniques (Wang et al., 2022; Xiong et al., 2024; Lyu et al., 2024), and probability-based techniques (Guo et al., 2017; Zhang et al., 2020; Malinin and Gales, 2021; Kuhn et al., 2023; Deng et al., 2023; Xiong et al., 2024).

Verbalization-based methods assess whether models can explicitly articulate their confidence, while consistency-based and probability-based methods calibrate output distributions, either without or with access to logits. Typically, multiple generations are sampled using decoding strategies such as temperature scaling (Lin et al., 2024; Desai and Durrett, 2020), beam search (Kuhn et al., 2023), or prompt variation (Tian et al., 2023; Xiong et al., 2024; Pedapati et al., 2024). Notably, initiated by Leng et al. (2024), a line of probability based methods for visual contrastive decoding (Wang et al., 2024; Favero et al., 2024; Zhang et al., 2024), contrasts token level probability distributions conditioned on the original image with those conditioned on a perturbed version of the input instruction or image to mitigate multimodal hallucination. Despite these advances, little work has systematically examined whether model responses can be calibrated through self-estimated confidence derived from question samples in multimodal contexts.

3 TACO: Verified Atomic Confidence Estimation

Unlike prior work that primarily validates object-existence hallucinations (Wang et al., 2023; Li et al., 2023), our goal is to detect and correct a broader range of object-level hallucinations in MLLM-generated responses, encompassing inaccuracies in object existence, attributes, and relations (Bai et al., 2024). Our approach consists of four stages: atomic query generation, query reformulation, confidence estimation, and response refinement. Given an image and a query, the MLLM first produces an initial response. We then evaluate its self-consistency or self-confidence on each extracted atomic question, and leverage an LLM to refine the response by resolving any identified hallucinations. An overview of this framework is illustrated in Figure 1.

3.1 Atomic Query Generation

To comprehensively address different categories of object hallucinations, it is essential to define the types of verification questions that can be generated. Inspired by (Cho et al., 2024), we introduce a taxonomy of question types in Table 1 and require that all verification questions satisfy two key criteria:

1. Each question must be atomic (Cho et al., 2024), i.e., it should capture the smallest possible semantic unit. This means focusing on a single atomic fact, as specified by the taxonomy in Table 1, and ensuring the question is self-contained and answerable without additional context.
2. Each question must be a positively framed binary question, enabling an unambiguous "yes" or "no" answer.

We employ an LLM to generate atomic questions that provide full semantic coverage of the initial response requiring refinement. To improve this process, we design a two-stage procedure: First, given the user's question and the MLLM's initial response, we extract atomic semantic tuples based on the taxonomy. For example, to verify the existence of a truck, the tuple entity–whole (truck) is instantiated from the "Entity–Whole" category. This tuple is then converted into a binary question such as "Is there a truck?" If the original VQA question is already atomic and binary, it is preserved directly as the output. Additional implementation details are provided in Section A.1.

3.2 Query Reformulation

For each atomic question, we apply question scaling to generate sampled responses for confidence estimation. A straightforward approach would be to use different decoding strategies, such as greedy decoding, beam search, top-p sampling, or temperature scaling. However, these perturbation methods fail to sufficiently explore the MLLM's output space when applied to binary questions. In contrast, question paraphrasing offers a more effective perturbation strategy, as MLLMs are highly sensitive to syntactic variations in text (Ismithdeen et al., 2025). This sensitivity allows paraphrasing to better expose overconfidence and improve calibration. Accordingly, we employ an LLM to paraphrase each atomic question into n variations and evaluate the resulting outputs. Additional implementation details are provided in Section A.2.

3.3 Confidence Estimation

The variance among multiple responses to a given question has been proposed as a proxy for model confidence (Xiong et al., 2024). In this step, given the original image as input, we prompt the MLLM to generate responses to each perturbed

Hallucination Type	Subcategory
Entity	Whole (entire entity, e.g., *boy*), Part (part of entity, e.g., *boy's arm*)
Attribute	State (e.g., *happy emoji*), Color (e.g., *white chalk*), Type (e.g., *aviator goggles*), Text rendering (e.g., *text "START"*), Material (e.g., *plastic bowl*), Shape (e.g., *round plate*), Size (e.g., *long bench*), Count (e.g., *three cars*), Texture (e.g., *flattened surface*), Style (e.g., *realistic photo*), Temporal (e.g., *old clock*)
Relation	Spatial (e.g., *A behind B*), Action (e.g., *A touches B*)

Table 1: Taxonomy used by the atomic query generator to guide the creation of targeted question types. The taxonomy defines core categories of object hallucination, covering entities, attributes, and relations, which can be extended to more comprehensive supersets for broader and benchmark-specific coverage.

atomic question. For each initial atomic question, we collect an answer set $\mathbf{a} = \{\hat{a}_1, \ldots, \hat{a}_n\}$, from which the majority answer is determined as $\bar{a} = \arg\max_{a \in \text{Yes,No}} \sum_{i=1}^{n} \mathbf{1}(\hat{a}_i = a)$ using aggregation functions over candidate answers.

We then propose two methods for estimating and aggregating the MLLM's self-confidence: *black-box assessment* and *gray-box assessment*. Each method produces a confidence score $\texttt{conf}(q, v, \bar{a})$ for atomic question q given visual input v, which is subsequently used to calibrate the reliability of the MLLM's prediction.

Black-Box Assessment We estimate self-confidence by measuring the self-consistency of candidate answers, without requiring access to the model's internal states or output logits. Following prior work that evaluates agreement between candidate responses $\hat{a}_i$ and the majority answer $\bar{a}$ (Wang et al., 2022; Lyu et al., 2024), self-consistency is defined as:

$$C_{\text{self-consistency}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{\hat{a}_i = \bar{a}\}. \tag{1}$$

Gray-Box Assessment When output probabilities are available, we can leverage them as uncertainty indicators. For each paraphrased binary question q_i and visual input v, we extract the probability $\hat{p}(\hat{a}_i \mid q_i, v)$ assigned to the predicted answer $\hat{a}_i \in \{\text{“Yes”}, \text{“No”}\}$. These probabilities are incorporated as weights in the aggregation function used to assess alignment with the majority answer. By default, we use the mean as the aggregation function, while alternative functions are explored in Section 6. The self-confidence score is then computed as:

$$C_{\text{self-confidence}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{\hat{a}_i = \bar{a}\} \cdot \hat{p}(\hat{a}_i \mid q_i, v). \tag{2}$$

3.4 Response Refinement

Guided by the atomic verification questions and their answers, we employ an LLM to refine the MLLM's initial response by incorporating the more confident answer to each atomic query and integrating them into a unified output. Notably, the LLM does not require visual inputs for this step, thereby reducing reliance on the external helper model. If only a single atomic query is generated (i.e., when the original question is already atomic), this refinement step is unnecessary. In our experiments, we use `Claude-3-Sonnet` and `Claude-3.7-Sonnet` as the LLMs. Additional implementation details are provided in Section A.3.

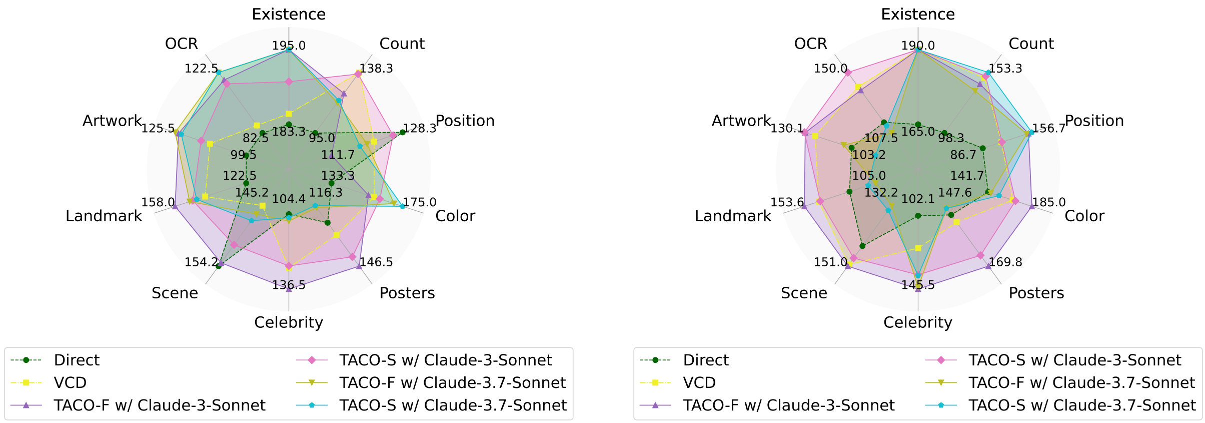
4 Experiment Setup

We evaluate our confidence calibration approach on two state-of-the-art MLLMs, `LLaVA-1.5-7B` and `CogVLM2`. Experiments are conducted across three discriminative benchmarks: POPE (Li et al., 2023), MME (Fu et al., 2023), and HallusionBench (Guan et al., 2023) as well as two generative benchmarks: AMBER (Wang et al., 2023) and MM-Hal (Sun et al., 2023). We compare TACO against the following baselines: (1) direct prompting of the MLLM, and (2) VCD, which applies contrastive decoding by comparing the model's responses with those generated from a perturbed image input (Leng et al., 2024). We refer to our self-consistency approach as TACO-S and our self-confidence approach as TACO-F. Further implementation details are provided in Section A.4.

POPE POPE (Li et al., 2023) introduces a polling-based query method to evaluate an MLLM's ability to answer object-existence questions in images. The task is framed as binary classification using yes/no questions, which improves both stability and flexibility. The benchmark is constructed from 500 MSCOCO images with an equal

Model	Approach	Accuracy			F1		
		Adversarial	Popular	Random	Adversarial	Popular	Random
LLaVA-1.5-7B	Direct	79.60	81.73	83.47	78.36	80.17	81.71
	VCD	81.90	84.83	86.83	81.30	83.84	85.66
	TACO-S w/ Claude-3-Sonnet	83.54	86.58	88.78	83.15	85.78	87.95
	TACO-F w/ Claude-3-Sonnet	84.88	87.93	89.29	84.47	87.21	88.48
	TACO-S w/ Claude-3.7-Sonnet	84.70	87.57	89.07	84.21	86.79	88.20
	TACO-F w/ Claude-3.7-Sonnet	84.80	87.70	89.33	84.40	86.99	88.53
CogVLM2	Direct	84.80	85.90	88.53	84.17	85.06	87.55
	VCD	85.87	87.20	89.97	85.11	86.37	89.04
	TACO-S w/ Claude-3-Sonnet	85.32	85.90	86.84	83.61	84.10	84.99
	TACO-F w/ Claude-3-Sonnet	85.64	86.53	87.42	84.00	84.84	85.70
	TACO-S w/ Claude-3.7-Sonnet	84.70	85.50	86.13	82.53	83.40	83.98
	TACO-F w/ Claude-3.7-Sonnet	84.77	85.67	86.43	82.67	83.64	84.39

Table 2: Results on POPE. *Direct* denotes the direct sampling baseline, and *VCD* refers to the visual contrastive decoding baseline (Leng et al., 2024).



(a) LLaVA-1.5-7B on MME.

(b) CogVLM2 on MME.

Figure 2: Results on the MME benchmark for both models. Subfigures show perception-task performance of (a) LLaVA-1.5-7B and (b) CogVLM2.

number of positive and negative ground-truth examples, yielding 9,000 questions. These are evenly distributed across three settings: *random*, *popular*, and *adversarial*, for generating negative examples. We report accuracy and F1 score.

MME The MME benchmark evaluates the perception and cognition capabilities of MLLMs across 14 tasks with 2374 examples (Fu et al., 2023). Consistent with our definition in Section 3.1, each question is binary, focusing on an atomic fact derived from an image. Following (Yin et al., 2023; Leng et al., 2024), we restrict our evaluation to perception tasks in order to specifically examine object hallucination. We report the score following Fu et al. (2023) based on accuracy.

HallusionBench HallusionBench is a benchmark designed to evaluate the failure modes of MLLMs (Guan et al., 2023). It contains 1,129 questions divided into two categories: *Visual Dependent* and *Visual Supplement*. *Visual Dependent* questions require visual information for accurate answers, while *Visual Supplement* questions can be answered without it, with the visual input serving only as additional context or correction. This design enables assessment of both visual reasoning ability and the interplay between language priors and visual content. All questions are binary under our definition. We report question-pair accuracy (qAcc), figure accuracy (fAcc), and overall accuracy (aAcc) as evaluation metrics.

AMBER AMBER is an LLM-free, multi-dimensional benchmark designed to evaluate hallucinations in MLLMs across both generative and discriminative tasks (Wang et al., 2023). In our work, we use its 1,004 generative questions to assess the effectiveness of our atomic question generation approach. Evaluation is based on four metrics: CHAIR, Cover, Hal, and Cog. CHAIR measures the frequency of hallucinated objects in responses; Cover measures object coverage; Hal measures the proportion of responses containing hallucinations; and Cog measures whether MLLM hallucinations resemble patterns observed in human cognition.

MM-Hal MM-Hal is a benchmark specifically designed to evaluate hallucinations in MLLMs (Sun et al., 2023). Unlike prior benchmarks that focus on general response quality or restrict evaluation to yes/no questions, MM-Hal employs open-ended, realistic questions to better capture hallucination phenomena in practical settings. It consists of 96 image–question pairs across 8 categories and 12 object topics, targeting common failure modes such as incorrect attributes, adversarial objects, counting, comparison, spatial relations, environment, and holistic descriptions. Evaluation is performed with GPT-4o as the judge, comparing model responses against human-written references to determine whether hallucinations are present.

5 Results

POPE The experimental results on POPE across the random, popular, and adversarial settings are shown in Table 2. Our approach consistently outperforms both baselines in terms of accuracy and F1 score across all subtasks on `LLaVA-1.5-7B`, while achieving comparable performance on `CogVLM2`. Importantly, self-confidence estimation consistently surpasses self-consistency estimation across all subtasks and models, demonstrating that leveraging the model's probabilistic distribution leads to more effective confidence calibration. The performances of `Claude-3-Sonnet` and `Claude-3.7-Sonnet` are largely comparable.

MME Figure 2 shows radar plots of evaluation results across MME perception task subsets. Unlike benchmarks focused solely on object existence, MME also evaluates attribute-level hallucinations, providing a broader assessment of model reliability. Our methods consistently reduce hallucinations across different subsets and MLLMs. Consistent with the POPE results, self-confidence estimation generally outperforms self-consistency in terms of overall accuracy. By contrast, the VCD approach yields only marginal gains.

HallusionBench Our evaluation shows that both of our approaches outperform baseline methods across all accuracy metrics. The self-consistency and self-confidence estimation methods achieve largely comparable results, while VCD does not yield clear improvements on this benchmark. This is because HallusionBench provides a more rigorous and adversarial evaluation of VQA dependencies on visual content. In such cases, VCD often fails when the perturbed image does not offer sufficiently informative contrastive cues to guide the model's attention.

AMBER The results for both models in Table 4 show that TACO outperforms the direct prompting baseline on the CHAIR, Hal, and Cog metrics, while maintaining comparable object coverage as measured by COVER. Notably, the hallucination rate decreases substantially, with minimal loss of useful information.

MM-Hal On MM-Hal in Table 5, we find that TACO improves performance for `LLaVA-1.5-7B` but not for `CogVLM2`. Notably, `CogVLM2` outperforms `LLaVA-1.5-7B` by a significant margin. We attribute this to the strong baseline performance of CogVLM2 on this task, where the involvement of an auxiliary language model may introduce negative effects when the original MLLM already possesses sufficient capability to answer the questions.

6 Discussion

Based on our experiments across both discriminative and generative benchmarks, we highlight two key findings: (1) TACO consistently outperforms both direct sampling and visual contrastive decoding baselines, demonstrating its effectiveness as a confidence calibration method. (2) Self-confidence estimation surpasses self-consistency estimation, showing that gray-box assessments provide deeper insight into an MLLM's calibrated confidence.

Below, we further analyze statistics from POPE and HallusionBench to shed light on the underlying behavior of our method.

Does the model exhibit greater consistency across reformulated queries when producing correct answers? To examine whether the dispersion of model outputs (i.e., disagreement among paraphrased responses) is linked to prediction reliability, we analyzed the per-question variance of binary answers on POPE. For each question, we computed the proportion of "yes" responses and

Model	Approach	qAcc (↑)	fAcc (↑)	Easy aAcc (↑)	Hard aAcc (↑)	aAcc (↑)
LLaVA-1.5-7B	Direct	17.58	19.36	42.42	43.72	49.60
	VCD	16.92	18.79	41.10	42.33	49.25
	TACO-S w/ Claude-3-Sonnet	19.56	21.39	46.37	46.28	50.49
	TACO-F w/ Claude-3-Sonnet	14.51	22.54	45.93	48.60	51.20
	TACO-S w/ Claude-3.7-Sonnet	20.88	25.14	49.67	48.60	55.09
	TACO-F w/ Claude-3.7-Sonnet	20.44	26.30	48.35	50.00	55.62
CogVLM2	Direct	21.98	21.68	50.55	42.33	53.06
	VCD	21.32	23.12	52.75	41.40	52.88
	TACO-S w/ Claude-3-Sonnet	30.11	32.08	53.85	56.28	60.50
	TACO-F w/ Claude-3-Sonnet	30.55	36.13	56.48	61.86	63.86
	TACO-S w/ Claude-3.7-Sonnet	24.84	27.46	50.33	55.35	59.26
	TACO-F w/ Claude-3.7-Sonnet	24.62	25.43	50.99	54.19	59.26

Table 3: Results on HallusionBench.

Model	Approach	CHAIR (↓)	COVER (↑)	Hal (↓)	Cog (↓)
LLaVA-1.5-7B	Direct	11.7	51.1	49.5	4.4
	TACO-S w/ Claude-3.7-Sonnet	6.5	48.5	29.1	1.9
	TACO-F w/ Claude-3.7-Sonnet	6.4	49.0	28.8	2.0
CogVLM2	Direct	11.3	61.9	50.9	4.1
	TACO-S w/ Claude-3.7-Sonnet	7.6	59.1	36.6	2.1
	TACO-F w/ Claude-3.7-Sonnet	7.7	59.0	37.1	2.2

Table 4: Results on AMBER. For all metrics except *COVER*, lower values indicate better performance.

Model	Approach	Average Score (↑)	Hallucination Rate (↓)
LLaVA-1.5-7B	Direct	1.85	0.69
	TACO-S w/ Claude-3.7-Sonnet	2.25	0.58
	TACO-F w/ Claude-3.7-Sonnet	2.53	0.52
CogVLM2	Direct	2.71	0.53
	TACO-S w/ Claude-3.7-Sonnet	2.59	0.50
	TACO-F w/ Claude-3.7-Sonnet	2.59	0.51

Table 5: Results on MM-Hal Bench. Lower values indicate better performance for hallucination rate.

derived the variance $p(1-p)$ as a measure of uncertainty. The majority vote was then compared against the ground truth to assess correctness.

We tested whether variance differed significantly between correctly and incorrectly answered questions using Welch's two-sample t-test, and verified robustness with the non-parametric Mann–Whitney U test. We also quantified the relationship between variance and correctness using the point-biserial correlation. Results show that incorrect predictions exhibit substantially higher mean variance than correct ones, with both statistical tests confirming significance and a negative correlation observed between variance and correctness.

These findings support the hypothesis that greater disagreement among paraphrased answers is predictive of reduced majority-vote accuracy. They further suggest that MLLMs are more sensitive to syntactic variation when less confident, consistent with our intuition.

Metric	LLaVA-1.5-7B			CogVLM2		
	Random	Popular	Adv.	Random	Popular	Adv.
T-Test p-value	3.1e-25	3.3e-21	3.0e-23	2.0e-21	6.8e-22	3.5e-22
Mann-Whitney U p-value	4.4e-88	2.1e-58	4.5e-53	3.3e-63	7.3e-61	3.1e-56

Table 6: Statistical significance tests comparing the variance $p(1-p)$ of atomic answer distributions between correctly and incorrectly answered questions on POPE. Reported values are p-values from Welch's two-sample t-test and the Mann–Whitney U test across random, popular, and adversarial settings for LLaVA-1.5-7B and CogVLM2. Lower p-values indicate stronger evidence that variances differ between correct and incorrect answers.

What bias does TACO correct? Prior work has shown that MLLMs often exhibit a bias toward answering "Yes," particularly in existence and attribute verification tasks (Guan et al., 2023). This tendency stems from poor confidence calibration and leads models to over-predict object presence or attributes even when they are absent.

To evaluate whether TACO mitigates this bias, we analyzed the ratio of "Yes" responses on HallusionBench, comparing our approach with baseline methods. Specifically, we measured the deviation in *Yes* prediction percentages from the reference values reported by Guan et al. (2023).

Our results show that TACO effectively reduces the systematic "Yes" bias observed in the baselines, producing more balanced predictions between "Yes" and "No." In particular, our methods substantially lower the internal bias of both models, as reflected in improvements on the Pct Diff and FP Ratio metrics. These findings indicate that TACO not only calibrates confidence more accurately but also helps correct the inherent response imbalance in MLLMs.

Approach	LLaVA-1.5-7B		CogVLM2	
	Pct Diff (~0)	FP Ratio (~0.5)	Pct Diff (~0)	FP Ratio (~0.5)
Direct	0.12	0.62	0.17	0.68
VCD	0.14	0.63	0.19	0.70
TACO-S	0.00	0.50	-0.07	0.41
TACO-F	0.02	0.52	-0.08	0.40

Table 7: Analysis of Yes/No prediction bias on HallusionBench. We report the *Yes Percentage Difference* (Pct Diff; closer to 0 indicates balanced Yes/No predictions) and the *False Positive Ratio* (FP Ratio; closer to 0.5 indicates reduced bias toward predicting "Yes"). Results show that both TACO-S and TACO-F substantially reduce the Yes-bias compared to the direct and VCD baselines for both LLaVA-1.5-7B and CogVLM2.

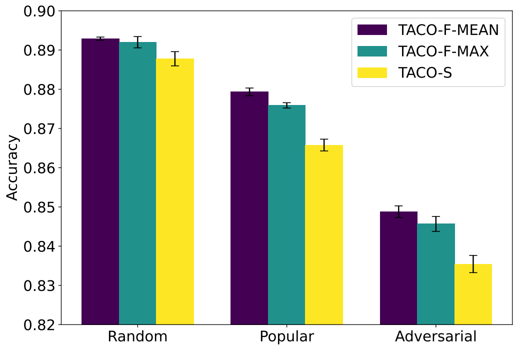


Figure 3: Comparison of aggregation functions for self-confidence estimation on POPE using LLaVA-1.5-7B.

Which aggregation function is better for self-confidence estimation: MEAN or MAX? We compare two aggregation strategies for self-confidence estimation: MAX and MEAN. The MAX function selects the larger probability between predicting Yes" and No" across all paraphrased atomic questions, while the MEAN function computes the average probability of each answer over all paraphrases. As shown in Figure 3, MEAN consistently provides more reliable confidence estimates than MAX, leading to superior calibration performance.

Would removing query reformulation decrease performance? We examine the effect of removing the query reformulation step while retaining atomic verification, and report the results in Table 8. Across both LLaVA-1.5-7B and CogVLM2 on POPE and HallusionBench, this modification consistently leads to performance degradation. For instance, on POPE, accuracy decreases across random, popular, and adversarial settings when reformulation is removed, and on HallusionBench the overall score for CogVLM2 drops from 63.9 to 60.7. Despite this decline, the variant without reformulation remains competitive and generally outperforms direct prompting, indicating that atomic verification provides the primary robustness improvement, while query reformulation further enhances stability and response quality.

Model & Dataset	Metric	Direct	TACO w/o Step 2	TACO
LLaVA-1.5-7B on POPE	Random	0.835	0.890	0.893
	Popular	0.817	0.867	0.879
	Adversarial	0.796	0.817	0.849
CogVLM2 on POPE	Random	0.885	0.864	0.874
	Popular	0.859	0.854	0.865
	Adversarial	0.848	0.843	0.856
LLaVA-1.5-7B on HallusionBench	qAcc	17.58	13.8	14.51
	fAcc	19.36	23.1	22.54
	Easy	42.42	46.3	45.93
	Hard	43.72	47.2	48.60
	Overall	49.60	50.6	51.20
CogVLM2 on HallusionBench	qAcc	21.98	25.5	30.6
	fAcc	21.68	35.0	36.1
	Easy	50.55	55.8	56.5
	Hard	42.33	54.4	61.9
	Overall	53.06	60.7	63.9

Table 8: Ablation study on removing the query reformulation step (Step 2). We use TACO-F w/ Claude-3-Sonnet. Results are reported for LLaVA and CogVLM2 on POPE and HallusionBench.

Can LLMs handle negative questions? During our experiments, we observed an interesting phenomenon: current MLLMs struggle with negative queries. Both LLaVA-1.5-7B and CogVLM2 frequently misinterpret questions such as "Isn't there a ball in the image?" or "Are there no balls in the image?", often producing inconsistent or random answers, even when they correctly answer the corresponding positive query. This suggests that their basic linguistic reasoning capabilities remain limited, highlighting the need for improved training strategies to strengthen textual understanding and better align language with visual inputs.

Would TACO introduce significantly higher latency? The computational cost depends primarily on the number of atomic queries generated from the original question and the number of paraphrases used per query. Suppose a question produces m atomic queries and we generate n paraphrases for each. The total number of model calls becomes: (1) one LLM call for atomic query generation, (2) m LLM calls for reformulation, (3) one LLM cal for answer refinement, and (4) $m \times n$ calls to the underlying MLLM for confidence estimation. Let t_{llm} denote the time for one LLM call and t_{mllm} the time for one MLLM call. The direct prompting baseline requires only t_{mllm}. Our pipeline requires

Dataset	Metric	Direct	TACO-S
POPE	Random F1	0.857	0.875
	Random Acc	0.873	0.888
	Adversarial F1	0.837	0.850
	Adversarial Acc	0.852	0.863
	Popular F1	0.847	0.871
	Popular Acc	0.863	0.882
MME	Score	1611	1711
AMBER	CHAIR (↓)	6.7	5.3
	Cover (↑)	63.3	61.9
	Hal (↓)	34.8	28.4
	Cog (↓)	2.0	1.5

Table 9: Comparison between Direct and TACO-S across POPE, MME, and AMBER benchmarks using `Qwen2.5-VL-7B-Instruct`.

approximately $t_{\text{llm}} \times (1 + m + 1) + t_{\text{mllm}} \times (mn)$. The dominant cost arises from the confidence estimation stage, where the MLLM is queried $m \times n$ times. In practice, we parallelize MLLM calls across multiple replicas so that the system can process reformulated queries in a first come first serve manner. Because API latency can vary significantly and different environments support different acceleration strategies, absolute latency numbers may vary depending on hardware, API conditions, and parallelization settings.

Would TACO generalize to stronger VLMs? To evaluate the generalization ability of TACO on stronger vision language models, we conduct additional experiments using `Qwen2.5-VL-7B-Instruct` (Bai et al., 2025) under the same experimental configuration on POPE, MME, and AMBER. The results are reported in Table 9. Across all three benchmarks, TACO-S consistently improves performance, demonstrating its effectiveness and robustness on more capable VLMs.

7 Conclusion

In this work, we presented a sampling-based confidence calibration framework to address object hallucinations in VQA tasks. Our four-step pipeline systematically generates atomic questions, reformulates them into multiple variations, estimates the MLLM's confidence, and refines its initial response through calibrated feedback. Experiments on multiple benchmarks and state-of-the-art MLLMs show that question paraphrasing is an effective strategy for sampling diverse generations, while LLM-assisted atomic fact extraction and question formulation substantially reduce hallucinations in generative settings. These results highlight the potential of verified atomic confidence estimation as a lightweight yet powerful approach for improving the faithfulness and reliability of MLLMs.

Limitations

While our study demonstrates the effectiveness of TACO in mitigating multimodal hallucinations, several limitations remain that open up directions for future work.

First, our framework is evaluated primarily on a selection of widely used benchmarks. Although these cover both discriminative and generative tasks, they may not fully capture the diversity of real-world multimodal scenarios, such as open-domain dialogue, video understanding, or interactive systems. Extending the evaluation to broader and more dynamic datasets would help verify the generalizability of our approach.

Second, our method currently focuses on hallucinations at the object level: existence, attributes, and relations. While these represent the most common and impactful error modes, hallucinations can also manifest in more abstract forms, such as commonsense reasoning or causal inference. Incorporating atomic query generation for higher-level semantic reasoning remains an interesting direction for future research.

Additionally, our framework relies on an external LLM to perform instruction following steps such as formatting atomic queries, paraphrasing, and rewriting. Although these operations are lightweight compared to visual expert based pipelines, they still require reliable adherence to structured output constraints and moderate few shot learning capability. The LLMs used in our experiment do not represent current frontier performance, and different LLM choices may influence the stability and quality of atomic query generation.

Finally, while we deliberately avoid reliance on external vision experts to keep the framework lightweight and generalizable, there may still be scenarios where integrating complementary signals from specialized vision modules could further enhance robustness. Exploring hybrid approaches that combine self-verification with external guidance in a controllable way may provide a balance between autonomy and reliability. Due to computational constraints, we leave such investigations to future work.

Ethical Considerations

Our work addresses the problem of hallucinations in MLLMs, with the goal of improving the faithfulness and reliability of model outputs. While this research aims to reduce the risks associated with inaccurate or misleading responses, mitigating hallucinations does not guarantee the elimination of all errors. Models calibrated with TACO may still produce inaccurate or biased outputs, particularly when operating on out-of-distribution data. Users should be aware that even with improved confidence estimation, MLLMs should not be blindly trusted in safety-critical domains such as healthcare, law, or autonomous systems without human oversight.

Our framework relies on large-scale pretrained MLLMs and LLMs, which themselves may encode societal biases present in their training data. Although TACO helps calibrate confidence and correct factual inconsistencies, it does not directly address biases or harmful stereotypes inherent to the underlying models.

Additionally, improving faithfulness may inadvertently increase user trust in MLLMs. While this is a desirable outcome for reliability, it also carries the risk of over-reliance, particularly if users assume that calibrated models are universally correct. It is important that system designers clearly communicate residual limitations and provide mechanisms for human-in-the-loop verification.

Acknowledgment

We sincerely thank Renato Negrinho, Manuel Mager, Neha Anna John, Yassine Benajiba, and Chao Shang for their support during the early stages of this project.

This material is based in part upon work supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B; by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645; by Schmidt Sciences SAFE-AI Grant; by the Frontier Model Forum and AI Safety Fund; by Coefficient Giving; by the Survival and Flourishing Fund; and by the Cooperative AI Foundation. The usage of OpenAI credits is largely supported by the Tübingen AI Center and Schmidt Sciences. Resources used in preparing this research project were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Neil Band, Xuechen Li, Tengyu Ma, and Tatsunori Hashimoto. 2024. Linguistic calibration of long-form generations. In *Forty-first International Conference on Machine Learning*.

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024a. Unified hallucination detection for multimodal large language models. *CoRR*, abs/2402.03190.

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024b. Unified hallucination detection for multimodal large language models. *arXiv preprint arXiv:2402.03190*.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. 2024. Davidsonian Scene Graph: Improving Reliability in Fine-Grained Evaluation for Text-to-Image Generation. In *ICLR*.

Souvik Chowdhury and Badal Soni. 2025. R-vqa: A robust visual question answering model. *Knowledge-Based Systems*, 309:112827.

Ailin Deng, Miao Xiong, and Bryan Hooi. 2023. Great models think alike: Improving model reliability via inter-model latent agreement. *arXiv preprint arXiv:2305.01481*.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *Preprint*, arXiv:2310.14566.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, and 1 others. 2024. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland. Association for Computational Linguistics.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. 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*, pages 13418–13427.

Mohamed Insaf Ismithdeen, Muhammad Uzair Khattak, and Salman Khan. 2025. Promptception: How sensitive are large multimodal models to prompts? *arXiv preprint arXiv:2509.03986*.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *ArXiv*, abs/2302.09664.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. 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 (CVPR)*, pages 13872–13882.

Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. 2024. Self-checker: Plug-and-play modules for fact-checking with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 163–181, Mexico City, Mexico. 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. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. Deductive verification of chain-of-thought reasoning. *arXiv preprint arXiv:2306.03872*.

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

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. 2024. Calibrating large language models with sample consistency. *arXiv preprint arXiv:2402.13904*.

Andrey Malinin and Mark John Francis Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Ning Miao, Yee Whye Teh, and Tom Rainforth. 2023. Selfcheck: Using llms to zero-shot check their own step-by-step reasoning. *arXiv preprint arXiv:2308.00436*.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34:15682–15694.

Tejaswini Pedapati, Amit Dhurandhar, Soumya Ghosh, Soham Dan, and Prasanna Sattigeri. 2024. Large language model confidence estimation via black-box access. *arXiv preprint arXiv:2406.04370*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Pritish Sahu, Karan Sikka, and Ajay Divakaran. 2024. Pelican: Correcting hallucination in vision-llms via claim decomposition and program of thought verification. *arXiv preprint arXiv:2407.02352*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the inference calibration of neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3070–3079, Online. Association for Computational Linguistics.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2550–2575, Singapore. Association for Computational Linguistics.

Penghao Wu and Saining Xie. 2024. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094.

Miao Xiong, Ailin Deng, Pang Wei W Koh, Jiaying Wu, Shen Li, Jianqing Xu, and Bryan Hooi. 2023. Proximity-informed calibration for deep neural networks. *Advances in Neural Information Processing Systems*, 36:68511–68538.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*.

KiYoon Yoo, Jangho Kim, Jiho Jang, and Nojun Kwak. 2022. Detection of word adversarial examples in text classification: Benchmark and baseline via robust density estimation. *arXiv preprint arXiv:2203.01677*.

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. 2024. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12944–12953.

Jize Zhang, Bhavya Kailkhura, and T Yong-Jin Han. 2020. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International conference on machine learning*, pages 11117–11128. PMLR.

Yi-Fan Zhang, Weichen Yu, Qingsong Wen, Xue Wang, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2024. Debiasing multimodal large language models. *arXiv preprint arXiv:2403.05262*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning over semantic-level graph for fact checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6170–6180, Online. Association for Computational Linguistics.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524, Singapore. Association for Computational Linguistics.

Mingyang Zhou, Lingyu Zhang, Sophia Horng, Maximillian Chen, Kung-Hsiang Huang, and Shih-Fu Chang. 2025. M^2-TabFact: Multi-document multimodal fact verification with visual and textual representations of tabular data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26239–26256, Vienna, Austria. Association for Computational Linguistics.

A Appendix

A.1 Atomic Query Generation Details

Atomic query generation consists of two steps. First, we prompt an LLM to generate tuples following the provided taxonomy. The prompt is shown below. We provide five-shot examples.

```
Task: Based on the example input
questions, the example output tuples,
and the provided tuple taxonomy below,
generate skill-specific tuples to help
verify and refine the answer of the
last input question.

Requirements:
1. Ensure the generated tuples fully
capture the factual information of the
input question, with each tuple
representing a distinct atomic and
positive statement. Subjective elements
in the initial answer should be
disregarded.
2. If the input question is irrelevant
to any category, output "None."
3. You must remove any negative words
including "not" and "no" from your
generation regardless of whether it
will result in the opposite meaning.
4. Do not generate trivial tuples about
the image itself such as "entity -
whole (image)".
5. Each tuple should be output in the
following format: id | tuple

Tuple taxonomy:
```
Entity relationships:
* entity - whole
* entity - part

Attribute relationships:
* attribute - state
* attribute - color
* attribute - type
* attribute - text rendering
* attribute - material
```
```

~~~
* attribute - shape
* attribute - size
* attribute - count
* attribute - texture
* attribute - style
* attribute - temporal

Relations:
* relation - spatial
* relation - action

Miscellaneous:
* other - other
```
~~~

Second, we prompt the same LLM to convert these tuples into atomic queries. The corresponding prompt is shown below. We conduct experiments with two LLMs, `Claude-3-Sonnet` and `Claude-3.7-Sonnet`. For both models, the temperature is set to 0 and the maximum output length to 1000 tokens. This decoding configuration is applied consistently across all uses of the LLM. We provide two shot examples.

```
Task: Given the example input
questions, skill-specific tuples, and
the example output of generated binary
questions, re-write each tuple from the
last example into a standalone,
positively framed natural language
binary question.

Requirements:
1. Each binary question should be
non-trivial for a vision model to
verify. Exclude trivial tuples that do
not help in verifying and refining the
initial answer.
2. Each binary question should be
self-contained and answerable
independently, without requiring
knowledge of other binary questions.
2. Generate one binary question only
for the two or more tuples sharing the
same meaning or the opposite meaning.
3. Ensure the generated questions fully
capture the factual information of the
input question. Create additional
binary questions if they are helpful
and complementary for refining the
initial answer.
4. Treat conditional statements or
given information in ”Question:” as
context that you don't need to ask
questions from.
5. You must generate positively framed
questions and remove any negative words
including ”not” and ”no” from your
generation regardless of whether it
will result in the opposite meaning.
For example, instead of generating ”is
this artwork not created by Jacob?”,
you should always ask its corresponding
positive question ”is this artwork
created by Jacob?”
output format: id | question
```

A.2 Query Reformulation Details

For each atomic query, we generate nine reformulated variations using the same LLM as above. The corresponding prompt is shown below. We provide two-shot examples.

~~~
Paraphrase the following question about
an image maintaining the exact same
meaning. You must keep the entity names
in the paraphrased questions the same
as in the input question to prevent any
ambiguity. Ensure each generated
question is easily understandable and
can be answered with ”yes” or ”no.”
Generate 10 distinct paraphrased
versions of the question.

Input question:
```
{question}
```

Directly provide your paraphrased
questions in a numbered list without
any explanations.
~~~

A.3 Response Refinement Details

The following prompt is used to guide the LLM in refining the MLLM’s answer.

```
Given a VQA question-answer pair,
refine the model's initial answer using
the context of verification questions
and their ground truth answers.
Preserve the model's answer if the
verification context confirms that the
final answer is correct, even if the
```

```
model's reasoning is flawed. Only
revise the model's answer if the
verification context provides highly
specific and directly relevant evidence
that the final answer itself is
incorrect. If no sufficiently relevant
verification questions are available,
return the initial answer as the
output. Ensure that all output text is
derived from the initial answer or the
provided context; do not generate any
new, unverified information.

Question: "{question}"

Model's initial answer: "{answer}"

Verification context:
```
{verification_qa}
```

Provide only the revised answer without
any explanation or additional text.
```

### A.4 Experiment Details

We conduct our experiments on 8 NVIDIA A100 80G GPUs, totaling approximately 1,000 GPU hours. For all experiments involving MLLMs, results are averaged over three random seeds, and we report the mean performance. We set the decoding temperature to 0.6 to reduce repetitions in text generation and limit the maximum number of new tokens to 1,024. For VCD, we use $\alpha = 1$, $\beta = 0.1$, and 500 noise steps. The total API cost for LLM usage was approximately $1000.

We use the official implementations of `LLaVA-1.5-7B`[1] and `CogVLM2`[2].

### A.5 Dataset details

All datasets used in this work are subject to their respective licenses and are employed strictly for research purposes, consistent with their original intended use. The datasets contain only VQA examples and do not include any personally identifiable information or offensive content.

### A.6 Usage of AI Assistants

We use AI assistants solely to correct grammar and improve clarity of writing.

[1] https://github.com/haotian-liu/LLaVA

[2] https://github.com/zai-org/CogVLM2