Reefknot: A Comprehensive Benchmark for Relation Hallucination Evaluation, Analysis and Mitigation in Multimodal Large Language Models

Kening Zheng[1*], Junkai Chen[1*], Yibo Yan[1,3], Xin Zou[1], Xuming Hu[1,2,3†]
[1]Hong Kong University of Science and Technology (Guangzhou)
[2]Guangxi Zhuang Autonomous Region Big Data Research Institute
[3]Hong Kong University of Science and Technology
{neok2zkn, junkai.chen.0917}@gmail.com
xuminghu@hkust-gz.com

Abstract

Hallucination issues continue to affect multimodal large language models (MLLMs), with existing research mainly addressing object-level or attribute-level hallucinations, neglecting the more complex relation hallucinations that require advanced reasoning. Current benchmarks for relation hallucinations lack detailed evaluation and effective mitigation, and their datasets often suffer from biases due to systematic annotation processes. To address these challenges, we introduce Reefknot, a comprehensive benchmark targeting relation hallucinations, comprising over 20,000 real-world samples. We provide a systematic definition of relation hallucinations, integrating perceptive and cognitive perspectives, and construct a relation-based corpus using the Visual Genome scene graph dataset. Our comparative evaluation reveals significant limitations in current MLLMs' ability to handle relation hallucinations. Additionally, we propose a novel confidence-based mitigation strategy, which reduces the hallucination rate by an average of 9.75% across three datasets, including Reefknot. Our work offers valuable insights for achieving trustworthy multimodal intelligence. The dataset and code are released at https://github.com/JackChen-seu/Reefknot.

1 Introduction

In recent years, large language models (LLMs) have revolutionized the AI field by expanding their training data to trillions of tokens and increasing their parameter counts to hundreds of billions (Brown et al., 2020; Achiam et al., 2023; Touvron et al., 2023). This has unlocked powerful emergent abilities, and seen widespread applications in diverse domains (Achiam et al., 2023; Yan et al., 2024b; Wang et al., 2023a). Recently, the community managed to combine visual backbones with powerful LLMs, resulting in multimodal large language models (MLLMs) (Liu et al., 2023b). While this has led to advancements in multimodal scenarios, it also presents challenges for MLLMs, notably their tendency to generate hallucinations (Zou et al., 2024). In LLMs, hallucinations occur when the model generates inaccurate or misleading factual information that cannot be supported by existing knowledge (Zhang et al., 2023). However, such issues become more complex in MLLMs, as hallucinations can manifest as responses containing references or descriptions of the input image that are incorrect (Bai et al., 2024; Huo et al., 2024a). Therefore, it is crucial to evaluate and mitigate these hallucinations to improve the trustworthy of MLLMs in real-world scenarios.

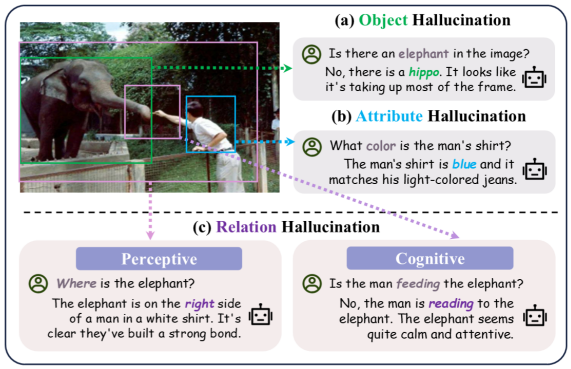


Figure 1: Comparison between the focus of Reefknot — relation hallucination with two categories (i.e., perceptive & cognitive) vs. object & attribute hallucinations.

Hallucination in MLLMs likely originates from knowledge biases between pre-training and fine-tuning, including statistical biases in the training data, over-reliance on parametric knowledge, and skewed representation learning, as suggested by previous research (Bai et al., 2024; Zhu et al., 2024). Specifically, the hallucination in MLLMs can be divided into three categories: ***object***, ***attribute*** and ***relation*** hallucinations (Bai et al., 2024).

*Equal Contribution.
†Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2025, pages 6193–6212
July 27 - August 1, 2025

Benchmarks	Dataset		Evaluation			Analysis	
	Source	Construction	Y/N	MCQ	VQA	Metric	Improv. Focus
POPE	COCO	Post-processed	✓	✗	✗	Acc.	Co-occur.
HaELM	MS-COCO	Post-processed	✓	✗	✗	Acc.	Attention
MME	Self-Sourced	Manual	✓	✗	✗	Acc.	-
AMBER	MS-COCO	Post-processed	✓	✗	✗	Acc.	-
MHaluBench	Self-Sourced	Post-processed	✓	✗	✗	Prec.	-
R-Bench	COCO	Post-processed	✓	✗	✗	Acc.	Co-occur.
MMRel	Visual-Genome	Multi-source	✓	✗	✗	Acc.	-
VALOR-EVAL	GQA	Post-processed	✗	✗	✗	LLM-based	Co-occur.
FAITHSCORE	MS-COCO & LLaVA-1k	Post-processed	✗	✗	✗	Faithscore	Atomic Facts.
FIHA	MS-COCO Foggy & VG	Ruled-based	✓	✗	✓	Acc.	Annotation Free
TIFA	Synthetic	Caption-based	✓	✓	✗	Acc.	T2I Evaluation
Reefknot (Ours)	Visual-Genome	Original	✓	✓	✓	R_{socre}	Confidence

Table 1: Comparsions of our proposed Reefknot benchmark with relevant benchmarks. POPE (Li et al., 2023c), HAELM (Wang et al., 2023b), MME (Fu et al., 2024), AMBER (Wang et al., 2024a), MHaluBench (Chen et al., 2024b), R-bench (Wu et al., 2024), MMRel (Nie et al., 2024), VALOR-EVAL (Qiu et al., 2024), FAITHSCORE (Jing et al., 2024), FIHA (Yan et al., 2024a), TIFA (Hu et al., 2023b)

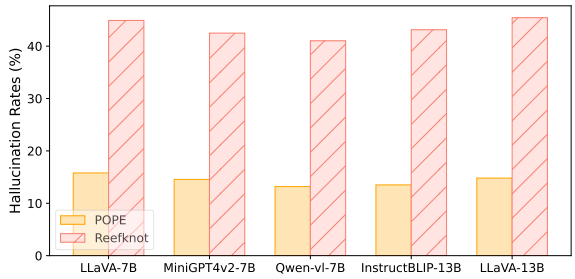


Figure 2: The hallucination rates on POPE, an object hallucination benchmark, and our Reefknot with a focus on relation hallucination (w/ same configuration).

As depicted in Figure 1 (a), object-level hallucination focuses on the model's discrimination of the existence of basic objects (Li et al., 2023c); while as shown in Figure 1 (b), attribute-level hallucination often focuses on whether the model can distinguish some properties of the object itself like color, number, shape and so on (Fu et al., 2024). Their commonality is that they only focus on the single object present in the image. There has been much work exploring the alleviation of these two types of hallucinations. For example, Woodpecker (Yin et al., 2023) used a post-processing method to correct hallucination after the generation process; VCD (Leng et al., 2023) proposed a contrast decoding strategy to mitigate object-level hallucination by adding noise to images.

Despite these efforts, *the community barely considers relation hallucination, which demands more complex reasoning capabilities from MLLMs*. Figure 1 (c) reveals that relation hallucination is related to at least two objects simultaneously in the image, through either perceptive or cognitive perspective. Furthermore, as outlined in Table 1, recent highly relevant benchmarks on hallucinations *lack thorough evaluation and effective mitigation strategies*.

Specifically, these works either adopted either simple Yes/No (Y/N) task to assess the models' accuracy/precision, or utilized LLM to score the performance, but none of them were able to give a comprehensive evaluation from both discriminative (Y/N MCQ) and generative (VQA) perspectives.

Moreover, previous benchmarks seldom proposed mitigation methods, with only a few focusing on co-occurrence or attention mechanisms to address these issues. In contrast, our paper analyzes token-level confidence at each layer to detect and promptly mitigate hallucinations.

To handle the aforementioned research gaps, we present a specific definition for relation hallucination and propose the first comprehensive benchmark **Reefknot** to evaluate the performance on *relation hallucination*. ***Relation hallucination*** *refers to the phenomenon in which MLLMs misinterpret the logical relationships between two or more objects in the corresponding image*. Unlike many benchmarks of MLLM that were constructed by automatic labeling technique, we construct the relation-based dataset based on semantic triplets retrieved from the scene graph dataset. Table 1 demonstrates that our triplets are from real-life scenarios, without any post-processing (e.g., segmentation and bounding box techniques), manual annotation, and synthetic method (e.g., diffusion-based generation). We categorize relation hallucinations into two types: perceptive, involving concrete rela-

tional terms like *"on", "in", "behind"* and cognitive, which encompasses more abstract terms such as *"blowing"* and *"watching"*. Second, we evaluate the mainstream MLLMs on Reefknot via three diverse tasks across two types of relation hallucinations. Figure 2 also illustrates that relation hallucination can be more severe than object hallucination in current MLLMs, highlighting the importance of our evaluation. Furthermore, we propose a simple relation hallucination mitigation method named Detect-then-Calibrate. This originates from the experimental observation that when relation hallucinations occur, the response probability drops significantly, hovering just above 50% in extreme cases compared to the usual nearly 90%. Our method achieves an average improvement of 9.75% across three relation hallucination benchmarks. In summary, Our main contributions are as follows:

1. We have constructed Reefknot, a benchmark comprising two types of relationships and three evaluation tasks to assess relation hallucinations comprehensively.

2. We have conducted a thorough evaluation of relation hallucinations across mainstream MLLMs, uncovering that these models are disproportionately susceptible to perceptual hallucinations in comparison to cognitive ones.

3. We investigated the mechanism of relation hallucination generation from the perspective of response confidence and identified a correlation between relation hallucination and high uncertainty from token levels.

4. We have proposed a novel Detect-then-Calibrate method to detect and mitigate hallucination. By analyzing token confidence scores, we established a threshold to identify hallucinations. Further, we applied a calibration strategy to mitigate hallucination at intermediate confidence levels. Extensive experiments on three relation hallucination datasets demonstrate the effectiveness of our approach.

2 Relation Hallucination Benchmark

In this section, we describe the dataset construction pipeline of Reefknot benchmark in Figure 3. Unlike object and attribute hallucinations that only involve one entity, relations involve two entities making it more difficult to handle. We first identify relation triplets from Visual Genome (VG) dataset (Krishna et al., 2016) (Phase a), and conduct triplet filtering (Phase b). Subsequently, we extract the semantic triplets (Phase c) and categorize their relations (Phase d). Then, a relation-based question set can be constructed into three types (Phase e). Finally, the quality of dataset is ensured by three rounds of expert-based validation (Phase f).

Category	Y/N	MCQ	VQA	Total
#Perception	5,440	4,800	2,150	13,260
#Cognition	4,300	2,150	2,720	8,600
#Total	9,740	6,950	4,870	21,880
Ratio of positive and negative samples				1:1
Number of perceptive relationship types				56
Number of cognitive relationship types				152
Number of images				11084

Table 2: Detailed statistical overview of the Reefknot benchmark dataset, including sample distribution across categories and task types.

Triplet Identification, Filtering and Extraction The dataset comprises of 11,084 images taken from VG dataset, a finely annotated scene graph dataset utilized by the research community (Tang et al., 2020; Liang et al., 2019). As indicated in Figure 3 (a), visual objects and their relations from VG dataset can be easily identified. Besides, we filter the triplets with redundant information, incorrect relationships, or noisy descriptions, as depicted in Figure 3 (b). Subsequently, we can extract semantic triplets by identifying subject-object pairs and the relationships between them, forming *(subject, relation, object)* triplets in Figure 3 (c).

Relation Categorization As depicted in Figure 3 (d), we categorize relationships into two categories based on deep semantic meanings: perceptive and cognitive. Perceptive relationships involve locational prepositions, such as <boy, behind, sofa >and <cup, on, table>; whereas cognitive relationships are expressed through action phrases indicating states, such as <boy, eating, food>and <girl, sleeping in, bed>. ChatGPT is employed to assist in this classification. Table 2 also indicates the sample numbers in different tasks and hallucination categories. The prompt we used for relation categorization is listed in Appendix A.5.

Relation-Relevant Question Construction As shown in Figure 3 (e), we construct three types of relation-relevant question sets to evaluate the state-of-the-art MLLMs' abilities in relation-level perception and reasoning.

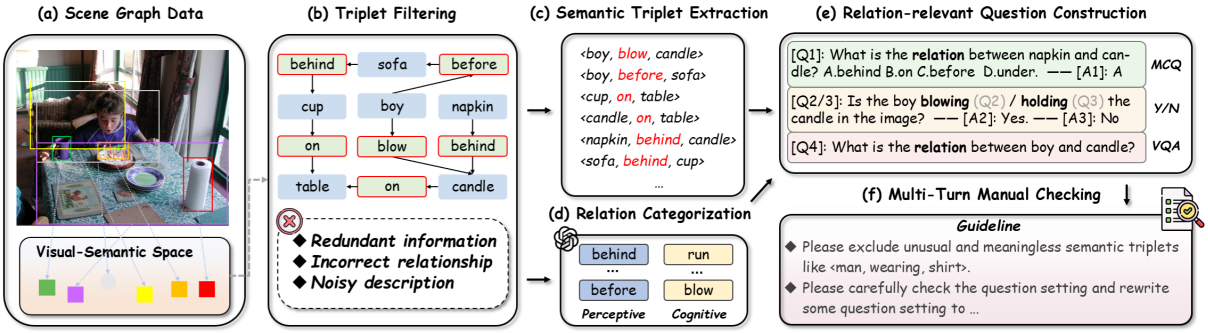


Figure 3: The data construction pipeline of our proposed Reefknot benchmark.

MLLMs	Size	Perception↓			Cognition↓			Total↑
		Y/N	MCQ	VQA	Y/N	MCQ	VQA	R_{score}
Phi-3-vision-128k-instruct (Abdin et al., 2024)	4.2B	39.88	57.07	50.98	33.97	21.35	49.45	60.30
Yi-VL (AI et al., 2024)	6B	33.56	47.53	71.02	33.16	16.33	74.96	55.81
LLaVA (Liu et al., 2023a)	7B	37.67	68.05	52.93	33.99	51.04	54.56	51.41
MiniGPT4-v2 (Chen et al., 2023)	7B	46.7	78.00	61.30	43.73	68.50	65.88	39.88
MiniCPM (Hu et al., 2024)	7B	**31.93**	48.65	**47.63**	27.65	16.71	**45.96**	**65.73**
Qwen-VL (Bai et al., 2023)	7B	42.21	56.7	72.47	33.53	21.88	73.01	52.55
Deepseek-VL (DeepSeek-AI et al., 2024)	7B	37.58	56.33	67.07	32.22	23.60	59.34	56.39
GLM4V (GLM et al., 2024)	9B	34.09	50.47	58.09	**27.08**	16.87	56.47	62.03
LLaVA (Liu et al., 2023a)	13B	40.7	59.35	48.93	34.19	29.19	54.45	57.47
CogVLM (Wang et al., 2024b)	19B	37.23	47.95	70.14	29.89	18.54	66.18	57.1
Yi-VL (AI et al., 2024)	34B	32.79	**44.19**	57.67	33.75	**14.85**	52.72	62.61
GPT-4o (OpenAI et al., 2024)	-	32.56	40.93	42.70	26.27	11.53	48.78	68.32

Table 3: **Evaluation of hallucination rates** on the different MLLMs. Additionally, we use **bold** to highlight the best performance of open-sourced MLLMs, and underline to emphasize the distinction between perception and cognition of LLaVA-13B. Note all experiments are done with the temperature to 0 to keep the reproducibility.

- For `Yes/No` (Y/N) questions, we employ an adversarial approach by introducing a negative sample within the same triplet, alongside the positive sample, to test the model's ability to correctly answer "No".
- `Multiple Choice Questions` (MCQ) are designed with one correct answer and three random options to evaluate the model's resistance to relation hallucinations within a controlled and limited vocabulary.
- `Visual Question Answering` (VQA) is an open-ended task that allows us to comprehensively assess a model's instruction-following capabilities and relation perception within an open-domain environment.

Multi-Turn Manual Checking Finally, we perform multi-turn manual verification to ensure the quality of the question sets (see Figure 3 (f)). Each question undergoes at least three rounds of review by four domain experts. We revise any inappropriate expressions and exclude meaningless questions, such as *"Is the window on the wall?"*, which lack informative value and can be answered without visual input. After rigorous screening process, our dataset comprises 21,880 questions across 11,084 images as shwon in Table 2.

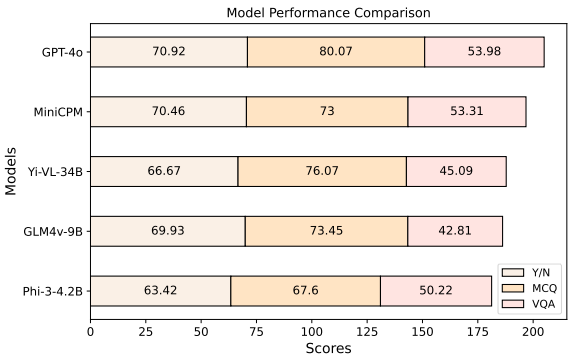


Figure 4: MLLMs with **top five best performance evaluated on Reefknot benchmark**. We report the sum of the respective metric across three tasks for reference.

3 Hallucination Evaluation

3.1 Task and Evaluation Metrics

As illustrated in Table 3, we conducted evaluations on mainstream MLLMs to evaluate relation-level hallucinations. All experiments of open-sourced MLLMs were conducted using 8 NVIDIA A100 GPUs. Each experiment ran three times, and we reported the average of these results. In our evaluation, we reported our results from two distinct categories: perception and cognition.

For discriminative Y/N and MCQ tasks, we reported the hallucination rate $Halr$ as a metric. Generative questions have always posed a challenge in evaluation of MLLMs. In our assessment of the generative VQA task, we employed the DeBERTa model using a bidirectional entailment approach for label match (Kuhn et al., 2023), and we denoted $Halr$ as metric for simplicity as well. For more details, we have listed in Appendix A.7. In general, we use the following metric R_{score} to comprehensively evaluate the overall performance across the three tasks, which is expressed in Formula 1.

$$R_{\text{score}} = \text{Avg}\left[\sum_{i=1}^{3}(1 - Halr_i)\right]. \quad (1)$$

3.2 Main results

Overall Performance Significant performance differences among the various models are evident. Table 3 shows significant differences in the performance of various models across different question types. For instance, Qwen-vl-chat excels in Y/N and MCQ settings but encounters serious hallucination issues in VQA tasks. A detailed review of the model's responses reveals that while Qwen can follow instructions accurately, many responses contain expressions that are completely unrelated to the labels. We presume that such a phenomenon is owing to over-fitting training during the instruction-tuning process. Among open-sourced models, MiniCPM (Hu et al., 2024) stands out, likely due to its adoption of the fine-grained alignment technique such as RLHF-V (Yu et al., 2023) to alleviate hallucinations.

Error Analysis Figure 4 presents a comparative analysis of the top five best models' performances on the Reefknot benchmark. The comparison reveals a clear hierarchy in performance, with GPT-4o outperforming the rest across all tasks. MiniCPM, Yi-VL-34B, and GLM4v-9B show a competitive edge, closely following GPT-4o, while Phi-3 lags behind relatively. In terms of task performance, Figure 4 indicates that while some models excel in one task, they may not perform as well in others. For example, Yi-VL-34B achieves the second-highest score in MCQ but underperforms in VQA. This underscores the significance of task-specific model tuning and the necessity of a balanced approach to enhance overall performance across all relation hallucination tasks.

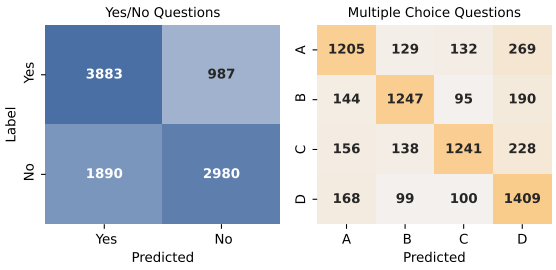


Figure 5: **Confusion matrices** of MiniCPM-7B's performance on Reefknot benchmark (Left: Y/N setting; Right: MCQ setting).

Perceptive vs. Cognitive Hallucination We find the occurrence of cognitive hallucinations is generally lower than that of perceptive hallucinations, which may diverge from intuition. Across all models and settings, the incidence of perceptive hallucinations is consistently 10% higher than that of cognitive hallucinations. In the most extreme case, such as LLaVA-13B model in the MCQ setting, the rate of perceptive hallucinations is 30.16% higher than that of cognitive hallucinations. This phenomenon may be caused by the fact that models often utilize large-scale image-caption datasets during the pre-training and fine-tuning processes. These datasets typically contain detailed visual descriptions, enabling models to perform better in cognitive relationships such as *running, eating,* etc. Conversely, these models may struggle when dealing with some perceptive relationships based on common sense because they were ignored in the annotation process of the original dataset.

To analyze the error cases quantitatively, we visualized the results of MiniCPM-7B, the best open-sourced model, across two discriminative settings, as illustrated in Figure 5. For Y/N questions, the model tends to favor positive responses (i.e., Yes). Specifically, among all misclassifications, the instances where a `No` label is incorrectly classified as `Yes` are twice as times as instances where a `Yes`

label is incorrectly classified as `No`. Besides, the model tends to answer `D` in MCQ settings. We suspect the tendency is likely due to the imbalance in the distribution of the training data.

4 Analysis of Relation Hallucination

To quantitatively compare the decision probability distribution when hallucinations occur, we calculate the average probability of an equal number of relation hallucination and non-hallucination examples, as shown in Table 4.

Dataset	Reefknot	MMRel	Rbench
LLaVA	0.67	0.76	0.80
MiniGPT4-v2	0.76	0.75	0.62

Table 4: The **average probability of all hallucination cases**. We evaluate LLaVA and MiniGPT4-v2 on the Reefknot dataset, along with two other representative relation hallucination benchmarks.

The table shows that when hallucinations occur, the confidence level in the answers is quite low. Specifically, experiments conducted on three relational-level datasets indicate that the overall probability of the answers is only about 70%. In contrast, under normal circumstances, large language models can achieve probability values of up to 95% when providing factual and truthful answers. Therefore, a straightforward approach to detecting relation hallucinations is to utilize the entropy $E(X)$ of the probability distribution.

$$E(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i). \tag{2}$$

Note that because MLLMs have an extensive vocabulary, it becomes challenging to discern meaningful patterns in entropy variation when predicting the next word across the entire distribution. Consequently, our analysis is restricted to observing the variation patterns of vocabulary within the range of potential answers x_i. We present the ratio of hallucination cases to non-hallucination cases among both object and relation hallucination benchmarks in Figure 6. When $E(X) > 0.6$, *relation hallucinations* occur to a significant degree, indicating the effectiveness of our method to detect *relation hallucination* via entropy. To investigate the mechanism behind hallucination generation, we conducted an in-depth analysis of the fluctuations in probability values across model layers. MLLMs consist of a vision encoder, a projection layer, and

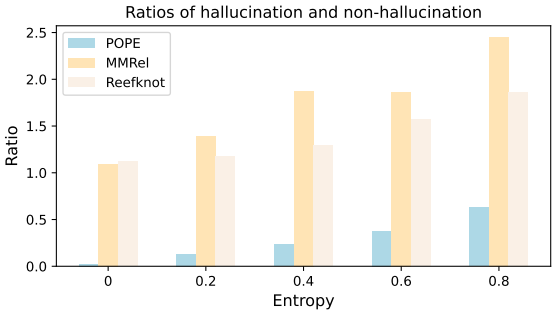


Figure 6: The respective **ratios between hallucination and non-hallucination** with different entropy values.

a strong LLM decoder, which is stacked by a set of $\mathcal{N}$ transformer layers, and an MLP layer $\phi(\cdot)$ for predicting the next-word probability of distribution auto-regressively. Given an image represented by vision encoder as $\mathcal{V}_t = \{v_1, v_2, \ldots, v_t\}$ and a text prompt of tokens $\mathcal{P}_t = \{p_1, p_2, \ldots, p_t\}$, they are processed as a sequence $\mathcal{H}_0 = \psi(\mathcal{V}_t, \mathcal{P}_t)$ through projection and concatenation function $\psi(\cdot)$. Thus $\mathcal{H}_0 = \{h_1^{(0)}, \ldots, h_{t-1}^{(0)}\}$, in which $h_i^{(k)}$ means the hidden states of i_{th} token in k_{th} language decoder layer. Then $\mathcal{H}_0$ would be processed by each of the transformer layers in the language decoder successively. We denote the output of the j-th layer as $\mathcal{H}_j$. In the normal forward process, $\mathcal{H}_j$ will be calculated by $\mathcal{N}$ times, then it will be passed through language model head layer $\phi(\cdot)$ to predict the probability of the next token r_t over the vocabulary set $\mathcal{X}$. In our set, we manually pass every $\mathcal{H}_n$ to explore the mechanism of hallucination with probability distributions and the next token to generate. For every layer, we can obtain the probability distribution $\mathbb{P}$ and the next word prediction $r_t^{(j)}$.

$$\mathbb{P}(\mathcal{H}_j \mid \mathcal{H}_{j-1}) = \mathrm{softmax}\big(\phi(\mathcal{H}_{j-1})\big), \quad j \in \mathcal{N}. \tag{3}$$

$$r_t^{(j)} = \arg\max \mathbb{P}(\mathcal{H}_j \mid \mathcal{H}_{j-1}^{(j)}), \quad r_t^{(j)} \in \mathcal{X}. \tag{4}$$

Using Equation 4, we visualized these changes during the forward propagation process, as illustrated in Figure 7. To avoid variability from individual instances, we reported the average values of all data involving hallucinations occurring in Reefknot. To ensure a fair comparison, we utilized the 32-layer MiniGPT4-v2-7B model (Chen et al., 2023) and the 40-layer LLaVA-13B model (Liu et al., 2023a).

It can be observed that in the shallow layers (0th-20th), the probability of possible answers does not increase. We hypothesize that this is because, in shallow layers, the model is aggregating information to generate an answer. Hallucination occurs in

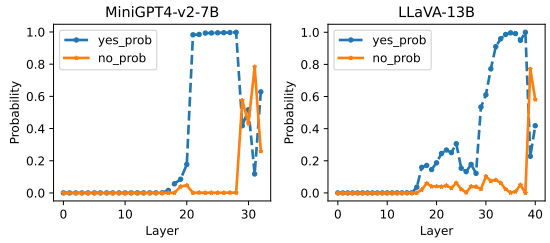


Figure 7: The **average probability across layers during hallucination occurrences**. Significant probability changes are observed in deeper layers. Results are reported for MiniGPTv2-7B (32 layers) and LLaVA-13B (40 layers) to demonstrate universality.

the deep layers. Since deep layers contain a vast amount of knowledge, we speculate that they may cause relation hallucinations.

5 Detect-Then-Calibrate Mitigation Method

The evaluation results indicate that MLLMs commonly suffer from severe *relation hallucinations*. As analyzed in Section 4, we found that these *relation hallucinations* primarily stem from the model's lack of confidence. The lack of confidence in the model's responses results in a relatively high entropy value. Therefore, we can detect the occurrence of *relation hallucinations* by monitoring changes in entropy. Initially, we set a specific entropy threshold γ to detect potential hallucinations in the model's output. If the entropy of the model's response exceeds γ, which suggests a significant lack of confidence, we infer a high probability that the model has generated hallucinations. In cases where the model is identified to hallucinate, we will utilize the hidden states from intermediate layers to calibrate the final outputs layers, which is inspired by Chuang et al. (2023). Note, unlike traditional contrastive decoding strategies (Li et al., 2023a; Leng et al., 2023; Huang et al., 2023), we do not calibrate all cases. Instead, we focus on mitigating potential hallucination cases to avoid altering non-hallucinatory cases into hallucination ones. Formula 5 show our calibration process.

$$r = \begin{cases} \arg\max log \frac{(1+\alpha)\cdot softmax(\phi(h_t^n))}{\alpha \cdot softmax(\phi(h_t^{n-\lambda}))}, & \text{if } E_t > \gamma. \\ \arg\max\big(softmax(\phi(\ h_t^n))\big), & \text{otherwise.} \end{cases} \tag{5}$$

Note that λ is the hyperparameter to control the degree of the intermediate layer; $\mathbf{r}$ is the token generated after calibration; α represents the degree of calibration operation. The algorithmic flow can be seen from the pseudo-code below:

Algorithm 1: Detect-Then-Calibrate Algorithm

Require: Image v_t; MLLM $\mathcal{M}(\cdot)$; Prompt p_t; Uncertainty Entropy threshold γ

$\mathcal{H}, r_0, \mathbb{P}, = \mathcal{M}(x_t, p_t)$
The entropy of generate token $E(r_0) = -\sum_{i=1}^{n} \rho_i \log \rho_i \ \rho \in \mathbb{P}$
if $E(r_0) \geq \gamma$ **then**
  Hallucination occurs! Calibrate $\mathcal{H}$ by first λ layer
  $h_\delta = \log \frac{(1+\alpha)\cdot softmax(\phi(h_t^n))}{\alpha \cdot softmax(\phi(h_t^{n-\lambda}))}$
  Calibrated response $\bar{r} = \arg\max h_\delta$
else
  Normal response $r = \arg\max\big(softmax(\phi(\ h_t^n))\big)$
end if

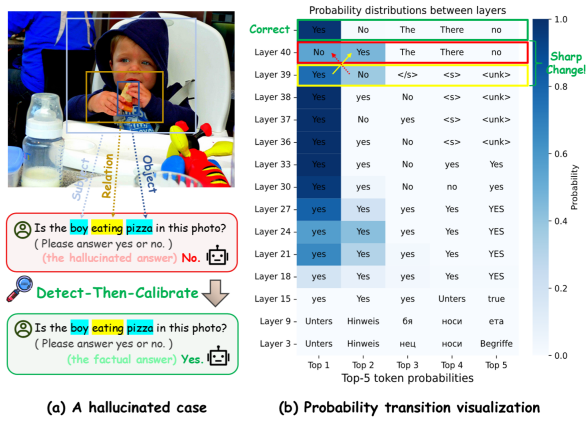
As illustrated in Table 5, we conducted experiments using LLaVA-13B. To demonstrate the robustness of our results, we employed two additional relation hallucination datasets, MMRel (Nie et al., 2024) and R-bench (Wu et al., 2024). During experiments, we set $\lambda = 2$,$\alpha = 0.1$, $\gamma = 0.9$. For a fair comparison, in addition to reporting the baseline model, we also report some training-free methods such as VCD (Leng et al., 2023), DoLa (Chuang et al., 2023), and OPERA (Huang et al., 2023). Our approach achieved improvements across all three relation hallucination datasets. Specifically, on the MMRel dataset, our model achieved a 19.7% improvement compared to the baseline setting.

Methods	**Reefknot**	**MMRel**	**R-bench**
Baseline	37.06	40.43	29.52
+ VCD (Leng et al., 2023)	38.32	41.96	22.05
+ DoLa (Chuang et al., 2023)	36.96	39.68	23.52
+ OPERA (Huang et al., 2023)	35.73	39.22	26.73
+ Detect-then-Calibrate (Ours)	**34.50**	**21.73**	**22.02**

Table 5: The **hallucination rates of LLaVA-13B and its variants**(lower is better) with hallucination mitigation methods across Reefknot and two other relation hallucination benchmarks (*i.e.*, MMRel & R-bench).

6 Case study

To intuitively demonstrate the generation process of hallucination, we visualized a real case from Reefknot in Figure 8 (Nostalgebraist, 2020). In the image, when we ask "`Is the boy eating pizza in the photo?`", MLLMs are unable to provide precise answers. To investigate this, we analyze the probabilities associated with the top five predicted tokens at each layer of the language model. As illustrated in Figure 8, the model typically converges



(a) A hallucinated case (b) Probability transition visualization

Figure 8: Part (a) illustrates a genuine case of relation hallucination derived from the Reefknot benchmark. The visualization of probability transitions across the layers of the language model is presented in Part (b). Subsequently, our refined results are distinctly highlighted in green, juxtaposed against significant probability variations, which are clearly demarcated within a red-bordered box.

on the correct answer as it progresses from shallow to deep layers. Notably, hallucinations didn't occur until the final layer. Additionally, we can observe that as the number of layers increases, the model increasingly focuses on narrowing down the range of answer choices, with the probability for both 'yes' and 'no' rising in deeper layers.

However, as depicted in the changes of colors, when it reaches the final decoder block (layer 38-40th), the model's choices suddenly become uncertain and the entropy rises, accompanied by the emergence of hallucinations. So it is instinctive to utilize the entropy to detect hallucination and use logits of the intermediate layers, which are less likely to hallucinate, to calibrate the final logits to correct the answer. Following our calibration utilizing Equation 5, we not only dispel potential relation-level hallucination but also significantly bolster the confidence of model's responses.

7 Related Work

Hallucination Benchmarks in MLLMs Large Language Models (LLMs) are powerful tools, and exploring their interpretability (Han et al., 2024; Huo et al., 2024b), operational mechanisms, and methods to ensure trustworthy responses(Hu et al., 2023a) has become a crucial area of research. In the realm of MLLMs, hallucinations are typically categorized into three distinct types: object, attribute, and relation, as outlined in prior studies (Yin et al., 2024; Liu et al., 2024; Guan et al., 2024). At the object and attribute levels, a considerable number of representative benchmarks such as POPE (Li et al., 2023c) and HaELM (Wang et al., 2023b) have been introduced by researchers. Many benchmarks are accompanied by a diverse array of evaluation criteria, with generative criteria including CHAIR (Li et al., 2023b), THRONE (Kaul et al., 2024) and various discriminative criteria being particularly notable. The existing relation hallucinations benchmarks (Wu et al., 2024; Nie et al., 2024) focus only on discriminative criteria. The gap of current benchmarks is to identify relationship pairs that accurately reflect the dynamic interactions.

Confidence Calibration Confidence estimation (Zou et al., 2023; Chen et al., 2024a) and calibration are essential for enhancing the reliability of LLMs such as GPT-3 (Brown et al., 2020). To assess the confidence associated with outputs from LLMs, Kuhn et al. (2023) have developed a method called semantic entropy that utilizes linguistic invariances reflecting shared meanings. However, this method relies on accessing token-level probabilities, which are often unavailable through current black-box APIs. Kadavath et al. (2022) have designed prompts that encourage the models to self-assess their responses and to explicitly calculate the probability that an answer is true; while Lin et al. (2022) have prompted LLMs to provide both an answer and an accompanying confidence level.

8 Conclusion

In conclusion, we propose a comprehensive benchmark called Reefknot to evaluate and mitigate relation hallucinations in MLLMs. We construct the dataset with over 20k data through a scene graph based construction pipeline, covering two discriminative tasks (Y/N and MCQ) and one generative task (VQA). Our in-depth evaluation highlights a substantial performance gap on relation hallucination in existing MLLMs, emphasizing the need for more sophisticated reasoning capabilities. Subsequently, we discover that relation hallucinations tend to occur when MLLMs respond with low confidence. Therefore, we propose a Detect-then-Calibrate method to mitigate the relation hallucination via entropy threshold, with an average reduction of 9.75% in the hallucination rate across Reefknot and two other representative relation hallucination datasets. In general, we anticipate that the Reefknot benchmark we propose will serve as

a critical foundation for future developments in the field of trustworthy multimodal intelligence, enabling more reliable and robust systems.

9 Limitation

Despite promising, our proposed approach focuses on mitigating basic discriminative hallucinations, but relation hallucinations in open domains are still challenging to quantitatively assess and mitigate. In future research, we will delve deeper into the underlying causes of hallucinations in open domains and investigate both the mechanisms and mitigation strategies. We anticipate that Reefknot will further improve the reliability and practical utility of MLLMs.

Acknowledgement

This work was supported by Open Project Program of Guangxi Key Laboratory of Digital Infrastructure (Grant Number: GXDIOP2024015); Guangdong Provincial Department of Education Project (Grant No.2024KQNCX028); Scientific Research Projects for the Higher-educational Institutions (Grant No.2024312096), Education Bureau of Guangzhou Municipality; Guangzhou-HKUST(GZ) Joint Funding Program (Grant No.2025A03J3957), Education Bureau of Guangzhou Municipality.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

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*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024a. M^3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *Preprint*, arXiv:2405.16473.

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. *Preprint*, arXiv:2402.03190.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Yang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jin Chen, Jingyang Yuan, Junjie Qiu, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruizhe Pan, Runxin Xu, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Size Zheng, T. Wang, Tian Pei, Tian Yuan, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Liu, Xin Xie, Xingkai Yu, Xinnan Song, Xinyi Zhou, Xinyu Yang, Xuan Lu, Xuecheng Su, Y. Wu, Y. K. Li, Y. X. Wei, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Zheng, Yichao Zhang, Yiliang Xiong, Yilong Zhao, Ying He, Ying Tang, Yishi Piao, Yixin Dong, Yixuan Tan, Yiyuan Liu, Yongji Wang, Yongqiang Guo, Yuchen Zhu, Yuduan Wang, Yuheng Zou, Yukun Zha, Yunxian Ma, Yuting Yan, Yuxiang You, Yuxuan Liu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhewen Hao, Zhihong Shao, Zhiniu Wen, Zhipeng Xu, Zhongyu Zhang, Zhuoshu Li, Zihan Wang, Zihui Gu, Zilin Li, and Ziwei Xie. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *Preprint*, arXiv:2405.04434.

S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Chi Han, Jialiang Xu, Manling Li, Yi Fung, Chenkai Sun, Nan Jiang, Tarek Abdelzaher, and Heng Ji. 2024. Word embeddings are steers for language models. *Preprint*, arXiv:2305.12798.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *Preprint*, arXiv:2404.06395.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. 2023a. Do large language models know about facts? *Preprint*, arXiv:2310.05177.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023b. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *Preprint*, arXiv:2303.11897.

Qidong Huang, Xiaoyi Dong, Pan zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2023. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *arXiv preprint arXiv:2311.17911*.

Jiahao Huo, Yibo Yan, Boren Hu, Yutao Yue, and Xuming Hu. 2024a. Mmneuron: Discovering neuron-level domain-specific interpretation in multimodal large language model. *arXiv preprint arXiv:2406.11193*.

Jiahao Huo, Yibo Yan, Boren Hu, Yutao Yue, and Xuming Hu. 2024b. Mmneuron: Discovering neuron-level domain-specific interpretation in multimodal large language model. *Preprint*, arXiv:2406.11193.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. FaithScore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5042–5063, Miami, Florida, USA. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, C. J. Taylor, and Stefano Soatto. 2024. Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. *Preprint*, arXiv:2405.05256.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Preprint*, arXiv:1602.07332.

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

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *Preprint*, arXiv:2311.16922.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023a. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Yuanzhi Liang, Yalong Bai, Wei Zhang, Xueming Qian, Li Zhu, and Tao Mei. 2019. Vrr-vg: Refocusing visually-relevant relationships. *Preprint*, arXiv:1902.00313.

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

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. *Preprint*, arXiv:2402.00253.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

Jiahao Nie, Gongjie Zhang, Wenbin An, Yap-Peng Tan, Alex C Kot, and Shijian Lu. 2024. Mmrel: A relation understanding dataset and benchmark in the mllm era. *arXiv preprint arXiv:2406.09121*.

Nostalgebraist. 2020. interpreting gpt: the logit lens.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin

Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Haoyi Qiu, Wenbo Hu, Zi-Yi Dou, and Nanyun Peng. 2024. Valor-eval: Holistic coverage and faithfulness evaluation of large vision-language models. *arXiv preprint arXiv:2404.13874*.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023a. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *Preprint*, arXiv:2310.07521.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. 2024a. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *Preprint*, arXiv:2311.07397.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, Jitao Sang, and Haoyu Tang. 2023b. Evaluation and analysis of hallucination in large vision-language models. *Preprint*, arXiv:2308.15126.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2024b. Cogvlm: Visual expert for pretrained language models. *Preprint*, arXiv:2311.03079.

Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. 2024. Evaluating and analyzing relationship hallucinations in large vision-language models. *Preprint*, arXiv:2406.16449.

Bowen Yan, Zhengsong Zhang, Liqiang Jing, Eftekhar Hossain, and Xinya Du. 2024a. Fiha: Autonomous hallucination evaluation in vision-language models with davidson scene graphs. *Preprint*, arXiv:2409.13612.

Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2024b. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM on Web Conference 2024*, pages 4006–4017.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *Preprint*, arXiv:2306.13549.

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. *Preprint*, arXiv:2310.16045.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2023. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

Junyi Zhu, Shuochen Liu, Yu Yu, Bo Tang, Yibo Yan, Zhiyu Li, Feiyu Xiong, Tong Xu, and Matthew B Blaschko. 2024. Fastmem: Fast memorization of prompt improves context awareness of large language models. *arXiv preprint arXiv:2406.16069*.

Xin Zou, Chang Tang, Xiao Zheng, Zhenglai Li, Xiao He, Shan An, and Xinwang Liu. 2023. Dpnet: Dynamic poly-attention network for trustworthy multimodal classification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3550–3559.

Xin Zou, Yizhou Wang, Yibo Yan, Sirui Huang, Kening Zheng, Junkai Chen, Chang Tang, and Xuming Hu. 2024. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*.

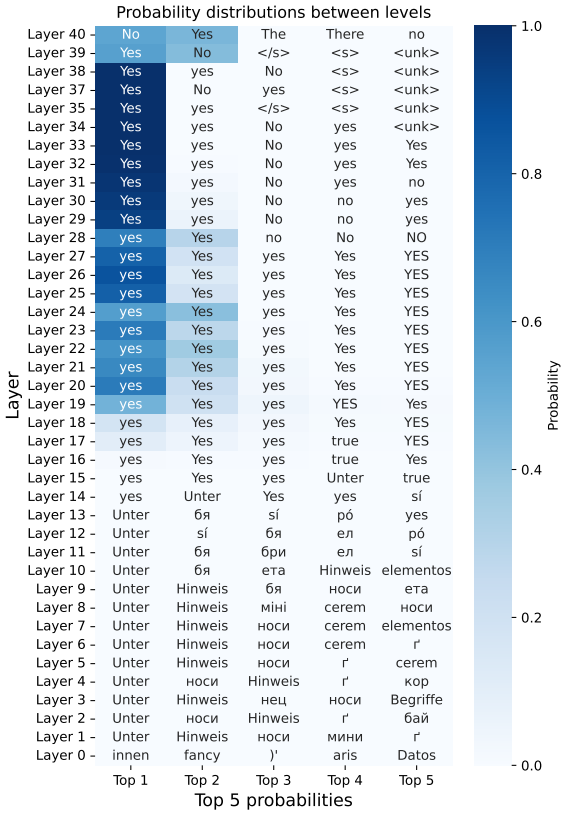


Figure 9: The **hallucination case** visualization of Top5-token probability transitions across the layers of MLLM.

A Appendix

A.1 Detailed Case Study

Due to space limitations, we have provided the complete set of changes in the Top-5 tokens across 40 layers within the Section **Case Study** in Figure 9.

A.2 Manual Checking Rules

In session of Triplet Filtering, the following types of relationships need to be avoided:

1. **Redundant and meaningless information**: Redundant and meaningless information refers to the relation are overly obvious or self-evident, often involving relationships or facts that are universally understood and require no further explanation. These types of information typically describe natural relation that does not need to be specially emphasized. For example, stating that 'cows have eyes' or 'people wear shirts' emphasises basic, well-known facts that don't effectively reflect the model's ability to recognise relationships, which is meaningless for our benchmarks.

2. **Data Bias** In the process of filtering relationships, the number of different relationships should be approximately the same. If the number of cases corresponding to each relationship in the dataset is not balanced, i.e., there are far more cases for some relationships than for others, this will result in a correspondingly higher overall score when the model performs well on a particular relationship. Such a scenario would mask the model's deficiencies in a small number of relationships, leading to distorted assessment results. Therefore, in order to ensure the fairness of the assessment and the reliability of the results, it is important to ensure that the number of cases for each relationship in the dataset is relatively balanced. Otherwise, the strengths and weaknesses of the model will be difficult to accurately reflect through the overall score.

In session of Multi-Turn Manual Checking, the annotator should follow these rules:

1. **Clarity and Precision** Ensure that the questions are directly related to verifiable content within the image, steering clear of ambiguity and multiple interpretations. Each question should unambiguously point towards a specific, correct answer.

   The following situations need to be avoided:

   (a) **Multiple Subjects**: Avoid situations where the presence of multiple subjects in the image leads to unclear references in the semantic triad, making it difficult to discern the subject of the question.

   (b) **Similar Relational Words**: In multiple-choice questions, ensure that the options provided do not include relational words with overlapping or similar meanings.

   (c) **Grammatical Accuracy**: Ensure all samples are free from grammatical errors to maintain the professionalism and clarity of the questions.

2. **Relevance Check** Verify that each question is directly related to and necessitates reference to the content of the image. Questions should focus on elements that are prominent and significant within the visual data.

   (a) **Minor Details**: Refrain from formulating questions about elements that are minor or indistinct, as they may not significantly influence the understanding of the image.

   (b) **Independence from Image**: Avoid questions that can be answered without referring to the image information, ensuring that the visual content is integral to solving the query.

Each case is assigned to three annotators for inspection, after which a voting mechanism is used to decide whether or not to keep it in our dataset.

A.3 Complete guideline for Manual Check

Next I will provide you with three types of questions and corresponding images. The types of questions are Multiple Choice Question (MCQ), Yes/No(Y/N), and Visual Question Answer (VQA). MCQ asks about the direct relationship between two objects that already exist in the picture, and the question presupposes four options, with only one option being the final correct answer; Y/N questions presuppose a relationship between the two objects in the image, with the objects already existing in the image; and the VQA is the same as the multiple-choice question setup, but with no options provided. Your task is to make judgments about the reasonableness of the question setup and the relevance of the question to the picture.

The degree of reasonableness of the question refers to the fact that the objects involved in the question set must be present in the picture and that the collocation between the two objects has a certain degree of immobility; typical fixed-type collocations are triplets like, <ear on face>, <hair on head>, and when you come across these kinds of relational words in the question set, you should filter them out straight away.

The relevance of the picture means that the objects appearing in the question should occupy the main part of the picture, if the picture appears in the minutiae of the question (e.g., when it comes to some very subtle information about the background), you should exclude the picture and the question.

You can modify, discard and keep the pairs you see, with the following precautions:

❶ *Please exclude unusual and meaningless semantic triplets like <man, wearing, shirt>.*

❷ *Please carefully check the question setting and rewrite some question setting to keep diversity.*

❸ *Please make a final judgment on each pair, you can choose to keep, change and keep, or delete.*

A.4 More Error Analysis

We listed more error analyses for Y/N and MCQ for MLLMs In Figure 10, it shows Yi-VL-34b-chat. Figure 11 shows GLM4v-9b-chat. Figure 12 shows Qwen-vl-chat. Figure 13 shows Phi-vision-128k. It can be observed that, with the exception of Phi-vision-128k, all models demonstrate identical distribution and preference trends.

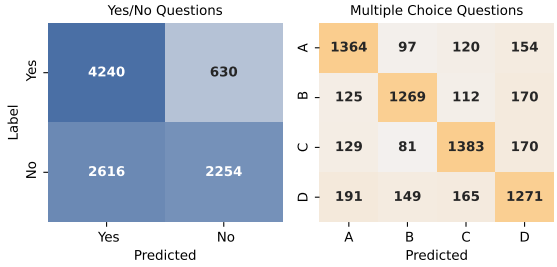


Figure 10: Confusion matrixes of Yi-vl-34b-chat on Reefknot benchmark (Left: Y/N setting; Right: MCQ setting).

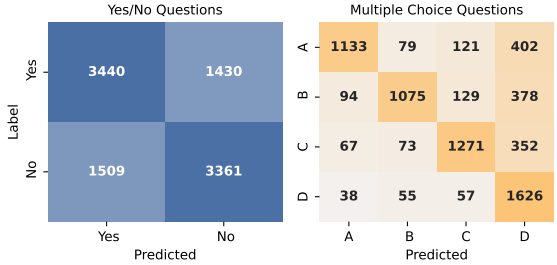


Figure 11: Confusion matrixes of GLM4v-9b-chat on Reefknot benchmark (Left: Y/N setting; Right: MCQ setting).

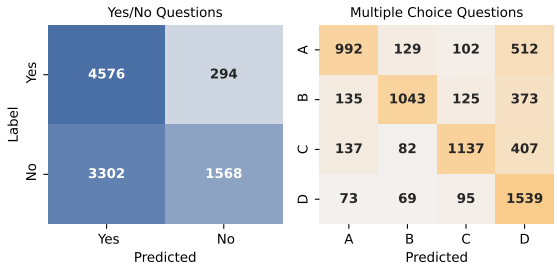


Figure 12: Confusion matrixes of Qwen-vl-chat on Reefknot benchmark (Left: Y/N setting; Right: MCQ setting).

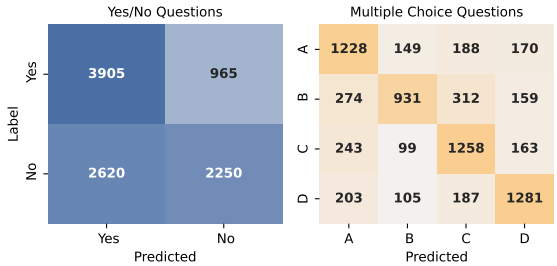


Figure 13: Confusion matrixes of Phi3-vision-128k on Reefknot benchmark (Left: Y/N setting; Right: MCQ setting).

A.5 Prompt to Categorize

In Figure 14, we show our prompt used to categorize the relation word used in our benchmark during phase(d) in the pipeline of Reefknot.

Prompt:

You are a relation term classification assistant. Please help me determine whether the following relational terms/phrases belong to perceptive relationships or cognitive relationships.
Perceptive relationships are defined as those involving locational or state-based prepositions, such as “on" and “behind".
Cognitive relationships are defined as some words that indicate an action or behavior, such as holding, watching, etc. Here I will give you some demonstrations for reference: {In-context examples}
Please answer with a single word in **<Perception, Cognition >**.
Input: {Input}

Figure 14: Prompt template for relation categorization.

A.6 Potential Concerns

❶ Why detect-than-calibrate method is specific to relation hallucination?

Our approach is specific to relation hallucinations, which uniquely correlate with high entropy, as shown in Table 4 and Figure 7. In contrast, other types of hallucinations do not display this trait. The Table 6 below shows results from the object-level hallucination benchmark POPE (Li et al., 2023c), where the average probability is over 90%, resulting in an average entropy of about 0.4, much lower than the $\lambda = 0.9$ set in the paper.

Model/Dataset	POPE	Entropy
LLaVA	0.932	**0.358**
MiniGPT4-v2	0.904	**0.456**

Table 6: The average probability and entropy of object-level hallucination instances.

❷ What is the difference between detect-than-calibrate and contrastive decoding based methods like (*i.e.*, vcd , dola)?

Compared to the **intuitive** contrastive decoding methods, our approach relies on analyzing relational hallucinations and the variation in hidden layers confidence when relation hallucinations occur. In contrast, the intuitive contrastive decoding method does not conduct a detailed analysis of the issues in the individual cases but instead applies a contrastive-decoding approach to all examples, regardless of whether they are hallucination cases. This may mitigate hallucination effects, while **leading to non-hallucination examples becoming hallucination ones**, which does not fully address the hallucination. In contrast, our method is based on a cause analysis of relation hallucinations from the perspective of model confidence. We first identified a correlation between relation-level hallucinations and high entropy values. Then, we filter hallucination cases based on entropy values before calibration operation, thereby avoiding the potential negative impacts of contrastive decoding methods on normal instances.

In a word, our approach differs from contrastive decoding methods in that we selectively correct potential hallucination examples based on the characteristics of relation hallucinations. The application of the sole contrastive methods tend to convert many non-hallucination cases into hallucination cases, resulting in no significant effects, as demonstrated in Table 5.

❸ Why no more experiments on other benchmarks *(i.e., POPE , CHAIR)*?

POPE (Li et al., 2023c) and CHAIR (Li et al., 2023b) are indeed well-known benchmarks for hallucination evaluation. However, POPE is specifically focused on **object-level hallucinations**, while CHAIR is a **generative metric** for evaluating **object-level hallucinations** (caption-based).

Our Detect-Than-Calibrate focuses on exploring the mechanisms and mitigation of **relation-level** hallucinations. In Figure 6, we demonstrate the hallucination ratio of POPE under different entropy distributions. As shown in Figure 6, POPE does not show an increase in hallucinations with higher entropy, making it unsuitable for evaluating our entropy-based Detect-than-Calibrate method. CHAIR, on the other hand, is a generative metric for object-level hallucinations, whereas our method targets **discriminative** relation-level hallucination mitigation. Therefore, it is not feasible to evaluate our approach using POPE and CHAIR.

To validate the effectiveness of our method, besides experiments on our Reefknot, we also conducted experiments on two benchmarks specifically targeting **relation-level hallucination**, R-bench (Wu et al., 2024) and MMrel (Nie et al., 2024), in addition to our own Reefknot. The results in Table 5 demonstrate the effectiveness of our approach. Additionally, in the paper, we explore the distinction between object-level and relation-level hallucinations in Figures 2, 6, and Table 6.

❹ The Proof Reliability of DebERTa

following the setting from Kuhn et al. (2023); Farquhar et al. (2024), we reformulated VQA task as an NLI task by appending the instruction "*Please answer in the following format: Subject is <relation> Object*" to the prompt. This raises two potential concerns: First, there is a concern about DeBERTa's ability to distinguish between MLLM's response and the ground truth; Second, there is a concern about how the introduction of answer templates might influence the results.

To address the concern regarding DeBERTa's capabilities, we selected a number of examples from the study for verification and had human annotators annotate these examples for comparison, as shown in Table 8. The results showed that over 95% of the cases yielded identical annotations between DeBERTa and the human annotators, demonstrating that DeBERTa is well-suited for this task.

To address the concern about the introduction of answer templates, we compared responses with and without templates. Responses generated with the template were structured as phrases, while those without the template were more open-ended and variable. We employed GPT-4o-mini to assess whether the open-ended responses were semantically equivalent to the ground truth. Table 7 shows that without a constrained template, MLLMs tend to show a higher hallucination rate, which proves the necessity of the template in prompts.

❺ Why do you set $E(X) > 0.6$ as a threshold to calibrate? The ratio seems unchanged between $E(X)$=0.4 and $E(X)$=0.6 in Figure 6.

The higher the entropy value, the lower the confidence of the response, and the greater the likelihood of relation hallucinations occurring. The $E(X)$ values we use serve as a threshold for our calibration operation. Calibration is performed only when the entropy exceeds this threshold. The benefit of this approach is that it helps avoid unnecessary calibration for cases that do not involve hallucinations. We chose $E = 0.6$ as a threshold to focus on cases where hallucinations are more likely to occur, ***ensuring that calibration is applied only to cases***

with a higher degree of certainty of hallucination, and try our best to reduce calibration on non-hallucination cases.

❻ What is the meaning of meaningless questions in manual verification when constructing Reefknot?

"Meaningless questions" are defined as questions like "Are there ears on a milk cow?" or "Are there eyes on a man?" These types of triplets are fixed combinations, and even without visual information, they can easily be inferred through common sense.

A.7 VQA Criterion

Model	Perception↓		Cognition↓	
	DeBERTa	GPT-assisted	DeBERTa	GPT-assisted
LLaVA	0.68	0.93	0.67	0.94
Deepseek	0.70	0.89	0.69	0.81
MiniGPT4	0.61	0.78	0.61	0.83
MinCPM	0.76	0.87	0.69	0.94

Table 7: Comparison of Hallucination Rates between DeBERTa-Based and GPT-Assisted Methods in VQA settings.

Here are the evaluation criteria for VQA questions. We use the DeBERTa[1] model to determine whether the models entail each other. Only when the label and response contain each other will it be judged as a correct reply. We show our key function code in Figure 15. To show potential bias associated with the form of constrained prompt and DeBERTa, we compared the evaluation results using a GPT-assisted method. In this setting, no restrictions were imposed on the prompt (e.g., "What is the relationship between A and B?"). The results, as shown in Table 7, demonstrate that when the prompt does not constrain the responses, the performance of MLLMs in VQA about relationships is significantly impaired. To further validate the effectiveness of the DeBERTa model, we compared the judgments made by the DeBERTa with those of human checkers, which demonstrate that DeBERTa can accurately assess the correctness of the model's responses in Table 8. Besides, we show a VQA case in Figure 17.

A.8 More Cases

Figure 16 presents examples of Yes/No, multiple-choice questions (MCQ) and visual question answering (VQA) tasks under perceptive and cognitive conditions, shown separately for comparison.

A.9 Visualization of Relation Word

In Figure 18, we present a word cloud that visualizes the proportion of relational terms within our dataset. It can be observed that, due to the use of semantic triples, our relational terms exhibit greater diversity.

[1] https://huggingface.co/microsoft/deberta-v2-xlarge-mnli

Ground truth	Response	DeBERTa	Manual check
sunlight is **shining on** train	sunlight is **illuminating** train	✓	✓
bear is **reading** book	bear is **sitting** on book	✗	✗
picture is **hanging** wall	picture is **on** wall	✓	✓
man is **lying on** couch	man is **lying** couch	✓	✓
dog is **barking** at stranger	dog is **at** stranger	✗	✗
rain is **falling on** roof	rain is **on** roof	✗	✗
car is **parked in** garage	car is **in** garage	✓	✓
fish is **swimming** pond	fish is **darting** water	✗	✗
dog is **barking** yard	dog is **howling** garden	✓	✓
fire is **burning** hearth	fire is **crackling** fireplace	✓	✓

Table 8: The comparison of results between DeBERTa and Manual Check for real cases from our Reefknot.

```
def are_equivalent(label,response , model, tokenizer, device):
        #label is groundtruth; response is the reply of VLM;
        def check_implication(label, response):
            inputs = tokenizer(label,response, return_tensors="pt").to(device)
            outputs = model(**inputs)
            logits = outputs.logits
            largest_index = torch.argmax(F.softmax(logits, dim=1))
            return largest_index.cpu().item()
        implication_1 = check_implication(label, response)
        implication_2 = check_implication(response, label)

        assert (implication_1 in [0, 1, 2]) and (implication_2 in [0, 1, 2])
        implications = [implication_1, implication_2]
        semantically_equivalent = (implications[0] == 2) and (implications[1] ==
            2)
        # only when both are 2, they are semantically equivalent
        return "yes" if semantically_equivalent else "no"
```

Figure 15: Function to check the semantic equivalence of response and our label.

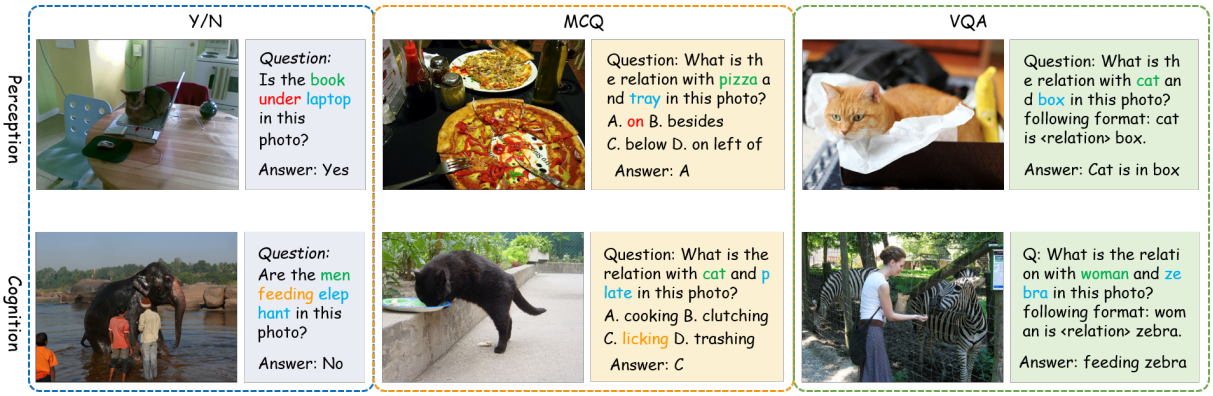


Figure 16: Real-world cases from our proposed Reefknof benchmark. We outline the questions for three types of tasks between perception and cognition perspectives as a reference.

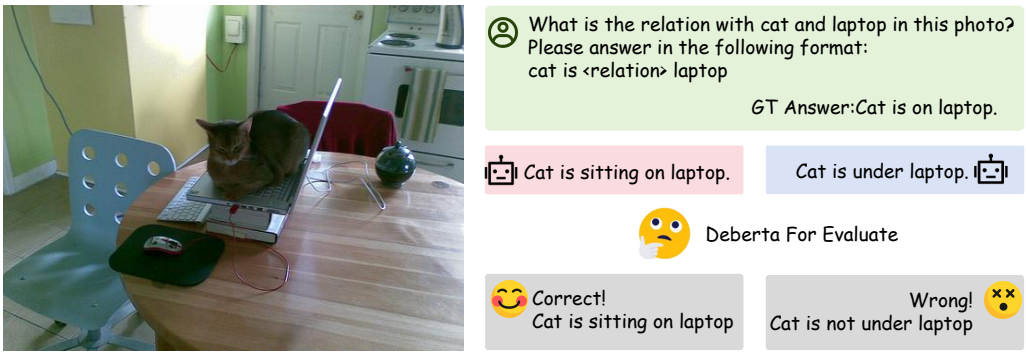


Figure 17: An example of using DeBERTa for evaluation

Figure 18: The word cloud of our relations on Reefknot.