FROM DIAGNOSIS TO MITIGATION: BENCHMARKING AND ROUTED MITIGATION FOR VISUAL AND KNOWLEDGE HALLUCINATIONS IN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) perform tasks that differ in the primary source of information required for correct answers. Some tasks are driven by directly verifiable visual evidence, while others rely more on knowledge and reasoning. We denote them as visual-dominant and knowledge-dominant tasks. Hallucination arises when an MLLM fails to properly balance these two sources of information. To diagnose and evaluate visual and knowledge hallucinations, we construct a new hallucination benchmark, **VKHalluciBench**. Compared with existing benchmarks, VKHalluciBench adopts a novel taxonomy, provides broad coverage across visual- and knowledge-dominant tasks, and incorporates carefully designed open-ended questions and counter-intuition samples that are prone to induce hallucinations. After a comprehensive evaluation, we observe that most mitigation strategies face a common dilemma, which manifests as a trade-off between mitigating visual hallucinations and knowledge hallucinations. To address this limitation, we further propose a plug-and-play routing-based mitigation framework to enable adaptive hallucination mitigation. Specifically, we design a Diagnostic Signal Probe to provide diagnostic results, and an Adaptive Hallucination Mitigation Router then selects suitable mitigation strategies. Experimental results demonstrate that our method achieves the best performance on both visual and knowledge tasks and introduces only 0.33M trainable parameters, resulting in an overall improvement of 12.1% over the baseline model.

1 INTRODUCTION

In recent years, Large Language Models (Yang et al., 2024; Guo et al., 2025; Team et al., 2025b) have progressed rapidly and have evolved into MLLMs (Liu et al., 2024c; Lu et al., 2024a; Team et al., 2025a; Bai et al., 2025a), which have achieved strong performance on multimodal tasks such as visual question answering and image understanding (Liu et al., 2023; 2024c; Bai et al., 2023; Huai et al., 2025; Park et al., 2025). As model capability and deployment scope grow, reliability becomes a central concern, and hallucination remains one of the most prominent failures (Liu et al., 2024b; Wu et al., 2025). Hallucinations in MLLMs mainly fall into two types, namely ① **visual hallucinations** that conflict with visual evidence and ② **knowledge hallucinations** that contradict internal knowledge (Bai et al., 2024). Such hallucinations weaken user trust and introduce potential risks in high-stakes scenarios (Liu et al., 2024d; Ding et al., 2025). Systematic evaluation of hallucination in MLLMs is therefore a prerequisite for safe deployment and reliable use.

To distinguish the two hallucination types discussed above, we categorize multimodal tasks by the primary information source required for correct answers. ① Visual-dominant tasks are driven by directly verifiable visual evidence in the input image. ② Knowledge-dominant tasks require integrating the input with world or domain knowledge, logical computation, or multi-step reasoning. When an MLLM handles visual-dominant tasks and internal priors conflict with visual evidence, visual hallucination emerges (Xie et al., 2024). When knowledge-dominant tasks are answered without adequate knowledge support or reliable reasoning, knowledge hallucination emerges (Bai et al., 2024). To evaluate these hallucinations under a unified view, we construct a **Halluci**nation **Bench**mark grounded in a taxonomy that distinguishes **V**isual-dominant and **K**nowledge-dominant tasks, called **VKHalluciBench**. Compared with existing hallucination benchmarks that focus mainly on object,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

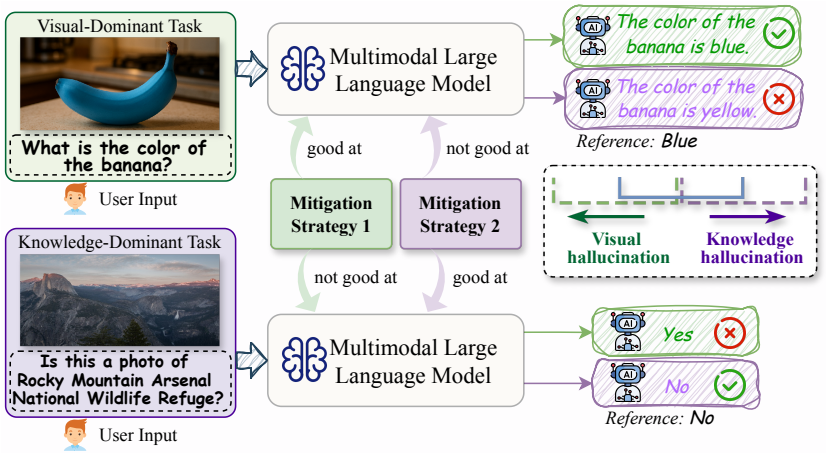


Figure 1: Mitigation strategies rarely excel on both task categories simultaneously, and a clear trade-off is observed between performance on visual and knowledge hallucinations.

attribute, and relation inconsistencies, VKHalluciBench covers a broader set of task scenarios and incorporates counter-intuition cases and carefully designed questions that are prone to induce hallucinations. VKHalluciBench enables quantitative assessment of both MLLM hallucination behavior and the effectiveness of mitigation strategies, and it supports analysis of failure causes. Extensive evaluations and comparisons across representative MLLMs and mitigation strategies are provided.

A key empirical finding is identified from the evaluation of a broad set of mitigation strategies on VKHalluciBench, as illustrated in Figure 1. ***Hallucination mitigation strategies exhibit a performance trade-off between these two task categories.*** They often deliver clear gains on either visual-dominant tasks or knowledge-dominant tasks, but provide limited benefits on the other category and, in some cases, even exacerbate hallucinations. To better exploit the strengths of mitigation strategies across task categories, we further propose the Adaptive Diagnostic Guided Mitigation Routing (ADGMR). ADGMR consists of a **Diagnostic Signal Probe** (DSP) and an **Adaptive Hallucination Mitigation Router** (AHMR). ✤ DSP infers whether an input query is visual-dominant or knowledge-dominant and outputs a diagnostic confidence score. ✤ AHMR selects a matched hallucination mitigation strategy based on the diagnostic output, which improves mitigation coverage by aligning the intervention with the dominant information source and enables more effective use of complementary strategies. Experimental results demonstrate that the proposed method achieves over 99% diagnostic accuracy and improves the performance of the baseline model by 12.1%, which substantially reduces hallucinations while maintaining strong generalization. An analysis of the fuzzy set further shows that ADGMR remains robust on uncertain samples with low diagnostic confidence. Our contributions are summarized as follows:

- ❶ We construct a new benchmark for MLLM hallucination, **VKHalluciBench**, and conduct a comprehensive quantitative evaluation and analysis of representative MLLMs and hallucination mitigation strategies.
- ❷ We propose ADGMR, which selects appropriate mitigation strategies based on accurate diagnosis of hallucination type. This *plug-and-play* design leverages the complementary strengths of different mitigation strategies, thereby achieving optimal effectiveness.
- ❸ Through system-level analysis, we reveal a pronounced trade-off between visual-dominant and knowledge-dominant hallucinations and show that explicit diagnosis and adaptive routing are necessary to avoid negative transfer caused by mismatched interventions.
- ❹ Experimental results show that ADGMR effectively reduces hallucinations while introducing only 0.33M trainable parameters. It improves the performance of the baseline model by 12.1% and achieves the best performance on both visual-dominant and knowledge-dominant tasks.

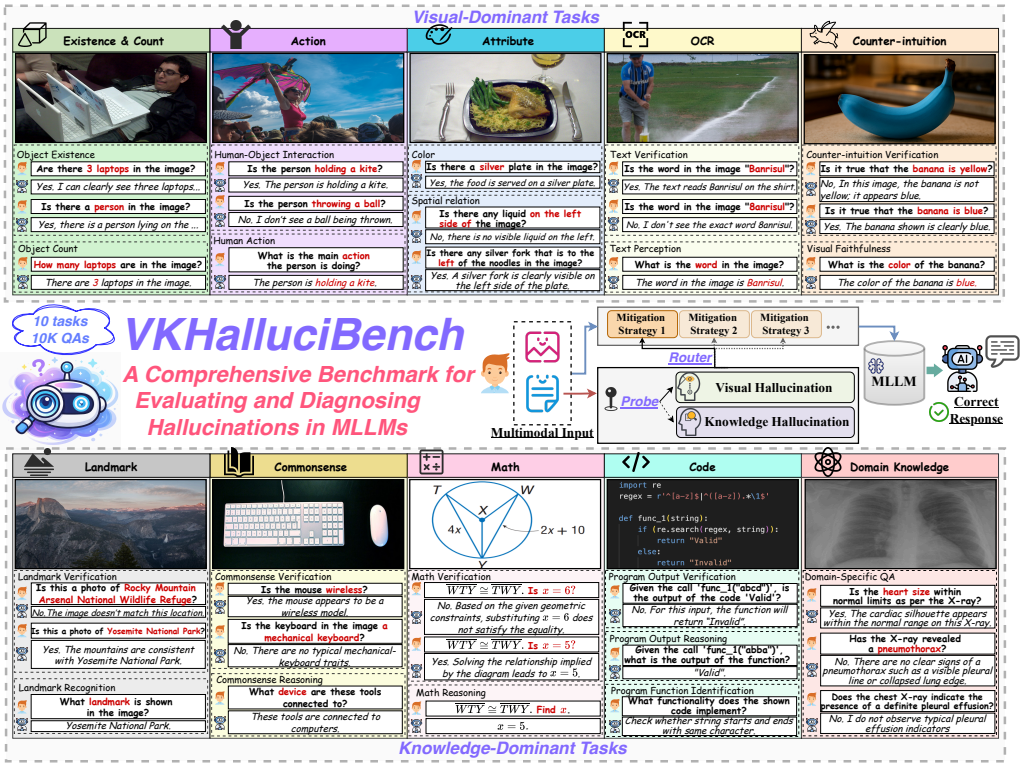


Figure 2: Illustration of our VKHalluciBench. The benchmark evaluates the performance of MLLMs and hallucination mitigation strategies from both visual and knowledge perspectives, covering a total of ten sub-tasks.

2 Related Work

2.1 MLLM Hallucination Benchmarks

Hallucinations in MLLMs have received increasing attention in recent years, motivating the development of diverse benchmarks for evaluation and analysis. MME (Fu et al., 2023) offers a comprehensive suite of coarse-grained and fine-grained tasks, in which the questions are relatively simple and provide a clear snapshot of overall MLLM performance. POPE (Li et al., 2023) formulates object hallucination evaluation as a binary classification task and has been widely adopted due to its simplicity and scalability. AMBER (Wang et al., 2023) covers hallucinations of existence, attributes, and relations, and it integrates CHAIR-based generative metrics and a discriminative F1 score into a unified measure. PhD (Liu et al., 2025a) constructs a visual hallucination evaluation dataset by prompting ChatGPT and frames the evaluation as large-scale, objective binary VQA. Hallusion-Bench (Guan et al., 2024) provides expert-curated visual inputs to analyze the reliance on visual context under different query types.

2.2 Hallucination Mitigation in MLLMs

Existing hallucination mitigation methods can be grouped into three categories. The first category is training-based methods, which typically construct dedicated datasets and then perform post-training via SFT or RL to reduce hallucinations. LRV (Liu et al., 2024a) builds a high-quality dataset with balanced positive and negative samples for SFT, while RLAIF-V (Yu et al., 2025) constructs high-quality feedback data and applies DPO. The second category is decoding-based methods, which reduce hallucinations by modifying the decoding strategy of MLLMs. VCD (Leng et al., 2024) feeds the MLLM with a clean image and a noise-perturbed image separately, and then subtracts the resulting logits to mitigate hallucinations. The third category is intervention-based methods, which aim to reduce hallucinations by intervening in the model's internal activations. VTI (Liu et al.,

Table 1: Comparison of hallucination benchmarks.

Benchmark	Type	QA Form	# Img. / QA	# Subtask	Counter-intuition
MME	General eval	Discriminative	1,187 / 1,457	14	No
POPE	Visual	Discriminative	500 / 3,000	1	No
AMBER	Visual	Discriminative + Generative	1,004 / 14,216	4	No
PhD	Visual	Discriminative	15,398 / 102,564	5	Yes
HallusionBench	Visual + Knowledge	Discriminative	346 / 1,129	9	No
VKHalluciBench	**Visual + Knowledge**	**Discriminative + Generative**	**2,000 / 10,725**	**10**	**Yes**

Table 2: Overview of VKHalluciBench.

Category	Subtask	Data Source	#Img. / #QA	#Total
Visual-dominant	Existence & Count	COCO (Lin et al., 2014)	200 / 1,262	1,000 / 5,462
	Attribute	GQA (Hudson & Manning, 2019)	200 / 1,200	
	OCR	TextOCR (Singh et al., 2021)	200 / 600	
	Action	HICO-DET (Chao et al., 2018)	200 / 1,200	
	Counter-intuition	Generated by GPT-Image	200 / 1,200	
Knowledge-dominant	Landmark	GLDv2 (Weyand et al., 2020)	200 / 1,000	1,000 / 5,263
	Commonsense	OK-VQAv2 (Reichman et al., 2023)	200 / 1,000	
	Math	MathVista (Lu et al., 2024b)	200 / 1,000	
	Code	MBPP (Austin et al., 2021), with code snippets rendered as images	200 / 1,427	
	Domain Knowledge	MMed-RAG (Xia et al., 2025)	200 / 836	

2025b) intervenes in the activations of both the ViT and the LLM. ICT (Chen et al., 2025) performs activation interventions at both the image level and the object level.

Different from existing strategies, our method considers the hallucination type of each query and selects a strategy that is more effective for that type to mitigate hallucinations.

3 VKHALLUCIBENCH

To enable diagnosis of hallucination types and comprehensive evaluation of MLLMs and mitigation strategies, we construct an MLLM hallucination benchmark named VKHalluciBench, as shown in Figure 2. The comparison of benchmarks is summarized in Table 1. VKHalluciBench systematically characterizes hallucination causes, identifies conditions under which hallucinations are more likely to occur, and evaluates the effectiveness of mitigation strategies. By quantitatively analyzing response tendencies, VKHalluciBench supports selection of mitigation strategies and reduces suboptimal performance caused by mismatched mitigation. Table 2 summarizes VKHalluciBench.

3.1 TAXONOMY

Visual-dominant tasks rely primarily on verifiable visual evidence in images to derive correct answers. This category includes five subtasks: *Existence & Count*, *Attribute*, *OCR*, *Action*, and *Counter-intuition*.

Existence & Count includes true-positive questions, random negatives, high-frequency but absent objects, and count perturbations around the ground-truth number. It is designed to evaluate hallucination tendencies of MLLMs in object existence verification and counting scenarios.

Attribute focuses on visual attributes such as color and spatial relations. Under nearly identical syntax and natural semantics, the attribute term is replaced with a semantically similar alternative to construct confounding questions that are likely to induce hallucinations.

OCR evaluates image text reading and fine-grained character recognition. It includes verification questions where a target string appears in the image, distractors with similar characters or similar words (e.g., STOP↔SHOP and 8↔B), and open-ended OCR questions. This subtask measures high-confidence hallucinations caused by character misreading.

Action covers human action recognition and human–object interaction recognition. It primarily adopts substitutions between similar actions and perturbations of interaction objects. It evaluates whether language priors lead to incorrect action predictions or incorrect interaction objects under visually similar cues.

Counter-intuition constructs samples that conflict with commonsense while preserving clear visual evidence. It includes four types of anti-commonsense cases, namely abnormal object attributes,

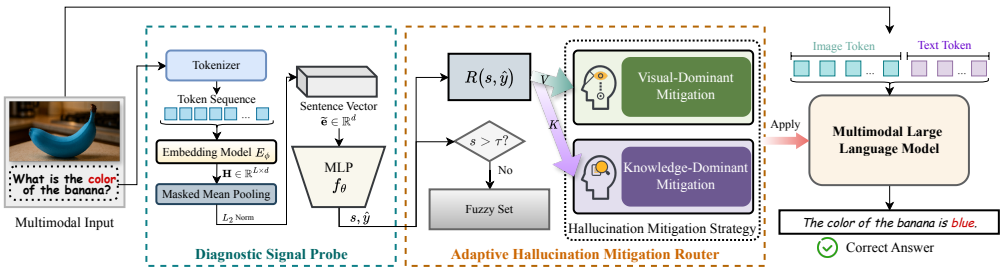


Figure 3: Illustration of ADGMR. The plug-and-play ADGMR consists of two components, DSP and AHMR. DSP first diagnoses the hallucination type from the text input and produces a diagnostic signal. Based on this signal, AHMR selects a matched mitigation strategy and applies it to the MLLM to obtain the final correct answer.

object misuse, role mismatch, and scene conflict. Counter-intuition images are generated using GPT-Image. This subtask evaluates whether models prioritize visual evidence under strong priors.

Knowledge-dominant tasks rely primarily on knowledge and reasoning. Correct answers require integrating visual evidence with world knowledge, domain knowledge, or logical computation. This category covers five subtasks: *Landmark*, *Commonsense*, *Math*, *Code*, and *Domain Knowledge*.

Landmark evaluates recognition of well-known buildings or scenes and the associated location-related knowledge. The settings include correct landmark names, confusable landmark distractors that are semantically or visually similar, and locally cropped images. These designs create cases with insufficient visual cues and test whether models generate overconfident responses under uncertainty.

Commonsense includes questions supported by the image, contrastive questions that are commonsense-plausible yet conflict with the image, and strongly prior-inducing prompts. This subtask captures hallucinations in which commonsense priors override visual evidence.

Math evaluates the ability to read key visual information and perform multi-step computation. Adjacent numeric answers are introduced as distractors to evaluate hallucinations caused by incorrect mathematical reasoning.

Code evaluates program understanding and execution-level reasoning using executable code snippets. The settings include predicting the program output for a given input, summarizing the program functionality, and answering verification questions with correct and confusable outputs.

Domain Knowledge focuses on specialized medical scenarios covering pathology, radiology, and ophthalmology. Since this work targets general MLLMs, this subtask adopts image-grounded question answering with both verification-style and open-ended questions to evaluate hallucinations caused by insufficient domain knowledge.

3.2 Evaluation Metrics

To evaluate model performance on our benchmark in a simple and consistent manner, two metrics are adopted, namely ACC-Q and ACC-I. ACC-Q represents the accuracy that is computed at the question level, while ACC-I is computed at the image level. The latter requires all questions associated with the same image to be answered correctly, which yields a stricter evaluation and better reflects whether a model follows visual evidence rather than relying on priors. For binary discriminative questions, the answer space is restricted to *yes* or *no*. For open-ended questions, LLM-assisted accuracy is used. To ensure reliability, we set the temperature to 0. Each answer is evaluated three times, and the final score is the mean of the three runs.

4 Adaptive Diagnostic Guided Mitigation Routing

We conduct a comprehensive evaluation of existing multimodal hallucination mitigation strategies, with results reported in Table 4. The results indicate a consistent performance trade-off between mit-

igating visual hallucinations and mitigating knowledge hallucinations. One plausible explanation is that, for visual-dominant tasks, aggressively enabling knowledge-intensive reasoning can introduce irrelevant facts and enlarge the space for fabrication. For knowledge-dominant tasks, overly strict visual constraints can reduce the required reasoning depth and knowledge utilization, thereby degrading accuracy. To address this issue, we propose ADGMR. As shown in Figure 3, the framework consists of two components. DSP first diagnoses the hallucination type, and AHMR then selects a better-matched strategy for hallucination mitigation. ADGMR supports sample-wise adaptive mitigation and reduces performance degradation caused by mismatched interventions.

4.1 Diagnostic Signal Probe

DSP is proposed to diagnose hallucination types using only text information from the input query. It outputs a probability vector that indicates the preference of the input toward one of the two categories and drives the subsequent router to select a better-matched mitigation strategy.

Specifically, given a text prompt $\mathbf{x}$, it is first tokenized into a sequence of length L with truncation and padding, together with an attention mask $\mathbf{m} \in \{0,1\}^L$, where $m_i = 1$ indicates a valid token. The token sequence is then fed into a pretrained encoder E_ϕ to produce hidden states $\mathbf{H} \in \mathbb{R}^{L\times d}$, where d is the hidden dimension. The encoder remains frozen and is used only for feature extraction. A fixed-length sentence vector is then obtained by masked mean pooling:

$$\mathbf{H} = E_\phi(\mathbf{x}), \quad \alpha_i = \frac{m_i}{\sum_{j=1}^{L} m_j}, \quad \mathbf{e} = \sum_{i=1}^{L} \alpha_i \, \mathbf{H}_i, \tag{1}$$

where $\mathbf{H}_i \in \mathbb{R}^d$ denotes the hidden state of the i-th token. The vector $\mathbf{e}$ is further normalized by L_2 normalization to improve stability and comparability:

$$\|\mathbf{e}\|_2 = \sqrt{\sum_{j=1}^{d} e_j^2}, \quad \tilde{\mathbf{e}} = \frac{\mathbf{e}}{\|\mathbf{e}\|_2}. \tag{2}$$

The resulting feature $\tilde{\mathbf{e}} \in \mathbb{R}^d$ is fed into a classifier head, implemented as a two-hidden-layer MLP f_θ with ReLU activation $\sigma(\cdot)$. The MLP produces logits $\mathbf{z} \in \mathbb{R}^2$ as

$$\mathbf{h}_1 = \sigma(W_1\tilde{\mathbf{e}} + b_1), \quad \mathbf{h}_2 = \sigma(W_2\mathbf{h}_1 + b_2), \quad \mathbf{z} = W_3\mathbf{h}_2 + b_3. \tag{3}$$

The class probabilities p are then obtained by applying softmax to $\mathbf{z}$:

$$p = \mathrm{softmax}(\mathbf{z}), \quad p_c = \frac{\exp(z_c)}{\sum_{c'\in\{v,k\}} \exp(z_{c'})}, \tag{4}$$

where c denotes the class label. During training, only θ is updated, while E_ϕ remains frozen. Given training samples $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$, the objective is a weighted cross-entropy loss:

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c\in\{v,k\}} w_c \, \mathbb{I}(y_n = c)\log p_{n,c}, \tag{5}$$

where $w_c = N/(2N_c)$ is the class-balancing weight, with N_c denoting the number of training samples belonging to class c.

4.2 Adaptive Hallucination Mitigation Router

Based on the diagnostic output of DSP, we design the AHMR to select appropriate mitigation strategies. Specifically, before the query is fed into the MLLM, DSP outputs the class probabilities $p = [p_v, p_k]$, the diagnostic label $\hat{y}$, and a diagnostic confidence score:

$$\hat{y} = \arg\max_{c\in\{v,k\}} p_c, \qquad s = \max(p_v, p_k), \tag{6}$$

which measures the certainty of the diagnostic decision. We observe that some inputs yield low diagnostic confidence. Therefore, inputs with confidence scores below a threshold are assigned to a fuzzy set for further analysis. AHMR uses a routing function $\mathcal{R}$ to select a mitigation strategy:

$$\mathcal{R}(s,\hat{y}) = \begin{cases} \mathcal{V}, & \hat{y} = v, \\ \mathcal{K}, & \hat{y} = k, \end{cases} \quad \mathcal{F} = \{\mathbf{x} \mid s(\mathbf{x}) < \tau\}, \tag{7}$$

Table 3: Performance comparison of MLLMs. The best and the second-best results are highlighted in **red** and **blue**, respectively.

Models	Visual-Dominant Tasks												Knowledge-Dominant Tasks											
	E&C		Attr.		Action		C-Int.		OCR		AVG.		Code		Common.		Domain		Land.		Math		AVG.	
	ACC-Q	ACC-I	ACC-Q	ACC-I	ACC-Q	ACC-I	ACC-Q	ACC-I	ACC-Q	ACC-I	ACC-Q	ACC-I	ACC-Q	ACC-I	ACC-Q	ACC-I	ACC-Q	ACC-I	ACC-Q	ACC-I	ACC-Q	ACC-I	ACC-Q	ACC-I
Closed-Source MLLMs																								
Claude-4.5	80.19	35.30	62.67	3.50	81.50	45.00	82.92	42.00	83.67	63.00	78.19	37.76	**82.62**	**34.50**	87.40	55.50	46.05	15.50	60.40	11.50	**79.50**	**30.00**	**71.19**	**29.40**
GPT-4o	83.52	42.00	67.58	5.50	**94.67**	**78.00**	**88.83**	**59.00**	**92.17**	**81.50**	**85.35**	**53.20**	73.72	26.00	**94.30**	**75.00**	22.13	4.00	**68.50**	**19.00**	65.10	17.50	64.75	28.30
Gemini3-Pro	**85.82**	**50.50**	**79.58**	**25.50**	90.08	65.00	**90.75**	**61.50**	**95.17**	**86.50**	**88.28**	**57.80**	**84.86**	**46.50**	**90.60**	**60.00**	54.67	**26.50**	**69.50**	**22.00**	**83.80**	**53.50**	**76.69**	**41.70**
Open-Source MLLMs																								
LLaVA-v1.5-7B	72.03	25.00	75.58	4.50	59.50	17.50	68.33	10.00	56.67	15.00	66.42	14.40	23.76	0.00	69.20	17.00	43.66	5.00	45.30	4.00	51.20	1.50	46.62	5.50
Qwen2.5-VL-3B	77.89	28.00	73.42	**19.50**	76.42	23.00	75.92	21.50	67.17	27.00	74.16	23.80	46.53	2.50	63.70	10.00	33.61	2.50	54.50	7.00	61.20	10.50	51.91	6.50
Qwen2.5-VL-7B	79.95	34.00	**76.25**	18.00	83.58	41.00	79.67	32.00	79.00	47.00	79.69	34.40	63.42	12.50	77.60	29.00	16.63	4.00	59.40	9.00	67.80	18.00	56.97	14.50
Mantis-8B	80.74	31.50	65.67	4.50	85.33	48.00	77.58	19.50	65.50	19.50	74.96	24.60	37.07	0.00	85.90	50.50	**70.45**	26.00	61.40	7.50	46.10	1.50	60.18	17.10
DeepSeek-VL-7B	21.79	0.00	11.17	0.00	29.00	0.00	53.92	15.00	27.00	0.00	28.58	3.00	14.51	0.00	75.40	27.00	21.29	2.00	30.10	0.00	18.90	0.00	32.04	5.80
Gemma3-4B	72.03	3.00	57.08	0.00	69.67	14.00	63.25	6.50	61.33	3.00	64.67	5.30	27.54	0.00	55.70	6.00	0.36	0.00	30.30	0.00	28.70	0.00	28.52	1.20
Idefics3-8B	79.16	32.00	66.58	6.50	74.67	19.50	74.33	14.00	53.00	10.00	69.55	16.40	35.25	0.00	65.90	9.00	49.16	9.00	55.50	3.50	54.80	7.00	52.12	5.70
InternVL3-8B	77.97	25.00	72.42	18.50	87.50	52.50	81.67	38.50	84.17	62.00	80.75	39.30	54.66	4.00	68.20	13.50	35.53	7.50	60.40	8.50	62.00	11.00	56.16	8.90
Qwen3-VL-8B	83.44	39.00	62.92	3.00	87.17	60.50	83.08	38.50	81.50	52.00	79.62	38.60	61.32	10.50	87.40	51.00	53.11	26.00	61.60	9.00	63.10	12.00	65.31	21.70
Qwen3-VL-32B	**84.31**	**43.00**	69.17	8.50	**90.75**	**68.00**	87.08	53.00	87.33	68.00	83.73	48.10	66.29	16.50	89.50	**60.00**	**57.42**	**31.50**	64.10	11.50	68.40	19.00	69.14	27.70

where $\mathcal{V}$ and $\mathcal{K}$ denote visual-dominant and knowledge-dominant mitigation strategies, respectively. $\mathcal{F}$ denotes the fuzzy set, with a threshold τ. AHMR ensures input-adaptive mitigation and avoids suboptimal performance caused by applying a single fixed strategy across tasks, thereby leveraging the strengths of different strategies to mitigate multimodal hallucinations to the greatest extent.

5 Experiments

5.1 Experimental Setups

Baseline MLLMs. We evaluate thirteen MLLMs on VKHalluciBench. The closed-source models include Claude-4.5 (Anthropic, 2025), GPT-4o (OpenAI, 2024), and Gemini3-Pro (Google DeepMind, 2025). The open-source models include LLaVA-v1.5-7B (Liu et al., 2024c), Qwen2.5-VL-3B/7B (Bai et al., 2025b), Mantis-8B (Jiang et al., 2024), DeepSeek-VL-7B (Lu et al., 2024a), Gemma3-4B (Team et al., 2025a), Idefics3-8B (Laurençon et al., 2024), InternVL3-8B (Zhu et al., 2025), and Qwen3-VL-8B/32B (Bai et al., 2025a).

Hallucination Mitigation Strategies. LLaVA-v1.5-7B serves as the base model for evaluating mitigation strategies. Three categories of strategies are applied, including training-based methods (HADPO (Zhao et al., 2023), HALVA (Sarkar et al., 2024), CCA-LLaVA (Xing et al., 2024), RLAIF-V (Yu et al., 2025), SENTINEL (Peng et al., 2025), SymMPO (Liu et al., 2025c), and SEVA (Zhu et al., 2024)), decoding-based methods (VCD (Leng et al., 2024), ONLY (Wan et al., 2025), DECO (Wang et al., 2024), and RITUAL (Woo et al., 2024)), and intervention-based methods (VHR (He et al., 2025), VAF (Yin et al., 2025), Causal-LLaVA (Zhou et al., 2025), VTI (Liu et al., 2025b), and ICT (Chen et al., 2025)).

5.2 Comparative Performance Analysis

Evaluation of MLLMs. As shown in Table 3, Gemini3-Pro achieves the best average performance on both visual-dominant and knowledge-dominant tasks and ranks first in 7 out of the 10 subtasks. GPT-4o demonstrates stronger visual perception, ranking second on visual-dominant tasks and attaining the best performance on Action, whereas Claude-4.5 ranks second on knowledge-dominant tasks, indicating a relative advantage in knowledge-intensive reasoning.

✤ **Obs 1**: Image-level consistency ACC-I remains a key bottleneck, even for top closed-source models.

As illustrated in Figure 4 (a), Gemini3-Pro lies in the upper-right region and achieves the highest scores on both ACC-Q and ACC-I, reflecting stronger holistic understanding and cross-question consistency. GPT-4o and Claude-4.5 follow closely, while a clear gap on ACC-I remains between them, which suggests that performance diverges more under the stricter requirement that all questions associated with the same image must be answered correctly. Open-source models mainly cluster in the middle-to-lower range. Among them, Qwen3-VL-32B is closer to the high-performing region, yet it still lags behind leading closed-source models, and the gap is more pronounced on ACC-I. These ob-

Table 4: Performance comparisons of mitigation strategies. The best and the second-best results are highlighted in red and blue, respectively.

Mitigation Strategies	Visual-Dominant Tasks												Knowledge-Dominant Tasks											
	E&C		Attribute		Action		C-Int.		OCR		AVG.		Code		Common.		Domain		Land.		Math		AVG.	
	ACC-Q	ACC-I	ACC-Q	ACC-I	ACC-Q	ACC-I	ACC-Q	ACC-I	ACC-Q	ACC-I	ACC-Q	ACC-I	ACC-Q	ACC-I	ACC-Q	ACC-I	ACC-Q	ACC-I	ACC-Q	ACC-I	ACC-Q	ACC-I	ACC-Q	ACC-I
Vanilla	72.03	25.00	75.58	4.50	59.50	17.50	68.33	10.00	56.67	15.00	66.42	14.40	23.76	0.00	69.20	17.00	43.66	5.00	45.30	4.00	51.20	1.50	46.62	5.50
Training-based																								
HADPO	71.79	24.00	72.50	1.00	57.33	16.50	68.00	12.50	49.50	12.50	63.82	13.30	22.99	0.00	69.50	19.50	42.82	4.50	42.40	3.00	51.70	0.00	45.88	5.40
HALVA	71.55	21.00	72.50	4.50	60.92	13.50	69.75	17.00	54.33	16.00	65.81	14.40	25.30	0.00	78.00	26.00	37.44	4.00	44.60	2.00	49.90	0.00	47.05	6.40
CCA-LLaVA	68.94	21.00	72.67	4.00	60.33	6.00	63.42	7.00	45.17	5.00	62.11	8.60	21.72	0.00	68.20	15.00	33.49	1.50	42.70	1.00	49.40	1.50	43.10	3.80
RLAIF-V	**77.50**	**45.50**	**82.83**	4.00	**66.75**	14.50	72.00	**27.00**	59.83	2.00	**71.78**	18.60	6.80	0.00	63.50	10.50	13.28	1.00	49.00	4.00	53.00	1.00	37.12	3.30
SENTINEL	73.22	25.50	76.08	4.50	61.50	17.50	70.67	14.00	49.67	12.00	66.23	14.70	24.25	0.00	71.30	19.00	42.70	5.50	45.00	1.50	49.50	1.00	46.55	5.40
SymMPO	77.10	24.50	73.42	2.50	61.17	13.00	69.08	22.50	60.17	**26.50**	68.19	17.80	**28.03**	0.00	**79.70**	**34.00**	31.10	5.00	47.10	2.00	**58.80**	1.50	48.95	8.50
SEVA	**78.53**	37.50	80.50	**9.00**	**66.83**	**34.00**	**80.42**	**28.50**	**63.50**	**26.50**	**73.96**	**27.10**	24.60	0.00	**80.90**	**37.50**	42.82	5.00	**53.90**	**8.50**	50.20	1.50	50.48	**10.50**
Decoding-based																								
VCD	72.35	31.50	76.67	3.00	60.17	19.00	74.00	15.50	**61.33**	2.00	68.90	14.20	23.20	0.00	74.80	26.50	47.73	6.00	41.50	2.00	51.10	**2.00**	47.67	7.30
ONLY	71.71	20.00	70.67	2.50	58.58	5.00	64.08	12.00	48.50	10.00	62.71	9.90	22.42	0.00	63.90	11.50	36.24	4.00	40.80	2.00	42.00	0.00	41.07	3.50
DECO	76.39	37.50	80.33	1.50	61.42	**24.50**	**77.92**	18.50	59.83	19.50	71.18	20.30	20.88	0.00	75.50	27.00	**55.14**	**10.50**	49.90	3.50	54.20	**2.00**	**51.12**	**8.60**
RITUAL	77.02	37.50	79.83	3.50	62.92	**24.50**	74.92	22.00	59.67	**26.00**	70.87	**22.70**	27.68	0.00	74.90	25.00	**54.78**	**9.00**	46.10	2.50	**59.50**	1.50	**52.59**	7.60
Intervention-based																								
VHR	75.99	**40.50**	**81.50**	5.00	60.08	23.50	76.42	18.00	57.33	14.50	70.26	20.30	**28.52**	0.00	74.00	22.50	53.95	8.00	48.30	2.00	50.20	1.50	50.99	6.80
Causal-LLaVA	75.20	33.00	80.42	**5.50**	65.00	22.50	75.33	17.00	60.33	21.00	71.26	19.80	26.91	0.00	74.30	27.00	43.90	2.50	**50.70**	**4.50**	52.60	**5.00**	49.68	7.80
VAF	74.41	25.50	75.75	4.50	59.75	17.00	69.17	17.00	53.83	15.50	66.58	15.90	22.92	0.00	71.90	20.50	43.06	3.00	46.30	3.00	50.50	**2.00**	46.94	5.70
VTI	74.56	25.50	75.08	1.50	57.92	16.50	70.33	19.50	54.67	16.50	66.51	15.90	22.99	0.00	70.10	18.00	41.99	3.50	40.20	1.00	50.30	1.00	45.12	4.70
ICT	70.76	28.00	75.83	2.00	58.08	17.00	68.92	10.00	51.50	12.50	65.02	13.90	22.70	0.00	67.00	14.00	38.40	5.50	43.90	1.00	46.80	0.50	43.76	4.20

servations further highlight image-level consistency as a key bottleneck that separates current model capabilities.

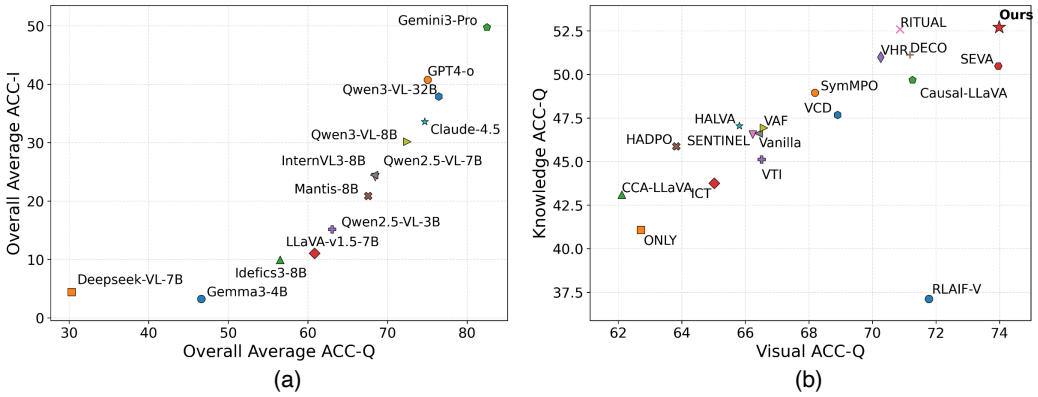


Figure 4: Scatter plot of performance on VKHalluciBench. (a) MLLMs. (b) Mitigation strategies.

Evaluation of hallucination mitigation strategies. On visual-dominant tasks in VKHalluciBench, Table 4 shows that training-based strategies are generally more effective, indicating that interventions that strengthen visual alignment and enforce visual evidence constraints can substantially reduce visually induced hallucinations. However, the advantage of training-based strategies does not transfer to knowledge-dominant tasks. RLAIF-V ranks among the top methods on visual-dominant tasks but degrades sharply on knowledge-dominant tasks, which suggests that an intervention tailored to visual hallucinations can become ineffective, or even harmful, when correct answers depend more on knowledge and reasoning. In contrast, decoding-based strategies tend to be more effective on knowledge-dominant tasks. RITUAL and DECO achieve stronger average performance, which indicates that these methods better support reasoning and knowledge utilization. Their gains on visual-dominant tasks remain smaller than those of the strongest training-based strategies.

✤ **Obs 2**: A trade-off exists between visual- and knowledge-dominant tasks, making it hard for a single strategy to achieve strong gains on both simultaneously.

The difficulty of balancing the two hallucination types becomes more evident in Table 5 and Figure 4. RLAIF-V improves the visual performance but substantially reduces the accuracy of knowledge-dominant tasks, which yields an overall ACC-Q below vanilla and reflects negative transfer. DECO, SEVA, and RITUAL each perform well on a specific dimension and improve the overall score,

Table 5: Overall comparison of mitigation strategies. The results are reported as ACC-Q.

Mitigation Strategies	VKHalluciBench			MME		
	Visual	Knowledge	Overall	Perception	Cognition	Overall
Vanilla	66.42	46.62	56.52	70.38	48.48	59.43
RLAIF-V	71.78	37.12	54.45	77.26	32.50	54.88
DECO	71.18	51.12	61.15	**80.87**	52.41	66.64
VTI	66.51	45.12	55.82	72.41	49.47	60.94
ICT	65.02	43.76	54.39	71.00	52.86	61.93
SEVA	73.96	50.48	62.22	79.84	52.59	66.22
RITUAL	70.87	52.59	61.73	76.41	**55.63**	66.02
Ours	**73.99**	**52.71**	**63.35**	79.70	55.54	**67.62**

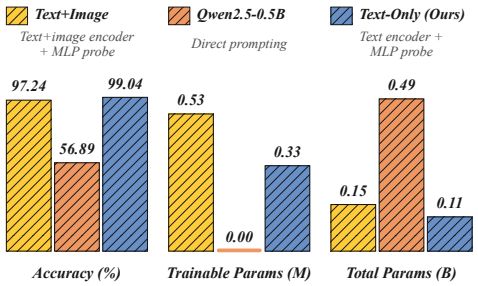


Figure 5: Comparison of different probe variants.

yet the results remain constrained by cross-type trade-offs. Our method achieves the best results on both visual and knowledge, and it attains the highest overall score, which indicates that ADGMR can cover both hallucination types and mitigate performance loss caused by strategy mismatch. To evaluate the generalization and adaptability of ADGMR, we directly apply ADGMR to MME. Single strategies exhibit clear trade-offs between perception and cognition, whereas our method achieves the best overall performance while maintaining competitive cognition results, demonstrating adaptability across benchmarks.

5.3 Analysis of ADGMR

Analysis of the Fuzzy Set. In the fuzzy set $\mathcal{F}$ identified by ADGMR, there are 21 low-confidence samples. To examine the sensitivity of the fuzzy set to routing decisions, we swapped the mitigation strategies assigned to the two branches. The predictions of $20/21$ samples remain unchanged, and only $1/21$ sample exhibits a different outcome. This pattern suggests that the performance bottleneck of fuzzy samples rarely depends on whether the visual branch or the knowledge branch is selected. Instead, it is more likely driven by shared underlying factors, including weakly discriminative evidence, ambiguous question formulations, or intrinsic model uncertainty on these inputs.

✤ **Obs 3**: ADGMR remains robust under diagnostic uncertainty, as the final outputs for fuzzy samples are largely invariant to the selected mitigation strategy.

A practical insight is that samples in $\mathcal{F}$ are largely insensitive to the specific branch strategy. For low-confidence inputs, the risk of negative transfer from routing mismatch is therefore substantially reduced, and both mitigation families tend to be similarly applicable. This finding indicates robustness of ADGMR under diagnostic uncertainty.

Choice of probe type. To obtain the most reliable diagnostic signal, three probe variants are evaluated. Text-Only takes the text modality as input and implements the probe as an MLP. Text+Image takes both the image and the text as inputs, encodes them with CLIP (Radford et al., 2021), concatenates the resulting embeddings, and feeds the fused representation into an MLP. The third variant directly uses Qwen2.5-0.5B (Yang et al., 2025) as the probe. As shown in Figure 5, Text-Only achieves the best accuracy of 99.04%. It is also more parameter-efficient than Text+Image, which makes it a better fit for a lightweight diagnostic module. Based on the above experiments, we select the Text-Only setting for DSP.

6 Conclusion

In this work, a large-scale and comprehensive hallucination benchmark is constructed based on visual-dominant and knowledge-dominant tasks. A broad set of representative open-source and closed-source MLLMs, together with three categories of hallucination mitigation methods, is evaluated on this benchmark through a unified quantitative protocol. The results show that existing mitigation methods are often effective for only one hallucination type, while offering limited benefits for the other and sometimes exhibiting negative transfer. To address this issue, we propose ADGMR to diagnose the hallucination type and to select mitigation strategies accordingly. Experiments confirm the effectiveness of ADGMR, which achieves the best performance on both task categories, surpasses existing mitigation methods, and demonstrates strong generalization and adaptability.

AI USAGE STATEMENT

In this work, generative AI tools were used to assist with benchmark construction, including the generation of counter-intuition images and the generation or refinement of selected questions, distractors, and textual annotations. We also used an LLM as an auxiliary judge for evaluating open-ended model responses. In addition, AI tools were used for language editing to improve grammar, sentence structure, and readability, as well as for limited coding assistance. Generative AI tools were not used to originate the core research ideas, methodology, experimental design, result analysis, or scientific conclusions of this work. All AI-assisted content and evaluation procedures were reviewed and verified by the authors. The authors take full responsibility for the accuracy and integrity of all content, experimental results, and conclusions presented in this work.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. The study does not involve newly recruited human participants or animal experimentation. The source datasets and open-source pretrained models used in our experiments are publicly available and were used in accordance with their respective licenses and terms of use. Proprietary models were accessed through their officially provided services or interfaces under the applicable terms of use.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the benchmark construction process, the proposed framework, experimental settings, and evaluation protocols used in this work. The paper also specifies the data sources, question construction procedures, model configurations, probe training settings, routing procedure, and the prompt used for LLM-assisted evaluation. The source datasets, open-source pretrained multimodal large language models, and evaluation benchmarks used in our experiments are publicly available. Proprietary models are identified by their exact model names and were accessed through officially provided services or interfaces. Additional details are provided in the appendix.

REFERENCES

Anthropic. System Card: Claude Sonnet 4.5, 2025.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, et al. Qwen2.5-vl technical report, 2025b.

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

Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision*, pp. 381–389. IEEE, 2018.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. Ict: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. In *CVPR*, pp. 4209–4221, 2025.

Yifan Ding, Matthew Facciani, Ellen Joyce, Amrit Poudel, Sanmitra Bhattacharya, Balaji Veeramani, Sal Aguinaga, and Tim Weninger. Citations and trust in llm generated responses. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 23787–23795, 2025.

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

Google DeepMind. Gemini 3 Pro model card, 2025.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, pp. 14375–14385, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. ISSN 1476-4687.

Jinghan He, Kuan Zhu, Haiyun Guo, Junfeng Fang, Zhenglin Hua, Yuheng Jia, Ming Tang, Tat-Seng Chua, and Jinqiao Wang. Cracking the code of hallucination in lvlms with vision-aware head divergence. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3488–3501, 2025.

Tianyu Huai, Jie Zhou, Xingjiao Wu, Qin Chen, Qingchun Bai, Ze Zhou, and Liang He. Cl-moe: Enhancing multimodal large language model with dual momentum mixture-of-experts for continual visual question answering. In *CVPR*, pp. 19608–19617, 2025.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pp. 6700–6709, 2019.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*, pp. 13872–13882, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2024a.

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

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pp. 26296–26306, 2024c.

Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. Phd: A chatgpt-prompted visual hallucination evaluation dataset. In *CVPR*, pp. 19857–19866, 2025a.

Sheng Liu, Haotian Ye, and James Zou. Reducing hallucinations in large vision-language models via latent space steering. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Wenqi Liu, Xuemeng Song, Jiaxi Li, Yinwei Wei, Na Zheng, Jianhua Yin, and Liqiang Nie. Mitigating hallucination through theory-consistent symmetric multimodal preference optimization. *arXiv preprint arXiv:2506.11712*, 2025c.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, pp. 386–403. Springer, 2024d.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024a.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*, 2024b.

OpenAI. GPT-4o system card, 2024.

Sung-Yeon Park, Can Cui, Yunsheng Ma, Ahmadreza Moradipari, Rohit Gupta, Kyungtae Han, and Ziran Wang. Nuplanqa: A large-scale dataset and benchmark for multi-view driving scene understanding in multi-modal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8066–8076, 2025.

Shangpin Peng, Senqiao Yang, Li Jiang, and Zhuotao Tian. Mitigating object hallucinations via sentence-level early intervention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 635–646, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Benjamin Z. Reichman, Anirudh Sundar, Christopher Richardson, Tamara Zubatiy, Prithwijit Chowdhury, Aaryan Shah, Jack Truxal, Micah Grimes, Dristi Shah, Woo Ju Chee, Saif Punjwani, Atishay Jain, and Larry Heck. Outside knowledge visual question answering version 2.0. In *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10096074.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. Mitigating object hallucination in mllms via data-augmented phrase-level alignment. *arXiv preprint arXiv:2405.18654*, 2024.

Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *CVPR*, pp. 8802–8812, 2021.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, et al. Gemma 3 technical report, 2025a.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report, 2025b.

Zifu Wan, Ce Zhang, Silong Yong, Martin Q Ma, Simon Stepputtis, Louis-Philippe Morency, Deva Ramanan, Katia Sycara, and Yaqi Xie. Only: One-layer intervention sufficiently mitigates hallucinations in large vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3225–3234, 2025.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. Mllm can see? dynamic correction decoding for hallucination mitigation. *arXiv preprint arXiv:2410.11779*, 2024.

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

Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *CVPR*, pp. 2575–2584, 2020.

Sangmin Woo, Jaehyuk Jang, Donguk Kim, Yubin Choi, and Changick Kim. Ritual: Random image transformations as a universal anti-hallucination lever in large vision language models. *arXiv preprint arXiv:2405.17821*, 2024.

Shengqiong Wu, Hao Fei, Liangming Pan, William Yang Wang, Shuicheng Yan, and Tat-Seng Chua. Combating multimodal llm hallucination via bottom-up holistic reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8460–8468, 2025.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. MMed-RAG: Versatile multimodal RAG system for medical vision language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13258–13273, 2024.

Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. Mitigating object hallucination via concentric causal attention. *Advances in neural information processing systems*, 37:92012–92035, 2024.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, et al. Qwen2 technical report, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, et al. Qwen2.5 technical report, 2025.

Hao Yin, Guangzong Si, and Zilei Wang. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *CVPR*, pp. 14625–14634, 2025.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, et al. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In *CVPR*, pp. 19985–19995, 2025.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Ke Zhu, Liang Zhao, Zheng Ge, and Xiangyu Zhang. Self-supervised visual preference alignment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 291–300, 2024.

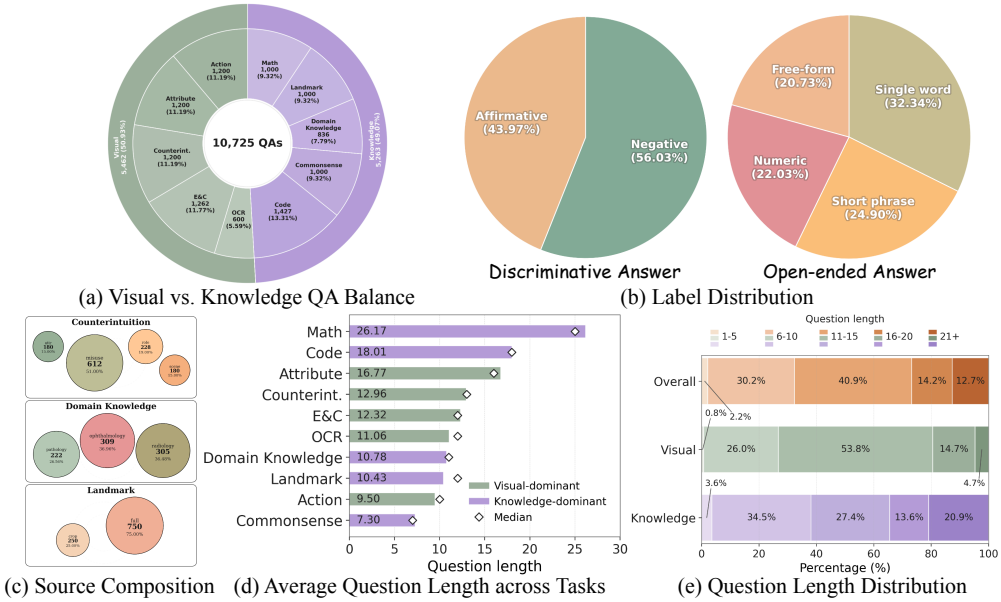
A VKHALLUCIBENCH CONSTRUCTION DETAILS

Table 6: Overview of question construction strategies in VKHalluciBench.

Subset	Source	Construction Focus	Question Design	Answer Type
Existence & Count	COCO	Combine POPE-style existence queries with annotation-derived counting targets; counting negatives are constructed with nearby but incorrect counts as hard foils.	Yes/no existence questions, plus one positive yes/no count question, one open-ended count question, and one negative yes/no count question.	yes/no + integer
Action	HICO-DET	Build questions around the main action and human-object interaction; use distractor actions and distractor interactions as hard negatives for verification.	One open-ended action question, one open-ended interaction question, and positive/negative yes/no questions for both action and interaction.	open-ended + yes/no
Attribute	GQA scene graphs + images	Construct positive/negative questions for color, position, and relation; negatives are constructed using wrong colors, flipped positions, and mutually exclusive spatial relations while reducing ambiguity.	One positive and one negative yes/no question for color attribute, absolute position, and spatial relation, respectively.	yes/no
OCR	TextOCR	Use the exact text string as the positive target and create hard negatives with character-level perturbations, especially visually confusable OCR alternatives.	One positive yes/no OCR question, one negative yes/no OCR question, and one open-ended text recognition question.	yes/no + string
Counter-Intuition	Generated by GPT Image 1	Build image-question pairs from counter-intuition concepts across misuse, scene, role, and attribute, and pair them with supportive and contradictory verification statements.	Four open questions on event, key property, scene, and object, plus one positive and one negative yes/no verification question.	open-ended + yes/no
Commonsense	OK-VQAv2	Extend original VQA supervision with commonsense-focused verification; negative questions are constructed as hard foils through negation, substitution, reversal, or mismatch.	Two open-ended questions and three yes/no questions, including one positive sample and two negative commonsense foils.	open-ended + yes/no
Landmark	GLDv2	Combine true landmark labels with semantically similar distractor landmarks, and include coarse semantic labels and type labels for multi-level recognition.	Landmark identification, positive and negative yes/no landmark verification, coarse natural/human-made classification, and type recognition.	open-ended + yes/no
Math	MathVista	Recast each problem into answer verification by constructing assertion-style questions; negative assertions use answer-specific foils such as nearby numbers or incorrect options.	One original open-ended math question, one positive yes/no assertion, and three negative yes/no assertions.	open-ended + yes/no
Code	MBPP	Turn code into screenshots and construct execution-based questions from function calls; negative verification uses plausible but incorrect outputs as hard distractors.	Multiple open-ended execution questions, one open-ended functionality question, and one positive plus two negative yes/no verification questions.	open-ended + yes/no
Domain Knowledge	MMed-RAG	Construct the subset by selecting 70 radiology samples, 70 pathology samples, and 60 ophthalmology samples from MMed-RAG.	Retain image-grounded source QA pairs and augment them with short open-ended questions and open-ended rewrites while preserving their semantic truth conditions.	open-ended + yes/no

We present the construction details of VKHalluciBench. We introduce the data sources and construction pipelines of all subsets, and describe how the benchmark is organized to support reliable and fine-grained evaluation of multimodal hallucinations. In particular, the Counter-Intuition subset follows a separate construction protocol, since it is based on deliberately designed counter-intuition

concepts and synthesized images instead of directly reusing samples from existing benchmarks. The details are summarized in Table 6. A detailed **benchmark analysis** is provided in Figure 6.



(a) Visual vs. Knowledge QA Balance (b) Label Distribution

(c) Source Composition (d) Average Question Length across Tasks (e) Question Length Distribution

Figure 6: Benchmark distribution and data analysis.

Existence & Count. **Data Source.** We construct the *Existence & Count* subset on top of COCO val2014 by combining POPE-style existence queries with COCO instance-level annotations. Specifically, the existence questions are collected from the *popular*, *adversarial*, and *random* POPE splits, while the counting supervision is derived from the official COCO object annotations.

Construction Pipeline. We first aggregate the POPE questions by image and sample 200 images from the POPE-covered COCO pool using a fixed random seed of 42 for reproducibility. To standardize the evaluation format, we append "Please answer yes or no." to each existence query and remove duplicated questions within each image while preserving the first occurrence. For counting, we compute the object categories and their instance counts for each sampled image, and then select the category with the largest instance count as the counting target. Let the corresponding ground-truth count be x.

Question Design. The subset jointly evaluates **existence recognition** and **instance counting**. For existence, we directly use the cleaned yes/no queries from POPE. For counting, we augment each image with three additional questions: **(1)** a positive yes/no question asking whether there are x instances of the target object, **(2)** an open-ended question asking for the exact number of instances, and **(3)** a negative yes/no question asking whether there are y instances, where $y = x + 1$ if $x \leq 1$, and otherwise y is chosen as either $x - 1$ or $x + 1$. This design enables the benchmark to assess both coarse-grained object presence and fine-grained quantity discrimination in a unified setting.

Answer Format. All yes/no questions are labeled with *Yes* or *No*, while open-ended counting questions are labeled with the corresponding integer string.

Action. **Data Source.** We build the *Action* subset from the HICO-DET test split together with its human-object interaction annotations. This subset is designed to evaluate visually grounded human action understanding and human-object interaction recognition.

Construction Pipeline. We first retain only images containing visible human-object interactions with valid human-object connections. From this candidate pool, we sample 200 images with a coverage-oriented strategy that prioritizes images introducing more previously uncovered interaction categories, and then fill the remaining slots by random sampling under a fixed seed of 42. In the initial stage, each image is associated with a primary interaction selected from its annotations.

To further improve naturalness and reduce rule-based bias, we then regenerate the final question set with a vision-language model, which produces structured outputs containing candidate actions, interactions, distractor actions, and distractor interactions. We parse these outputs and apply conservative fallback rules whenever any field is missing, ensuring that every sampled image can be converted into a complete benchmark entry.

Question Design. The subset evaluates four related capabilities: **action recognition**, **interaction understanding**, **action-level distractor rejection**, and **interaction-level distractor rejection**. For each image, we construct six mutually independent questions: **(1)** an open-ended question asking for the main action, **(2)** an open-ended question asking for the main human-object interaction, **(3)** a positive yes/no action question, **(4)** a negative yes/no action question using a distractor action, **(5)** a positive yes/no interaction question, and **(6)** a negative yes/no interaction question using a distractor interaction. This design makes each question self-contained and suitable for large-scale batched evaluation.

Answer Format. The first two questions require short free-form answers describing the main action and interaction, respectively. The remaining four yes/no questions are labeled as *Yes*, *No*, *Yes*, and *No*, respectively.

Attribute. Data Source. We construct the *Attribute* subset from GQA scene graphs and their corresponding images. The scene graphs provide structured annotations including object names, bounding boxes, attributes, and relations, which enable us to derive visual facts for attribute understanding.

Construction Pipeline. We scan the scene graphs image by image and only retain images for which three subtypes can all be successfully instantiated: attribute, position, and relation. The final subset contains 200 images, and each retained image is paired with a corresponding text file storing all associated questions. For attribute construction, we extract color terms from object attributes and use them to form positive and negative samples. For position construction, we use the horizontal center of each object box to determine whether the object lies on the left or right side of the image, while discarding objects in the middle region to avoid ambiguity. For relation construction, we extract stable spatial relations from the scene graph and keep only those that admit reliable negation. Before including an image in the final subset, we further apply explicit disambiguation rules to reduce referential ambiguity caused by repeated object categories or overly similar alternatives.

Question Design. The subset evaluates three forms of visually grounded attribute understanding: **color attribute recognition**, **absolute position recognition**, and **spatial relation recognition**. For each image, we construct six yes/no questions in total, including one positive and one negative sample for each subtype. For attribute questions, the negative sample is created by replacing the true color with a wrong color selected from a predefined vocabulary while avoiding identical colors, same-name object conflicts, and near-color ambiguities. For position questions, the negative sample is obtained by flipping the true side label between left and right. For relation questions, the negative sample is generated by replacing the true spatial relation with its mutually exclusive counterpart, provided that the fabricated triplet does not already exist in the scene graph. These filtering and negation rules make the final subset more visually grounded and less ambiguous.

Answer Format. All questions in the *Attribute* subset are formulated as yes/no questions.

OCR. Data Source. We construct the *OCR* subset from TextOCR annotations and their corresponding original images. The goal of this subset is to evaluate whether a model can accurately recognize the text appearing in an image, reject visually confusing alternatives, and directly output the exact word or number shown in the scene.

Construction Pipeline. To reduce ambiguity in open-ended OCR evaluation, we only retain images satisfying four conditions: **(1)** the image contains exactly one annotated text instance, **(2)** the corresponding text annotation is non-empty, **(3)** the text contains at least one alphanumeric character, and **(4)** the original image is available. This single-instance filtering is critical because the benchmark uses the full original image rather than cropped text patches; without it, the open OCR question could be underdetermined when multiple text regions coexist in one image. After filtering the candidate pool, we randomly sample 200 images using a fixed seed to ensure reproducibility. Each selected image is then paired with a text file with the same name.

Question Design. Each image is associated with three OCR-oriented questions. **(1)** A positive yes/no question asks whether the text in the image matches the ground-truth string. **(2)** A negative yes/no question replaces the ground-truth string with a visually confusing alternative generated by a small character-level perturbation. **(3)** An open-ended question asks the model to directly output the word or number appearing in the image. To construct reliable negative samples, we prioritize substitutions between commonly confused OCR characters, such as *O/0*, *I/1*, *S/5*, *B/8*, and *Z/2*. When such substitutions are not applicable, we use a simple fallback perturbation, such as deleting the last character or appending a new character, while always ensuring that the negative string is non-empty and different from the ground truth.

Answer Format. The two verification questions are labeled with *Yes* and *No*, respectively, while the open OCR question is labeled with the exact string. This design allows the subset to jointly evaluate binary verification and precise text recognition under the same image-question-answer interface.

Counter-Intuition. Concept Taxonomy. We construct the *Counter-Intuition* subset to evaluate whether a model follows the visual evidence when the image conflicts with default commonsense expectations. Rather than relying on naturally collected images, this subset is built from deliberately designed counter-intuition concepts in which an object appears in an unusual usage pattern, scene, social role, or attribute state. The current taxonomy includes four major types: **misuse**, where an object is used in an unconventional way; **scene**, where an object appears in an implausible location; **role**, where an object is associated with an unexpected user or profession; and **attribute**, where an object exhibits an unusual color, shape, or appearance.

Prompt-to-Image Generation. Each sample is first defined by a structured concept specification that records the central object, the intended counter-intuition condition, the target scene, and the positive and negative verification statements. Based on this specification, we generate one image for each sample using `GPT Image 1`. Some concepts are intentionally scene-flexible and therefore do not bind the object to a single pre-fixed location. In such cases, the placeholder scene is materialized into a concrete place under a fixed random seed before image generation, so that both the visual prompt and the final textual supervision refer to a specific location rather than an underspecified placeholder. The final subset contains 200 image-text pairs.

Question Protocol. Each sample contains six question-answer pairs, including four open questions and two verification questions. The four open questions target: **(1)** the overall event description, **(2)** the key counter-intuition property, **(3)** the scene or location, and **(4)** the salient object. The last two questions are a positive and a negative verification pair that explicitly test whether the model accepts the image-grounded statement and rejects the contradictory one. Among these questions, the second open question is category-dependent. For *misuse* samples, it asks how the object is being used; for *scene* samples, it asks where the object is located; for *role* samples, it asks who is using the object; and for *attr* samples, it asks about the unusual color, shape, or appearance of the object.

Answer Normalization. Because open-ended answers can easily become noisy or overly verbose, we apply explicit normalization to the key open-answer field. For attribute-based samples, the answer is simplified to the attribute value itself, such as a color or shape term. For scene-based samples, the answer is reduced to a short location phrase. For role-based samples, the answer is normalized to the user or profession word. For misuse-based samples, the answer keeps the core usage relation phrase. In addition, the construction pipeline supports post-hoc text-only refinement of the question wording and normalized answers without regenerating the images, which helps maintain consistency across samples after the image generation stage is complete.

Final Format. Each sample consists of one generated image and one text file with the same name storing all six question-answer pairs in a line-by-line format. This organization keeps the subset compatible with standard VLM evaluation pipelines while making it possible to analyze performance separately on event recognition, key counter-intuition reasoning, scene grounding, object identification, and verification-based hallucination resistance.

Commonsense. Data Source. We construct the *Commonsense* subset from the validation split of OK-VQAv2, including the paired images, questions, and human-annotated answers. This subset is designed to evaluate whether a model can answer image-grounded questions that require commonsense reasoning, rather than relying only on literal object existence or shallow visual matching.

Construction Pipeline. We first align each question with its corresponding image and answer annotations, and compute a majority-vote answer from the ten human responses. To improve answer reliability for open-ended evaluation, we prioritize samples with higher answer agreement and progressively relax the majority threshold only when necessary to reach the target scale. The candidate questions are then shuffled under a fixed random seed, and we select 200 images for the final subset. For each retained image, we generate a fixed set of five question-answer pairs, where each image is paired with a text file with the same name.

Question Design. Each image is associated with **two open-ended questions** and **three yes/no questions**. The first open-ended question directly reuses the original OK-VQAv2 question, with the majority-vote answer as the label, so that the benchmark remains connected to the original VQA distribution. The second open-ended question is preferably generated by a vision-language model from the image content itself; when generation fails or is unavailable, it is replaced with a fallback reformulation of the original question. For yes/no questions, we target a balanced structure of **one positive sample** and **two negative foils**. These yes/no questions are designed to emphasize commonsense reasoning rather than pure existence detection, and thus explicitly avoid standard existence-oriented patterns such as asking whether an object is simply present in the image. The negative foils are constructed to induce commonsense errors through strategies such as negation, near-neighbor substitution, role reversal, causal reversal, function mismatch, scene mismatch, and normative misunderstanding. When direct model generation is unavailable or invalid, a rule-based fallback mechanism is used to ensure that each image still yields a complete and evaluable question set.

Answer Format. The two open-ended questions are labeled with short textual answers, while the three verification questions are labeled with *Yes* or *No*. This design allows the subset to jointly evaluate image-grounded factual answering, commonsense-based verification, and hallucination resistance under a unified benchmark interface.

Landmark. **Data Source.** We construct the *Landmark* subset from the index split of GLDv2 and its associated landmark metadata. The benchmark is designed to evaluate visually grounded landmark recognition, coarse semantic categorization, and fine-grained discrimination between visually similar landmarks.

Construction Pipeline. We first filter the candidate pool to keep only images with valid landmark annotations, decodable landmark names, available coarse property labels, and usable type metadata. To avoid overly obscure categories that are difficult even for strong vision-language models, we further apply a popularity constraint based on landmark frequency in the source index split. The final version contains 200 samples, consisting of **150 original images** and **50 cropped variants**. To encourage class diversity, we limit the number of original images drawn from the same landmark category and use a fixed random seed for reproducibility. The cropped samples are obtained by applying a center crop and resizing it back to the original resolution, which makes landmark recognition more challenging by reducing the amount of surrounding contextual evidence.

Question Design. For original images, we construct five questions: **(1)** a positive yes/no question asking whether the image shows the true landmark, **(2)** a negative yes/no question asking whether the image shows a distractor landmark, **(3)** an open-ended identification question asking for the landmark name, **(4)** a binary semantic question asking whether the landmark is *natural* or *human-made*, and **(5)** a type question asking for the best semantic type of the landmark. The landmark name is normalized from the source category title to provide a clean textual answer. For the type question, we preferentially use a normalized hierarchical type label, and fall back to a finer-grained type description only when necessary. The distractor landmark in the negative yes/no question is selected to be challenging rather than random: we first prefer landmarks sharing the same normalized type, then landmarks sharing a similar superordinate category, and only use a broader fallback when no closer alternative is available. For cropped images, we retain the same five-question protocol while replacing the original visual input with its cropped variant, thereby increasing recognition difficulty under reduced contextual evidence.

Answer Format. The yes/no verification questions are labeled with *Yes* or *No*. The landmark identification question is labeled with the normalized landmark name, the coarse semantic question is labeled with *natural* or *human-made*, and the type question is labeled with the selected semantic

type string. In this way, the subset evaluates both fine-grained landmark recognition and higher-level semantic understanding within a unified image-question-answer format.

Code. Data Source. We construct the *Code* subset from MBPP, a benchmark of short Python programming problems with reference implementations and unit tests. Rather than presenting the original textual prompt as input, we convert the reference code of each problem into a code-only screenshot and evaluate whether a multimodal model can reason about program behavior directly from the visualized code.

Construction Pipeline. We first shuffle the MBPP samples with a fixed random seed and retain only problems that satisfy three requirements: the reference program contains at least one valid function definition, the unit tests are non-empty, and the function calls appearing in the tests can be successfully parsed and executed. To reduce lexical leakage from semantically informative function names, we anonymize the entry function invoked in the first test case by replacing its original name with a neutral placeholder. We then render the anonymized Python code into a clean screenshot using a monospaced font on a white background, without including the natural-language prompt or the unit tests in the image itself. During benchmark construction, the reference code is executed in a controlled environment, and each test call is evaluated to obtain the corresponding GT. We continue this process until 200 valid code-image pairs are collected.

Question Design. The subset evaluates three complementary capabilities: **program execution reasoning**, **functional understanding**, and **hallucination-resistant output verification**. For each executable test call associated with a sample, we generate one open-ended question asking for the output of the function under that call. We further add one open-ended question asking what functionality the shown code implements, where the answer is a compressed functional phrase derived from the original programming instruction rather than a copied prompt sentence. Finally, based on the first executable test call, we construct one positive yes/no verification question using the true output and two negative yes/no verification questions using perturbed incorrect outputs. The negative outputs are not random strings; instead, they are generated by type-aware perturbations so that they remain plausible but incorrect.

Answer Format. Each sample contains a variable number of question-answer pairs because the number of execution questions depends on the number of valid test calls in the source problem. Open execution questions are labeled with the exact evaluated output string, the functionality question is labeled with a short functional phrase, and the verification questions are labeled with *Yes* or *No*.

Math. Data Source. We construct the *Math* subset from MathVista, using its images, original math questions, and reference answers. We further use auxiliary annotations when needed to support the construction of harder textual distractors. This subset is designed to evaluate image-grounded mathematical reasoning across arithmetic, geometry, chart understanding, symbolic interpretation, and visually grounded word problems.

Construction Pipeline. We first remove samples without valid answers and exclude questions whose original answer is already *yes* or *no*, since our benchmark reformulates answer correctness itself as a yes/no verification task. We then perform coverage-aware sampling to ensure diversity across answer types, question types, and languages, and finally fill the remaining slots by random sampling under a fixed seed until 200 images are selected. For each retained image, we keep the original MathVista question-answer pair and then construct additional verification questions around the same problem. The resulting benchmark preserves the original visual context while recasting answer evaluation into a unified assertion-based format.

Question Design. Each sample contains five question-answer pairs. The first directly reuses the original MathVista question and its ground-truth answer. The second is a positive yes/no assertion that explicitly asks whether the answer to the original question is the correct answer. The remaining three are negative yes/no assertions built from carefully designed foils. The foil construction strategy depends on the answer type. For numerical answers, we use nearby numerical perturbations, digit swaps, or scale changes while preserving units whenever possible. For multiple-choice questions, we select incorrect options from the original choice set. For textual answers, we construct distractors using confusion pairs, generic candidate pools, and auxiliary semantic annotations. Importantly, each assertion explicitly repeats the original question so that every line can be evaluated independently without relying on previous context.

Answer Format. The original question is labeled with the reference answer string. The positive assertion is labeled with *yes*, and the three foil assertions are labeled with *no*. This design jointly evaluates direct mathematical answering and answer-verification robustness under a unified image-question-answer interface.

Domain Knowledge. Data Source. We construct the *Domain Knowledge* subset from **MMed-RAG**, focusing on image-question pairs from three medical domains: **radiology**, **pathology**, and **ophthalmology**. This subset is designed to test whether a model can answer domain-specific questions from image evidence itself, rather than hallucinating unsupported clinical or scientific knowledge.

Construction Pipeline. We first collect the medical visual question-answer pairs and corresponding images from MMed-RAG, and then build a domain-balanced benchmark containing 200 images in total, including 70 radiology samples, 70 pathology samples, and 60 ophthalmology samples. We apply a filtering stage to remove questions that are weakly grounded in the image, such as report-style references, explicitly non-visual metadata, demographic descriptions, family-history cues, and other leakage-heavy clinical context. Starting from the filtered MMed-RAG annotations, we further enrich the benchmark with image-grounded open-ended questions and rewrite selected retained yes/no questions into open-ended questions while preserving their semantic truth conditions. This design makes the subset more suitable for analysis under open-ended evaluation.

Question Design. The subset jointly evaluates **medical visual grounding**, **open-ended domain reasoning**, and **knowledge hallucination resistance**. Each image is paired with multiple question-answer lines rather than a single prompt. These lines include filtered original medical QA pairs, additional image-grounded open questions, and rewrite-based open questions derived from retained source annotations. The generated open questions are constrained to remain short, visually answerable, and semantically consistent with the source supervision, while avoiding demographic leakage and purely report-dependent phrasing. Finally, the benchmark probes whether the model can remain faithful to the visual evidence even in specialized domains where strong prior knowledge may otherwise dominate the response.

Answer Format. Each medical image is paired with a text file containing multiple `question<TAB>answer` entries. Original verification-style questions retain their normalized labels, while generated open-ended questions use short textual answers rather than *Yes*/*No*. This organization allows the subset to support both strict image-grounded evaluation and richer open-ended hallucination analysis.

B Implementation Details

We adopt MiniLM-L12-v2 as the embedding model. f_θ is an MLP with two hidden layers, which takes a 384-dimensional embedding as input, uses hidden dimensions of 512 and 256, and produces two-class logits. For DSP training, 50 images are randomly sampled from each subtask, and all associated questions are included as training instances. f_θ is optimized with Adam for 50 epochs with a learning rate of 0.001. For visual-dominant tasks, SEVA is applied as the mitigation strategy, while RITUAL is applied for knowledge-dominant tasks. The fuzzy-sample threshold τ is set to 0.6.

C Details of LLM-Assisted Evaluation

To improve evaluation efficiency and reduce the cost of manual annotation, we use an LLM as an auxiliary judge to determine whether a model prediction is consistent with the reference answer under the given image and question. The judge takes four inputs: the image, the question, the reference answer, and the predicted answer. It is instructed to inspect the visual content carefully and compare the predicted answer with the reference answer in context. If the predicted answer does not conflict with the reference answer, the judge returns `correct`; otherwise, it returns `incorrect`. The output is restricted to JSON format to support automatic parsing and post-processing.

This protocol is designed to evaluate answer-level consistency rather than open-ended preference. More specifically, the judge does not assess style or fluency. Instead, it determines whether the predicted answer is semantically consistent with the reference answer given the visual evidence and the

question. This design is well suited to our benchmark because the main objective is to test whether a response remains faithful to the image-question pair and avoids hallucinated or contradictory content.

To improve robustness, we use a strict prompt that explicitly instructs GPT-4o to examine the image, the question, and both answers before making a binary decision. The JSON-only output format further reduces ambiguity in downstream evaluation and allows the judgment to be directly converted into correctness indicators for quantitative analysis. The prompt template is shown below.

LLM-Assisted Evaluation Prompt

```
You are given an image, a question, a reference answer, and a
generative answer. Carefully inspect the image and read all
text inputs before making a judgment.
Question: {question}
Reference Answer: {reference_answer}
Generative Answer: {generative_answer}
Determine whether the generative answer is consistent with the
reference answer under the given image and question.
If the generative answer does not conflict with the reference
answer, output correct. If the generative answer conflicts
with the reference answer, output incorrect.
Return the result in JSON format:
{
''judgment'': ''correct or incorrect'',
''reason'': ''brief explanation''
}
```

D Performance Comparisons on Each Subtask

To complement the overall results in the main text, we further present performance comparisons at the subtask level. Figures 7 and 8 report the per-subtask performance of representative MLLMs. Figures 9 and 10 further examine the effects of mitigation strategies across different subtasks. Across all subtasks and models, ACC-Q remains consistently higher than ACC-I, indicating that, compared with achieving answer-level correctness alone, satisfying the stricter requirement of evidence consistency remains the main bottleneck in evaluating hallucination-related behavior.

For the visual-dominant subtasks, Existence&Count and Attribute generally achieve better performance and show relatively smaller differences across models, suggesting that coarse-grained visual localization and attribute recognition are comparatively stable. By contrast, OCR and Counter-intuition usually exhibit larger ACC-Q/ACC-I gaps and more pronounced variation across models, which indicates that fine-grained text perception and counter-intuition reasoning are more likely to induce unsupported outputs. Action presents intermediate difficulty. Models often attain reasonably strong ACC-Q on this subtask, yet still struggle to maintain high ACC-I under strict consistency criteria. For the knowledge-dominant subtasks, Commonsense is generally more robust, whereas Landmark and Domain Knowledge impose higher demands on external knowledge and evidence alignment, leading to more obvious declines in ACC-I. Code is the most challenging subtask overall, with the lowest ACC-I in most cases and the largest gap between ACC-Q and ACC-I, which suggests that, without external tool support, code-related reasoning and verification are especially prone to hallucination.

E Pseudo-Algorithm for ADGMR

Algorithm 1 summarizes the full pipeline of ADGMR. ADGMR is a plug-and-play framework that performs *sample-wise* mitigation routing at inference time. Given a multimodal query (I, x), DSP determines whether the *text query* x is visual-dominant or knowledge-dominant, and outputs both a predicted label $\hat{y} \in \{v, k\}$ and a confidence score s. AHMR then routes the query to the corresponding mitigation strategy, namely visual-dominant mitigation $\mathcal{V}$ or knowledge-dominant mitigation $\mathcal{K}$,

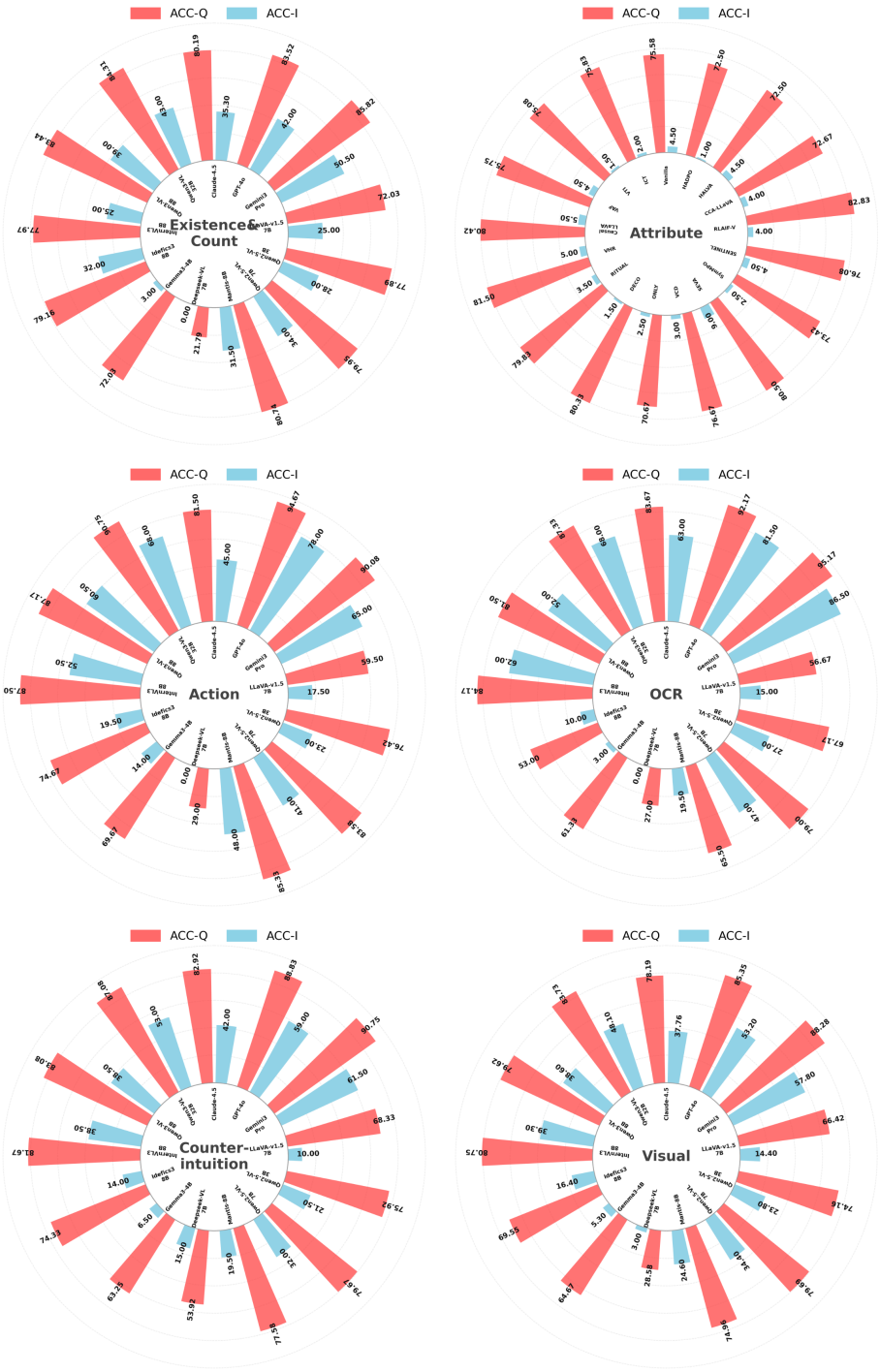


Figure 7: Subtask-wise performance comparisons of MLLMs on visual-dominant tasks.

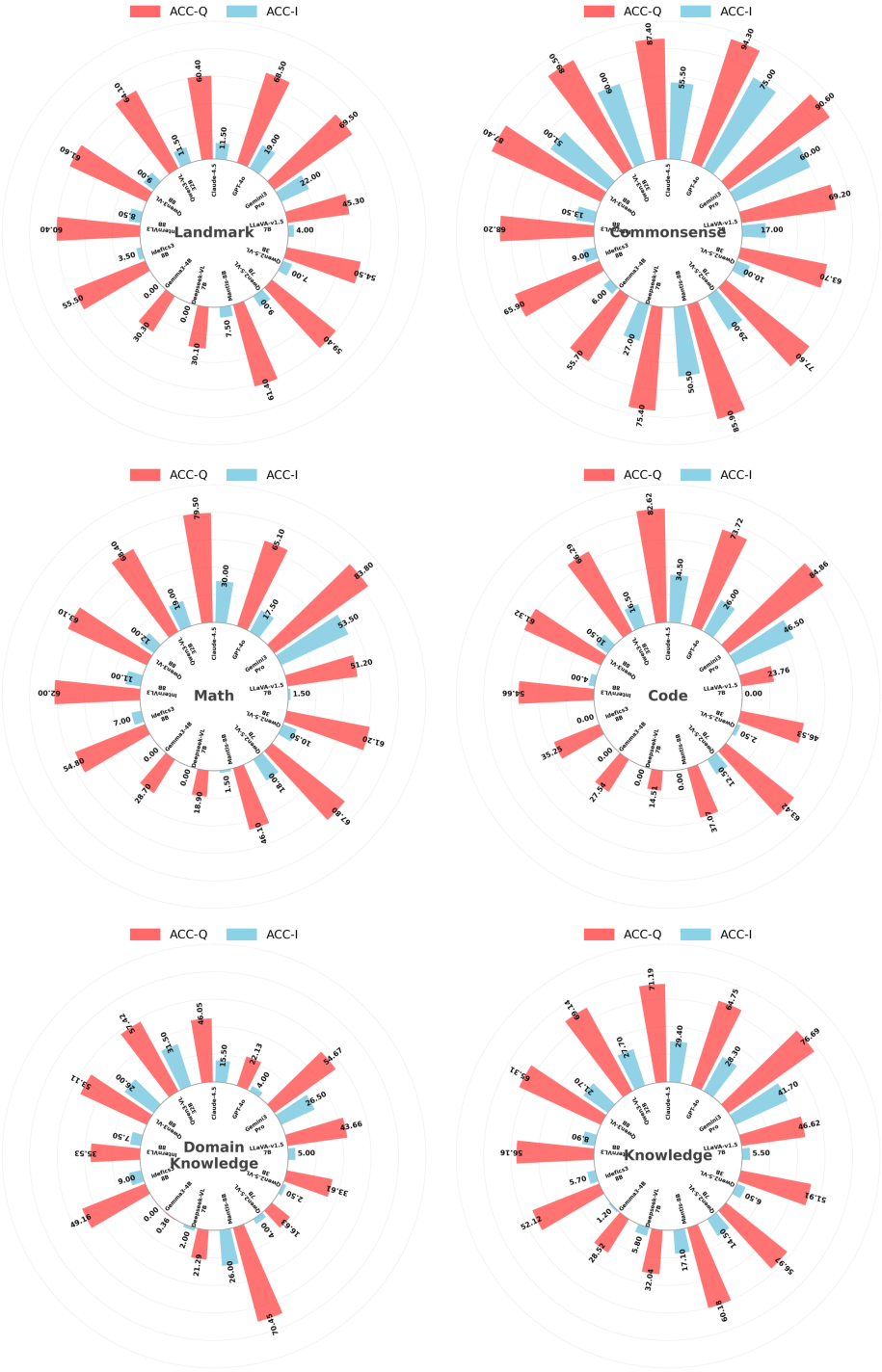


Figure 8: Subtask-wise performance comparisons of MLLMs on knowledge-dominant tasks.

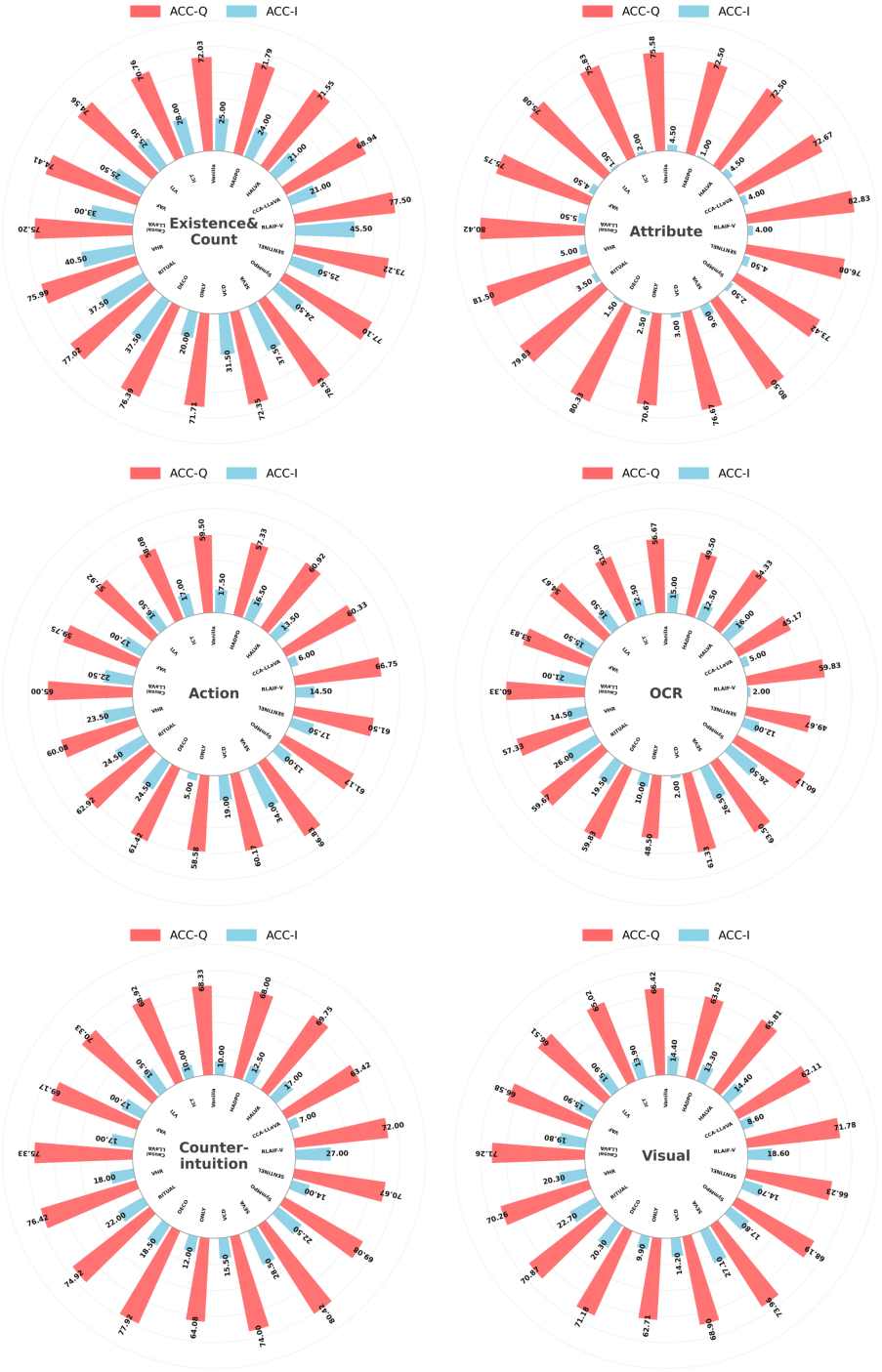


Figure 9: Subtask-wise performance comparisons of mitigation strategies on visual-dominant tasks.

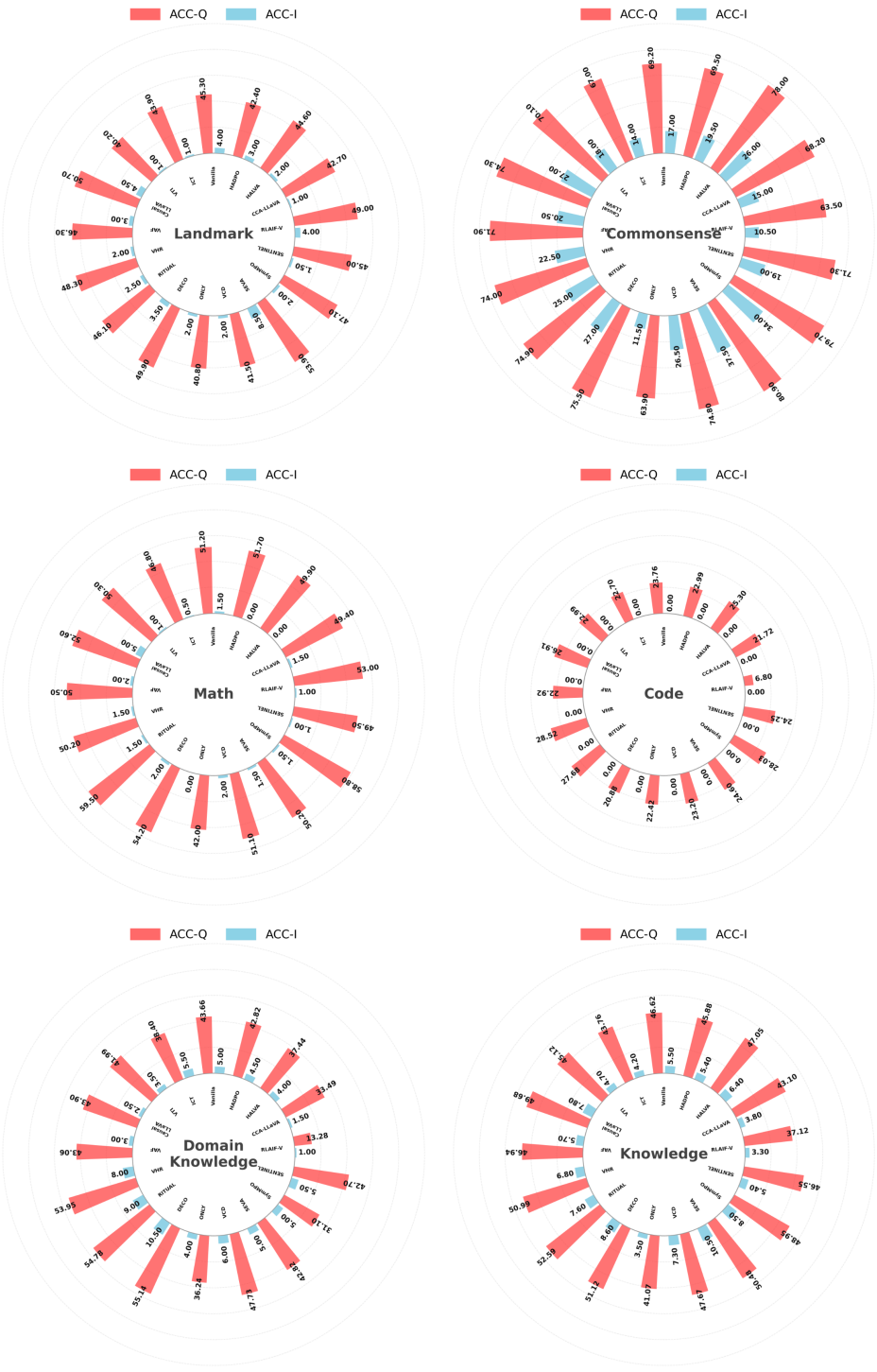


Figure 10: Subtask-wise performance comparisons of mitigation strategies on knowledge-dominant tasks.

before the query is passed to the target MLLM. For low-confidence samples with $s < \tau$, ADGMR records them in a fuzzy set $\mathcal{F}$ for further analysis, while the routing decision is still determined by $\hat{y}$.

Algorithm 1 Adaptive Diagnostic Guided Mitigation Routing

Input: Multimodal query (I, x); target MLLM $\mathcal{M}$; frozen text encoder E_ϕ; DSP classifier head f_θ; visual-dominant mitigation $\mathcal{V}$; knowledge-dominant mitigation $\mathcal{K}$; confidence threshold τ.
Output: Final prediction $\hat{a}$; diagnostic tuple $(\hat{y}, s, p)$; fuzzy-set indicator $\mathbb{I}[x \in \mathcal{F}]$.

// DSP: diagnose the query type from the text-only input x
Tokenize x into a token sequence $\mathbf{x}$ with attention mask $\mathbf{m} \in \{0,1\}^L$.
$\mathbf{H} \leftarrow E_\phi(\mathbf{x}) \in \mathbb{R}^{L \times d}$
$\alpha_i \leftarrow \dfrac{m_i}{\sum_{j=1}^{L} m_j}$; $\mathbf{e} \leftarrow \sum_{i=1}^{L} \alpha_i \mathbf{H}_i$
$\tilde{\mathbf{e}} \leftarrow \dfrac{\mathbf{e}}{\|\mathbf{e}\|_2}$
$\mathbf{z} \leftarrow f_\theta(\tilde{\mathbf{e}}) \in \mathbb{R}^2$; $p \leftarrow \mathrm{softmax}(\mathbf{z}) = [p_v, p_k]$
$\hat{y} \leftarrow \arg\max_{c \in \{v,k\}} p_c$; $s \leftarrow \max(p_v, p_k)$
// AHMR: route the query to the matched mitigation strategy
$\mathbb{I}[x \in \mathcal{F}] \leftarrow \mathbb{I}[s < \tau]$; $\mathcal{F} = \{x \mid s(x) < \tau\}$
if $\hat{y} = v$ **then**
  $\mathcal{S} \leftarrow \mathcal{V}$
else
  $\mathcal{S} \leftarrow \mathcal{K}$
end if
// Apply mitigation and generate the answer
$\mathcal{MS} \leftarrow \mathrm{Apply}(\mathcal{S}, \mathcal{M})$
$\hat{a} \leftarrow \mathcal{MS}(I, x)$
return $\hat{a}$, $(\hat{y}, s, p)$, $\mathbb{I}[x \in \mathcal{F}]$

F Case Analysis on VKHalluciBench

To complement the quantitative results, we further present representative cases for each subtask to provide more concrete qualitative evidence. In each case, we highlight the hallucinated span directly in red within the incorrect response and briefly analyze the corresponding failure mode, thereby clarifying how the error arises in context.

Case 1: Existence & Count

Input



Question: How many instances of stop sign are in the image?
Reference Answer: 1

Ritual

Response: There are two instances of stop signs in the image.
Analysis: The model overcounts the queried object and incorrectly predicts two stop signs instead of one.

Ours

Response: There is one stop sign in the image, located underneath the street sign.
Analysis: The response correctly identifies the number of stop signs and remains consistent with the reference answer.

Case 2: Action

Input



Question: What is the main human-object interaction?
Reference Answer: touching sheep

Ritual

Response: The main human-object interaction in the image is a man kissing a wooden fence.
Analysis: The model mistakes the target object and attributes the interaction to the fence rather than the sheep, resulting in an object substitution hallucination.

Ours

Response: The main human-object interaction in the image is a man kissing or touching the nose of a sheep through a fence.
Analysis: The response correctly identifies the human-object interaction and remains consistent with the reference answer.

Case 3: Attribute

Input



Question: Is there any blue hat in the image? Please answer yes or no.
Reference Answer: Yes

Ritual

Response: No, there is no blue hat in the image.
Analysis: The model incorrectly flips the binary visual attribute judgment, leading to a direct contradiction with the reference answer.

Ours

Response: Yes, there is a blue hat in the image.
Analysis: The response correctly identifies the queried attribute and aligns with the ground-truth answer.

Case 4: OCR

Input



Question: What is the word (or number) in the image?
Reference Answer: HUAWEI

Ritual

Response: The word (or number) in the image is "5".
Analysis: The model fails to recognize the brand text and hallucinates a digit instead, reflecting a clear OCR decoding error.

Ours

Response: The word in the image is "Huawei." It is presented on ...
Analysis: The response correctly identifies the word in the image and remains consistent with the reference answer.

Case 5: Counter-intuition

Input



Question: What is the most salient object involved in the scene? Answer with one noun phrase.
Reference Answer: fork

Ritual

Response: Hairbrush
Analysis: The model defaults to a visually plausible but incorrect everyday object, missing the intentionally counter-intuition usage in the scene.

Ours

Response: Fork
Analysis: The response correctly identifies the salient object involved in the counter-intuition scene.

Case 6: Landmark

Input



Question: Is this a photo of Rijksmuseum Amsterdam? Please answer yes or no.
Reference Answer: Yes

SEVA

Response: No
Analysis: The model fails to recognize the landmark identity and directly contradicts the correct binary label.

Ours

Response: Yes
Analysis: The response correctly identifies the landmark and matches the reference answer.

Case 7: Commonsense Reasoning

Input



Question: What are the metal objects within the tire called?
Reference Answer: spokes

SEVA

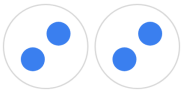
Response: The metal objects within the tire of the bicycle are called the clips and the framework, or in this case, the kickstand.
Analysis: The model introduces unrelated bicycle parts and fails to identify the correct concept, resulting in a clear commonsense recognition error.

Ours

Response: The metal objects within the tire are called spokes.
Analysis: The response correctly identifies the queried object and is fully consistent with the reference answer.

Case 8: Math

Input



Question: Fill in the blank to describe the model. The model has 4 dots divided into 2 equal groups. There are (_) dots in each group.
Reference Answer: 2

SEVA

Response: The model has 4 dots divided into 2 equal groups. There are 3 dots in each group.
Analysis: The model makes a basic arithmetic reasoning error and outputs an inconsistent group size.

Ours

Response: The model has 4 dots divided into 2 equal groups. There are 2 dots in each group.
Analysis: The response correctly completes the quantitative reasoning required by the question.

Case 9: Code

Input

```
def func_1(A, x):
    (left, right) = (0, len(A) - 1)
    result = -1
    while left <= right:
        mid = (left + right) // 2
        if x == A[mid]:
            result = mid
            right = mid - 1
        elif x < A[mid]:
            right = mid - 1
        else:
            left = mid + 1
    return result
```

Question: Given the call `func_1([2, 5, 5, 5, 6, 6, 8, 9, 9, 9], 5)`, what is the output of the function?
Reference Answer: 1

SEVA

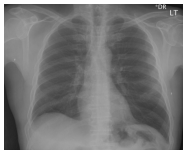
Response: The output of the function would be the `result_mid` (5) because it is specified in the function definition.
Analysis: The model hallucinates the execution behavior and incorrectly returns an input-related value instead of the true function output.

Ours

Response: The output of the function is `1`.
Analysis: The response correctly follows the code logic and produces the correct output.

Case 10: Domain Knowledge

Input



Question: Are the lungs of the patient appearing well-inflated on the X-ray?
Reference Answer: Yes

SEVA

Response: The lungs of the patient appear to be partially collapsed on the X-ray.
Analysis: The model introduces a medically unsupported finding that directly contradicts the reference answer.

Ours

Response: Yes, the lungs appear well-inflated on the X-ray.
Analysis: The response correctly answers the question and remains consistent with the reference label.