

ClueTracer: Question-to-Vision Clue Tracing for Training-Free Hallucination Suppression in Multimodal Reasoning

Gongli Xi¹, Kun Wang², Zeming Gao¹, Huahui Yi²,
Haolang Lu¹, Qianshan Wei³, Ye Tian¹, Wendong Wang¹,

¹Beijing University of Posts and Telecommunications, ²Nanyang Technological University,

³Institute of Automation, Chinese Academy of Sciences

Correspondence: Ye Tian, Kun Wang

Abstract

Large multimodal reasoning models solve challenging visual problems via explicit long-chain inference: they gather visual clues from images and decode clues into textual tokens. Yet this capability also increases hallucinations, where the model generates content that is not supported by the input image or the question. To understand this failure mode, we identify *reasoning drift*: during clue gathering, the model over-focuses on question-irrelevant entities, diluting focus on task-relevant cues and gradually decoupling the reasoning trace from visual grounding. As a consequence, many inference-time localization or intervention methods developed for non-reasoning models fail to pinpoint the true clues in reasoning settings. Motivated by these insights, we introduce CLUERECALL, a metric for assessing visual clue retrieval, and present CLUETRACER, a training-free and architecture-agnostic plugin for hallucination suppression. CLUETRACER starts from the question and traces how key clues propagate along the model’s reasoning pathway (question → outputs → visual tokens), thereby localizing task-relevant patches while suppressing spurious attention to irrelevant regions. Remarkably, **without any additional training**, CLUETRACER improves all **reasoning** architectures (including R1-OneVision, Ocean-R1, MM-Eureka, etc.) by $1.21\times$ on reasoning benchmarks. Our code is available at [this link](#).

1 Introduction

Recent MLLMs have evolved from image description toward cross-modal reasoning, giving rise to Multimodal Large Reasoning Models (MLRMs) that produce multi-step reasoning traces over visual and textual inputs (Liu et al., 2023b; Huang et al., 2025). These reasoning abilities are typically enhanced through supervised fine-tuning or reinforcement learning on pretrained multimodal

backbones (Zhao et al., 2024; Yao et al., 2024; Xu et al., 2025; Yu et al., 2024; Guo et al., 2025).

However, stronger reasoning can also exacerbate hallucination (Lu et al., 2025; Zhou et al., 2024; Bai et al., 2024; Wei et al., 2025), especially when long language-dominant deliberations encourage models to collect excessive or irrelevant clues (Jiang et al., 2025; Dong et al., 2025a; Yi and Shang, 2025; Wang et al., 2025d), causing attention to drift away from visually decisive regions. We refer to this failure mode as *reasoning drift* (Fig. 1 (b)). As illustrated in Fig. 1 (c), constraining the model toward task-critical regions can recover the correct answer, suggesting that the bottleneck lies not in model capacity but in ungrounded clue collection.

In practice, localizing task-critical visual clues at inference time faces two coupled difficulties: (i) **long-horizon output control** and (ii) **generalization beyond object-centric tasks**. These difficulties stem from two implicit assumptions in existing mitigation strategies about where **clues** come from. First, many methods are *token-role-agnostic*: they apply global inference-time biasing, such as contrastive decoding, logit steering, or attention reallocation, without identifying which generated tokens require visual evidence (Leng et al., 2024; Im et al., 2025; Li et al., 2025; Tu et al., 2026; Chen et al., 2025). This can benefit short answers but may accumulate bias over long reasoning traces and miss clues that emerge only at intermediate steps (Zheng et al., 2025). Second, other methods are *task-agnostic*: they seek visual priors from richer captions or descriptions, assuming that the necessary clues can be obtained from the image alone (Ghosh et al., 2024; Yi and Shang, 2025). Yet for complex reasoning tasks such as math (Wang et al., 2024a), the question and intermediate reasoning often define which visual regions are actually decisive. Thus, prior inference-time methods can struggle to isolate the *decisive* clues required for long-chain multimodal reasoning.

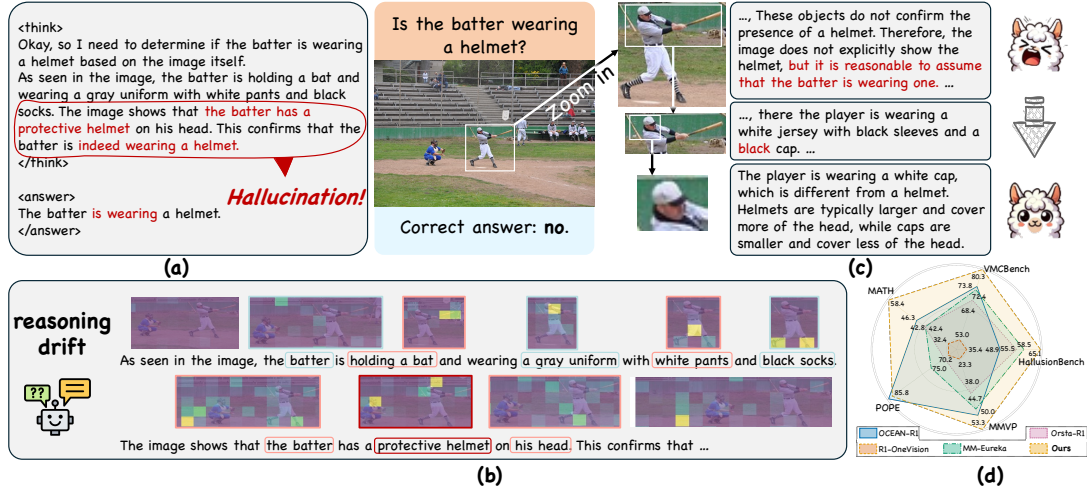


Figure 1: (a) Example outputs from a reasoning model; hallucinated content is highlighted in red. (b) Reasoning drift in a multimodal reasoning model, where attention during inference is allocated to task-irrelevant regions. (c) An intuitive illustration: by progressively zooming in on task-critical regions, hallucinations (in red) diminish and the answers become correct. (d) Radar-chart comparisons among different reasoning models.

To address these mismatches, we introduce CLUERECALL, a metric for measuring visual clue retrieval, and propose CLUETRACER, a training-free inference plugin that suppresses hallucinations by tracing the flow of task-critical clues. Rather than imposing a global visual bias or relying on image-only priors, CLUETRACER explicitly follows a **query** $\rightarrow$ **output** $\rightarrow$ **vision** clue path. It first uses the query to identify task constraints, then leverages generated tokens to determine which parts of the reasoning process need visual support, and finally traces these tokens back to the minimal supporting visual regions. The traced regions are used to re-anchor subsequent reasoning to task-critical evidence, thereby mitigating hallucination without parameter updates or task-specific training.

Our contributions are as follows:

- **(CLUERECALL) Layer-wise Average Clue Recall.** We introduce CLUERECALL to quantify layer-wise visual clue retrieval and reveal a consistent middle-to-late layer peak across architectures and scales.
- **(CLUETRACER) Hallucination suppression in reasoning architectures.** We propose CLUETRACER, which traces native attention along **query** $\rightarrow$ **output** $\rightarrow$ **vision** to localize task-relevant evidence. It improves HallusionBench by $1.25\times$ and VMCBench by $1.17\times$.
- **Transferability to non-reasoning MLLMs.** CLUETRACER also transfers to non-reasoning MLLMs, yielding an average **+8.7** accuracy improvement on non-reasoning settings.

2 Related Work

Multimodal Large Reasoning Models (MLRMs).

Recent MLLMs extend chain-of-thought reasoning to multimodal settings through supervised fine-tuning, reinforcement learning, search, reflection, self-refinement, and long-chain data scaling (Zhao et al., 2024; Zhang et al., 2024a; Xu et al., 2025; Liu et al., 2023b; Thawakar et al., 2025; Yao et al., 2024). To improve reasoning reliability, preference-based and rule-based RL methods optimize models with human feedback, closed-loop refinement, reasoning trace comparisons, or GRPO-style rewards (Yu et al., 2024; Zhang et al., 2024d; Dong et al., 2025b; Guo et al., 2025; Liu et al., 2025b; Zhang et al., 2024b; Huang et al., 2025; Wang et al., 2025b; Meng et al., 2025). Existing MLRMs mainly follow two paradigms: two-stage SFT+RL pipelines and direct large-scale RL training for emergent reasoning (Yang et al., 2025; Tan et al., 2025; Zhang et al., 2025a; Ming et al., 2025; Wang et al., 2025c), with recent work further exploring unified reasoning-perception frameworks and RL-enhanced generative reasoning (Ma et al., 2025b; Duan et al., 2025). However, how to keep long-chain multimodal reasoning grounded in task-critical visual evidence remains underexplored.

Hallucination in MLRMs. Long reasoning traces can amplify hallucination by encouraging models to rely on language priors over visual evidence (Bai et al., 2024; Liu et al., 2025a). Existing inference-time mitigation methods either apply token-role-agnostic grounding, including contrastive decod-

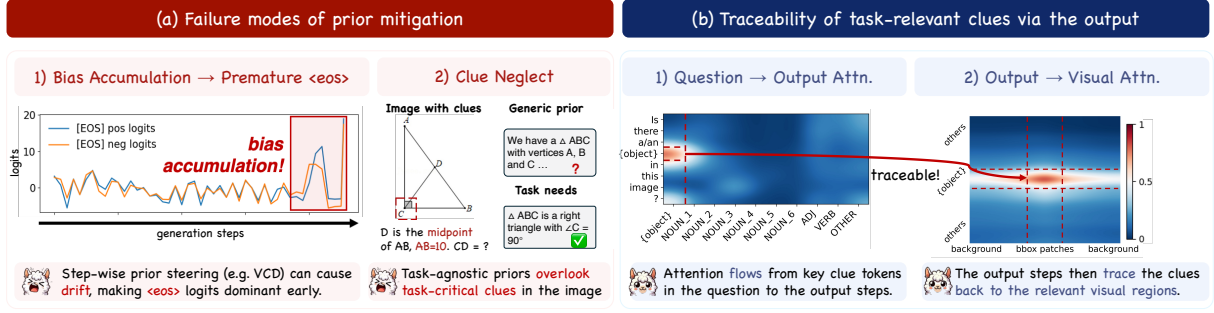


Figure 2: Our Motivations. From Failure Modes to Output-Guided Clue Tracing.

ing, logit steering, and attention control (Leng et al., 2024; Kim et al., 2024a; Im et al., 2025; An et al., 2025a; Li et al., 2025; Kang et al., 2025; Tu et al., 2026; Chen et al., 2025; Ye et al., 2025), or introduce task-agnostic visual priors through generic descriptions or captions (Ghosh et al., 2024; Yi and Shang, 2025). The former may accumulate bias over long traces and miss decisive intermediate evidence (Zheng et al., 2025), while the latter often fails to capture question-conditional constraints required by complex reasoning tasks such as multimodal math (Wang et al., 2024a). Although recent efforts improve hallucination mitigation through training-time alignment and datasets (Dong et al., 2025a; Jiang et al., 2025; Zhang et al., 2025c; Fan et al., 2025; Zou et al., 2024; Wu et al., 2025; Wang et al., 2025a), a step-aware, question-conditional inference-time mechanism for isolating *decisive* visual clues is still lacking.

3 Preliminaries

Multimodal Large Reasoning Models (MLRMs). We consider decoder-style MLRMs that connect a pretrained visual encoder to a language model through a projector. The visual encoder and projector convert an image into visual tokens $\mathbf{X}_v$. Given system tokens $\mathbf{X}_s$ and question tokens $\mathbf{X}_q$, the multimodal prefix is $\mathbf{X}_c = [\mathbf{X}_s, \mathbf{X}_v, \mathbf{X}_q]$, where N_s , N_v , and N_q denote their lengths, and $N_c = N_s + N_v + N_q$. At decoding step t , the model predicts $y_t \sim p(y_t | \mathbf{X}_c, \mathbf{y}_{<t})$, where $\mathbf{y}_{<t} = \{y_i\}_{i=1}^{t-1}$. Reasoning-oriented MLLMs typically generate explicit reasoning traces, often enclosed by `<think>` and `</think>` tags, before producing the final answer (Liu et al., 2023c; Zhang et al., 2023).

Attention in MLRMs. For a Transformer decoder with L layers, we denote by $\mathbf{A}_{t,l} \in [0, 1]^{N_c+t-1}$ the attention distribution at decoding step t and layer l over all available context tokens, including the multimodal prefix $\mathbf{X}_c$ and previously

generated tokens $\mathbf{y}_{<t}$. Thus, each element $A_{t,l,i}$ measures how much the current decoding step attends to the i -th context token, and $\sum_i A_{t,l,i} = 1$. This attention tensor allows us to aggregate token-level attention across layers and decoding steps.

4 Motivations

We conduct three targeted study to understand the thinking process of MLRMs and distill three observations that directly motivate our design in CLUE-TRACER. Each observation is supported by a controlled study and empirical evidence.

Obs 1: Bias accumulation and clue neglect in prior inference-time methods

As shown in Fig. 2 (a), directly steering output logits can encourage an MLRM to rely more on visual signals, but in long-chain reasoning it may induce *bias accumulation*, which can prematurely trigger the `<eos>` token. For example, in VCD (Leng et al., 2024), constructing a noisy negative counterpart and applying step-wise contrast can cause decoding to drift after a certain step, eventually making the `<eos>` logits dominant. Meanwhile, task-agnostic prior methods (e.g., generic captioning) still risk overlooking task-critical clues that are present in the image.

Obs 2: Task-relevant clues are traceable across streams via the output

To examine whether task-relevant clues can be traced across streams, we use category-bbox annotations to identify key visual regions and visualize the attention flow mediated by output tokens. Specifically, we compute attention heatmaps over 2,000 samples by tracing the chain from **question** to **output** to **vision**; implementation details are provided in Appendix E. As shown in Fig. 2 (b), the

heatmaps reveal that decisive visual regions can be localized through this cross-stream chain, supporting the traceability of key clues via the output.

Obs3: Key question tokens show larger **attn. variance** along the output dimension.

Extending **Obs2**, key question tokens exhibit localized attention peaks along the output axis; consequently their attention traces show larger variance than non-key tokens, providing a efficient, training-free signal for key-token detection.

These observations provide a pathway to suppress hallucinations in MLRMs: using the output as a conduit from the question to the visual tokens.

5 Method

In this paper, we present CLUETRACER, an inference-time approach that leverages an MLRM’s long reasoning chain to isolate visual clues critical for answering the question. We first introduce CLUERECALL, a layer-wise visual-grounding metric that identifies the decoder layer where clue retrieval is most reliable (Sec. 5.1). Building on this, CLUETRACER traces key clues across the model’s three streams, progressively localizing pivotal patch tokens via question $\rightarrow$ output $\rightarrow$ vision computations (Sec. 5.2). Finally, we describe how the traced patches are organized into reusable clue regions for downstream reasoning (Sec. 5.3). As shown in Fig. 3, CLUETRACER uses stream-aware clue tracing to pinpoint the minimal visual support required for problem solving, thereby re-grounding generation and mitigating hallucinations.

5.1 CLUERECALL: Assessing Visual Clue Retrieval

Perception is a prerequisite for reliable reasoning. The goal of CLUERECALL is to assess, for each layer of an MLRM, how well its visual attention captures question-relevant visual clues. A single image typically contains multiple dense visual clusters, but only a subset is truly informative for answering a given question (e.g., when asked “Is the batter wearing a helmet?”, the decisive evidence should lie on the batter rather than the catcher). As a precursor to our method, CLUERECALL runs an automatic pipeline that measures how much attention each layer allocates to task-critical regions in three steps:

Query construction. For each image I , we obtain its set of bounding boxes and category labels cat from either annotations (e.g., COCO (Lin et al., 2014)) or a segmentation model (e.g., Mask R-CNN (He et al., 2017)). For each category, we instantiate a query using the fixed template $q_{\text{cat}} = \text{“Is there a/an \{cat\} in this image?”}$, and tokenize it into question tokens $\mathbf{X}_{q_{\text{cat}}}$.

Query-aligned regions. For each (I, q_{cat}) pair, we treat the corresponding bounding box of category cat as the question-relevant clues. We encode I into visual tokens $\mathbf{X}_v$ with the visual encoder and projector, and map the bounding box to its covered visual-token indices, denoted by $\mathbf{X}_{\text{bbox}} \subseteq \mathbf{X}_v$.

CLUERECALL. Collecting all such perception-labeled instances yields a set $\mathcal{X}_{\text{perc}} = \{(\mathbf{X}_{q_{\text{cat}}}, \mathbf{X}_v, \mathbf{X}_{\text{bbox}}, \text{cat})\}$. For each item in $\mathcal{X}_{\text{perc}}$, we prepend the system prompt $\mathbf{X}_s$ to form the complete input $\mathbf{X}_c = [\mathbf{X}_s, \mathbf{X}_v, \mathbf{X}_{q_{\text{cat}}}]$, feed it into the reasoning model $\mathcal{RM}$, and obtain output tokens $\{y_1, \dots, y_T\}$ and the attention tensor $\mathbf{A} \in [0, 1]^{T \times L \times N_c}$. Let $\mathcal{V}$ denote the visual-token indices within $\mathbf{X}_c$, and let $\mathbf{A}_{t,l,\mathcal{V}} \in [0, 1]^{N_v}$ be the restriction of $\mathbf{A}_{t,l}$ to these visual indices. We evaluate each layer l by how well its visual attention, at the exact output steps where the category token is mentioned, retrieves the ground-truth region $\mathbf{X}_{\text{bbox}}$. For each layer $l \in \{1, \dots, L\}$, we define

$$\text{CLUERECALL}(l) = \frac{1}{|\mathcal{X}_{\text{perc}}|} \sum_{\mathbf{X}_c \in \mathcal{X}_{\text{perc}}} \frac{|\text{TopK}_{|\mathbf{X}_{\text{bbox}}|} \left(\sum_{t: y_t = \text{cat}} \mathbf{A}_{t,l,\mathcal{V}} \right) \cap \mathbf{X}_{\text{bbox}}|}{|\mathbf{X}_{\text{bbox}}|}. \quad (1)$$

which directly measures a layer’s ability to retrieve object-aligned patches when the object is explicitly mentioned in the model’s output. Accordingly, we set $L_{\text{max}} \triangleq \arg \max_{l \in \{1, \dots, L\}} \text{CLUERECALL}(l)$ and use L_{max} as the reference layer for all subsequent clue-tracing computations.

5.2 CLUETRACER: Task-Critical Token Identification

Building on CLUERECALL, we have identified the layer with the strongest *clue retrieval*, denoted by L_{max} . In CLUETRACER, we move to a more general setting: we ask *how to recover task-critical visual regions when neither bbox annotations nor fixed question templates are available?* To this end, we leverage the observations in Sec. 4 and use the

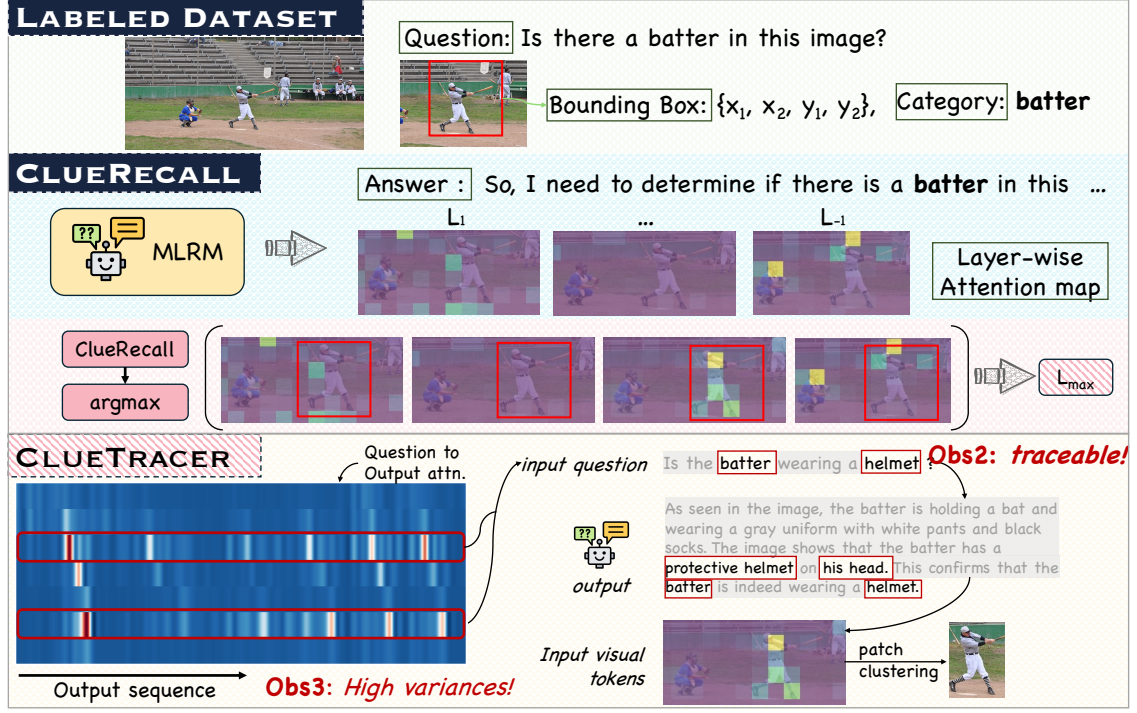


Figure 3: Overview of our method. **Top**: CLUERECALL uses bbox/category annotations from COCO or lightweight detectors, enabling extension to arbitrary datasets. **Middle**: layer-wise CLUERECALL identifies $L_{\max}$, the layer with strongest visual clue retrieval. **Bottom**: CLUETRACER traces the question $\rightarrow$ output $\rightarrow$ vision pathway to localize task-relevant regions. Best viewed in color.

pathway from question $\rightarrow$ output $\rightarrow$ visual tokens to progressively mine task-relevant clues.

Key query tokens via output-axis variability. In *Obs 3*, we find that when the model’s output is tightly aligned with the question, the decoder repeatedly allocates attention to a small set of key query tokens. As a result, the attention trajectory of these key tokens along the output axis becomes highly non-uniform, exhibiting larger variance. We therefore collect query-side attention at layer $L_{\max}$ across the entire output axis and retain query tokens whose standardized variability is high. Let $\mathcal{Q} \subseteq \{1, \dots, N_c\}$ denote the index set of query tokens within $\mathbf{X}_c$. For each query index $n_q \in \mathcal{Q}$ and each decoding step $t \in \{1, \dots, T\}$, we write

$$A_{t,n_q} = A_{t,L_{\max},n_q}, \mathbf{a}_{n_q} = (A_{1,n_q}, \dots, A_{T,n_q}) \in \mathbb{R}^T \quad (2)$$

where $\mathbf{a}_{n_q}$ is the attention trajectory of the n_q -th query token x_{n_q} over the output steps. We compute the variance $\text{Var}(\mathbf{a}_{n_q})$ for each $n_q \in \mathcal{Q}$, and select key query tokens as

$$\mathcal{X}_q^* = \{x_{n_q} \in \mathbf{X}_q : \text{zscore}(\text{Var}(\mathbf{a}_{n_q})) \geq \tau_q\}, \quad (3)$$

where $\text{zscore}(\cdot)$ denotes the standard score obtained by subtracting the mean and dividing by

the standard deviation (Walpole et al., 2011). The indices in $\mathcal{X}_q^*$ correspond to query tokens whose attention trajectories exhibit high output-axis variability and are thus treated as key query clues.

Tracing visual clues via Query-Output-Vision propagation. Having identified the key query clues $\mathcal{X}_q^*$, we seek the *task-supporting* visual evidence: a subset of visual tokens $\mathcal{X}_v^* \subset \mathbf{X}_v$ that best *grounds* these query constraints and thus suffices to solve the task. We formalize this as selecting $\mathcal{X}_v^*$ to maximize the query-conditioned support $P(\mathcal{X}_v^* | \mathcal{X}_q^*)$. However, under autoregressive (causally-masked) decoding, this dependency is not directly observable and is effectively mediated by the intermediate outputs, motivating a posterior signal. To bridge this, CLUETRACER leverages the model’s intermediate reasoning chain $\mathbf{y}_{1:T}$ as latent variables. We formulate the relevance of a visual token x_v to the query constraints as a probabilistic trace, marginalizing over the output steps:

$$P(x_v | \mathcal{X}_q^*) = \sum_{t=1}^T \underbrace{P(x_v | y_t)}_{\text{Visual Grounding}} \cdot \underbrace{P(y_t | \mathcal{X}_q^*)}_{\text{Query Alignment}}. \quad (4)$$

In this formulation, $P(y_t | \mathcal{X}_q^*)$ represents how strongly step t aligns with the query constraints,

Table 1: Comparison on reasoning-oriented hallucination benchmarks with and without CLUETracer. qAcc, fAcc, hAcc, and aAcc denote, respectively, the accuracy *per question pair*, *per figure*, *on hard questions*, and the *overall average*. Case studies are provided in Fig. 9-12.

Model	HallusionBench (Accuracy %)				VMCBench (Accuracy %)						Average Δ Acc
	qAcc	fAcc	hAcc	aAcc	Gen.	Reason.	OCR	Math	Doc&Chart	Overall	
R1-OneVision	8.57	12.43	31.86	35.43	56.64	45.14	65.39	32.44	52.35	52.99	+22.20%
+CLUETracer	18.18	30.79	41.16	49.16	76.11	61.81	89.97	51.21	78.88	73.91	
Ocean-R1	19.34	27.76	40.00	48.89	77.74	61.69	88.91	46.25	76.58	73.75	+10.60%
+CLUETracer	30.11	43.35	54.88	63.51	80.31	62.72	96.24	52.00	81.07	77.02	
MM-Eureka	23.07	31.50	49.30	58.46	76.31	58.79	90.46	42.44	76.21	72.44	+7.26%
+CLUETracer	33.41	47.11	56.98	65.10	83.56	67.26	94.05	58.38	85.93	80.31	
ORSTA-R1	21.76	28.61	45.81	55.45	71.55	55.62	85.06	42.82	72.60	68.39	+7.30%
+CLUETracer	27.91	36.71	53.49	60.05	81.52	64.21	94.41	51.11	84.63	78.39	

and $P(x_v | y_t)$ represents the visual support required by that step. In practice, we instantiate this probabilistic flow using the attention map at layer $L_{\max}$, which CLUERecall identifies as the most reliable layer for *clue retrieval*. We define the *Trace Score* $\mathcal{T}_{\mathcal{RM}}(x_v)$ under the reasoning model $\mathcal{RM}$ by instantiating the probabilities with normalized attention weights:

$$\begin{aligned}
 P(x_v | \mathcal{X}_q^*) &\propto \mathcal{T}_{\mathcal{RM}}(x_v) \\
 &= \sum_{t=1}^T \underbrace{\left(\sum_{x_{nq} \in \mathcal{X}_q^*} A_{t, L_{\max}, nq} \right)}_{\text{Query Alignment}} \cdot \underbrace{A_{t, L_{\max}, v}}_{\text{Vis. Grounding}}.
 \end{aligned} \quad (5)$$

Eq. (5) effectively creates a continuous flow of importance from the query, through the reasoning bottleneck, to the image. High scores are assigned to visual tokens that are repeatedly attended to by reasoning steps that are themselves highly sensitive to the question. Finally, to suppress background noise, we apply a statistical filtering function $\psi(\cdot)$ (implemented via *z*-score normalization (Walpole et al., 2011)) to select the final task-critical set:

$$\mathcal{X}_v^* = \{x_v \in \mathbf{X}_v \mid \psi(\mathcal{T}(x_v)) \geq \tau_v\}. \quad (6)$$

This traced set $\mathcal{X}_v^*$ represents the minimal visual clues sufficient to support the query constraints, derived not from heuristics but from the model’s dynamics.

5.3 Evidence-Region Construction and Inference-Time Use

Organization. Starting from the selected visual tokens $\mathcal{X}_v^*$, we map each token $x_v^* \in \mathcal{X}_v^*$ to its image-plane center $\phi(x_v^*) \in \mathbb{N}^2$, cluster the point set

$\{\phi(x_v^*)\}_{x_v^* \in \mathcal{X}_v^*}$ with DBSCAN (Ester et al., 1996) to obtain clusters $\{C_r\}_{r=1}^R$, and enclose each cluster by an axis-aligned rectangle $R_r = \text{BBox}(C_r)$. The resulting region set $\mathcal{R} = \{R_1, \dots, R_R\}$ defines crops of the original image. This construction preserves the model’s native rectangular interface while consolidating fragmented evidence into a small number of spatially coherent *visual clues*.

Input modes. We use these crops in two complementary ways. *Offline*: precompute and cache $\mathcal{R}$ for repeated evaluation or families of similar images. *Online*: a two-stage inference pipeline—**Stage 1** derives $\mathcal{R}$ from the model; **Stage 2** re-invokes the model on $(I; \mathcal{R})$ to refine reasoning with focused evidence.

6 Experiments

In this section, we conduct experiments to address the following research questions:

- **RQ1.** How well does CLUETracer reduce hallucinations and improve accuracy across diverse reasoning architectures? Which layer achieves the highest CLUERecall?
- **RQ2.** During generation, can CLUETracer mitigate reasoning drift exhibited in Fig. 1, maintaining alignment between the question focus and produced content?
- **RQ3.** How do results change when key settings varied, and specifically, does CLUETracer’s output-mediated patch selection outperform baselines designed for non-reasoning MLLMs?

6.1 Experimental Setup

We begin with a concise overview of models, datasets, and evaluation metrics; full implementation details are deferred to the Appendix C.

Table 2: Per-layer CLUERECALL (left) and accuracy on POPE/MMVP with and without CLUETRACER (right).

Model	CLUERECALL per Layer (%)						POPE (%)		MMVP (%)	
	0	6	12	18	24	27	w/o CT	w/ CT	w/o CT	w/ CT
R1-OneVision	30.78%	33.68%	43.11%	50.60%	47.28%	44.11%	70.22%	76.97%	23.33%	46.00%
Ocean-R1	31.05%	32.64%	48.56%	53.79%	50.92%	49.23%	86.77%	85.77%	47.33%	50.00%
MM-Eureka	29.30%	30.57%	45.45%	51.06%	51.38%	48.11%	75.00%	81.38%	44.67%	50.00%
ORSTA-R1	31.14%	32.38%	47.55%	51.66%	55.32%	50.56%	71.36%	82.98%	38.00%	53.33%

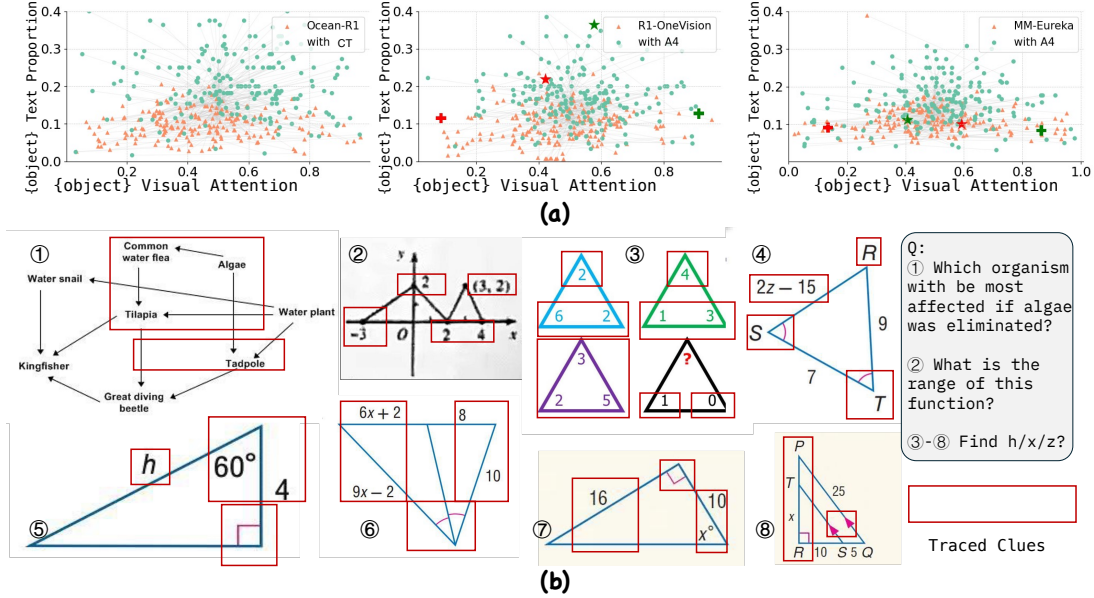


Figure 4: Visualization of reduced reasoning drift and traced visual clues. (a) Comparative scatter plot of attention allocation versus output mentions for task-relevant objects. (b) Case studies showing that CLUETRACER traces implicit visual clues required by the question, with red boxes highlighting the discovered conditions in the image.

Models. CLUETRACER is compatible with existing hallucination-mitigation schemes for ML-RMs. Accordingly, we apply it to several reasoning models—R1-OneVision (Yang et al., 2025), Ocean-R1 (Ming et al., 2025), Orsta-R1 (Ma et al., 2025b), and MM-Eureka (Meng et al., 2025)—to assess generality. To test transferability, we also port CLUETRACER to *non-reasoning* MLLMs, including LLaVA-1.6 (Liu et al., 2024) and Qwen2.5-VL (Bai et al., 2025), as discussed in Appendix D.2.

Datasets & Evaluation Metrics. We evaluate on four widely used hallucination benchmarks, grouped as follows: *reasoning-oriented*: 1) **VM-CBench** (Zhang et al., 2025d), a unified multiple-choice questions drawn from 20 VQA datasets, e.g., *MathVision* (Wang et al., 2024a), *ScienceQA* (Lu et al., 2022), and 2) **HallusionBench** (Guan et al., 2024) benchmarks image-context reasoning where language hallucinations and visual illusions are entangled; and *perception-oriented*: 3) **POPE** (Li et al., 2023) and 4) **MMVP** (Tong et al., 2024). We

follow each dataset’s official evaluation protocol.

6.2 Performance with CLUETRACER on Reasoning Models (RQ1)

To assess hallucination in *reasoning* models, we evaluate each model *before* and *after* applying CLUETRACER on hallucination benchmarks; the results are reported in Table 1 and Table 2. Additional results on **other variants** and **larger models** are provided in Appendix D. We can draw the following findings:

Finding 1: Consistent gains across models and datasets. CLUETRACER improves *all* reasoning models across the four hallucination benchmark. Specifically, on the reasoning-oriented hallucination datasets—**HallusionBench** and **VM-CBench**—CLUETRACER achieves a **maximum** accuracy gain of **22.20%** and an **average** accuracy gain of **11.84%**.

Finding 2: Models with different training architectures exhibit similar perceptual capacity. From Table 2, the CLUERECALL curves for all models peak around **50%–75%** of the network

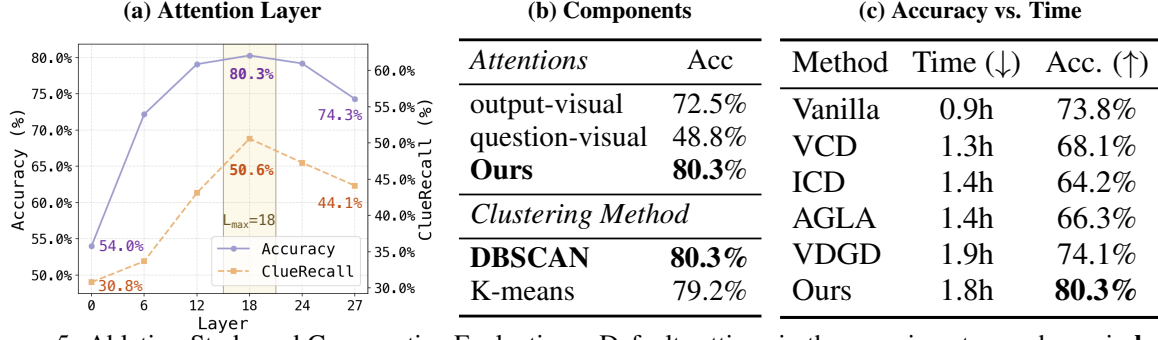


Figure 5: Ablation Study and Comparative Evaluations. Default settings in the experiments are shown in **bold**. Additional ablations on hyperparameters (e.g., DBSCAN and z-score thresholds) are provided in the Appendix F.

depth and follow a rise-then-fall pattern. Appendix F further shows that this phenomenon remains consistent across a broader range of architectures and model scales.

6.3 Hallucination Analysis (RQ2)

To further verify the hallucination-suppression ability of CLUETRACER, we compute two key indicators on the POPE dataset: (1) *{object} visual attention*, i.e., the attention mass that falls inside the *{object}* bounding box; and (2) *{object} text proportion*, i.e., the fraction of noun tokens in the model output that correspond to *{object}*. We further provide MathVista (Lu et al., 2023) examples in Fig. 4 (b), where many decisive clues, such as parallelism and perpendicularity, must be grounded in the image. Our key findings are as follows:

Finding 3: CLUETRACER mitigates reasoning drift. As shown in Fig. 4 (a), after models are equipped with CLUETRACER, the *{object} text proportion* in the generated text increases markedly across all models. This indicates that the reasoning process pays more attention to task-critical content, thereby reducing reasoning drift. Fig. 4 (b) further provides MathVista examples, showing that CLUETRACER can localize implicit visual clues required by the question, such as *parallelism*, *perpendicularity*, and *equal angles*, even when the question is as simple as “find x .”

6.4 Ablation and Comparison (RQ3)

As discussed in pervious sections, CLUETRACER effectively identifies task-relevant visual evidence. In this section, we examine the *module sensitivity* of CLUETRACER and explain why analogous approaches developed for *non-reasoning* MLLMs do not carry over to *reasoning* models. We conduct ablation studies and comparative evaluations on Ocean-R1 using VMCBench. For the comparative setting, we benchmark against prior contrastive-

decoding and description-prior methods:

VCD (Leng et al., 2024) contrasts decoding distributions from the original image and a corrupted view; **ICD** (Wang et al., 2024b) contrasts standard vs. disturbance-instruction decoding to subtract hallucinated concepts; **AGLA** (An et al., 2025b) assembles global generation with prompt-relevant local cues via an augmented (masked) view; **VDGD** (Ghosh et al., 2024) grounds decoding with a self-generated visual description prefix to better support reasoning. The results are summarized in Fig. 5. Our findings are:

Finding 4: Performance trends align with expectations. Based on Figure 5 (a–b), the accuracy trend over attention layers mirrors the CLUECALL pattern in Table 2. For **clustering**, DBSCAN is preferable to K-means, as it is less sensitive to noise and can naturally discard isolated noisy tokens when the attention map contains spurious activations; see Appendix D.4 for a detailed discussion. **Finding 5: CLUETRACER outperforms prior inference-time baselines on reasoning settings.** As shown in Fig. 5 (c), CLUETRACER consistently improves over baselines originally developed for non-reasoning settings, which are less effective under the long-context dynamics of reasoning models. In particular, contrastive-decoding methods can suffer from *bias accumulation* over long chains, causing decoding to drift and even reducing accuracy below the vanilla model, while description-prior approaches often fail to provide task-specific priors that are actually useful for solving the question.

7 Conclusion

In this paper, we introduce CLUETRACER, a training-free approach for suppressing hallucination in MLRLMs. CLUETRACER traces the question→output→vision attention pathway to localize task-relevant visual patches and guide inference toward decisive evidence.

Limitations

While CLUETRACER efficiently mitigates hallucination in MLRMs, several limitations merit acknowledgment and point to concrete remedies. **(i) Two-stage inference.** Owing to current reasoning-model architectures, CLUETRACER operates in two stages (attention mining then focused re-inference). A learned controller that adaptively decides *when* to trigger CLUETRACER and *how much* context to generate (e.g., output length for Stage 1) could reduce overhead and retain quality in a single pass. **(ii) Rectangular crops.** Our evidence regions are axis-aligned; objects, however, are often non-rectangular, so rectangular cropping may include extraneous background.

Future work. **(i) Single-pass integration.** Post-train a lightweight policy to fuse CLUETRACER into one forward pass, enabling adaptive, example-dependent activation and scope (triggering and output length) within the model’s native decoding; reinforcement learning or adaptive-compute objectives are natural fits. **(ii) Beyond rectangles.** Explore mask- or token-level selection: derive instance/part masks with promptable segmentation to form shape-accurate evidence, or re-inject only selected visual tokens (retaining their original positional encodings) while dropping or down-weighting others, thus avoiding commitment to a particular crop geometry. **(iii) Compute-network-aware deployment.** Given its training-free and plug-and-play nature, CLUETRACER provides a promising direction for resource-constrained or collaborative edge-cloud settings (Cheng et al., 2023). Specifically, Stage 1 could be replaced by a lightweight clue-tracing model deployed on edge nodes, which efficiently identifies task-relevant visual evidence and forwards the extracted clues to a cloud-based large model for reasoning. Such a design can enable flexible compute-network resource orchestration across heterogeneous infrastructures, effectively reducing inference costs while preserving or even enhancing reasoning accuracy in distributed deployment scenarios.

ACKNOWLEDGMENTS

This work was supported in part by the National Key Research and Development Program of China under Grant 2024YFB2906705 and the National Natural Science Foundation of China under Grant 62676038.

References

- Wenbin An, Jiahao Nie, Feng Tian, Haonan Lin, Yaqiang Wu, QianYing Wang, Xiaoqin Zhang, Shijian Lu, and 1 others. 2025a. Boosting knowledge utilization in multimodal large language models via adaptive logits fusion and attention reallocation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. 2025b. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29915–29926.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuezhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Zeichen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.
- Feiqi Cao, Soyeon Caren Han, Siqu Long, Changwei Xu, and Josiah Poon. 2022. [Understanding attention for vision-and-language tasks](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3438–3453, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.
- Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. 2025. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. *arXiv preprint arXiv:2503.01773*.
- Jiujun Cheng, Huiyu Sun, Zhangkai Ni, and Aiguo Zhou. 2023. A dynamic evolution model for decentralized autonomous car clusters in a highway scene. *IEEE Transactions on Computational Social Systems*, 11(3):3792–3802.
- Bowen Dong, Minheng Ni, Zitong Huang, Guanglei Yang, Wangmeng Zuo, and Lei Zhang. 2025a. Mirage: Assessing hallucination in multimodal reasoning chains of mllm. *arXiv preprint arXiv:2505.24238*.
- Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. 2025b.

- Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9062–9072.
- Chengqi Duan, Rongyao Fang, Yuqing Wang, Kun Wang, Linjiang Huang, Xingyu Zeng, Hongsheng Li, and Xihui Liu. 2025. Got-r1: Unleashing reasoning capability of mllm for visual generation with reinforcement learning. In *arXiv preprint arXiv:2505.17022*.
- Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. [A density-based algorithm for discovering clusters in large spatial databases with noise](#). In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD’96)*, pages 226–231, Portland, OR, USA. AAAI Press.
- Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. 2025. Grit: Teaching mllms to think with images. *arXiv preprint arXiv:2505.15879*.
- Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Utkarsh Tyagi, Oriol Nieto, Zeyu Jin, and Dinesh Manocha. 2024. Visual description grounding reduces hallucinations and boosts reasoning in lvlms. *arXiv preprint arXiv:2405.15683*.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. [DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning](#). *Nature*, 645(8081):633–638.
- Danna Gurari, Qing Li, Andrew J. Stangl, Jianlong Guo, Airi Lin, and Kristen Grauman. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3608–3617.
- Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.
- Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.
- Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6700–6709.
- Nikolai Ilinykh and Simon Dobnik. 2021. [What does a language-and-vision transformer see: The impact of semantic information on visual representations](#). *Frontiers in Artificial Intelligence*, Volume 4 - 2021.
- Nikolai Ilinykh and Simon Dobnik. 2022. [Attention as grounding: Exploring textual and cross-modal attention on entities and relations in language-and-vision transformer](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4062–4073. Association for Computational Linguistics.
- Eun Woo Im, Muhammad Kashif Ali, and Vivek Gupta. 2025. Self-augmented visual contrastive decoding. *arXiv preprint arXiv:2510.13315*.
- Chaoya Jiang, Yongrui Heng, Wei Ye, Han Yang, Haiyang Xu, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. 2025. Vlm-r³: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. *arXiv preprint arXiv:2505.16192*.
- Simeon Junker and Sina Zarieß. 2023. Keeping an eye on context: Attention allocation over input partitions in referring expression generation. In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 20–27.
- Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. See what you are told: Visual attention sink in large multimodal models. *arXiv preprint arXiv:2503.03321*.
- Aniruddha Kembhavi, Michael Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. [A diagram is worth a dozen images](#). In *European Conference on Computer Vision (ECCV)*.
- Junho Kim, Hyunjun Kim, Kim Yeonju, and Yong Man Ro. 2024a. Code: Contrasting self-generated description to combat hallucination in large multi-modal models. *Advances in Neural Information Processing Systems*, 37:133571–133599.
- Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. 2024b. Tablevqa-bench: A visual question answering benchmark on multiple table domains. *arXiv preprint arXiv:2404.19205*.

- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.
- Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. 2024. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. [Evaluating object hallucination in large vision-language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.
- Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhen-ting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N Metaxas. 2025. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. *arXiv preprint arXiv:2502.03628*.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. [Microsoft coco: Common objects in context](#). In *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755, Cham. Springer.
- Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. 2025a. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *arXiv preprint arXiv:2505.21523*.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. [Improved baselines with visual instruction tuning](#). *Preprint*, arXiv:2310.03744.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. [Llava-next: Improved reasoning, ocr, and world knowledge](#).
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. [Visual instruction tuning](#). *arXiv preprint arXiv:2304.08485*.
- Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. 2025b. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. In *arXiv preprint arXiv:2503.06520*.
- Haolang Lu, Yilian Liu, Jingxin Xu, Guoshun Nan, Yuanlong Yu, Zhican Chen, and Kun Wang. 2025. Auditing meta-cognitive hallucinations in reasoning large language models. *arXiv preprint arXiv:2505.13143*.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.
- Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Taffjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *arXiv preprint arXiv:2209.09513*.
- Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 2025a. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6924–6934.
- Yan Ma, Linge Du, Xuyang Shen, Shaoxiang Chen, Pengfei Li, Qibing Ren, Lizhuang Ma, Yuchao Dai, Pengfei Liu, and Junjie Yan. 2025b. One rl to see them all: Visual triple unified reinforcement learning. *arXiv preprint arXiv:2505.18129*.
- Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. [Chartqa: A benchmark for question answering about charts with visual and logical reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2022*.
- Minesh Mathew, Viraj Bagal, Rubén Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2582–2591.
- Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.
- Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. 2025. [Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning](#). *Preprint*, arXiv:2503.07365.
- Filippo Merlo, Ece Takmaz, Wenkai Chen, and Albert Gatt. 2026. Common objects out of context (cooco): Investigating multimodal context and semantic scene violations in referential communication. *Transactions of the Association for Computational Linguistics*, 14:1163–1185.

- Lingfeng Ming, Yadong Li, Song Chen, Jianhua Xu, Zenan Zhou, and Weipeng Chen. 2025. Ocean-r1: An open and generalizable large vision-language model enhanced by reinforcement learning. <https://github.com/VLM-RL/Ocean-R1>. Accessed: 2025-04-03.
- Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *International Conference on Document Analysis and Recognition (ICDAR)*.
- Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.
- Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*.
- Anand Mishra Singh, Ali Furkan Biten, Minesh Mathew, Devi Parikh, and 1 others. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. 2025. Reason-rft: Reinforcement fine-tuning for visual reasoning. In *arXiv preprint arXiv:2503.20752*.
- Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, and 1 others. 2025. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*.
- Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *Preprint*, arXiv:2401.06209.
- Chongjun Tu, Peng Ye, Dongzhan Zhou, Lei Bai, Gang Yu, Tao Chen, and Wanli Ouyang. 2026. Attention reallocation: Towards zero-cost and controllable hallucination mitigation of mllms. *International Journal of Computer Vision*, 134(1):22.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.
- Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, and Keying Ye. 2011. *Probability and Statistics for Engineers and Scientists*, 9th edition. Pearson.
- Jiaqi Wang, Kevin Qinghong Lin, James Cheng, and Mike Zheng Shou. 2025a. Think or not? selective reasoning via reinforcement learning for vision-language models. *arXiv preprint arXiv:2505.16854*.
- Ke Wang, Juntong Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with MATH-Vision dataset. *arXiv preprint arXiv:2402.14804*.
- Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, Lewei Lu, Haodong Duan, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025b. Visualprm: An effective process reward model for multimodal reasoning. In *arXiv preprint arXiv:2503.10291*.
- Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024b. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*.
- Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. 2025c. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. In *arXiv preprint arXiv:2504.07934*.
- Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. 2025d. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*.
- Qianshan Wei, Tengchao Yang, Yaochen Wang, Xinfeng Li, Lijun Li, Zhenfei Yin, Yi Zhan, Thorsten Holz, Zhiqiang Lin, and XiaoFeng Wang. 2025. A-memguard: A proactive defense framework for llm-based agent memory. *arXiv preprint arXiv:2510.02373*.
- Jiulong Wu, Zhengliang Shi, Shuaiqiang Wang, Jizhou Huang, Dawei Yin, Lingyong Yan, Min Cao, and Min Zhang. 2025. Mitigating hallucinations in large vision-language models via entity-centric multimodal preference optimization. *arXiv preprint arXiv:2506.04039*.
- xAI. 2024. Realworldqa. <https://huggingface.co/datasets/xai-org/RealworldQA>. Benchmark dataset for real-world multimodal understanding.
- Haotian Xu, Xing Wu, Weinong Wang, Zhongzhi Li, Da Zheng, Boyuan Chen, Yi Hu, Shijia Kang, Jiaming Ji, Yingying Zhang, and 1 others. 2025. Redstar: Does scaling long-cot data unlock better slow-reasoning systems? *arXiv preprint arXiv:2501.11284*.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

- Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. 2025. [R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization](#). *Preprint*, arXiv:2503.10615.
- Huanjin Yao, Jiaying Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, and Dacheng Tao. 2024. Mulberry: Empowering vision-language models with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*.
- Zekai Ye, Qiming Li, Xiaocheng Feng, Libo Qin, Yichong Huang, Baohang Li, Kui Jiang, Yang Xiang, Zhirui Zhang, Yunfei Lu, and 1 others. 2025. Claim: Mitigating multilingual object hallucination in large vision-language models with cross-lingual attention intervention. *arXiv preprint arXiv:2506.11073*.
- Shixin Yi and Lin Shang. 2025. Corgi: Verified chain-of-thought reasoning with visual grounding. *arXiv preprint arXiv:2508.00378*.
- Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and 1 others. 2024. RLhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.
- Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. MM-Vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.
- Xingyu Yue, Gaole Qu, Xiuyu Chen, and 1 others. 2023. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. *arXiv preprint arXiv:2311.16502*.
- Di Zhang, Jianbo Wu, Jingdi Lei, Tong Che, Jia-tong Li, Tong Xie, Xiaoshui Huang, Shufei Zhang, Marco Pavone, Yuqiang Li, and 1 others. 2024a. Llama-berry: Pairwise optimization for o1-like olympiad-level mathematical reasoning. *arXiv preprint arXiv:2410.02884*.
- Hongxuan Zhang, Zhining Liu, Yao Zhao, Jiaqi Zheng, Chenyi Zhuang, Jinjie Gu, and Guihai Chen. 2024b. Fast chain-of-thought: A glimpse of future from jacobi decoding leads to answers faster. In *arXiv preprint arXiv:2311.08263*.
- Jingyi Zhang, Jiaying Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025a. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. In *arXiv preprint arXiv:2503.12937*.
- Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2024c. [Lmms-eval: Reality check on the evaluation of large multimodal models](#). *Preprint*, arXiv:2407.12772.
- Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2024d. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*.
- Xiaofeng Zhang, Yihao Quan, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. 2025b. [From redundancy to relevance: Information flow in LVLMS across reasoning tasks](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2289–2299, Albuquerque, New Mexico. Association for Computational Linguistics.
- Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and 1 others. 2025c. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl. *arXiv preprint arXiv:2505.15436*.
- Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Aklilu, Alejandro Lozano, Anjiang Wei, Ludwig Schmidt, and Serena Yeung-Levy. 2025d. Automated generation of challenging multiple-choice questions for vision language model evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29580–29590.
- Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. [Multi-modal chain-of-thought reasoning in language models](#). *arXiv preprint arXiv:2302.00923*.
- Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. Marco-o1: Towards open reasoning models for open-ended solutions. *arXiv preprint arXiv:2411.14405*.
- Ge Zheng, Jiaye Qian, Jiajin Tang, and Sibe Yang. 2025. Why lvlms are more prone to hallucinations in longer responses: The role of context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4101–4113.
- Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. 2024. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. *arXiv preprint arXiv:2410.04780*.
- Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Ken-ting Zheng, Sirui Huang, Junkai Chen, Peijie Jiang,

Jia Liu, Chang Tang, and 1 others. 2024. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*.

A LLM Usage

We used a large language model [ChatGPT 5.4] strictly for writing assistance limited to spelling, grammar, and minor stylistic polishing.

B Notations

The notation used in this paper is summarized in Table 3.

C Experiment Details

C.1 Baselines

Models. Because CLUETRACER is training-free and architecture-agnostic, we integrate it at inference into four recent multimodal reasoning models without altering their weights or decoding. (i) **R1-OneVision** (Yang et al., 2025)—a cross-modal formalization pipeline that converts visual content into structured textual representations before language reasoning, trained with SFT and RL. (ii) **Ocean-R1** (Ming et al., 2025)—an open RL-enhanced VLM whose public artifacts indicate a two-stage recipe (reasoning first, then visual perception) with released weights/data. (iii) **Orsta** (Ma et al., 2025b)—a family of VLMs post-trained via the *V-Triune* unified RL system to improve both perception and reasoning. (iv) **MM-Eureka** (Meng et al., 2025)—a rule-based (R1-style) reinforcement learning approach extended to multimodal reasoning with an open pipeline. To assess transferability, we further deploy the CLUETRACER-derived visual cropping prior on two *non-reasoning* MLLMs: (v) **Qwen2.5-VL** (Bai et al., 2025)—a general-purpose vision-language model series (3B/7B/72B) with an efficient ViT and official instruction-tuned checkpoints; and (vi) **LLaVA-1.6-Mistral** (Liu et al., 2024)—an instruction-tuned VLM pairing a vision encoder with Mistral-7B, documented with improved training data and dynamic high-resolution support.

Datasets. We evaluate hallucination on four benchmarks. **POPE** (Li et al., 2023) targets object-level hallucination via yes/no “object presence” queries and contains three standard splits (popular/adversarial/random). **VMC-Bench** (Zhang et al., 2025d) unifies 20 VQA datasets into a single multiple-choice suite (9,018 questions) spanning diverse domains. **MMVP** (Tong et al., 2024) focuses on nine basic visual patterns and “CLIP-blind pairs,” probing failure modes that induce incorrect answers or overconfident, hallucinatory ex-

planations. **HallusionBench** (Guan et al., 2024) is an image-context reasoning benchmark diagnosing entangled language hallucination and visual illusion, consisting of 346 images and 1,129 expert-crafted questions with control structures for analysis. For **POPE** we report accuracy using the *lmms-eval* (Zhang et al., 2024c) implementation. For **VMC-Bench**, **MMVP**, and **HallusionBench** we follow each benchmark’s official evaluation pipeline.

License. All models evaluated in this work are publicly available open-weight models. Specifically, R1-OneVision, MM-Eureka, LLaVA-1.6, and Qwen2.5-VL are released under the Apache-2.0 license, while Orsta-R1 is released under the MIT license. Ocean-R1 is publicly released through its official Hugging Face and GitHub repositories, but we did not identify a separate license statement for its model weights. We use all models only for academic evaluation and do not redistribute model weights.

C.2 Evaluation setup details

Prompt templates. For *yes/no* questions (POPE, MMVP, HallusionBench) and *multiple choice* questions (VMCBench), we use the following templates:

You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think></think> tags. The final answer MUST BE a single word enclosed in <answer></answer> tags. Let’s think more.
{question}

You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think></think> tags. The final answer must be an option letter from the given choices, enclosed in <answer></answer> tags. Let’s think more.
{question}

Evaluation. During evaluation, we parse the model output with a regular expression to extract the content enclosed in <answer>...</answer> and compare it with the ground truth.

Table 3: **Notations used throughout CLUERECALL and CLUETRACER.**

Notation	Description
$\mathcal{RM}, L$	Multimodal reasoning model (MLRM) and its number of Transformer layers
I, q	Input image and textual que
cat, bbox	Category of an entity in I and its bounding box in the image plane
$\mathbf{X}_c, \mathbf{X}_s, \mathbf{X}_q, \mathbf{X}_v$	Complete input, system prompt tokens, query tokens, and visual tokens
N_c, N_s, N_q, N_v	Lengths of $\mathbf{X}_c, \mathbf{X}_s, \mathbf{X}_q$, and $\mathbf{X}_v$
$\mathbf{X}_{q_{\text{cat}}}, \mathbf{X}_{\text{bbox}}$	Query tokens for category cat ; visual tokens within the bounding box bbox
$\mathcal{X}_{\text{perc}}$	Labeled dataset of $(\mathbf{X}_{q_{\text{cat}}}, \mathbf{X}_v, \mathbf{X}_{\text{bbox}}, \text{cat})$ for computing CLUERECALL
$\mathbf{y}_{1:T}$	Model output token sequence $\{y_1, \dots, y_T\}$ of length T
$\mathbf{A}$	Attention tensor over inputs, $\mathbf{A} = (A_{t,l,n}) \in [0, 1]^{T \times L \times N_c}$
$\mathcal{X}_q^*, \mathcal{X}_v^*$	Key query tokens and task-critical visual tokens traced by CLUETRACER
$\mathcal{T}(x_v)$	Trace score representing the probability flow $P(x_v \mathcal{X}_q^*)$ via output outputs
$\psi(\cdot), \phi(\cdot)$	Filtering function (implemented as z -score); mapping from a region to its bounding box

Generate config. We use the same decoding settings across all datasets.

Parameter	Value
min_new_tokens	50
max_new_tokens	1024
do_sample	True

D More Experimental Results

D.1 More results on POPE and VMC-Bench

Evaluation Setup. We report detailed results across multiple benchmarks following each dataset’s official protocol. For POPE, we report accuracy and F1 on the *Random*, *Popular*, and *Adversarial* splits (Li et al., 2023). For VMCBENCH, we report per-subset and overall accuracy (Zhang et al., 2025d). For MMVP, we follow the authors’ evaluation on CLIP-blind visual patterns (Tong et al., 2024).

VMCBENCH Subsets. We adopt VMCBENCH’s four capability-oriented subsets and their official composition (Zhang et al., 2025d): **(1) General:** *VQAv2* (Goyal et al., 2017), *GQA* (Hudson and Manning, 2019), *VizWiz* (Gurari et al., 2018), *OK-VQA* (Marino et al., 2019), *A-OKVQA* (Schwenk et al., 2022); **(2) Reasoning:** *MMMU* (Yue et al., 2023), *MathVista* (Lu et al., 2023), *MathVision* (Wang et al., 2024a), *MMStar* (Chen et al., 2024); **(3) OCR:** *TextVQA* (Singh et al., 2019), *OCR-VQA* (Mishra et al., 2019); **(4) Doc & Chart:** *DocVQA* (Mathew et al., 2021), *TableVQA-Bench* (Kim et al., 2024b), *InfographicVQA* (InfoVQA) (Mathew et al., 2022),

AI2D (Kembhavi et al., 2016), *ChartQA* (Masry et al., 2022).

D.2 Transfer CLUETRACER to Non-Reasoning MLLMs

Reasoning subset members. *MMMU*: college-level, multi-discipline problems requiring deliberative reasoning (Yue et al., 2023). *MathVista*: mathematical reasoning in visual contexts (charts, plots, geometry, etc.) (Lu et al., 2023). *MathVision*: math problem solving on real-world photos and synthetic scenes (Wang et al., 2024a). *MMStar*: broad, multi-skill evaluation of multimodal reasoning (Chen et al., 2024).

General subset members. *VQAv2*: balanced open-ended VQA over everyday photos (Goyal et al., 2017). *GQA*: compositional visual reasoning with scene-graph grounding (Hudson and Manning, 2019). *VizWiz*: real-world VQA from blind photographers; noisy images and conversational questions (Gurari et al., 2018). *OK-VQA/A-OKVQA*: knowledge-based VQA requiring external/world knowledge (Marino et al., 2019; Schwenk et al., 2022).

OCR subset members. *TextVQA*: reading text in the wild for VQA (Singh et al., 2019). *OCR-VQA*: VQA by reading textual content in images (book covers, storefronts, etc.) (Mishra et al., 2019).

Doc & Chart subset members. *DocVQA*: document understanding for VQA (forms, invoices, pages) (Mathew et al., 2021). *TableVQA-Bench*: multi-domain table QA benchmark for visual tables (Kim et al., 2024b). *InfographicVQA*: QA

Table 4: Accuracy (%) on selected VMCBENCH. Including: MMMU, MathVista, MMStar, AI2D, ScienceQA, SEEDBench, MM-Vet, MathVision, TableVQABench (TVQA), RealWorldQA (RWQA). Rows are paired as baseline vs. +CLUETRACER for each model.

Model	MMMU	MathVista	MMStar	AI2D	ScienceQA	SEEDBench	MM-Vet	MathVision	TVQA	RWQA
R1-OneVision	36.30	36.63	40.38	46.70	61.54	54.81	48.12	28.99	48.31	46.79
+CLUETRACER (Ours)	50.49	67.00	55.96	68.06	82.35	72.35	74.29	36.41	73.79	54.62
Ocean-R1	56.28	60.89	58.81	74.94	80.09	73.58	73.85	32.35	69.02	60.09
+CLUETRACER (Ours)	55.42	69.50	62.53	79.68	84.39	80.99	79.44	36.97	79.09	66.67
MM-Eureka	51.20	58.42	56.06	71.07	79.64	72.84	73.38	27.42	68.02	54.59
+CLUETRACER (Ours)	60.58	74.75	63.66	80.41	83.48	77.78	85.61	41.78	80.41	56.42
ORSTA-R1	45.43	51.49	51.31	66.06	71.04	70.37	67.67	33.17	67.65	57.57
+CLUETRACER (Ours)	58.89	68.50	62.71	74.94	79.41	76.30	85.82	34.22	77.03	63.07

Table 5: Accuracy (%) on the remaining VMCBENCH datasets: TextVQA, InfoVQA, DocVQA, OCRVQA, VizWiz, ChartQA, GQA, A-OKVQA, OKVQA, VQAv2. Rows are paired as baseline vs. +CLUETRACER for each model.

Model	TextVQA	InfoVQA	DocVQA	OCRVQA	VizWiz	ChartQA	GQA	A-OKVQA	OKVQA	VQAv2
R1-OneVision	64.72	56.36	58.56	66.06	64.18	56.65	61.86	56.71	63.95	66.44
+CLUETRACER	90.56	77.14	94.44	89.38	84.42	82.11	80.93	80.24	81.98	84.03
Ocean-R1	91.01	71.19	92.00	87.31	86.86	75.92	80.68	81.65	86.67	83.56
+CLUETRACER	97.75	82.16	100.00	96.63	90.91	84.63	88.75	89.41	92.35	88.89
MM-Eureka	88.54	73.73	88.42	91.71	83.09	79.36	82.40	79.06	85.68	83.56
+CLUETRACER	94.83	82.75	96.88	93.26	90.44	88.99	87.04	86.82	93.33	87.96
ORSTA-R1	83.37	70.43	82.20	86.79	82.99	76.15	72.86	72.94	75.80	81.02
+CLUETRACER	95.28	86.71	97.97	93.52	89.18	86.70	81.66	80.71	88.89	87.27

over complex infographics requiring reading and reasoning (Mathew et al., 2022). *AI2D*: diagram understanding and diagram-based QA (Kembhavi et al., 2016). *ChartQA*: QA over charts requiring numerical and logical reasoning (Masry et al., 2022).

Additional sets in VMCBENCH. *MM-Vet*: comprehensive capability evaluation of MLLMs, frequently used as a generalization stress test (Yu et al., 2023). *SEED-Bench*: multi-domain, fine-grained evaluation for perception and reasoning (Li et al., 2024). *RealWorldQA*: real-world spatial understanding; images largely from vehicles with verifiable Q&A (xAI, 2024).

Thinking length. We measure the average output token length on the “general” and “reasoning” subsets of VMC-Bench. With CLUETRACER, the average output length on the reasoning subset increases from 197.22 to 302.88 tokens (a 53.57% increase), while on the general subset it only increases from 134.43 to 136.10 tokens (a 1.24% increase). This further supports that CLUETRACER reduces the perceptual burden and lets the model use more capacity for reasoning.

Transfer to Non-Reasoning MLLMs. To verify the transferability of CLUETRACER, we evaluate MMVP on *non-reasoning* MLLMs. Concretely, we adopt the *offline* setting: attention-guided crops produced by Ocean-R1 are fed to Qwen2.5-VL and LLaVA-1.6-Mistral. Results are summarized in Fig. 6. Our findings are: MMVP is designed to probe visual defects in multimodal models. As shown in Fig. 6, CLUETRACER consistently improves performance; in particular, it moves several models—LLaVA-1.6, R1-OneVision, and ORSTA-R1—from near random guessing toward the GPT-4V level. In the qualitative examples of Fig. 6, without CLUETRACER the models barely attend to the key region of the question, whereas with CLUETRACER they shift focus to that region.

D.3 Results on RWQA and 3DSRBench

We evaluate CLUETRACER on RWQA (xAI, 2024) and 3DSRBench (Ma et al., 2025a) using three base models. As shown in Table 7, CLUETRACER consistently improves accuracy on both datasets for all backbones.

Table 6: POPE: accuracy (%) and F1 by split (Adversarial / Popular / Random / Average).

Model	Adversarial		Popular		Random		Average	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
R1-OneVision	69.20%	0.6629	70.33%	0.6711	71.12%	0.6769	70.22%	0.6703
+CLUETRACER	80.90%	0.7731	81.60%	0.7825	82.43%	0.7881	81.64%	0.7805
Ocean-R1	85.53%	0.8481	86.80%	0.8592	87.97%	0.8707	86.77%	0.8593
+CLUETRACER	84.43%	0.8271	86.13%	0.8458	86.70%	0.8496	85.76%	0.8408
MM-Eureka	74.17%	0.7107	74.21%	0.7108	76.63%	0.7308	75.00%	0.7174
+CLUETRACER	80.73%	0.7726	81.17%	0.7761	82.25%	0.7860	81.38%	0.7782
ORSTA-R1	68.31%	0.6445	72.00%	0.6754	73.77%	0.6866	71.36%	0.6688
+CLUETRACER	81.86%	0.7895	83.39%	0.8061	83.67%	0.8071	82.98%	0.8009

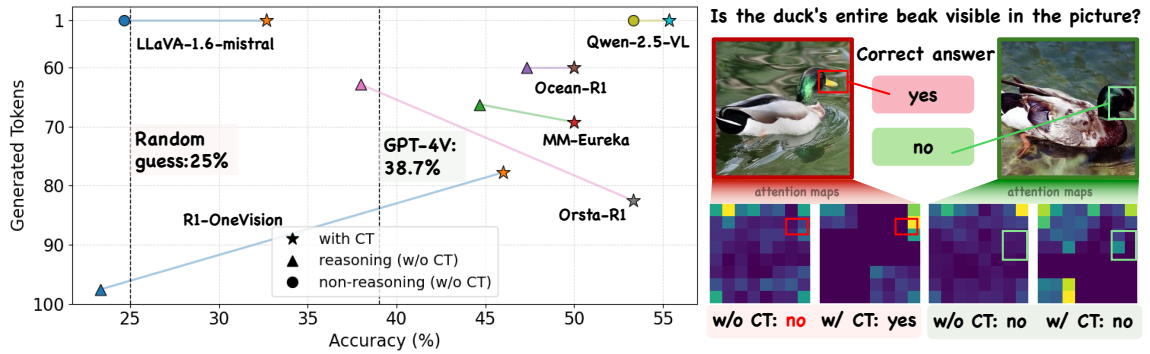


Figure 6: **Left:** Accuracy versus output-token length on MMVP across models; **Right:** attention allocation with and without CLUETRACER.

Table 7: Results on RWQA and 3DSRBench with and without CLUETRACER.

Dataset	Setting	Ocean-R1	ORSTA	MM-Eureka
RWQA	w/o CLUETRACER	63.27%	65.23%	60.39%
RWQA	with CLUETRACER	70.33%	67.84%	65.49%
3DSRBench	w/o CLUETRACER	37.66%	43.44%	56.25%
3DSRBench	with CLUETRACER	40.06%	45.14%	59.24%

D.4 Ablation Result Discussions

In this section, we discuss the ablation results in detail.

Clustering methods. We compare K-MEANS and DBSCAN at the patch clustering step. K-MEANS is a partition-based method, forcing every patch to be assigned to some cluster; this makes it sensitive to noise. In contrast, DBSCAN is density-based and permits points outside high-density regions to remain as noise. Practically, K-MEANS tends to absorb noisy patches into clusters, yielding overly large crops, whereas DBSCAN better isolates compact evidence regions. An example is provided in Fig. 7: K-MEANS nearly crops the entire image, while DBSCAN focuses on the key area.

Attention choices. We compare CLUETRACER with two baselines: *output*→*visual* attention (common in generative pipelines) and *question*→*visual* attention (input–input cross-attention, common in discriminative settings); see Fig. 8. Due to *reasoning drift*, *output*→*visual* alone can be distracted by complex scenes and fail to attend to the question-relevant region. Conversely, *question*→*visual* underutilizes the model’s training signal, which is primarily aligned to fitting the output sequence, leading to a mismatch and degraded localization.

D.5 Case Study

We present side-by-side comparisons across multiple reasoning models *with* and *without* CLUETRACER. As shown in Fig. 9 and Fig. 10, red boxes denote the visual crops. CLUETRACER consistently localizes task-relevant regions and produces more accurate answers than the baseline. We also report failure cases in Fig. 11 and Fig. 12, where Fig. 11 illustrates failures induced by visual semantics and Fig. 12 illustrates failures induced by query semantics.

E q2o and v2o Heatmaps

Setup and notation. We adopt the standard Transformer attention view for decoder-style ML-RMs (Vaswani et al., 2017): at output step t , the model attends over the input-side question tokens $\mathbf{X}_q = \{x_{q,i}\}_{i=1}^{N_q}$ and visual tokens $\mathbf{X}_v = \{x_{v,j}\}_{j=1}^{N_v}$, producing head-aggregated, key-normalized weights $A_{t,i}^{q \rightarrow o} \in [0, 1]$ and $A_{t,j}^{v \rightarrow o} \in [0, 1]$ with $\sum_i A_{t,i}^{q \rightarrow o} = 1$ and $\sum_j A_{t,j}^{v \rightarrow o} = 1$. Unless otherwise noted we average across layers/

heads (a common practice in attention-based attribution *post hoc* analyses). COCO categories and bounding boxes (x, y, w, h) are denoted by *cat* and *bbox* and follow the dataset’s instance annotations (Lin et al., 2014).

q2o heatmap (question → output). Given an object string *cat*, we first detect *object mentions* in the generated output $\{y_t\}_{t=1}^T$ by letters-only matching (case-insensitive) with a leniency margin (token length $\leq |\text{cat}|+2$) and POS tag NOUN. Let the set of indices of these mentions be $\mathcal{T}_{\text{obj}}$. For every output token t with POS NOUN, define its distance to the nearest object mention $d(t) = \min_{u \in \mathcal{T}_{\text{obj}}} |t - u|$. We form seven noun buckets $B_d = \{t : \text{POS}(t) = \text{NOUN}, d(t) = d\}$ for $d \in \{0, \dots, 6\}$, and three POS buckets over all outputs $B_{\text{ADJ}}, B_{\text{VERB}}, B_{\text{OTHER}}$. The $N_q \times 10$ matrix M used for the q2o heatmap is

$$M_{i,k} = \frac{1}{|B_k|} \sum_{t \in B_k} A_{t,i}^{q \rightarrow o}$$

for $k \in \{0, \dots, 6, \text{ADJ}, \text{VERB}, \text{OTHER}\}$.

with empty buckets contributing zeros. We visualize M as a diverging heatmap (high → red, low → blue) with smooth interpolation; the x-axis shows different buckets, and the y-axis lists question tokens.

v2o heatmap (visual → output around a bbox). For each COCO bounding box *bbox* of category *cat* present in the output, we locate the nearest occurrence index t_{cat} of *cat* (letters-only matching; multiword matched as a contiguous phrase). We extract a temporal window $W = \{t_{\text{cat}} - 10, \dots, t_{\text{cat}} + 10\} \cap [1, T]$ (21 positions) and the corresponding rows of the visual-to-output attention, giving $A_{W,:}^{v \rightarrow o} \in \mathbb{R}^{21 \times N_v}$. Using the patch grid (t_h, t_w) from the resized image, we map *bbox* to its covered patch set $P_{\text{bbox}} \subseteq \{1, \dots, N_v\}$ by $\text{IoU} > 0.5$ between *bbox* and each patch cell (pixel-aligned rectangles). Let $r_{\min} : r_{\max}$ and $c_{\min} : c_{\max}$ be the row/column spans of P_{bbox} on the $(t_h \times t_w)$ grid. For lateral offsets $s \in \{-5, \dots, 5\}$ we define expanded regions

$$P_{\text{bbox},s} = \{(r, c) : r \in I_r, c \in I_{c,s}\} \cap [0, t_h - 1] \times [0, t_w - 1],$$

$$I_r = [r_{\min}, r_{\max}],$$

$$I_{c,s} = [c_{\min} + \min(0, s), c_{\max} + \max(0, s)].$$

flattened to patch indices. The 21×11 matrix V for *bbox* is

$$V_{r,s+5} = \frac{1}{|P_{\text{bbox},s}|} \sum_{j \in P_{\text{bbox},s}} A_{t_r,j}^{v \rightarrow o}, \quad t_r \in W,$$

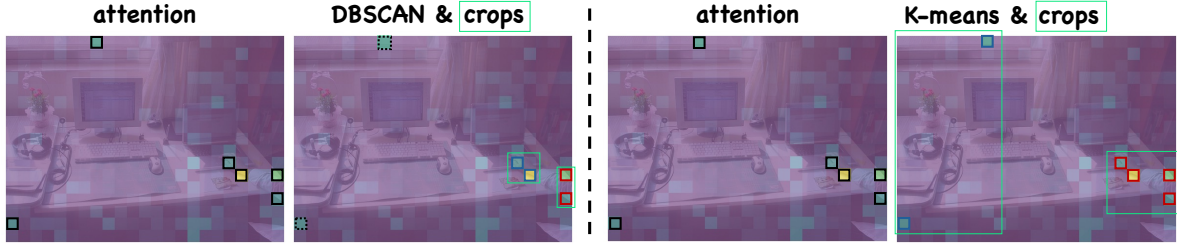


Figure 7: **Case study on clustering methods.** Each subfigure has two parts. *Left*: CLUETracer attention, i.e., a visualization of the attention map, where the CLUETracer-selected visual patches $\mathcal{V}_k^*$ are outlined with black boxes. *Right*: clustering results from different methods. Because K-means is sensitive to noise, the resulting crops include large amounts of question-irrelevant content induced by only two noisy points, whereas DBSCAN avoids this issue.

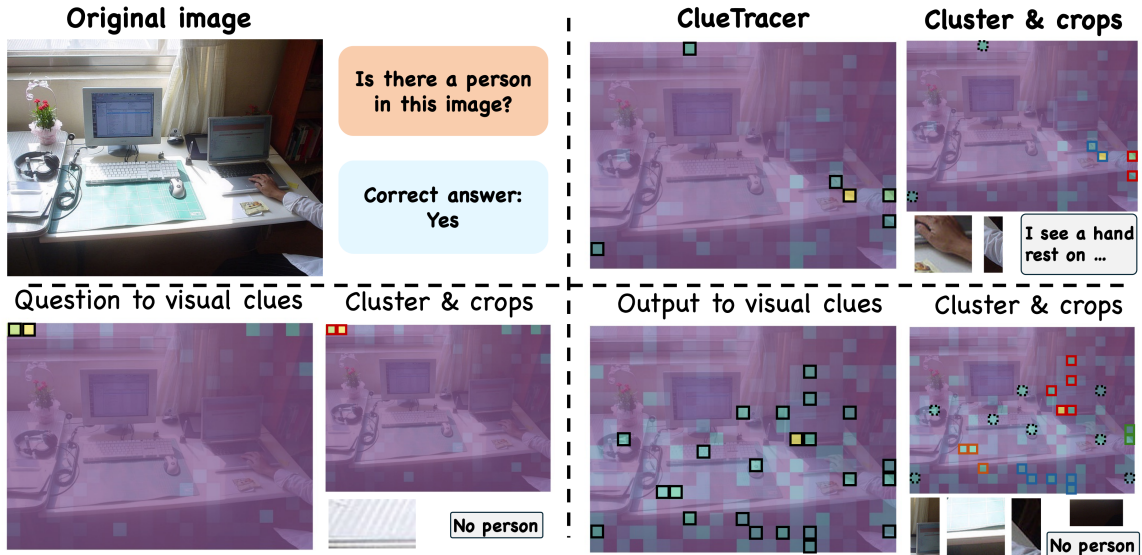


Figure 8: **Case study on attention choices.** Attention identifies question-relevant tokens and correctly crops the hand, thereby guiding the model to the correct answer. *Question to visual clues* denotes cross-attention between the input question and input visual tokens; because the model is not trained for this alignment, the signal is weak. *Output to visual* directly leverages generation-side attention, which is mainstream in MLLMs; however, in MLRMs, reasoning drift introduces numerous distractors.

summarizing how attention over visual patches varies near the bounding box versus its lateral background. We render V with the same diverging colormap and smooth interpolation; the y-axis uses three ticks (others–{cat}–others) centered at t_{cat} , and the x-axis uses three ticks (background–bbox patches–background) centered at $s=0$.

F Sensitivity, scalability, and deployment cost analysis

We further analyze the sensitivity of CLUETRACER to the thresholds τ_q, τ_o, τ_v and the clustering parameter $\text{dbscan-}\epsilon$. As shown in Tables 8–11. We also summarize latency and throughput in Table 12, and report layer-wise AttnRecall on larger 32B models and LLaVA architecture (Liu et al., 2023a) in Table 13 and Table 14.

Sensitivity Analysis. As summarized in Tables 8–11, the accuracies vary only mildly across a wide range of τ_q, τ_o, τ_v , and $\text{dbscan-}\epsilon$, indicating that CLUETRACER is not overly sensitive to these hyperparameters. In all main experiments, we therefore fix $\tau_q = 1, \tau_o = 1, \tau_v = 1.5$, and $\text{dbscan-}\epsilon = 2.9$.

Deployment Cost Analysis. We conduct a deployment cost analysis of CLUETRACER and report end-to-end costs under different Stage 1 output lengths. All experiments are run with a 7B model on a single NVIDIA GeForce RTX 3090 (20G). As shown in Table 12, longer Stage 1 sequences provide richer priors to Stage 2 but also incur higher inference cost, which is consistent with our expectation.

Scalability Analysis. As shown in table 13 and 14, on both 32B variants and models with different underlying architectures, the layer-wise CLUERECALL exhibits the same rise then fall pattern as in 7B models, and the peak perceptual layers still lie in the later part of the network. For both Orsta-32B and MM Eureka-32B, the global maximum of CLUERECALL appears around layer 40 out of 64, which corresponds to roughly 70% of the total depth. This indicates that our perception layer finding scales to larger parameter regimes.

G Potential Risks

This work aims to reduce hallucinations in multimodal reasoning models, but it may also introduce

potential risks if over-relied upon in high-stakes applications. Since CLUETRACER improves visual grounding by tracing model attention, its outputs may give users a stronger sense of reliability even when the underlying model still makes incorrect judgments.

H Discussion with Prior Attention-based Studies

H.1 Attention Applications in VLMs

Prior works have explored visual attention in vision-language models for model interpretation, visual grounding, and attention analysis (Ilinykh and Dobnik, 2021; Cao et al., 2022). For example, visual attention was used to inspect image-captioning models by revealing attended image regions during generation (Xu et al., 2015), while later studies analyzed textual and cross-modal attention to understand how multimodal transformers ground entities and relations (Ilinykh and Dobnik, 2022). Related studies in referring expression generation further investigated how models allocate attention to target objects, spatial information, and contextual regions during visual language generation (Junker and Zarrieß, 2023; Merlo et al., 2026).

Different from these studies, which mainly analyze attention patterns in conventional vision-language generation or object-level grounding, our work focuses on multimodal reasoning models and studies how intermediate outputs mediate visual evidence retrieval. Specifically, we introduce a question→output→vision tracing mechanism to identify task-critical clues during long-chain reasoning, enabling the characterization and mitigation of reasoning drift.

H.2 Compatibility with Gradient-based Attribution

Although CLUETRACER uses native attention for efficient online inference, its query→output→vision evidence-tracing formulation is not restricted to attention-based attribution and can be integrated with gradient-based methods such as Grad-CAM (Selvaraju et al., 2017). Existing Grad-CAM applications to LVLMs typically attribute the final answer token to visual regions (Zhang et al., 2025b). This is suitable for short-form generation, but in long-chain reasoning, decisive visual evidence may be retrieved earlier in the reasoning trajectory, while the final answer can depend primarily on

Table 8: Sensitivity of accuracy to τ_q on HallusionBench and POPE.

Method	0.5	1	1.5	2	3
HallusionBench + MM-Eureka	63.51	65.10	63.51	63.15	61.65
POPE + MM-Eureka	79.86	81.38	80.32	79.13	80.92
POPE + Orsta	83.32	83.67	85.01	83.12	82.91

Table 9: Sensitivity of accuracy to τ_o on HallusionBench and POPE.

Method	0.5	1	1.5	2	3
HallusionBench + MM-Eureka	66.43	65.10	63.86	63.42	63.77
POPE + MM-Eureka	82.32	81.38	79.22	79.62	80.62
POPE + Orsta	84.42	83.67	85.02	83.12	79.02

Table 10: Sensitivity of accuracy to τ_v on HallusionBench and POPE.

Method	0.5	1	1.5	2	3
HallusionBench + MM-Eureka	64.22	65.46	65.10	63.95	63.45
POPE + MM-Eureka	82.92	82.41	81.38	80.92	79.12
POPE + Orsta	81.82	83.22	83.67	85.81	83.22

Table 11: Sensitivity of accuracy to DBSCAN- ϵ on HallusionBench and POPE.

Method	1.1	1.9	2.1	2.9	3.1	3.9
HallusionBench + MM-Eureka	63.31	64.57	64.48	65.10	65.20	65.37
POPE + MM-Eureka	78.52	79.62	80.62	81.38	81.32	81.22
POPE + Orsta	84.52	84.62	85.41	83.67	84.52	83.02

Table 12: Latency and throughput under different online and offline settings.

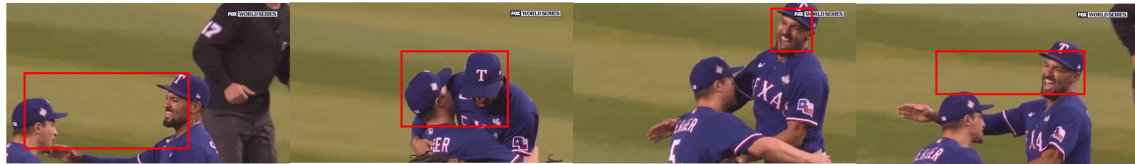
Mode / sequence length	offline/–	1	20	50	128	256
Latency (hours / 1000 items)	0.91	0.97	1.08	1.36	1.62	1.79
Throughput (tokens/s)	39.80	35.52	31.24	27.71	20.91	18.08

Table 13: Layer-wise CLUERECALL on 32B models.

Model	0	12	24	30	36	40	42	48	54	63
MM-Eureka-32B	0.3060	0.3377	0.4978	0.5358	0.5781	0.6271	0.5474	0.4854	0.3992	0.3099
Orsta-32B	0.3165	0.3397	0.4922	0.5381	0.5669	0.6350	0.5596	0.4961	0.4059	0.3172

Table 14: Layer-wise CLUERECALL on LLaVA-1.5-7B and LLaVA-Mistral-7B.

model	Arch	0	6	12	17	18	24	31
LLaVA-1.5-7B	LLaVA-1.5	33.81%	38.00%	43.04%	50.59%	48.29%	47.03%	39.02%
LLaVA-Mistral-7B	LLaVA-NeXT	10.37%	23.95%	46.00%	49.04%	49.72%	41.10%	35.95%



Q: They are growing apart. According to the positive sequence of the images, are they in the correct order?

Correct answer: no.

With ClueTracer: They are growing apart. According to the positive sequence of the images, they are not in the correct order. Answer: no

W/o ClueTracer: In the sequence of images frames, the players are growing apart, indicating to their they are in the correct order. Answer: yes



Q: They are skating to right. According to the positive sequence of the images, are they in the correct order?

Correct answer: no.

With ClueTracer: They are skating to right. According to the positive sequence of the images, they are not in the correct order. The third image should be before the second image. Answer: no

W/o ClueTracer: In the sequence of images frames, the players are growing apart, indicating to their they are in the correct order. Answer: yes

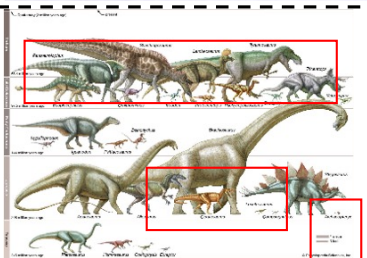


Q: Are all the items here engineering tools?

Correct answer: no.

With ClueTracer: In the image, there are a wrench, screwdriver, chisel and the carrot. Therefore, it can be seen that not all the items here are engineering tools. Answer: no

W/o ClueTracer: The image shows a variety of tools types, including a carrot for scale. The tools on the left appears to be a wrench, the next few tools appear to be various types of screwdrivers, which are used in construction and assembly engineering projects. Therefore, all the items here are engineering tools. Answer: yes

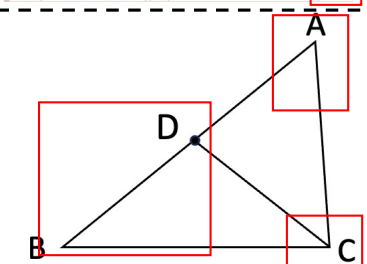


Q: Are there any hens in this figure?

Correct answer: no.

With ClueTracer: The image is a infographic showing different types of dinosaurs. There are no hens present in the image. Therefore, the answer is no. Answer: no

W/o ClueTracer: The image is a colorful chart illustrating various types of dinosaurs. Each dinosaur is labeled with its name scientific name or common name. The chart includes different kinds of dinosaurs such as sauropods, ornithomorphs, ceratopsians, and theropods. It provides a visual comparison of these different types, including showing their their the size, shape, and features of each one. Answer: yes



Q: If angle B is equal to 45 degree, is AB larger than $\sqrt{2}AC$?

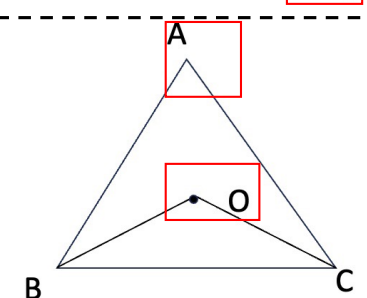
Correct answer: no.

With ClueTracer: Let's analyze the triangle step by step:

1. Since B is 45 degrees, and the sum of angles in a triangle is 180 degrees, we can infer that triangle ABC is a right triangle with a 45-degree angle, making it an isosceles right triangle.
2. In an isosceles right triangle, the two legs are equal, and the hypotenuse is $\sqrt{2}$ times the length of each leg.

Since the statement asks if AB is larger than $\sqrt{2}AC$, and we have established that $AB = \sqrt{2}AC$, the answer is no. Answer: no

W/o ClueTracer: In the triangle ABC, if angle B is 45 degrees, we can use the Pythagorean theorem to analyze the relationship between the $\sqrt{2}$. Answer: yes



Q: The two interior angle bisectors OB and OC of triangle ABC intersect at point O. If angle A is 110 degree, then is the angle BOC equal to 145 degree?

Correct answer: yes.

With ClueTracer: 1. We are given that angle A = 110.

2. OB and OC are the angle bisectors of angles ABC and ACB, respectively.

3. angle OBC = $\frac{1}{2}$ times angle ABC and angle OCB = $\frac{1}{2}$ times angle ACB.

4. Substituting the value of angle A, we get angle ABC + angle ACB = 180 - 110 = 70

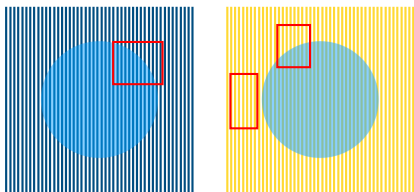
5. Since OB and OC are angle bisectors, OBC + OCB = $\frac{1}{2}$ (ABC + ACB) = 35

6. Therefore, angle BOC = 180 - (angle OBC + angle OCB) = 180 - 35 = 145

Therefore, the angle BOC is indeed 145. Answer: yes

W/o ClueTracer: elementGuidId\nGLEBE6I\nRDKp\nGiven the problem, we need need to determine if the angle BOC is equal to 145 degrees given that angle A is 110 degrees and the angle bisectors OB and OC intersect at point O.1. Since angle A is 110 degrees, the sum of the other other angles at B and C is $180 - 110 = 70$. Answer: no

Figure 9: Case study (Part One). The visual crops selected by CLUETracer are highlighted with red bounding boxes in the figure.

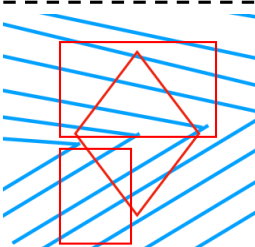


Q: The image shows a color illusion. Are two circles in the image different color?

Correct answer: **no.**

With ClueTracer: Looking at the first image, the circle appears blue due to color subtraction, where the blue color of the circle subtracts the surrounding blue lines, making the circle appear lighter. In the second image, the circle appears to be a lighter shade of blue, which is actually the same color as the circle in the first image. Therefore, the two circles in the image are the same color. **Answer: no**

W/o ClueTracer: The image shows a color illusion where the circle on the left appears to be a different color than the circle on the right due to the surrounding stripes. **Answer: yes**

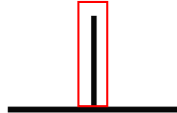


Q: Whether the red shape in the image is a square?

Correct answer: **no.**

With ClueTracer: The second image is not a square because it has one longer side than the others, making it an irregular quadrilateral. **Answer: no**

W/o A*4: The red shape in the image is a square. It has four equal sides and four right angles. **Answer: yes**

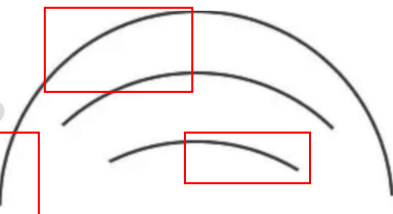


Q: Is the vertical line longer than the horizontal line?

Correct answer: **no.**

With ClueTracer: Both lines appear to be the same length. The vertical line is aligned with the top part of the horizontal line, suggesting they are equal in length. **Answer: no**

W/o ClueTracer: To determine if the vertical line is longer than the horizontal line, we need need to compare their lengths of the two lines. **Answer: yes**

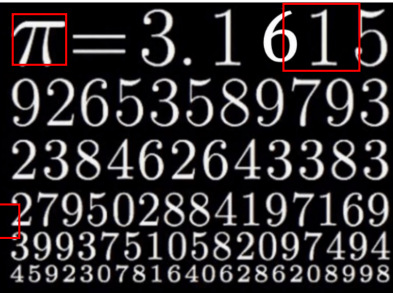


Q: According to the Torsion dystonia illusion, are the diameters of the three circles different?

Correct answer: **no.**

With ClueTracer: The image shows three circles that appear to be the same size due to the Torsion dystonia illusion. However, the illusion makes it seem like the middle circle is larger than the other two, which is a visual trick rather than a physical difference. Therefore, the diameters of the three circles are actually the same. **Answer: no**

W/o ClueTracer: The image depicts the Torsion dystonia illusion, which is a visual illusion that can make make circles appear to be of different sizes when they are in fact identical. The three arcs in the image are arranged in a way way that makes it difficult to perceive their their the circles as being being the same same size. **Answer: yes**



Q: According to the image, does π range from 3.1615926 to 3.1615927?

Correct answer: **yes.**

With ClueTracer: According to the image, π is given as 3.16159265358979323846264338327950288419716939937510582097494592307816406286208998. This value falls between 3.1615926 and 3.1615927. Therefore, the statement " π range from 3.1615926 to 3.1615927" is correct. **Answer: yes.**

W/o ClueTracer: The image shows the value of π as 3.16159265358979323846264338327950288419716939937510582097494592307816406286208998. This π is known to be e approximately 3.14159265358975. Therefore, the value given in the image is not within the range of 3.1615926 to 3.1615922. Thus, the statement is false. **Answer: no.**



Q: According to the image, does the Oyashio Current flow from the south to the north, while the Kuroshio Current flows from the north to the south?

Correct answer: **No.** Oyashio Current flows from the north to the south and the Kuroshio Current flows from the south to the north.

With ClueTracer: The Oyashio Current is shown flowing from north to south, while the Kuroshio Current is shown flowing from south to north. Therefore, the statement is incorrect. **Answer: no.**

W/o ClueTracer: The image shows the Oyashio Current flowing from the north to the south, indicated on the map. **Answer: yes.**

Medals winter olympics 2022

Country	gold	Silver	Bronze	Total
Norway	9	4	2	15
Germany	12	10	5	27
China	16	8	13	37

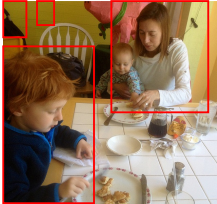
Q: According to the table, is the number of all medals of China 37 in winter Olympics in 2022?

Correct answer: **Yes.**

With ClueTracer: Yes, the image on the right side shows a table with the heading "Medals winter olympics 2022" and a column labeled "Total" which contains the number 37 for China. This indicates that the total number of medals China won in the 2022 Winter Olympics is indeed 37. **Answer: yes.**

W/o ClueTracer: According to the table, the number of all medals of China in the Winter Olympics 2022 is 37, not 37. Therefore, the statement is incorrect. **Answer: no.**

Figure 10: Case study (Part Two). The visual crops selected by CLUETRACER are highlighted with red bounding boxes in the figure.



Q: Is there a person in the image?

Correct answer: **yes.**

With ClueTracer: No, there is no person in the image. The image shows a dining table with plates of food, a fork and a knife. There is a white chair and a black bag in the background. The yellow paint on the wall adds a warm touch to the scene. **Answer: no**

W/o ClueTracer: In the image, there is a woman and two children sitting at a table. The woman appears to be feeding or assisting one of the children with food while the other child is seated across from her, engaged in eating or playing. **Answer: yes**

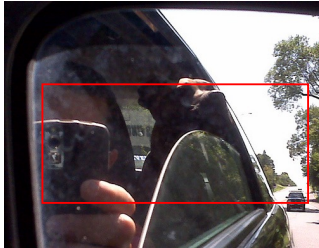


Q: Is there a truck in the image?

Correct answer: **no.**

With ClueTracer: This image seems to be taken from inside the passenger seat of a vehicle, showing the view through the windshield. The windshield has decorations, including flowers hanging from the top. Outside the windshield, there are other vehicles. The truck is partially visible, with its front end and part of the side showing. It's a large vehicle with a boxier shape, which is characteristic of trucks. **Answer: yes**

W/o ClueTracer: The image shows a view from inside a vehicle, likely a car, looking at another vehicle outside. The other. The vehicle in the background appears to be a truck, as indicated as it has a boxy. **Answer: no**



Q: Is there a truck in the image?

Correct answer: **yes.**

With ClueTracer: The image shows a person taking a picture from the inside of a car, with a dog in the back seat looking out the window. The background features a road with trees and a clear sky. The road is relatively straight and there are no other vehicles visible in the image. **Answer: no**

W/o ClueTracer: In the image, there appears to be a reflection and views outside. The reflection of shows the inside of the vehicle where someone's hand, holding a phone, can be reflecting in the driver side's side view mirror. The outside view in the side right side of shows vehicles driving on a road with trees lining the side. However, the image is not clear enough to distinguish specific features like the shape or features of a truck. **Answer: yes**



Q: Is there a hadbag in the image?

Correct answer: **no.**

With ClueTracer: There is a person standing near the storefront with a handbag hanging from their shoulder. This can be inferred from the typical way people carry handbags when shopping or walking, and the position of the person relative to the display. **Answer: yes**

W/o ClueTracer: The image shows a street vendor's stall with various items for sale, including flags and sports attire. The vendor appears to be a man standing behind the stall, and there are a few people people standing in front of the stall. The stall is located outside the HSBC bank building, and the wet street suggests recent rain. Given this information, it is likely that there are a handbag in the image, although it is not explicitly visible, but this can be a reasonable assumption. **Answer: no**

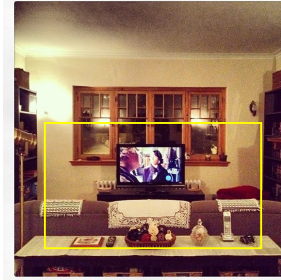


Q: Is there a backpack in the image?

Correct answer: **yes.**

With ClueTracer: No, there is no backpack in the image. The image shows a market stall with a man operating a machine that roasts peanuts. There are a few people in the background, and some boxes and crates are visible. There is no visible backpack in the image. **Answer: no**

W/o ClueTracer: The image shows an outdoor market scene with several people gathered under tents. In the foreground, there is a man in overalls standing behind a table with various items, including a sign reading \"TRESH ROASTED PEANuts\". There is a small green cooler to the left of the table, and a wooden box to the right. The table has bags of peanuts, and other another box and some other a small cooler to left in the table, and a wooden box to the ri. Some items including the right. In the background, there there are people are some are some people a small cooler to the left of the table, and a wooden box to the right. Some items to the right. There is a person in the right background holding what appears to be a baby, and some others standing and talking. The grass is green, and there the background, there there there **Answer: yes**



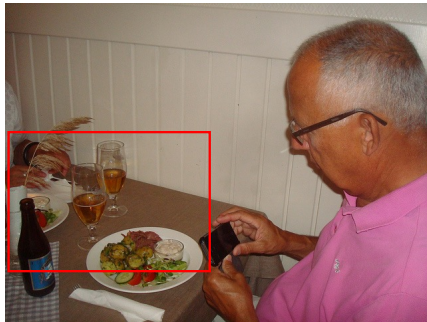
Q: Is there a person in the image?

Correct answer: **yes.**

With ClueTracer: No, there is no person in the image. The images appear to be identical, showing a living room with a couch, a table in front of it, and a television displaying a scene. The image is a close-up of a textured surface that does not contain a person. Therefore, the presence of a person cannot be determined from the image. **Answer: no**

W/o ClueTracer: The image shows a living room with a television that appears to be displaying a scene with people a person is on the screen. Therefore, there is a person in the image. **Answer: yes**

Figure 11: **Failure case study** (Part One). The visual crops selected by CLUETRACER are highlighted with red and yellow bounding boxes. We observe that, in most cases, CLUETRACER crops the correct regions, but it fails to do so in the third example.



Q: Is there a cup in the image?

Correct answer: **no**.

With ClueTracer: Yes, there is a cup in the image. The image shows a glass of beer on the table, which can be considered a cup. Therefore, there is a cup in the image.

Answer: yes

W/o ClueTracer: The image shows a man sitting at a table with a variety of items. There are a plate of food in front of him, including a bottle of beer to his left, and some two glasses of what appears to be beer. He is holding a smartphone in his hand, and appears to be taking a picture of his food. The glasses on the table have a shape and structure consistent with wine glasses, which are typically used for beer servings, but cannot be confirmed as such without further context. Therefore, it is not possible to definitively conclude that the glasses are contain beer, but they are be suggestive of that be beer. **Answer: no**

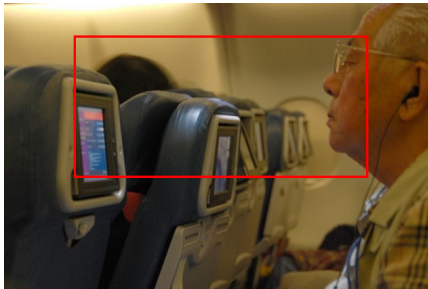


Q: Is there a vase in the image?

Correct answer: **no**.

With ClueTracer: Yes, there is a vase visible in the image. It is located on a white surface, likely a shelf or countertop, and contains a yellow flower arrangement. The vase has a rectangular shape and is positioned in the lower right corner of the image. The white surface it's on appears to be part of a bathroom setting, with tiles visible in the background. The overall scene suggests a clean and organized space, with the vase and flower adding a touch of natural beauty to the room. **Answer: yes**

W/o ClueTracer: The image shows a bathroom with a toilet, a sink, and a bathtub. The bathroom is well-lit and clean. **Answer: no**



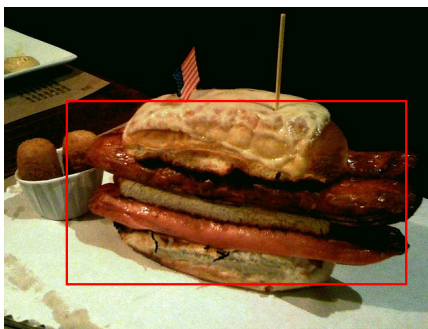
Q: Is there a chair in the image?

Correct answer: **yes**.

With ClueTracer: The image shows a row of seats on an airplane, with a man sitting in one of them. The seats have personal entertainment systems on the backrests, which is a common feature on planes. The man appears to be using the in-flight entertainment system, possibly playing a video or watching a movie. There is no chair in this image.

Answer: no

W/o ClueTracer: The image shows the interior of an airplane, focusing on the row of seats with personal entertainment systems embedded in the headrests. The seats are clearly visible, with their backrests and the screens in the front of each seat. This man is seated in one of these seats, watching looking ahead. Therefore, it is reasonable to conclude that the image does indeed contain chairs. **Answer: yes**



Q: Is there a dog in the image?

Correct answer: **yes**.

With ClueTracer: The image contains food of some sort, possibly a sandwich or burger, and there is a small flag visible on top of it. Upon further analysis, it appears that the food item is a hot dog, which is a type of sandwich. Therefore, there is no visible dog in the image. **Answer: no**

W/o ClueTracer: The image shows a large sandwich with various ingredients, including a hot dog. The top part of the sandwich appears to be a hot dog bun with a slice of cheese and possibly a sausage. the bottom part is another slice of bread, and there the sausage appears to be the main component of the sandwich. Therefore, the image does contain a dog, which is the sausage. **Answer: yes**

Figure 12: **Failure case study** (Part Two). The visual crops selected by CLUETRACER are highlighted with red/yellow boxes. We observe that these failures are driven by object-level semantics: in these examples, the model recognizes semantically similar but incorrect concepts (e.g., *glass* vs. *cup*, *seat* vs. *chair*).

Method	Ocean-R1	MM-Eureka
Standard Grad-CAM	0.267	0.274
CLUETRACER + Grad-CAM	0.377	0.390

Table 15: Compatibility of CLUETRACER with Grad-CAM. Incorporating question-relevant reasoning tokens substantially improves visual clue localization over standard final-token attribution.

previously generated text. We therefore extend Grad-CAM with our clue-tracing strategy: instead of attributing only the final answer token, we identify question-relevant reasoning tokens and aggregate their Grad-CAM maps. We evaluate localization using the same bounding-box recall metric.

These results further support that the core contribution of CLUETRACER lies in the output-mediated evidence-tracing formulation, which can be instantiated with different attribution mechanisms.