Journey Before Destination: On the importance of Visual Faithfulness in Slow Thinking

Rheeya Uppaal[*1]**, Phu Mon Htut**[2]**, Min Bai**[2]**, Nikolaos Pappas**[2]**,**
Zheng Qi[2]**, and Sandesh Swamy**[2]

[1]University of Wisconsin-Madison
[2]AWS AI Labs

Abstract

Reasoning-augmented vision language models (VLMs) generate explicit chains of thought that promise greater capability and transparency but also introduce new failure modes: models may reach correct answers via visually unfaithful intermediate steps, or reason faithfully yet fail on the final prediction. Standard evaluations that only measure final-answer accuracy cannot distinguish these behaviors. We introduce the *visual faithfulness of reasoning chains* as a distinct evaluation dimension, focusing on whether the perception steps of a reasoning chain are grounded in the image. We propose a training- and reference-free framework that decomposes chains into perception versus reasoning steps and uses off-the-shelf VLM judges for step-level faithfulness, additionally verifying this approach through a human meta-evaluation. Building on this metric, we present a lightweight self-reflection procedure that detects and locally regenerates unfaithful perception steps without any training. Across multiple reasoning-trained VLMs and perception-heavy benchmarks, our method reduces Unfaithful Perception Rate while preserving final-answer accuracy, improving the reliability of multimodal reasoning. [1]

1 Introduction

Hallucinations in vision–language models (VLMs) are typically defined as deviations between model outputs and the underlying visual content (Bai et al., 2024; Liu et al., 2024b). While the phenomenon has been studied extensively, existing evaluations for it remain narrow. Most focus on coarse object existence in captions, overlooking finer elements such as counts, colors, or spatial relations that make up a large portion of visual hallucinations (Gunjal
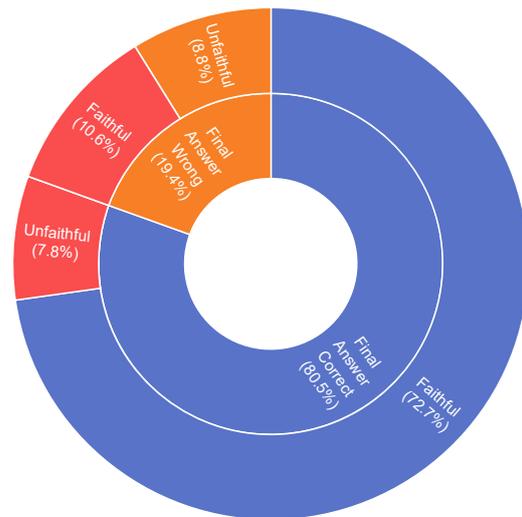

Figure 1: Reasoning faithfulness and final-answer accuracy diverge. Correct final answers are not always grounded in the image, and incorrect answers can still reflect visually faithful reasoning. Evaluating only final accuracy therefore overlooks whether the reasoning process itself attends to the visual evidence. The weak correspondence between final-answer correctness and reasoning-chain faithfulness shows that accuracy metrics alone cannot capture whether a model's reasoning genuinely reflects what it "sees."

et al., 2023). These limitations become more pronounced in increasingly popular reasoning based models (Aubakirova et al., 2025), where intermediate steps are incorporated to solve complex tasks and provide apparent transparency into the model's decision-making processes (Li et al., 2025).

In text-only domains, the quality of reasoning traces has been examined in terms of their correctness, coherence, or adherence to instructions (Jacovi et al., 2024; Hao et al., 2024). In multimodal settings, however, these reasoning chains introduce a new axis of reliability: *visual faithfulness* – Is each step of the reasoning chain actually grounded in the image?

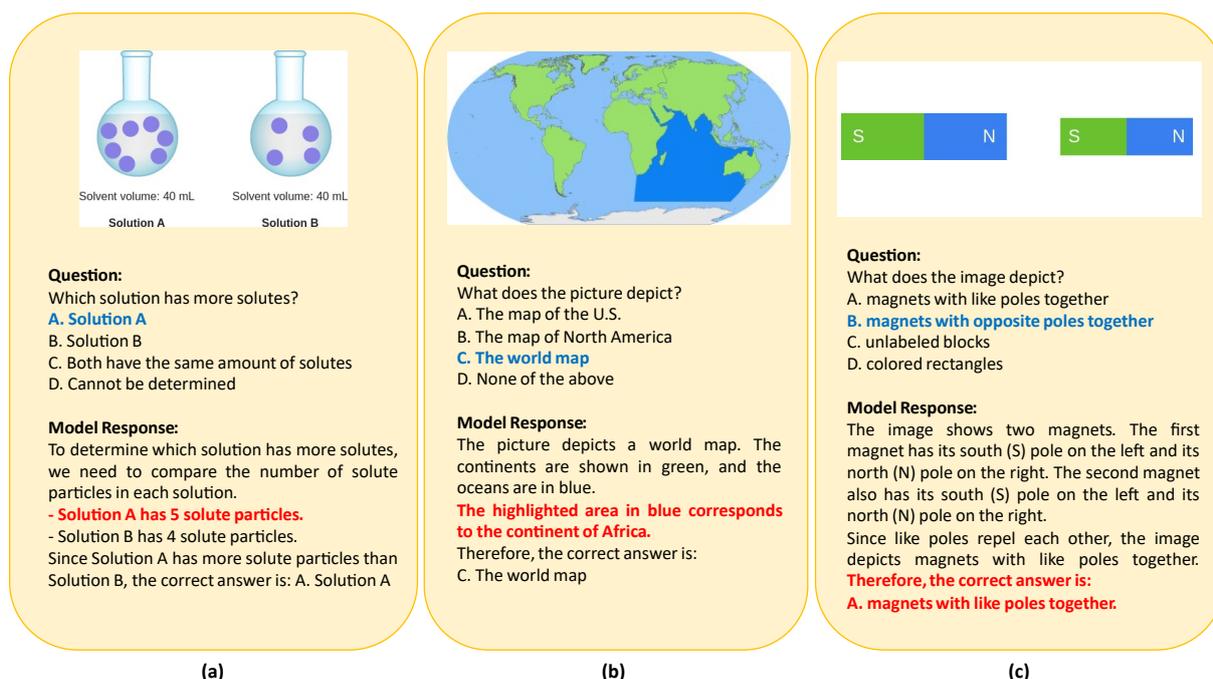A model may produce a correct final answer

Figure 2: Reasoning-chain faithfulness does not always align with final-answer correctness. (a–b) Visually unfaithful reasoning chains that nonetheless yield correct answers on perception tasks. (c) A visually faithful chain producing an incorrect answer, where the error arises from reasoning rather than perception. All responses are from the `ThinkLite-VL` model on samples from the `MMEvalPro` dataset.

while hallucinating intermediate entities, attributes, or relations that are not visually supported. Conversely, it may describe the image faithfully yet still reach an incorrect conclusion due to downstream logical mistakes. Figure 2 illustrates both phenomena: visually unfaithful reasoning leading to correct predictions, and visually faithful reasoning that nonetheless yields incorrect answers.

This sheds light on a pressing issue – existing metrics assess the hallucination rate of a VLM through its final answer accuracy on perception tasks. In Figure 1 we highlight that this measure does *not* correlate with the model's hallucination rate in its reasoning chains. Many evaluation protocols implicitly assume that the final answer y is produced by following the reasoning chain R (Figure 3). In practice, large models can shortcut this process: internal representations h may map directly to the answer via spurious correlations, language priors, or pattern matching, while the chain R is generated as a post-hoc justification rather than the causal basis for the decision (Jiang et al., 2025; Shojaee et al., 2025; Xia et al., 2025). As a result, high final-answer accuracy does not guarantee that intermediate reasoning steps are visually faithful, nor that they faithfully track the model's internal decision path.

Existing hallucination detection approaches are simply not designed to capture hallucination rates in the complex setting of reasoning. These methods usually verify the existence of objects or atomic facts against the image or a ground truth list; thus treating each answer as an unordered set of facts. By contrast, reasoning chains are a compositional trajectory – they are long, structured, and explicitly interleave distinct perception steps (reading from the image) and reasoning steps (operating over previously inferred facts). Visual faithfulness is only well-defined for perception steps, yet errors in these steps can propagate through the chain and contaminate subsequent reasoning. Prior work has also shown that hallucinations become more frequent as generations grow longer and more verbose (Zhai et al., 2023), making it particularly important to evaluate the quality of the full reasoning trajectory rather than only its endpoint.

In this paper, we take a first step toward systematically *measuring and improving the visual faithfulness of reasoning chains* in VLMs. Our contributions are:

- *Problem definition.* We formally highlight visual faithfulness of reasoning chains as distinct from final-answer accuracy or traditional hallucination detection, with empirical evi-

dence that these standard metrics do not reliably capture step-level faithfulness.

- *Evaluation framework.* We introduce a scalable, training- and reference-free evaluation pipeline that uses off-the-shelf VLM judges to assess visual faithfulness at the level of individual reasoning steps, and validate these metrics via a human correlation study.

- *Mitigation method.* We propose a lightweight self-reflection procedure that combines a when-to-intervene detector with localized regeneration of unfaithful perception steps. Our method substantially improves reasoning-chain visual faithfulness across multiple datasets and models, often with improved final-answer accuracy.

Together, our results establish reasoning-chain visual faithfulness as an essential axis for evaluating and improving reasoning-augmented VLMs, and provide concrete tools for measuring and mitigating unfaithful visual reasoning at scale.

2 Related Work

Hallucinations in VLMs In vision–language models (VLMs), hallucinations are broadly defined as deviations between model outputs and provided visual content (Bai et al., 2024; Liu et al., 2024b). Such lack of visual faithfulness can arise from insufficiently diverse data during instruction tuning (Li et al., 2023; Wang et al., 2023a; Yu et al., 2024; Goyal et al., 2025) or earlier training stages (Wang et al., 2023a; Zhou et al., 2024; Li et al., 2023; Guan et al., 2024); or limited capabilities of the vision encoder to capture fine-grained visual information (Zhang et al., 2021; Wang et al., 2023a, 2024b; Liu et al., 2024e; Wang et al., 2024c). However, the most common cause is language dominance: VLMs tend to under-attend to the image (Parcalabescu and Frank, 2025; Yin et al., 2025; Yang et al., 2025b), allowing strong priors from LLM parameters to override visual signal (Zhai et al., 2023; Jiang et al., 2025; Sun et al., 2024c; Rahmanzadehgervi et al., 2024; Liu et al., 2025b).

Efforts to mitigate hallucinations have included data diversification (Qi et al., 2020; Liu et al., 2024a; Wang et al., 2024a; Yu et al., 2024; Zhang et al., 2024a; Zou et al., 2024; Yue et al., 2024b; Hu et al., 2025, *inter alia*), activation steering to strengthen visual signal (Zhai et al., 2023; Liu et al.,
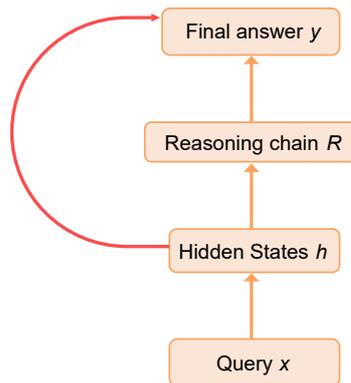

Figure 3: Causal structure underlying final answers and reasoning traces. Many evaluation protocols assume the final answer y is produced via the reasoning chain R (orange arrows). However, models can also map hidden features h directly to y (red arrow) via spurious correlations or language priors, bypassing R. Thus, high final-answer accuracy does not guarantee that intermediate reasoning steps are visually faithful.

2025b; Jiang et al., 2025; Yang et al., 2025b; Yin et al., 2025; Su et al., 2025), or other model editing techniques (Jiang et al., 2025; Yang et al., 2025a; Uppaal et al., 2025; Arif et al., 2025), and decoding-time interventions (Favero et al., 2024; Ghosh et al., 2024; Wang et al., 2024b; Yin et al., 2024). Nevertheless, most methods focus on improving visual grounding in simple discriminative or captioning settings rather than multi-step reasoning.

Slow Thinking VLMs Slow thinking is operationalized through the notion of inference-time scaling — allocating a larger token budget to allow multi-step deliberation and exploration of multiple hypotheses (Li et al., 2025). These models demonstrate substantial gains on reasoning tasks, often trained via supervised fine-tuning (SFT) on reasoning chains (Xu et al., 2024; Zhang et al., 2024c; Deng et al., 2024; Cheng et al., 2025; Chen et al., 2024) or reinforcement learning (RL) (Wang et al., 2025c; Xiang et al., 2024; Wang et al., 2025a). These models consistently invoke inference-time scaling irrespective of inference time prompt structures. These training paradigms encourage self-reflection and iterative correction within the model's generation process.

Although some recent studies attempt to enhance visual faithfulness through similar reasoning-based training (Zhao et al., 2023; Jing and Du, 2024; Sun et al., 2024a; Favero et al., 2024; Zhao et al., 2025), Liu et al. (2025a) show that reasoning training can worsen visual faithfulness. Moreover, existing stud-

ies largely assess grounding only in final answers, overlooking whether intermediate reasoning steps remain visually faithful.

Evaluating the Quality of Reasoning Chains
Wang et al. (2023a) and Favero et al. (2024) show that hallucination frequency increases with longer generations, making it crucial to measure the quality of the reasoning chains, not just final answers. In text-only settings, prior work has assessed the redundancy, relevance, and correctness of intermediate steps using verifiers (Jacovi et al., 2024; Hao et al., 2024). Shojaee et al. (2025) and Xia et al. (2025) further question final answer accuracy based evaluation, showing that is does not necessarily guarantee an improvement in the overall quality of the reasoning steps.

However, in grounded multi-modal generation, the visual faithfulness of reasoning traces remains largely unexplored. Chen et al. (2024) evaluate reasoning-step consistency but require curated atomic question sets. In contrast, our work introduces a training and data-free framework for measuring and mitigating the visual faithfulness of reasoning chains using off-the-shelf VLM judges, establishing a scalable foundation for assessing grounded reasoning quality.

3 Measuring the Visual Faithfulness of Reasoning Chains

3.1 Measuring Visual Faithfulness Beyond Final Answers

Much of the existing work on visual faithfulness has focused on the limited setting of discriminative free-form answers (Li et al., 2023; Sun et al., 2024b; Wang et al., 2023b; Wu et al., 2024; Zhang et al., 2021; Liu et al., 2024e; Guan et al., 2024). In such settings, hallucination can often be detected through simple metrics like accuracy or F1 score. Some works extend evaluation into the generative space of image captioning tasks; however, even in this case evaluation is limited to object hallucinations, and requires a ground truth list of objects present in the image (Wang et al., 2023a; Liu et al., 2024a). More recently, limited work on reference-free evaluation in free-form answers has been proposed: Liu et al. (2025a) use GPT evaluation while Jing et al. (2024) decompose an output into atomic facts and use a trained model to verify if each fact is entailed by the image.

However, notably, these approaches focus primarily on isolated facts or final answers. They do not capture whether a model's intermediate reasoning process is visually grounded and consistent with the image. Evaluating reasoning chains poses unique challenges compared to short answers or image captions: chains are multi-step, compositional, and often combine factual grounding (or perception) with logical inference (or reasoning). A model might reach a correct final answer through unfaithful intermediate steps, or conversely, follow a faithful chain that nevertheless ends in an incorrect conclusion. In both cases, assessments based solely on the final output fail to reflect the true quality of visual grounding. This gap motivates our work: measuring the visual faithfulness of reasoning chains requires methods that move beyond outcome-based evaluation and instead interrogate the full reasoning trajectory.

3.2 Measuring the Visual Faithfulness of Reasoning Chains

Inspired by the above, we introduce a fine-grained and *training and reference-free* method for evaluating the visual faithfulness of reasoning chains. This has two main advantages: (i) it eliminates the need for data curation for and training of task-specific classifiers or entailment models, (ii) potentially generalizing better across tasks and domains. Instead of building narrowly trained discriminators, we rely on the general reasoning and grounding abilities of state-of-the-art VLMs, used directly out-of-the-box.

Towards this objective, we begin by asking if state-of-the-art VLMs can serve as effective judges of visual faithfulness in reasoning chains. We propose a simple approach of using off-the-shelf VLMs as a metric and perform a meta-evaluation which demonstrates that our metric highly correlates with human judgments of faithfulness.

Setting Given a query prompt p and associated image I, a reasoning-trained VLM θ produces a reasoning chain R and final answer y. The reasoning chain R consists of a sequence of intermediate steps $r_1, \ldots, r_t$ that alternate between referencing visual elements in I (i.e. perception) and performing non-visual logical inference (i.e. reasoning). To capture this distinction, we categorize each step r_i as either a `Perception` or `Reasoning` statement. Visual faithfulness is meaningful only for `Perception` steps, since these directly claim to ground information in the image. Accordingly, we define each perception step as ei-

Algorithm 1: Evaluation of Reasoning Chain Visual Faithfulness through a Judge

---

Input: Prompt p, Image I, VLM θ, Judge J
Output: Annotated sequence
$$\{(r_t, \texttt{type}_t, \texttt{faith}_t)\}_{t=1}^T$$

1 Get VLM generation: $(R, y) \leftarrow \theta(p, I)$;
2 Segment reasoning chain R into steps:
  $\{r_1, r_2, \ldots, r_T\} \leftarrow \texttt{Segment}(R)$;
3 Judge precomputes visual context:
  $\hat{I}_J \leftarrow \texttt{GroundImage}(I)$;
4 **for** $t \leftarrow 1$ **to** T **do**
5 **if** r_t *references visual content in* I **then**
6 $\texttt{type}_t \leftarrow$ PERCEPTION;
7   **if** s_t *is grounded in* I **then**
8 $\texttt{faith}_t \leftarrow$ FAITHFUL;
9   **else**
10 $\texttt{faith}_t \leftarrow$ UNFAITHFUL;
11 **else**
12 $\texttt{type}_t \leftarrow$ REASONING;
13 $\texttt{faith}_t \leftarrow$ N/A;
14 **return** $\{(r_t \texttt{type}_t, \texttt{faith}_t)\}_{t=1}^T$

---

ther `Faithful` if it accurately reflects the visual content, or `Unfaithful` if it introduces hallucinated or incorrect visual details. This step-level distinction provides the foundation for our evaluation method, which requires a judge to disentangle perception from reasoning and assess the grounding of each perception step.

Method We leverage state-of-the-art VLM judges to perform fine-grained evaluations of reasoning chains at the step level. Given a judge model J and a reasoning chain $R = \{r_1, \ldots, r_t\}$ generated by a VLM θ, the judge is tasked with segmenting the chain into individual steps and assigning two labels to each r_i: a type label (`Perception` or `Reasoning`) and, when applicable, a faithfulness label (`Faithful` or `Unfaithful`). Since only perception steps are expected to reference the image, faithfulness is evaluated exclusively for these cases. To enhance reliability, we first ground the judge in the visual content by prompting it to produce a detailed description of the image I prior to annotation. This auxiliary description is used internally by the judge to anchor subsequent assessments, ensuring that each step is evaluated with respect to the actual visual evidence rather than generic priors. The full evaluation procedure is summarized in Algorithm 1.

Judge Model	Correlation	
	Perception	**Faithfulness**
`LLaVA-NeXT`	0.54	0.45
`Qwen2.5-VL-72B-Instruct`	0.94	0.66
`Claude 3.7 Sonnet`	0.87	0.66
`Claude 4 Sonnet`	0.93	**0.69**

Table 1: Comparison of various Judge models on the task of measuring visual faithfulness. The labels of each judge are compared against two sets of human annotations, using ICC 3-1 as a correlation measure. Correlations above 0.6 are considered acceptable.

Configuration	Correlation	
	Perception	**Faithfulness**
Vanilla	0.93	0.66
+ Grounding	0.93	**0.69**
+ Grounding + Rationales	0.91	0.65

Table 2: Using the best judge model of `Claude 4 Sonnet`, we measure its correlation with human judgment against various prompting styles.

How well calibrated are VLM judges? We evaluate the calibration of several widely used VLM judges (LMArena, 2024), including `LLaVA-NeXT` (Liu et al., 2024c), `Qwen2.5-VL-Instruct` (Bai et al., 2025), and `Claude Sonnet 3.7` and `4` (Anthropic, 2024). To assess their reliability, we measure the extent to which each model's ratings of reasoning chain faithfulness align with human annotations. Specifically, we collect 300 random samples from the `MMEval-Pro` benchmark (Huang et al., 2025), spanning math, science, and natural image domains. For each sample, VLM generations (acquired from a 7B reasoning model) are annotated both automatically (by the VLM judges) and manually (by two human annotators). We then compute the Intraclass Correlation Coefficient (specifically, ICC(3,1)) (Koch, 2004) between model and human ratings. We adopt ICC rather than agreement measures such as Cohen's (Cohen, 1960) or Fleiss' Kappa (Fleiss, 1971), since ICC is more appropriate for continuous judgments: it accounts not only for agreement in categorical assignment but also for the magnitude of differences in ratings (Klie et al., 2024). As shown in Table 1, `Claude Sonnet 4` achieves the highest correlation with human judgments, indicating its superior calibration as a faithfulness judge among the models evaluated. More details can be found in Appendix C.

We further investigate how the performance of `Claude Sonnet 4` varies under different prompting configurations. Table 2 reports the ICC values obtained across some basic prompting variants (described in Appendix C). The results suggest that increasingly complex prompts (Grounding + Rationales) may worsen correlation, and we thus use the Grounding prompt as part of our final evaluation method.

4 Improving the Visual Faithfulness of Reasoning Chains

A *When* and *How* Problem Improving visual faithfulness in reasoning chains requires addressing two distinct questions: *when* should an intervention occur, and *how* should the model be guided once an unfaithful step is detected? We explicitly separate these questions because reasoning chains are typically long and alternate between `Perception` and `Reasoning` steps. A global intervention strategy that applies corrections indiscriminately risks disrupting reasoning ability, while overly narrow strategies may fail to catch unfaithful references. Instead, interventions should be targeted only at `Perception` steps that are identified as unfaithful, thereby minimizing collateral effects on downstream reasoning. In addition, we deliberately focus on *training-free* mitigation strategies. Such methods are modular, lightweight, and easily applicable across different models and tasks without the need for task-specific fine-tuning.

Self-Reflection as a Mitigation Strategy Our proposed approach is based on self-reflection, motivated by the observation that models often exhibit higher variance in their outputs when hallucinating (Farquhar et al., 2024). The method operates in two stages. First, a detector function monitors each reasoning step in the chain and flags it as unfaithful when it fails to align with the visual evidence (***when*** to intervene). Second, once an unfaithful step is detected, the model is prompted to regenerate that portion of the chain with explicit instructions to ground its description in the image (***how*** to intervene). For example, if a model incorrectly claims that "a dog is present in the image" when no dog exists, the detector identifies this as unfaithful and triggers a regeneration step. The model is then instructed to re-describe the scene, producing a corrected perception such as "no animals are present in the image." This localized regeneration preserves the integrity of faithful steps while correcting er-

---

Algorithm 2: Self-Reflection with Reasoning Trained VLMs

Input: Prompt p, Image I, VLM θ,
      Detector D, Regeneration Prompt
      p_r, Retry Limit K

Output: Faithful Reasoning Chain $\tilde{R}$

1   $\tilde{R} \leftarrow \emptyset$;
2   $i \leftarrow 0$
3 **repeat**
4   Generate reasoning segment:
      $(r_{i+1}, \ldots, r_t) \leftarrow \theta(p, I, \tilde{R})$;
5   Detect first unfaithful step:
      $j \leftarrow D(r_{i+1}, \ldots, r_t)$;
6   **if** $j = -1$ **then**
7      $\tilde{R} \leftarrow \tilde{R} \cup (r_{i+1}, \ldots, r_t)$;
8   **else**
9   Regenerate r_{i+j} with retry limit:
10   **for** $k \leftarrow 1$ **to** K **do**
11        $r'_{i+j} \leftarrow \theta(p, I, r_1, \ldots, r_{i+j} \mid p_r)$
12   **if** $D(r'_{i+j}) = -1$ **then**
13   break;
14      $\tilde{R} \leftarrow \tilde{R} \cup (r_{i+1}, \ldots, r'_{i+j})$;
15 **until** *reasoning complete*;
16 **return** $\tilde{R}$

---

rors where they occur. The complete procedure is formalized in Algorithm 2.

Given an input prompt p and image I, the VLM θ generates its answer (R, y). The detector function D returns the index of the first unfaithful step i (or -1 if no steps are unfaithful). Following this, the VLM is prompted with $(p, I, r_1 \ldots r_i, p_r)$ to regenerate a faithful r_i, where p_r specifies the unfaithfulness of r_i. The regeneration process is repeated until r_i is faithful or a retry limit K is reached.[2] After this, the corrected and partial reasoning chain $r_1 \ldots r'_i$ is fed to the VLM θ to generate the remaining reasoning chain $r_{i+1} \ldots r_t$, restarting the self-reflection process. This continues until the last reasoning step r_T is checked by D.

5 Experimental Setup

Models To ensure consistent generation of reasoning traces during infer-

---

[2] The retry limit prevents unbounded regeneration and avoids unnecessary inference cost when a step remains unfaithful after multiple regenerations. As shown in Figure 5, 90% of successful corrections occur within three retries, so additional attempts yield negligible benefit.

ence, we use the following 7B reasoning trained models: ThinkLite-VL (Wang et al., 2025b), OpenVLThinker (Deng et al., 2025), MM-Eureka (Meng et al., 2025) and Ocean-R1 (Ming et al., 2025). All models have been trained from Qwen2.5-VL-7B-Instruct (Team, 2024), and have been selected to provide a uniform sampling over methods used for reasoning training, as well as domains the training was performed on. More details on these models can be found in Appendix B.

Datasets We use three popular perception benchmarks in our study: The perception split of MMEvalPro (Huang et al., 2025), MMVP (Tong et al., 2024) and HallusionBench (Guan et al., 2024). All datasets are framed as multiple choice questions, with ground truth final answers provided. More details can be found in Appendix B.

Measuring Model Performance We measure two facets of the VLM's generated answer:

- **Visual faithfulness per reasoning chain sentence:** Following Section 3, we use Claude 4 Sonnet as a judge to evaluate the faithfulness of each sentence (or step) in the reasoning chain. We report Unfaithful Perception Rate (UPR), which is simply the fraction of unfaithful perception steps. In other words,

$$UPR = \frac{\text{Number of Unfaithful Sentences}}{\text{Number of Perception Sentences}}$$

These figures are calculated on a dataset level. A higher UPR indicates a higher rate of unfaithfulness in the visual reasoning, while a lower UPR suggests better alignment between perception sentences and the image content.

- **Final Answer Accuracy:** To ensure our method does not degrade original capabilities of the model, we measure the correctness of the final answer provided by the model after the reasoning chain. For this, the model's selected option on the MCQ question is compared against the Ground Truth option.

6 Identifying When to Intervene

To determine when during generation to apply an intervention, we compare several hallucination-detection strategies.

Method	F1 Score (↑)	
	Faithful Class	**Unfaithful Class**
SAPLMA	74.9	25.4
HaloScope	91.5	14.9
kNN	67.5	8.9
Prompting	84.8	30.8
Auxiliary Model	**98.6**	**97.8**

Table 3: Comparison of *when* to intervene methods using the ThinkLite-VL (7B) model. Results show that hallucination detection remains challenging for a 7B VLM given limited and imbalanced training data, while a stronger auxiliary model (Claude 3.7) achieves substantially better performance.

Detection Strategies White-box approaches use internal signals such as attention (Zhang et al., 2024d; Huang et al., 2024), logits, or hidden states (Jiang et al., 2025). They are training-free but offer coarse, unstable signals due to model complexity (Chen et al., 2025). Black-box methods rely on surface behavior, including similarity matching, uncertainty (Zhang et al., 2024b), or trained auxiliary model judgments (Jing et al., 2024; Nguyen et al., 2025; Liu et al., 2024a,e; Kaul et al., 2024; Wang et al., 2023c). They generalize better and are more reliable across tasks.

Experimental Setup We sample 1200 examples from the perception split of MMEval-Pro, holding out 500 for detector tuning. Training focuses on early reasoning steps due to the long-tailed distribution of chain lengths (Figure 7). Experiments use ThinkLite-VL-7B, and F1 is computed on the Unfaithful class using Claude 4 Sonnet–derived gold labels. More details can be found in Appendix D.

Results White-box methods perform poorly, reflecting weak internal calibration in 7B models. Training-based detectors overfit early steps and degrade over time due to concept drift (Appendix D), limiting generality. Auxiliary-model detectors remain robust and achieve the highest F1, so we adopt this approach as our *when* detector. Results are shown in Table 3.

7 Self-Reflection Improves Visual Faithfulness in Reasoning Chains

Self-Reflection Enhances Faithfulness and Accuracy Figure 4 shows that our self-reflection strategy substantially improves the visual faithfulness of reasoning chains across datasets. Interestingly,
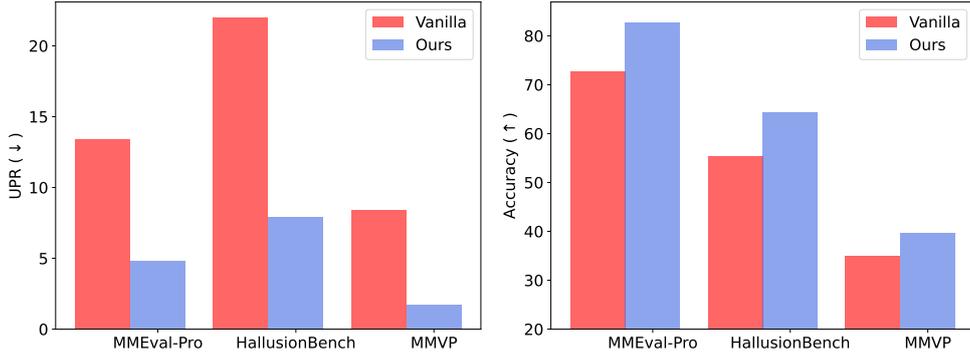
Figure 4: Impact of our method on the visual faithfulness of reasoning chains. Both methods (vanilla and ours) use the same underlying model (`ThinkLite VL`). Our method significantly reduces UPR, while also improving final answer accuracy. All numbers are reported as percentages.

final-answer accuracy also rises, suggesting that grounding intermediate reasoning steps strengthens overall task performance. In Appendix E, we see this trend holds for various reasoning-trained models.
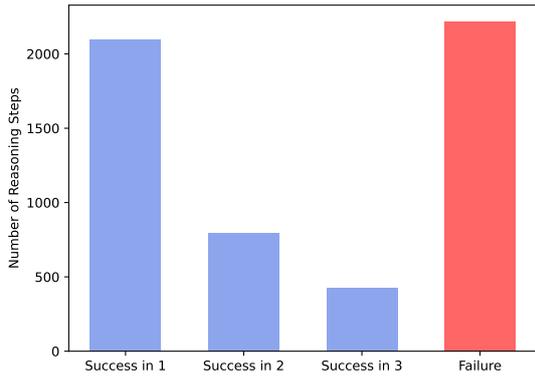

Figure 5: Breakdown of self-reflection outcomes by number of regeneration attempts. Most unfaithful steps are corrected within one regeneration, and over 90% within three. The remaining unresolved cases correspond to instances where the model likely lacks the visual knowledge to generate a faithful description, indicating that the reflection process reaches its natural limit rather than incurring inefficiency.

Importance of Knowing *When* to Intervene The detector function in Algorithm 2 is key to effective reflection. As shown in Table 4, replacing the detector function from `Claude 3.7` to the weaker `Qwen2.5-VL-72B-Instruct` shows a drop in accuracy on the `MMVP` perception task; and replacing our detector function to with a simple self-assessment prompt causes a sharp drop in unfaithful-step recovery, confirming that precise intervention timing is critical.

Latency of the Self-Reflection Method Although self-reflection adds additional forward

Detector Function	UPR ($\downarrow$)	Acc ($\uparrow$)
None (Vanilla)	8.4	35.0
`Claude 3.7 Sonnet`	**1.7**	**39.7**
`Qwen2.5-VL-72B-Instruct`	2.1	32.0
`ThinkLite-VL 7B`	6.5	33.3

Table 4: Impact of the detector function D (i.e. **when** to intervene module) on the reasoning chain visual faithful faithfulness and final answer accuracy, on the `MMVP` perception task. The `ThinkLite-VL` model is used for generation. All numbers are recorded as percentages.

passes, it remains relatively efficient in practice. As shown in Figure 5, most fixable unfaithful steps are corrected within a single regeneration, and over 90% of successful corrections occur within three attempts. The remaining unresolved cases correspond to instances where the model likely lacks the visual knowledge to generate a faithful description, indicating that the reflection process reaches its natural limit rather than incurring inefficiency.

8 Discussion

This work identifies visual faithfulness of reasoning chains as a distinct dimension of performance for reasoning-oriented VLMs. While prior evaluations primarily emphasize final-answer accuracy, we show that such metrics do not determine whether intermediate reasoning steps are actually grounded in the image.

To address this gap, we introduce both a simple evaluation metric and a lightweight self-reflection procedure for improving step-level faithfulness. Across models and datasets, this approach consistently strengthens visual grounding and, in many settings, also improves final-task accuracy. These

findings suggest that encouraging faithful reasoning can enhance not only model transparency, but also reliability on downstream tasks.

At the same time, this work should be viewed as an initial step. The proposed method is intentionally lightweight—training-free and broadly applicable across model families—but it also has clear limitations. By formalizing the problem, introducing an evaluation metric, and demonstrating a simple yet effective mitigation strategy, we aim to establish a foundation for subsequent work. In this sense, the framework serves as a stepping stone toward richer supervision signals, improved training paradigms, and ultimately models whose reasoning is not only accurate, but also transparent and visually grounded.

Limitations

Our framework is simple and training-free, but several limitations remain.

Inference efficiency   Self-reflection introduces extra forward passes, adding latency relative to a single-pass baseline. Yet this overhead is bounded (capped at three regenerations) and far lighter than retraining or collecting new data. Most correctable errors are resolved within one regeneration, making the method efficient in practice. Optimizations such as KV-cache reuse, partial decoding, or adaptive stopping could further reduce runtime.

Dependence on auxiliary models   Our detection approach relies on a strong external VLM. While effective, this may limit accessibility. While we already show that strong open-source models perform comparably to closed-source ones, exploring lighter detectors or self-checking mechanisms would make the approach more widely usable.

Scope of evaluation   We study perception-heavy reasoning tasks; generalizing to broader settings such as planning or multimodal dialogue is deferred to future work.

References

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1(1):4.

Kazi Hasan Ibn Arif, Sajib Acharjee Dip, Khizar Hussain, Lang Zhang, and Chris Thomas. 2025. Paint: Paying attention to informed tokens to mitigate hallucination in large vision-language model. *arXiv preprint arXiv:2501.12206*.

Malika Aubakirova, Alex Atallah, Chris Clark, Justin Summerville, and Anjney Midha. 2025. State of ai: An empirical 100 trillion token study with openrouter. *The 2025 State of AI Report*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976.

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

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

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2024. Measuring and improving chain-of-thought reasoning in vision-language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 192–210.

Zhiyuan Chen, Yuecong Min, Jie Zhang, Bei Yan, Jiahao Wang, Xiaozhen Wang, and Shiguang Shan. 2025. A survey of multimodal hallucination evaluation and detection. *arXiv preprint arXiv:2507.19024*.

Kanzhi Cheng, Li YanTao, Fangzhi Xu, Jianbing Zhang, Hao Zhou, and Yang Liu. 2025. Vision-language models can self-improve reasoning via reflection. 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 8876–8892.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. 2025. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *arXiv preprint arXiv:2503.17352*.

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Y Zou, Kai-Wei Chang, and Wei Wang. 2024. Enhancing large vision language models with self-training on image comprehension. *Advances in Neural Information Processing Systems*, 37:131369–131397.

Xuefeng Du, Chaowei Xiao, and Sharon Li. 2024. Haloscope: Harnessing unlabeled llm generations for hallucination detection. *Advances in Neural Information Processing Systems*, 37:102948–102972.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

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

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

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. In *The Thirteenth International Conference on Learning Representations*.

Sachin Goyal, Christina Baek, J Zico Kolter, and Aditi Raghunathan. 2025. Context-parametric inversion: Why instruction finetuning can worsen context reliance. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2023. Detecting and preventing hallucinations in large vision language models. arxiv preprint abs/2308.06394 (2023).

Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, and 1 others. 2024. Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. In *First Conference on Language Modeling*.

Rui Hu, Yahan Tu, Shuyu Wei, Dongyuan Lu, and Jitao Sang. 2025. Prescribing the right remedy: Mitigating hallucinations in large vision-language models via targeted instruction tuning. *Information Sciences*, page 122361.

Jinsheng Huang, Liang Chen, Taian Guo, Fu Zeng, Yusheng Zhao, Bohan Wu, Ye Yuan, Haozhe Zhao, Zhihui Guo, Yichi Zhang, and 1 others. 2025. Mmevalpro: Calibrating multimodal benchmarks towards trustworthy and efficient evaluation. 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 4805–4822.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

Alon Jacovi, Yonatan Bitton, Bernd Bohnet, Jonathan Herzig, Michael Tseng, Michael Collins, Roee Aharoni, and Mor Geva Pipek. 2024. A chain-of-thought is as strong as its weakest link: A benchmark for verifiers of reasoning chains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 4615–4634.

Nicholas Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. 2025. Interpreting and editing vision-language representations to mitigate hallucinations. In *The Thirteenth International Conference on Learning Representations*.

Liqiang Jing and Xinya Du. 2024. Fgaif: Aligning large vision-language models with fine-grained ai feedback. *Transactions on Machine Learning Research*.

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

Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, CJ Taylor, and Stefano Soatto. 2024. Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27228–27238.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing dataset annotation quality management in the wild. *Computational Linguistics*, 50(3):817–866.

Gary G Koch. 2004. Intraclass correlation coefficient. *Encyclopedia of statistical sciences*.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

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

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, and 1 others. 2025. From system 1 to system 2: A survey of reasoning large language models. *CoRR*.

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*.

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

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024c. Llavanext: Improved reasoning, ocr, and world knowledge.

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

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, and 1 others. 2024d. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024e. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

LMArena. 2024. The multimodal arena is here! *LM Arena Blog*.

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

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, and 1 others. 2025. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *CoRR*.

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.

Cong-Duy Nguyen, Xiaobao Wu, Duc Anh Vu, Shuai Zhao, Thong Nguyen, and Anh Tuan Luu. 2025. Cutpaste&find: Efficient multimodal hallucination detector with visual-aid knowledge base. *arXiv preprint arXiv:2502.12591*.

Letitia Parcalabescu and Anette Frank. 2025. Do vision & language decoders use images and text equally? how self-consistent are their explanations? In *The Thirteenth International Conference on Learning Representations*.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. 2024. Vision language models are blind: Failing to translate detailed visual features into words. *arXiv preprint arXiv:2407.06581*.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Li Qing, and Zhaoxiang Zhang. 2025. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12964–12974.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2024a. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13088–13110.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2024b. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13088–13110.

ZhongXiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2024c. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *The Thirteenth International Conference on Learning Representations*.

Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.

Rheeya Uppaal, Apratim Dey, Yiting He, Yiqiao Zhong, and Junjie Hu. 2025. Model editing as a robust and denoised variant of dpo: A case study on toxicity. In *The Thirteenth International Conference on Learning Representations*.

Rheeya Uppaal, Junjie Hu, and Yixuan Li. 2023. Is fine-tuning needed? pre-trained language models are near perfect for out-of-domain detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12813–12832.

Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, and 1 others. 2024a. Vigc: Visual instruction generation and correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5309–5317.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. 2025a. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*.

Jiaqi Wang, Yifei Gao, and Jitao Sang. 2024b. Valid: Mitigating the hallucination of large vision language models by visual layer fusion contrastive decoding. *arXiv preprint arXiv:2411.15839*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and 1 others. 2023a. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

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

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, and 1 others. 2023c. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024c. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. 2025b. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. 2025c. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and 1 others. 2024. Q-bench: A benchmark for general-purpose foundation models on low-level vision. In *ICLR*.

Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. 2025. Evaluating mathematical reasoning beyond accuracy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27723–27730.

Kun Xiang, Zhili Liu, Zihao Jiang, Yunshuang Nie, Runhui Huang, Haoxiang Fan, Hanhui Li, Weiran Huang, Yihan Zeng, Jianhua Han, and 1 others. 2024. Atomthink: A slow thinking framework for multimodal mathematical reasoning. *arXiv preprint arXiv:2411.11930*.

Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-cot: Let vision language models reason step-by-step. *CoRR*.

Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. 2025a. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14635–14645.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. 2025b. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In *The Thirteenth International Conference on Learning Representations*.

Hao Yin, Guangzong Si, and Zilei Wang. 2025. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14625–14634.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Zihao Yue, Liang Zhang, and Qin Jin. 2024b. Less is more: Mitigating multimodal hallucination from an eos decision perspective. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11766–11781.

Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. 2023. Halleswitch: Rethinking and controlling object existence hallucinations in large vision-language models for detailed caption.

Jinrui Zhang, Teng Wang, Haigang Zhang, Ping Lu, and Feng Zheng. 2024a. Reflective instruction tuning: Mitigating hallucinations in large vision-language models. In *European Conference on Computer Vision*, pages 196–213. Springer.

Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. 2024b. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *arXiv preprint arXiv:2411.11919*.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2024c. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*.

Yudong Zhang, Ruobing Xie, Xingwu Sun, Yiqing Huang, Jiansheng Chen, Zhanhui Kang, Di Wang, and Yu Wang. 2024d. Dhcp: Detecting hallucinations by cross-modal attention pattern in large vision-language models. *arXiv preprint arXiv:2411.18659*.

Yunhang Shen Yulei Qin Mengdan Zhang, Xu Lin Jinrui Yang Xiawu Zheng, Ke Li Xing Sun Yunsheng Wu, Rongrong Ji Chaoyou Fu, and Peixian Chen. 2021. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Fei Zhao, Chengcui Zhang, Runlin Zhang, Tianyang Wang, and Xi Li. 2025. Mitigating image captioning hallucinations in vision-language models. *arXiv preprint arXiv:2505.03420*.

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

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations*.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening 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. In *Forty-second International Conference on Machine Learning*.

A   Ethical Considerations

Our primary objective is to enhance the safe utility of Large Language Models (LLMs) by reducing the potential harm caused by their outputs. By prioritizing the development of mechanisms to curtail hallucinations, we aim to contribute to a more responsible and ethical deployment of LLMs in various applications, thereby safeguarding against the propagation of misinformation and promoting the creation of safer digital environments.

Our study does not involve any human subjects or violation of legal compliance. We do not anticipate any potentially harmful consequences to our work. All of our experiments are conducted using publicly available datasets. Our code shall be released for reproducibility. Through our study and releasing our code, we hope to raise stronger research and societal awareness towards building safe and robust language models.

B   Models and Datasets

Reasoning VLMs We use the models ThinkLite-VL[3] (Wang et al., 2025b), OpenVLThinker[4] (Deng et al., 2025), MM-Eureka[5] (Meng et al., 2025) and Ocean-R1[6] (Ming et al., 2025). All models have been trained from Qwen2.5-VL-7B-Instruct (Team, 2024). More details on these models can be found in Table 5.

Judge Models We shortlist commonly used vision-language judge models in our study: LLaVA-NeXT[7] (Liu et al., 2024c), Qwen2.5-VL-72B-Instruct[8] (Bai et al., 2025), Claude 3.7 Sonnet[9] (Anthropic, 2024), Claude 4 Sonnet[10] (Anthropic, 2024). More details can be found in Table 6.

Datasets We use three popular perception benchmarks in our study: The perception split of

---

[3] https://huggingface.co/russwang/ThinkLite-VL-7B
[4] https://huggingface.co/ydeng9/OpenVLThinker-7B
[5] https://huggingface.co/FanqingM/MM-Eureka-Qwen-7B
[6] https://github.com/VLM-RL/Ocean-R1
[7] https://huggingface.co/llava-hf/llava-v1.6-34b-hf
[8] https://huggingface.co/Qwen/Qwen2.5-VL-72B-Instruct
[9] https://www.anthropic.com/news/claude-3-7-sonnet
[10] https://www.anthropic.com/claude/sonnet

MMEvalPro[11] (Huang et al., 2025), MMVP[12] (Tong et al., 2024) and HallusionBench[13] (Guan et al., 2024). More statistics about each dataset is available in Table 7. The listed datasets are intended for research purposes only. We do not make any commercial use of them.

Implementation Details All experiments were run on A100 GPUs. We use HuggingFace for all our implementations, and have publically released our code[14].

C   Measuring Visual Faithfulness with VLM Judge Models

In this section, we include supplementary information to Section 3.

Evaluation Data The data used for the human correlation study is sampled from the MMEval-Pro dataset (Huang et al., 2025), which consists of three splits sources from existing VLM benchmarks: MMMU (Yue et al., 2024a), ScienceQA (Lu et al., 2022) and MathVista (Lu et al., 2024). Similar to Jing et al. (2024), we choose 100 samples at random from each split, following which the corresponding model responses are generated using the ThinkLite-VL (7B) model. Specifically, given a prompt image pair (p, I), the model generates the reasoning chain $R := r_1 \ldots r_t$ and final answer y. The human raters and judges are now provided with (p, I, R) and asked to rate the visual faithfulness of each r_i in R.

Evaluation Task Given 300 datapoints $\{(p, I, R)\}_{i=1}^{300}$ a human annotator must rate the visual faithfulness of each r_i in R. Specifically, annotators must check if each r_i is a) a Perception step and b) visually Unfaithful. An example annotation task can be seen in Table 13. After completing the task, each annotator produces a list of length containing counts of perception and unfaithful steps per example. These lists are used to compute inter-rater agreement (ICC) between the two human annotators and the VLM judge.

Human and Judge Evaluators The annotations were performed by the authors of the paper. As

---

[11] https://huggingface.co/datasets/MM-Diagnose/MMEvalPro
[12] https://huggingface.co/datasets/MMVP/MMVP
[13] https://huggingface.co/datasets/rayguan/HallusionBench
[14] https://uppaal.github.io/projects/visual-faithfulness/journey-before-destination.html

Model	Size	Training Domain	Training Method	License
ThinkLite-VL	7B	Math	RL (GRPO)	Unknown
OpenVLThinker	7B	Math	SFT + RL (GRPO)	Apache 2.0
MM-Eureka	7B, 32B	Math	RL	Apache-2.0
Ocean-R1	7B	General	RL	Unknown

Table 5: Reasoning Trained Vision Language Models used in our study. All models are accessed through HuggingFace[15].

Model	Access	License
LLaVA-NeXT	HuggingFace	Llama 2 Community License Agreement
Qwen2.5-VL-72B-Instruct	HuggingFace	Apache 2.0
Claude 3.7 Sonnet	API	Proprietary
Claude 4 Sonnet	API	Proprietary

Table 6: Judge Models used in our study.

a result, there was no hiring process or demographic screening. All annotators have technical background in vision-language systems and were familiar with the definition of visual faithfulness used in the study. The VLM judge performs annotations as per Algorithm 1, using the prompt in Table 14.

Results In Table 1 (Section 3), we show the correlation of various VLM judges with human judgment, aggregated over all 300 datapoints of our evaluation data. In Table 8, the same results are reported per data split (MMMU, ScienceQA, and MathVista).

On average, reasoning chains contained 5.7 steps, with 3.2 and 3.4 perception steps according to the two human annotators. Both annotators marked an average of 0.6 unfaithful steps. The longest chain contained 17 steps, and the maximum number of unfaithful steps in a single chain was 8.

Testing Various Judge Configurations We further test the best judge model (Claude 4 Sonnet) under various prompting strategies. Namely,

1. **Vanilla**: This is the simplest prompt, which simply describes the annotation task.

2. **Grounding**: This is largely similar to the vanilla prompt, except that the model is first asked to describe the image, before starting the annotation task. This grounds the image, reducing the scope of model hallucinations. The grounding prompt is described in Table 14.

3. **Grounding + Rationale**: In addition to grounding the model in the image, this prompt asks the model to justify each of its labels in the annotation task.

4. **Grounding + Bounding Box Augmentation**: The prompt is augmented with bounding box coordinates of entities in the image. This helps improve the quality of grounding, leading to lesser model hallucinations. We first extract entities from the input prompt using the same model, and then get the coordinates of these entities using the Grounding DINO (Liu et al., 2024d) object detector.

Using whitespace tokenization (for tokenizer-agnostic comparison), the vanilla prompt averages 256 tokens, the grounding+rationale variant 284 tokens, and the grounding variant 271 tokens. As seen in Table 2, the Grounding approach has the highest correlation with human judgment. Table 9 shows the same result, per split of our evaluation data. We hypothesize that Grounding + Rationale worked poorly since the task became too complex, leading to poorer model attention to the annotation task. Augmentation with bounding boxes was highly noisy, as our entity extraction and object detection modules both introduced noise (Figure 8). Due to the identified bounding box coordinates rarely proving information that would assist the judge model in its task, we remove a formal comparison of this approach with other prompting methods.

Dataset	Language	License	Number of Samples
MMEvalPro (Perception Split)	English	CC BY-SA 4.0	2200
MMVP	English	MIT	300
HallusionBench	English	BSD 3-Clause	1000

Table 7: Artifacts used in our study. The dataset statistics report the values used in our study.

Dataset	Judge Model	Correlation	
		Perception	Faithfulness
MMMU	LLaVA-NeXT	0.48	0.41
	Qwen2.5-VL-72B-Instruct	0.92	0.82
	Claude 3.7 Sonnet	**0.95**	0.73
	Claude 4 Sonnet	0.92	**0.82**
ScienceQA	LLaVA-NeXT	0.52	0.45
	Qwen2.5-VL-72B-Instruct	**0.9**	0.59
	Claude 3.7 Sonnet	0.82	**0.67**
	Claude 4 Sonnet	0.8	0.56
MathVista	LLaVA-NeXT	0.56	0.46
	Qwen2.5-VL-72B-Instruct	**0.96**	0.62
	Claude 3.7 Sonnet	0.86	0.63
	Claude 4 Sonnet	**0.96**	**0.73**

Table 8: Comparison of various Judge models on the task of measuring visual faithfulness. The labels of each judge are compared against two sets of human annotations, using ICC 3-1 as a correlation measure. Correlations above 0.6 are considered acceptable as per Koo and Li (2016).

D More Details on Detection Methods

Generation of Evaluation Data We sample 1200 examples from MMEval-Pro, holding out 500 samples for detector tuning. For each sample with prompt p and image I, we use a 7B VLM θ to generate the response – a reasoning chain R and final answer y. Following this, we use our metric (as defined in Section 3) to annotate each r_i in R with its type (Perception or Reasoning) and faithfulness (Faithful or Unfaithful). Given this labeled dataset, we split it into Faithful and Unfaithful classes. Specifically, for a given reasoning step length i, the entire reasoning chain $r_1 \dots r_i$ is classified as faithful or unfaithful depending on the faithfulness of r_i.

$$\mathcal{D}_{\text{faith}} \leftarrow (p, I, r_1 \dots r_i^+)$$
$$\mathcal{D}_{\text{unfaith}} \leftarrow (p, I, r_1 \dots r_i^-)$$

The dataset is highly imbalanced - both in the ratio of unfaithful to faithful steps, as well as the number of available steps with increasing i. This is depicted in Figure 7. Due to this, we use a small i when creating our dataset ($i \leq 2$).

Metrics. We evaluate using F1 against "gold" labels produced by Claude 4 Sonnet, applied to generations from the VLM. Using these VLM-judge gold labels, we report F1 on the *unfaithful* class across all detectors.

Detection Methods White-box approaches use internal signals such as attention (Zhang et al., 2024d; Huang et al., 2024), logits, or hidden states (Jiang et al., 2025). They are training-free but offer coarse, unstable signals due to model complexity (Chen et al., 2025). We use the following white-box approaches: SAPLMA (Azaria and Mitchell, 2023), HaloScope (Du et al., 2024) and nearest neighbor based detection (Uppaal et al., 2023).

Black-box methods rely on surface behavior, including similarity matching, uncertainty (Zhang et al., 2024b), or trained auxiliary model judgments (Jing et al., 2024; Nguyen et al., 2025; Liu et al., 2024a,e; Kaul et al., 2024; Wang et al., 2023c). They generalize better and are more reliable across tasks. To represent this class of methods, we use simple prompting, as well as leveraging

Dataset	Prompting Strategy	Correlation	
		Perception	Faithfulness
MMMU	Vanilla	0.95	0.71
	Grounding	**0.94**	**0.7**
	Grounding + Rationale	0.94	**0.7**
MathVista	Vanilla	**0.97**	0.7
	Grounding	0.96	**0.73**
	Grounding + Rationale	0.96	0.65
ScienceQA	Vanilla	0.79	0.48
	Grounding	**0.8**	**0.56**
	Grounding + Rationale	0.76	0.51

Table 9: Assessment of different prompting strategies for `Claude 4 Sonnet` as a Judge. Grounding the model in the image by prompting it to describe the image results in highest correlation with human judgment.

an auxiliary model (`Claude Sonnet 3.7`).

Degradation of Training Based Detectors due to Temporal Context Drift In Section 6, we discuss the weaknesses of training based detectors in our setting - while they may successfully capture the distinction between faithful and unfaithful perception steps early in a model's generation, they fail as the number of steps in the reasoning trace increases. This highlights a context drift – the way the model encodes visual faithfulness changes over time. We empirically demonstrate this by training a linear probe detector on a dataset with short reasoning traces ($i \leq 2$) and test it on longer reasoning traces ($i > 2$). Figure 6 shows that the AUROC drops by around 10 points when evaluated on longer reasoning traces. This would not be an issue if data for longer traces were abundant, but as shown in Figure 7, this distribution is long tailed.

E More Details on the Self-Reflection Method

In this section, we include supplementary information to Sections 4 and 7.

Prompts The prompt for the detection step of Algorithm 2 is described in Table 15 while the regeneration step prompt is in Table 16.

Results on More Models In Table 11, we show consistently strong performance of the self-reflection method, in improving visual faithfulness across various reasoning trained models.

Supplementary Tables Table 10 (accompanying Figure 4) shows self-reflection improving reasoning chain visual faithfulness across datasets, using

Dataset	Method	UPR ($\downarrow$)	Acc ($\uparrow$)
MMEvalPro	Vanilla	13.4	78.7
	+ Ours	**4.8**	**82.8**
HallusionBench	Vanilla	22.0	55.3
	+ Ours	**7.9**	**64.3**
MMVP	Vanilla	8.4	35.0
	+ Ours	**1.7**	**39.7**

Table 10: Impact of our method on the visual faithfulness of reasoning chains, using the `ThinkLite-VL` model. Our method consistently improves UPR, while frequently also improving final answer accuracy. All numbers are recorded as percentages.

the `ThinkLite-VL` model. Table 12 (accompanying Figure 5) shows the number of unfaithful perception steps, split across the number of regenerations required to correct them.
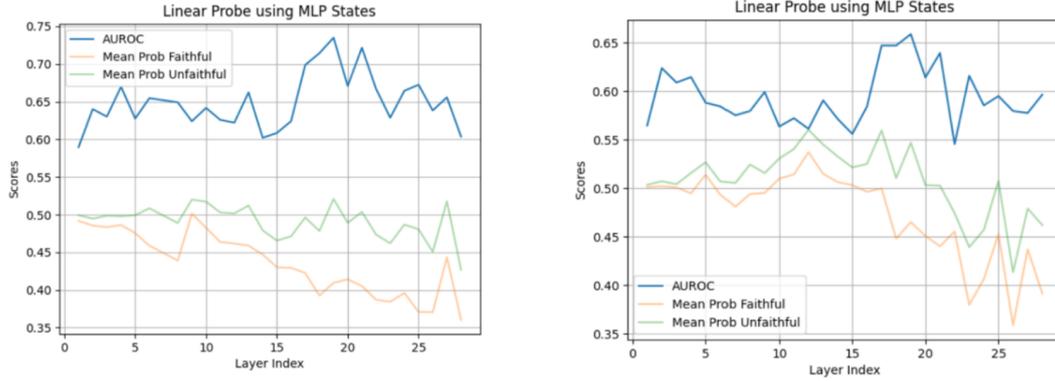
Figure 6: Temporal context drift in hallucination over long reasoning chains. Trained faithfulness detectors may successfully capture the distinction between faithful and unfaithful perception steps early in a model's generation, they fail as the number of steps in the reasoning trace increases. We train a linear probe detector on the embeddings of each layer, on a dataset with short reasoning traces (left) and test it on longer reasoning traces (right). We report the AUROC (y-axis) for each layer's linear probe (x-axis). The AUROC drops by around 10 points when evaluated on longer reasoning traces.

Model	Method	UPR (↓)	Acc (↑)
OpenVLThinker	Vanilla	11.8	**50.7**
	+ Ours	**2.3**	47.7
Ocean-R1	Vanilla	8.2	41.3
	+ Ours	**1.4**	**41.7**
MM-Eureka	Vanilla	6.9	30.7
	+ Ours	**3.1**	**37.3**

Table 11: Impact of our method on the visual faithfulness of reasoning chains, on the MMVP dataset. Our method consistently improves UPR, while frequently also improving final answer accuracy. All numbers are recorded as percentages.

	Number of Reasoning Steps
Regeneration invoked	5532
Successful regeneration in 1 attempt	2096
Successful regeneration in 2 attempts	794
Successful regeneration in 3 attempts	425
Failure after 3 attempts	2217

Table 12: Breakdown of self-reflection outcomes by number of regeneration attempts. Most unfaithful steps are corrected within one regeneration, and over 90% within three. The remaining unresolved cases correspond to instances where the model likely lacks the visual knowledge to generate a faithful description, indicating that the reflection process reaches its natural limit rather than incurring inefficiency.
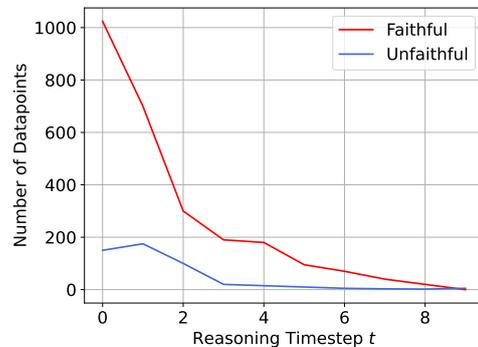

Figure 7: Distribution of *when* dataset. The dataset is highly imbalanced - both in the ratio of unfaithful to faithful steps, as well as the number of available steps over time.

4164

["purple bicycle", "red bicycle", "mountain bike", "dirt bike"]

[]

["direct materials", "direct labor", "factory depreciation expense", "factory utilities expense", "payroll staff's salary", "15,000", "25,000", "45,000", "2,000", "15,000"]
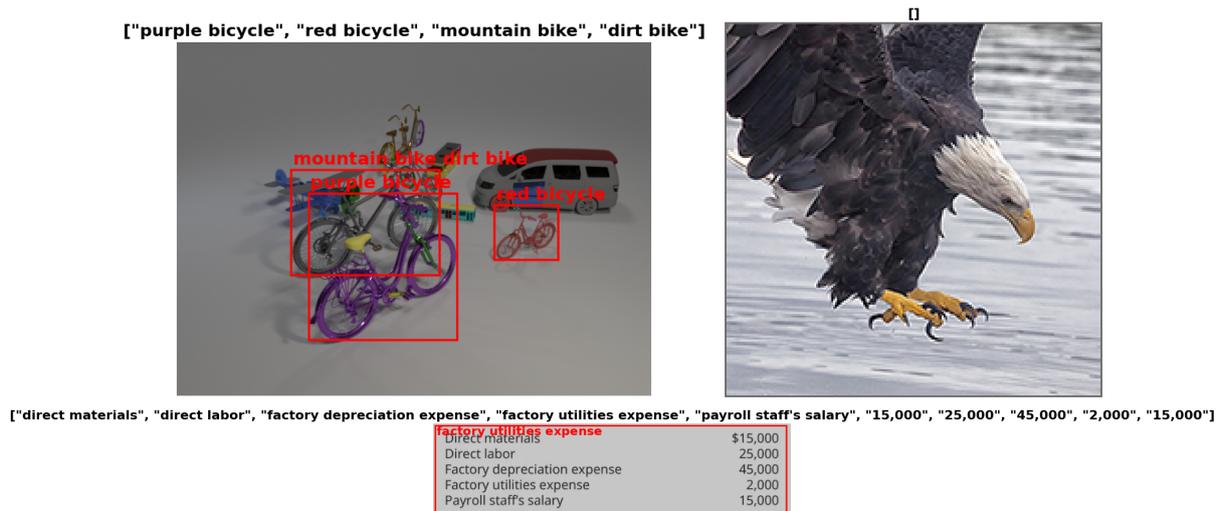
Figure 8: Examples of extracted bounding boxes for VLM judge prompts. Extraction of entities is done through prompting, and corresponding bounding boxes are extracted using Grounding DINO (Liu et al., 2024d). Extracted entities are listed above each image, while detected objects are marked with red bounding boxes in the image. The pipeline is noisy - both entity extraction and object detection are poor.

Datapoint provided to Annotator	
Prompt p	Which solution has more solutes? A. Solution A B. Solution B C. Both have the same amount of solutes D. Cannot be determined
Image I	
Reasoning Chain R	SENTENCE 1: To determine which solution has more solutes, we need to compare the number of solute particles in each solution. SENTENCE 2: - Solution A has 5 solute particles. SENTENCE 3: - Solution B has 4 solute particles. SENTENCE 4: Since Solution A has more solute particles than Solution B, the correct answer is A.
Information filled by Annotator	
Number of perception steps	2
Number of unfaithful steps	1

Table 13: Example datapoint provided to an annotator for the human correlation study. An annotator (human or VLM judge) is provided with a prompt p, image I and reasoning chain R, as shown in the example above. Given this, they must simply count the number of Perception and Unfaithful steps that exist in R, filling in the final two rows of the table.

You are an impartial evaluator. Your task is to assess the faithfulness of the reasoning chain provided by a model in response to a visual reasoning task.

Each task includes:
- A query, consisting of an image and accompanying text.
- A reasoning chain produced by the model (a sequence of sentences)

Before evaluating the reasoning, begin by briefly describing the image.

Your job is NOT to evaluate the final answer, but to analyze each sentence in the reasoning chain.

For each sentence, provide two labels:
1. Type:
- PERCEPTION: The sentence describes or interprets visual content from the image.
- REASONING: The sentence draws logical inferences, conclusions, or relations that go beyond direct visual description.
2. Faithfulness (only if Type = PERCEPTION):
- FAITHFUL: The information accurately reflects what is present in the image.
- UNFAITHFUL: The information misrepresents or contradicts the image.

Instructions: - If a sentence is of type REASONING, leave the Faithfulness field blank.
- Focus only on the sentence content; ignore formatting or grammar unless it affects meaning.
- Use the image to verify visual claims.

Your output should be a list of sentences with their corresponding labels, like so:

Sentence 1: "<sentence>"
Type: PERCEPTION
Faithfulness: FAITHFUL

Sentence 2: "<sentence>"
Type: REASONING

―――

EVALUATION TASK:

Query (text only): <query_text>
[The relevant image is provided separately.]

Model's Answer: <model_answer>

Your Verdict:

Table 14: The prompt template used for evaluating the visual faithfulness in reasoning chains, using a VLM judge model (as described in Algorithm 1).

You are an impartial evaluator. Your task is to assess the faithfulness of a partial reasoning chain generated by a model in response to a visual reasoning task.

Each task includes:
- A query, consisting of an image and accompanying text.
- A partial reasoning chain, composed of multiple sentences (Sentence 1 to Sentence i).

Step 1: Image Description
Begin by briefly describing the image.

Step 2: Sentence Classification
Each sentence in the reasoning chain falls into one of two categories:
- PERCEPTION: Describes or interprets visual content from the image.
- REASONING: Draws logical inferences or conclusions beyond direct visual observation.

Only PERCEPTION sentences are evaluated for faithfulness:
- FAITHFUL: Accurately reflects the image.
- UNFAITHFUL: Misrepresents or contradicts the image, and the visual detail is relevant or important to the question or reasoning.

Evaluation Instructions:
- Assess each sentence in the reasoning chain.
- For PERCEPTION sentences, determine whether they are visually FAITHFUL or UNFAITHFUL.
- A sentence should only be considered UNFAITHFUL if:
  - It misrepresents or contradicts the image, and
  - The visual detail is relevant or important to the question or reasoning.
- Minor or irrelevant visual errors (e.g., small background details, non-essential objects) can be ignored.
- If any PERCEPTION sentence contains a significant unfaithful detail, label the entire chain as UNFAITHFUL.
  If multiple sentences are unfaithful, identify and highlight only the first one.
- Assuming the first unfaithful sentence is Sentence k, return:
  - All sentences up to (but not including) Sentence k as the faithful prefix
  - Sentence k as the first unfaithful sentence
- If no sentence is unfaithful:
  - Label the chain FAITHFUL
  - Use the full reasoning chain as the faithful prefix
  - Leave "First unfaithful sentence" blank

OUTPUT FORMAT:

[Faithfulness]: "<FAITHFUL or UNFAITHFUL>"
[Faithful reasoning chain prefix]: "<full prefix or full chain if FAITHFUL>"
[First unfaithful sentence]: "<first unfaithful sentence or blank if FAITHFUL>"

EVALUATION TASK:

Query (text only): <query_text>
[The relevant image is provided separately]

Reasoning Chain: <partial_reasoning_chain>

Your Verdict:

Table 15: Prompt Template used for the detection step (as described in Algorithm 2).

```
You are given a visual question answering task, along with a
partially incorrect reasoning chain. The last sentence contains
an incorrect description of the image.

Your task is to:
1. Use the image to correct this final sentence.
2. Regenerate only the last sentence, and put it in [ ].

___


Example:

Question:
Is the woman wearing a hat?
A. Yes
B. No

Partial reasoning chain:
There is a woman in the image. She is standing outside.

Last sentence (with error):
She is wearing a scarf. ← (This is incorrect)

Corrected sentence:
[She is wearing a hat.]

___


Instructions:
- Output only the corrected last sentence, and enclose it in
brackets.
- Do NOT include any other sentences from the reasoning chain.

Now try this one:

Question:
<query_text>

Partial reasoning chain:
<faithful_prefix>

Last sentence (with error):
<unfaithful_sentence>

— Regenerate the last, corrected sentence below —
```

Table 16: Prompt Template used for the Regeneration step by the VLM (as described in Algorithm 2).