AIM-CoT: Active Information-driven Multimodal Chain-of-Thought for Vision-Language Reasoning

Xiping Li[1], **Jianghong Ma**[2]*

[1]The Chinese University of Hong Kong
[2]Harbin Institute of Technology (Shenzhen)
lihsiping@gmail.com, majianghong@hit.edu.cn

Abstract

Interleaved-Modal Chain-of-Thought (I-MCoT) advances vision-language reasoning, such as Visual Question Answering (VQA). This paradigm integrates specially selected visual evidence from the input image into the context of Vision-Language Models (VLMs), enabling them to ground their reasoning logic in these details. Accordingly, the efficacy of an I-MCoT framework relies on identifying *what* to see (evidence selection) and *when* to see it (triggering of insertions). However, existing methods fall short in both aspects. First, for selection, they rely on attention signals, which are unreliable—particularly under severe granularity imbalance between the brief textual query and the informative image. Second, for triggering, they adopt static triggers, which fail to capture the VLMs' dynamic needs for visual evidence. To this end, we propose a novel I-MCoT framework, **A**ctive **I**nformation-driven **M**ulti-modal **C**hain-**o**f-**T**hought (**AIM-CoT**), which aims to improve both evidence selection and insertion triggering via: (1) **Context-enhanced Attention-map Generation (CAG)** to mitigate granularity imbalance via textual context enhancement; (2) **Active Visual Probing (AVP)** to proactively select the most informative evidence via an information foraging process; and (3) **Dynamic Attention-shift Trigger (DAT)** to precisely activate insertions when VLM's attention shifts from text to visual context. Experiments across three benchmarks and four backbones demonstrate AIM-CoT's consistent superiority. Our code is available at `https://anonymous.4open.science/r/AIMCoT`.

1 Introduction

Chain-of-Thought (CoT) prompting, initially established for Large Language Models (LLMs) (Wei et al., 2022; Wang et al.; Zhang et al., b; Suzgun et al., 2023; Li et al., 2025; Wang et al., 2023; Diao et al., 2024; Chu et al., 2024), has been naturally adapted to Vision-Language Models (VLMs) (Zhang et al., 2025; Chen et al., 2024b; Cheng et al., 2025; Xu et al., 2024), empowering them to tackle vision-related tasks via a series of intermediate reasoning steps. Visual Question Answering (VQA) serves as a representative scenario targeted in this study. In a typical VQA scenario, the VLM is presented with a multimodal question comprising a textual query and an associated image, and is then prompted to answer the question.

Early efforts in this domain (Mitra et al., 2024; Zheng et al., 2023; Lei et al., 2025; Zhang et al., a) focused on generating text-only CoT. A pivotal advancement is the paradigm of **Interleaved-Modal Chain-of-Thought (I-MCoT)** (Gao et al., 2025). Unlike text-only methods, this paradigm aims to provide fine-grained visual evidence for model reasoning. Specifically, it first selects the salient sub-regions from the input image, and then inserts them as visual tokens into the context of reasoning chain.

However, to construct an effective I-MCoT framework, two critical questions should be well addressed: **(1) What to see?** (Selection: identifying the specific image regions that support the current reasoning step) and **(2) When to see it?** (Triggering: determining the precise moment visual insertion is needed).

Existing research (Gao et al., 2025) falls short in these two aspects due to its reliance on passive and static heuristics. Regarding **selection**, it adopts attention-based selection, which relies on the quality of cross-attention maps of VLM layers. Specifically, the visual regions that receive the most attention from tokens in the text context are selected, assumed to be most helpful for the subsequent generation. However, as revealed in Section 3.1, the raw VLM attention fails to accurately identify the salient regions, particularly when there is a granularity imbalance. Specifically, in the input context, the rich details in the image overwhelm the

*Corresponding Author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 26656–26681

brief textual query. Therefore, with limited semantic anchors, the query cannot effectively steer the cross-attention towards the crucial visual regions. Regarding **triggering**, prior works often insert visual information at fixed, predefined moments (e.g., upon generating a newline character (Gao et al., 2025)). As analyzed in Section 3.3, such a static mechanism fails to align with the model's dynamic cognitive need for fine-grained visual information.

To overcome these limitations, we propose a novel I-MCoT framework, **A**ctive **I**nformation-driven **M**ulti-modal **C**hain-**o**f-**T**hought (**AIM-CoT**), which aims to address both *what to see* and *when to see it* by shifting VLM reasoning from passive, static perception to active, dynamic exploration. Grounded in Information Foraging Theory (Pirolli and Card, 1999; Broadbent, 2013), AIM-CoT detects the VLM's momentary need for visual cues and proactively forages for the most informative evidence. Integrated into the context window, this evidence improves subsequent generation. AIM-CoT realizes this shift through three synergistic components: (1) **Context-enhanced Attention-map Generation (CAG):** To refine the VLM's cross-modal attention distribution, CAG elicits a query-conditioned description of the image and appends it to the textual context. Rather than serving as an auxiliary caption, CAG provides semantic anchors that align textual and visual granularities. (2) **Active Visual Probing (AVP):** To address the "What to see" problem, AVP quantifies the information gain of candidate regions, integrating those that provide the highest utility into the context for subsequent reasoning and generation. (3) **Dynamic Attention-shift Trigger (DAT):** To address the "When to see it" problem, DAT dynamically and precisely triggers visual insertion when a significant shift from textual to visual focus is detected, indicating the VLM's cognitive demand for visual evidence.

Our contributions are summarized as follows:

- We introduce AIM-CoT, a unified system (CAG, AVP, DAT) that enables VLMs to proactively forage for informative visual evidence and dynamically capture the critical triggering moments.
- Our analyses demonstrate that AIM-CoT (1) selects truly salient visual evidence via information gain, (2) withstands noisy attention and descriptions, and (3) remains deployment-friendly.
- Experiments conducted on four backbones across three VQA benchmarks demonstrate the consistently superior performance of AIM-CoT over state-of-the-art baselines.

2 Related Work

Multimodal CoT has been widely adopted in VLM research. Early text-only efforts enhance reasoning by decomposing questions (Zheng et al., 2023), generating intermediate scene graphs (Mitra et al., 2024), or overlaying coordinate grids for spatial referencing (Lei et al., 2025).

A pivotal advancement is the I-MCoT paradigm, which integrates visual evidence directly into reasoning chains to improve subsequent reasoning/generation. The leading approach, ICoT (Gao et al., 2025), selects the visual patches receiving the highest attention scores from the text tokens in the context. Then, this visual evidence is integrated when the VLM outputs a newline token.

However, our analysis in Section 3 highlights the limitations of existing research. Therefore, while we adopt the established I-MCoT paradigm (similar to ICoT (Gao et al., 2025)), our contributions lie in shifting from the passive selection and static trigger to the proactive information-foraging selection and dynamic attention-shifting trigger.

3 Motivation

We focus on two pivotal questions for I-MCoT: **What** and **When** to see. First, we expose the limitations of passive attention-based selection in Section 3.1. To bridge this, we explore an active paradigm shift, investigating the potential of information gain for selection (Section 3.2) and dynamic attention shifts for triggering (Section 3.3).

3.1 Moving Beyond Attention Maps: The Reliability Gap in Passive Visual Selection

Existing I-MCoT methods adopt attention-based visual selection. We examine the reliability of attention **for visual selection** by asking:

(1) **Sufficiency:** Do highly attended patches always benefit question answering?

(2) **Necessity:** Do truly crucial patches always receive high attention?

To investigate these two questions, we randomly sample 500 instances from M3CoT and 500 from ScienceQA, respectively.

Regarding sufficiency, we mask the top-K_{mask} most attended regions and evaluate the performance drop. Notably, the masking is only for evaluating the importance of these regions as visual evi-

dence, not for AIM-CoT itself. As shown in Table 8 (Appendix C.1), masking highly attended regions leads to only minor degradation. Although information redundancy may explain this, our upcoming necessity check suggests that attention peaks rarely coincide with truly salient regions, failing to be sufficient signals for evidence selection.

Regarding necessity, we examine how well the most attended regions R_{attn} cover the truly crucial regions R_{true} in Section 5.5, which is quantified by Intersection over Union (IoU). Across 1,000 instances, over 75% have an IoU even below 50%. This shows that attention peaks rarely align with truly critical visual evidence, i.e., crucial regions do not necessarily emerge as the most attended ones. Moreover, Appendix C.2 suggests that alleviating the text-vision granularity imbalance is an effective way to improve this alignment.

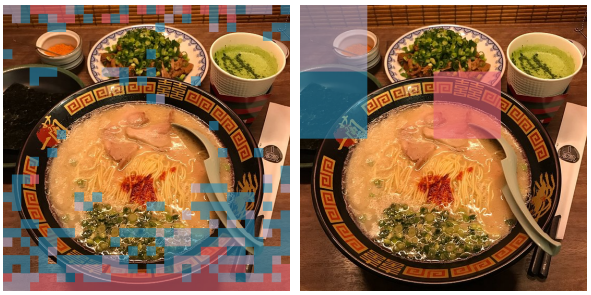
In conclusion, attention alone is unreliable for evidence selection. Therefore, we do not treat attention as a selection rule. However, improving **attention distribution** remains crucial, as it (1) underlies autoregressive generation in VLMs and (2) provides candidate regions for subsequent information-driven selection. Accordingly, we introduce CAG, which assists VLMs to attend to critical visual content precisely by enriching brief textual queries with semantic anchors (Section 4.2).

3.2 Identifying What to See: From Semantic Alignment to Information Gain

The empirical limitations observed in Section 3.1 can be attributed to a fundamental misalignment: cross-attention maps primarily capture the *semantic correlations* among tokens, whereas the ultimate goal of the selection phase in I-MCoT is to identify visual evidence that provides *rich visual information* for subsequent reasoning and generation.

This insight motivates a shift from attention-based selection to an information-centric approach. However, two fundamental questions arise regarding the design of such a mechanism: (1) **Metric**: What constitutes a theoretical measure of "information" in the context of VLM reasoning? (2) **Process**: How should the selection mechanism mimic the cognitive process of information gathering, as opposed to static Top-K selection?

To answer these, we adopt **Information Foraging Theory (IFT)** (Pirolli and Card, 1999; Broadbent, 2013; Oaksford and Chater, 1994; Friston, 2010) as a guiding principle for our framework design. First, regarding the *metric*, IFT suggests that information is valuable only when it reduces the agent's uncertainty. This motivates us to ground our selection criterion in uncertainty reduction (i.e., information gain), which can be quantified by the entropy of the VLM's predictive distribution over its vocabulary. Second, regarding the *process*, IFT characterizes foraging as a sequential trajectory where each step depends on previous knowledge. This suggests that the optimal strategy is not a one-off Top-K selection, but a dynamic, iterative process where the model updates its belief state after every glimpse, i.e., the selected regions are inserted into VLM's context in the form of visual tokens.



(a) Regions selected by attention-based strategy. (b) Regions selected by information-driven strategy.

Figure 1: A comparison of regions selected by different strategies. Details are provided in Appendix D.1.

These two theoretical insights directly shape the architecture of our AVP module (Section 4.3). Figure 1 compares AVP-selected regions with those from attention-based method, which is further detailed in Appendix D.1.

3.3 Identifying When to See: From Static Heuristics to Dynamic Attention Shifts

The fundamental goal of I-MCoT is to ground the VLM's reasoning in inserted fine-grained visual evidence. Therefore, the insertions should align with the model's fluctuating need for visual evidence across reasoning steps. However, existing methods rely on static triggers like newlines (Gao et al., 2025), failing to capture critical moments when the VLM actively seeks visual evidence to support subsequent generation.

To address this, we propose that the attention shift from textual to visual contexts (i.e., the text-to-vision attention shift) serves as a superior, dynamic indicator. Intuitively, a significant pivot in attention towards the visual modality suggests the VLM's cognitive demand for visual grounding. Unlike static triggers, monitoring these shifts allows for determining the precise "when" based on the

model's real-time internal state.

We empirically validate this intuition through an in-depth analysis of the baseline model on the LLaVA-W benchmark (detailed in Appendix F). Our investigation correlates the timing of visual insertions with model performance, revealing two key observations: **(1) Correlation analysis:** Synchronizing visual insertions with salient text-to-vision attention shifts strongly correlates with better generation quality. **(2) Group analysis:** This dynamic shift pattern distinguishes high-quality responses from low-quality ones.

In conclusion, although attention is unreliable as a decision rule for selection, its shift is still a **reliable diagnostic signal** revealing when the model is seeking visual information. Motivated by this, we present DAT, an attention-shift trigger that dynamically activates visual insertions (Section 4.4).

4 AIM-CoT

In this section, we begin by briefly reviewing the background of vision-language reasoning. Then, based on the motivations detailed in Section 3, we present AIM-CoT. As a training-free framework, AIM-CoT models the interleaved-modal reasoning as an information-foraging process with three synergistic components: (1) **CAG**, which pre-processes the input to generate a fine-grained description, mitigating text-vision granularity imbalance. (2) **AVP**, which proactively selects regions that maximize information gain whenever activated. (3) **DAT**, which triggers AVP to insert visual evidence into the CoT precisely when the model's cognitive focus shifts from text to vision. A visualization of AIM-CoT is shown in Figure 2.

It is important to distinguish AIM-CoT from attribution methods that rely on masking or deleting parts of the input image. AIM-CoT operates on a **frozen VLM** and adopts a "Trigger-Select-Insert" paradigm: although sub-regions are integrated as visual evidence, the input context, including the original image is fully preserved. This ensures the VLM operates on a coherent, augmented context rather than a disrupted one (Khorram et al., 2021).

4.1 Preliminaries

Due to space constraints, we defer the definitions of important concepts to Appendix B, including (1) the Vision-Language Models (VLMs), (2) the Context Window of a VLM, (3) Patches and Regions (i.e., Visual Tokenization), and (4) I-MCoT.

4.2 Context-enhanced Attention-map Generation (CAG)

As revealed in Section 3.1, attention is an unreliable basis for visual selection, particularly under severe text-vision granularity imbalance. Although AIM-CoT shifts from passive attention-based selection to proactive information-driven selection, we argue that refining the VLM's attention distribution remains meaningful for the interleaved-modal reasoning process. This is primarily for two reasons: (1) a more reliable attention distribution benefits reasoning, since it serves as the basis of model generation; and (2) crucially, the attention map serves as a source of candidate regions for subsequent selection (Section 4.3).

To this end, we propose **CAG** to improve the VLM's cross-attention distribution by mitigating the text-vision granularity imbalance. Specifically, before the VQA process begins, the VLM is elicited to carefully generate an explanatory description of the input image conditioned on the textual query. Rather than appending a caption, this process enriches the brief query with more semantic anchors that encode visual information within the image, thereby alleviating the granularity imbalance. Formally, this is expressed as follows:

$$\mathcal{D}_{\text{CAG}} = \text{VLM}(I, x, \mathcal{P}_{\text{CAG}}), \quad (1)$$

$$x' = \text{concat}(x, \mathcal{D}_{\text{CAG}}), \quad (2)$$

where $\mathcal{P}_{\text{CAG}}$ is the prompt for generating the description. x' is the updated textual query that integrates the description $\mathcal{D}_{\text{CAG}}$. A template of $\mathcal{P}_{\text{CAG}}$ is provided in Appendix A.3.

To ensure the output $\mathcal{D}_{\text{CAG}}$ serves as a reliable textual anchor while preventing error propagation to downstream modules, it is vital to mitigate potential hallucinations within it. To this end, the prompt $\mathcal{P}_{\text{CAG}}$ is meticulously designed with *negative constraints*: it explicitly instructs the model to prioritize visible evidence and adopt a rigorous, cautious stance—skipping uncertain details rather than speculating. Experiments in Appendix L.1 validate the effectiveness of negative constraints.

4.3 Active Visual Probing (AVP)

Drawing on the analyses in Section 3.2, we propose **AVP**, an information-driven mechanism designed to proactively select the regions with valuable information. Inspired by Information Foraging Theory (IFT), AVP operates through three systematic steps: (1) constructing a diverse candidate region set, (2)

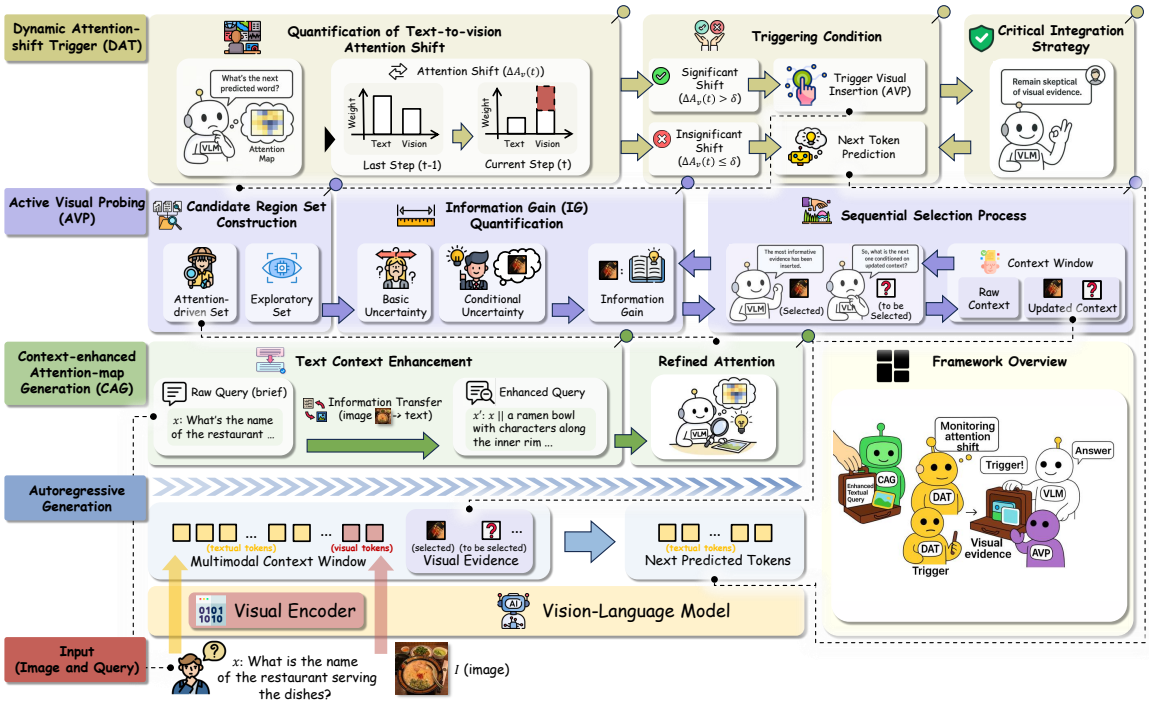


Figure 2: An overview of the AIM-CoT framework. CAG initially enhances the query to refine attention distributions, thereby supporting reasoning and supplying reliable candidates for AVP. During reasoning, AVP executes visual insertion by iteratively selecting the most informative regions as evidence conditioned on the current context. This insertion is triggered once DAT (the trigger) detects a significant text-to-vision shift in the VLM's internal attention state. See Appendix D.2 for a concise reading guide.

quantifying the information gain of each candidate region, and (3) executing a sequential, greedy selection process.

Step 1: Candidate Region Set Construction. As introduced in Appendix B, each input image is partitioned into multiple regions with a fixed partitioning method. Although the partition for a given input image is deterministic, evaluating the information gain of each of these regions could be computationally prohibitive. Therefore, it is necessary to construct a candidate pool, which is a subset of the raw partition.

We propose that although the candidate region set is a subset of the raw partition, it should remain sufficiently informative for subsequent selection. To this end, we derive candidate regions from diverse sources, covering both the VLM's current focus and potential blind spots, as follows:

- **Attention-driven Set (C_{attn}):** We select the top-N regions from the CAG-enhanced cross-attention map (A'), as this subset is a real-time reflection of VLM's internal state in the latest reasoning step. Although the attention-based signals can inherently carry noise, analysis in Appendix I.3 suggests that our subsequent information-driven selection effectively filters out non-informative high-attention regions (i.e., noise rejection).
- **Exploratory Set (C_{exp}):** To mitigate the "tunnel vision" of attention maps, we extract M exploratory regions by uniformly sampling from all the partitioned regions. Despite its simplicity, this effective strategy outperforms complex alternatives (Appendix I.1) and provides essential regions overlooked by the attention-driven set C_{attn} (Appendix I.2).

Formally, the total candidate set C is constructed as follows:

$$C_{attn} = \{R_1, R_2, \ldots, R_N\},$$
$$C_{exp} = \{R_{N+1}, R_{N+2}, \ldots, R_{N+M}\},$$
$$C = C_{attn} \cup C_{exp},$$

where each item denotes a specific image sub-region. For $i \leq N$, R_i is the i-th most attended region in A'.

Step 2: Information Gain Quantification. As analyzed in Section 3.2, IFT motivates us to identify the most informative regions. Therefore, it is necessary to quantify the information gain of each candidate region in C.

To quantify *Information Gain (IG)* of a candidate region, we compare the VLM's information

content before and after integrating this specific candidate (as visual tokens) into its current context. These pre- and post-insertion information contents are termed basic uncertainty and conditional uncertainty, respectively. Intuitively, they measure the VLM's uncertainty in predicting the next token (textual) during autoregression under these two states, respectively. Formally, *Basic Uncertainty* (U_B) is defined as the entropy of the VLM's next-token distribution given the current context, and *Conditional Uncertainty* ($U_{C,i}$) as the entropy after explicitly introducing candidate region R_i.

$$\begin{aligned} U_B &= H(Y|I, x, y_{<t}) \\ &= -\sum_{y \in V} P(y|I, x, y_{<t}) \log_2 P(y|I, x, y_{<t}), \quad (3) \\ U_{C,i} &= H(Y|I, x, y_{<t}, R_i) \\ &= -\sum_{y \in V} P(y|I, x, y_{<t}, R_i) \log_2 P(y|I, x, y_{<t}, R_i), \end{aligned}$$

where V is the vocabulary, and $y_{<t}$ represents the generated tokens. The information gain is thus derived as:

$$IG(\{R_i\}) = U_B - U_{C,i},\ i = 1, \ldots, N+M. \quad (4)$$

Step 3: Sequential Selection Process. Finally, as analyzed in Section 3.2, drawing on the iterative nature of information foraging, where each step depends on previous knowledge, we frame region selection as a sequential trajectory. Specifically, in each step, we propose to select the region that provides the maximum information gain conditioned on the context updated in the last step, as outlined in Algorithm 1. This is because this greedy algorithm (1) intuitively moves maximally toward providing the VLM with the most visual information at each step, and besides, (2) it has theoretical support: regarding such a subset selection problem, it yields an approximation to the global optimum (i.e., the maximum information gain) for the VLM. We provide comprehensive analyses of this theoretical property in Appendix I.4.

4.4 Dynamic Attention-shift Trigger (DAT)

As identified in Section 3.3, although attention remains an unreliable basis **for selection**, its shift serves as a reliable diagnostic signal **for triggering**. Motivated by this, we propose **DAT**, the trigger component of AIM-CoT. DAT tracks the VLM's text-to-vision attention shifts across the autoregressive steps, as they reflect the VLM's dynamic need for visual evidence. Then, based on the magnitude of each shift, DAT determines whether the visual evidence should be introduced via AVP at that step.

Algorithm 1: Greedy Algorithm for Optimal Region Selection

Input : Total candidate set C, Target selection size K
Output : Optimal selection set S

$R^* \leftarrow \varnothing$;
for $k \leftarrow 1$ **to** K **do**
  Compute Basic Uncertainty: $U_B \leftarrow H(Y|I, x, y_{<t}, R^*)$;
  for $R_i \in C \setminus R^*$ **do**
    Compute Conditional Uncertainty: $U_{C,i} \leftarrow H(Y|I, x, y_{<t}, R^* \cup \{R_i\})$;
    Calculate Gain: $IG(\{R_i\}) \leftarrow U_B - U_{C,i}$;
  Select Best Region: $R_{next} \leftarrow \text{argmax}_{R_i \in C \setminus R^*} \{IG(\{R_i\})\}$;
  Update Selected Set: $R^* \leftarrow R^* \cup \{R_{next}\}$;
$S \leftarrow R^*$;
return S

Quantification of Text-to-vision Attention Shift. The text-to-vision attention shift is determined by the difference in attention distributions between steps. Intuitively, it tracks the flow of attention from the text context in the previous step to the visual context in the current step. Furthermore, since the sum of attention scores over these two parts (i.e., text and vision) is fixed in the VLM's normalized attention map, it is sufficient to focus on the attention shift in the visual part. Formally, this can be quantified as follows:

$$\Delta A_{vision}(t) = A_{vision}(t) - A_{vision}(t-1), \quad (5)$$

where $A_{vision}(t)$ and $A_{vision}(t-1)$ represent the sum of attention scores allocated to all the visual tokens in the context by the currently predicted token (the t-th token) and the preceding token (the $(t-1)$-th token), respectively.

Triggering Condition. DAT determines if a given shift $\Delta A_{vision}(t)$ is significant enough to warrant the provision of visual evidence via a qualitative criterion. In the criterion, a hyper-parameter $\delta \in \mathbb{R}$ is introduced to serve as the threshold. The AVP module is triggered to perform visual insertion after the t-th generation step if and only if $\Delta A_{vision}(t) > \delta$.

In Appendix G, we provide a detailed sensitivity analysis of δ and introduce an adaptive thresholding strategy to set δ automatically.

Critical Integration Strategy. A potential challenge inherent to the I-MCoT paradigm is the difficulty in eliminating noise that visual evidence might introduce. To address this, DAT does not naively trigger the visual evidence insertion. Instead, it employs a critical integration mechanism: the inserted regions are accompanied by a **safety instruction**, which cues the VLM to treat the visual evidence as "supplementary references", encouraging the model to verify semantic consistency with the existing textual context. This effectively filters out irrelevant or hallucinated visual cues.

We validate the effectiveness of this safety instruction in Appendix L.2.

5 Experiments

5.1 Evaluation Setup

Benchmark. We evaluate AIM-CoT on three widely used VQA benchmarks: M3CoT (Chen et al., 2024a), ScienceQA (Saikh et al., 2022), and LLaVA-W (Liu et al., 2024). Detailed descriptions are provided in Appendix A.1.

Baselines. We compare against several text-only baselines, including vanilla VLM without CoT (No-CoT), DDCoT (Zheng et al., 2023), MM-CoT (Zhang et al., a), CCoT (Mitra et al., 2024), and SCAFFOLD (Lei et al., 2025). On top of these, the leading I-MCoT framework ICoT (Gao et al., 2025), is also included. More details are available in Appendix A.2. When possible, we directly report figures from prior work.

Backbones. The implementation is built on four mainstream VLM backbones with different architectures and scales: early-fusion Chameleon-7B (Team, 2024) and Janus-Pro-7B (Chen et al., 2025), and late-fusion Qwen2-VL-7B (Wang et al., 2024) and Qwen2.5-VL-32B (Bai et al., 2025). We conduct experiments in both 0- and 1-shot settings, using the prompt template from the open-source ICoT implementation (Gao et al., 2025).

Hyper-parameter. Hyper-parameter settings are listed in Appendix A.4.

5.2 Performance Comparison

Table 1 reports the experimental results. AIM-CoT consistently outperforms all baselines (both text-only and I-MCoT) across all benchmarks and backbones, under both 0- and 1-shot settings. This demonstrates the strong advantage of AIM-CoT for complex vision-language reasoning.

More specifically, both I-MCoT methods (AIM-CoT and ICoT) surpass the text-only baselines, confirming the efficacy of explicit incorporation of visual evidence. Moreover, although AIM-CoT and ICoT follow the same I-MCoT paradigm, AIM-CoT achieves stronger performance. This gain comes from three key improvements: (1) AIM-CoT uses CAG to directly refine the VLM's cross-attention distribution, instead of grounding in the unreliable signal used by ICoT; (2) it replaces fragile attention-based heuristics with information-driven selection (AVP); and (3) it replaces a static trigger with a dynamic and more precise trigger (DAT).

Finally, the performance gains vary across backbones. In Appendix K, we analyze this effect in detail and attribute it to the interaction between model architecture and model scale.

5.3 Ablation Study

In this section, we conduct the ablation study to verify the efficacy of each component within AIM-CoT. The detailed settings are as follows:

- **w/o CAG**: The VLM operates under the image I and the raw query x, instead of the CAG-enhanced query x'.
- **w/o AVP**: The information-driven selection (AVP) is replaced by attention-based selection.
- **w/o DAT**: DAT is substituted with a static trigger (i.e., visual evidence is inserted whenever the VLM outputs a newline character).

The results on Chameleon-7B are shown in Table 2 (more results and analyses on other backbones are deferred to Appendix J). First, refining the VLM's attention distribution via CAG can strongly improve VLM's performance on VQA tasks. This is because the attention is not only the basis of model generation but also one source of candidate evidence. Furthermore, (1) even without CAG, AIM-CoT still significantly outperforms ICoT (Table 1); and (2) removing AVP and DAT causes a larger performance drop than removing CAG. These results highlight the importance of tackling the *what to see* and *when to see it* questions. In response to these two issues, we propose AVP and DAT, respectively.

To provide a comprehensive understanding of AIM-CoT, we go beyond merely validating component efficacy. In Sections 5.5, 5.6, and Appendix H, we present both quantitative and visualized qualitative analyses of AVP, elucidating why information-driven regions are more effective than attention-

Table 1: Performance comparison results on three VQA benchmarks and four backbones. The best performances are shown in bold. We report Accuracy (ACC.) for M3CoT and ScienceQA, and ROUGE-L for LLaVA-W.

Method	M3CoT		ScienceQA		LLaVA-W	
	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot
Chameleon-7B						
No-CoT	29.1	28.4	47.7	48.5	13.1	23.9
DDCoT	28.6	29.8	49.8	49.2	20.2	23.1
MMCoT	28.5	30.6	49.0	50.7	20.4	20.6
CCoT	29.4	31.4	50.2	51.3	22.1	24.5
SCAFFOLD	29.6	31.1	48.5	47.5	21.7	24.7
ICoT	29.8	32.3	51.0	53.4	25.2	27.6
AIM-CoT (ours)	**31.4**‡	**32.8**†	**53.1**‡	**54.5**†	**29.8**‡	**32.0**‡
Improv.	*+5.4%*	*+1.5%*	*+4.1%*	*+2.1%*	*+18.3%*	*+15.9%*
Janus-Pro-7B						
No-CoT	36.5	37.8	53.9	61.2	29.8	30.9
DDCoT	36.8	38.6	53.2	61.6	29.0	30.4
MMCoT	35.2	36.6	50.8	58.0	28.2	28.9
CCoT	36.4	38.2	54.1	61.5	27.8	31.5
SCAFFOLD	36.1	38.0	52.4	60.5	29.7	30.8
ICoT	37.6	39.4	55.1	62.5	32.5	33.6
AIM-CoT (ours)	**39.7**‡	**41.5**‡	**56.9**‡	**64.9**‡	**35.5**‡	**36.6**‡
Improv.	*+5.6%*	*+5.3%*	*+3.3%*	*+3.8%*	*+9.2%*	*+8.9%*

Method	M3CoT		ScienceQA		LLaVA-W	
	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot
Qwen2-VL-7B						
No-CoT	43.6	45.4	56.3	64.4	32.7	33.5
DDCoT	42.6	45.7	55.2	64.9	31.2	32.8
MMCoT	40.1	42.5	51.3	58.3	30.7	31.4
CCoT	43.3	44.1	56.4	63.8	29.4	33.9
SCAFFOLD	41.7	44.9	53.7	62.5	31.8	33.1
ICoT	44.1	46.0	56.8	65.4	34.2	35.7
AIM-CoT (ours)	**44.7**†	**46.6**†	**57.4**†	**66.3**†	**36.3**†	**37.3**†
Improv.	*+1.4%*	*+1.3%*	*+1.1%*	*+1.4%*	*+6.1%*	*+4.5%*
Qwen2.5-VL-32B						
No-CoT	55.2	56.8	73.5	77.2	40.5	41.8
DDCoT	54.8	57.5	72.9	77.8	39.6	41.3
MMCoT	53.4	54.7	69.8	74.5	38.9	39.2
CCoT	55.3	56.9	73.4	77.1	37.8	42.4
SCAFFOLD	54.2	56.7	71.6	76.3	40.1	41.5
ICoT	56.9	59.1	75.1	79.2	43.4	44.7
AIM-CoT (ours)	**58.7**‡	**61.2**‡	**76.8**‡	**81.3**†	**46.5**‡	**49.1**‡
Improv.	*+3.2%*	*+3.6%*	*+2.3%*	*+2.7%*	*+7.1%*	*+9.8%*

Statistical significance: $^{\dagger}p < 0.05$, $^{\ddagger}p < 0.01$ for AIM-CoT compared with the second-best method under the same backbone/setting (McNemar for ACC. on M3CoT/ScienceQA; Wilcoxon signed-rank for ROUGE-L on LLaVA-W). Full results are reported in Appendix E.

Table 2: Ablation study of AIM-CoT conducted on Chameleon-7B under 0-shot setting.

Dataset	AIM-CoT	w/o CAG	w/o AVP	w/o DAT
M3CoT (ACC.)	31.4	30.5 (-0.9)	30.6 (-0.8)	30.8 (-0.6)
ScienceQA (ACC.)	53.1	52.8 (-0.3)	52.3 (-0.8)	52.7 (-0.4)
LLaVA-W (ROUGE-L)	29.8	26.8 (-3.0)	26.2 (-3.6)	27.3 (-2.5)

Table 3: Ablation study of the baseline model (BM) on Chameleon-7B under 0-shot setting.

Dataset	BM	BM w/ CAG	BM w/ AVP	BM w/ CAG, AVP
M3CoT (ACC.)	29.8	30.1 (+0.3)	30.2 (+0.4)	30.8 (+1.0)
ScienceQA (ACC.)	51.0	51.5 (+0.5)	51.9 (+0.9)	52.7 (+1.7)
LLaVA-W (ROUGE-L)	25.2	25.8 (+0.6)	26.4 (+1.2)	27.3 (+2.1)

based regions.

5.4 Interplay between CAG and AVP

In this section, we investigate the interaction between CAG and AVP. This is achieved by sequentially adding them to a basic model (BM) (i.e., AIM-CoT stripped of all its components).

As shown in Table 3, while CAG and AVP yield individual gains, their combination exhibits a clear phenomenon of **super-additivity**, i.e., the joint improvement consistently exceeds the sum of their separate contributions. Specifically, on M3CoT, the joint gain ($+1.0\%$) significantly surpasses the linear sum of individual gains ($0.3\% + 0.4\% = 0.7\%$). This trend holds across all benchmarks (e.g., $+1.7\% > 1.4\%$ on ScienceQA).

This non-linear boost validates the interplay between the modules: CAG refines the VLM's attention distribution, which amplifies AVP's effectiveness; the AVP, in turn, precisely mines the salient visual evidence from the attention-driven candidates (C_{attn}) enhanced by CAG.

5.5 Quantitative Analysis: Semantic Relevance

We further examine whether the visual regions selected by AIM-CoT match human-centered perception and the needs of the task. Our premise is that informative evidence should correspond to distinct and complete semantic entities rather than repetitive background patterns. We therefore use the Segment Anything Model (SAM) (Kirillov et al., 2023) as a proxy to assess whether the selected regions align with meaningful object concepts.

We conduct this evaluation on M3CoT and ScienceQA, randomly sampling 500 instances from each benchmark. Results are averaged over the four backbones: Chameleon-7B, Janus-Pro-7B, Qwen2-VL-7B, and Qwen2.5-VL-32B. To build a reference for visually relevant evidence, we use SAM to generate segmentation masks for the key entities mentioned in the question and ground-truth answer. We then define the **Semantic Hit Rate (SHR)** as the percentage of instances in which the selected regions substantially overlap with the corresponding

Table 4: Comparison of SHR on M3CoT and ScienceQA (subset of 500 samples each). The results are averaged across four backbones.

Method	Selection Strategy	SHR ↑	
		M3CoT	ScienceQA
ICoT (Gao et al., 2025)	Attention-driven Top-K	18.7%	24.3%
AIM-CoT (Ours)	**Info. Gain-driven AVP**	**65.2%**	**71.8%**

SAM masks, measured by IoU greater than 0.5.

Table 4 shows that the attention-driven Top-K strategy used by ICoT struggles to recover semantically relevant evidence, with SHR dropping to 18.7% on M3CoT. This supports the observation that raw attention maps often drift toward high-contrast background patterns or generic salient objects, rather than the specific evidence required for reasoning. In contrast, AIM-CoT achieves substantially higher SHR scores, reaching 65.2% on M3CoT and 71.8% on ScienceQA. These results indicate that our active information-seeking paradigm filters out distractors and selects regions that are both structurally coherent and semantically aligned with the question, making them closer to human judgments of relevant evidence.

5.6 Quantitative Analysis: Alignment with Human Intuition

We next evaluate whether the visual regions selected by AIM-CoT better align with human intuition than those selected by the attention-driven baseline. Following prior work, we use GPT-4v (Achiam et al., 2023) as an expert proxy judge, as strong multimodal judges have been shown to correlate well with human preferences (Huang et al., 2025; Gera et al., 2025; D'Souza et al., 2025; Findeis et al., 2025; Zhen et al., 2025).

We evaluate 500 randomly sampled instances from the M3CoT and LLaVA-W benchmarks. For each instance, GPT-4v is given the textual query, the original image, and two anonymized sets of selected regions: one produced by ICoT (Top-K) and the other by AIM-CoT (AVP). We randomize the order of the two sets to eliminate position bias. GPT-4v then compares them along three criteria: semantic relevance, object completeness, and overall helpfulness. The evaluation prompt is shown in Table 5.

As shown in Table 6, AIM-CoT is strongly preferred across both benchmarks. It wins on 76.4% of M3CoT samples and 81.2% of LLaVA-W samples, while only a small fraction favor ICoT. These results suggest that AIM-CoT more consistently identifies complete and semantically informative evidence that matches human intuition. GPT-4v's qualitative judgments further indicate that ICoT often focuses on noisy regions, whereas AIM-CoT more reliably localizes the full objects needed for reasoning.

Table 5: The core prompt used for GPT-4v blind pairwise comparison.

System Prompt: You are an expert judge. Compare two sets of image crops (Set A and Set B) extracted from the Original Image. Decide which set is more helpful for a human to answer the Question.
Criteria:
1. **Relevance:** Does the set contain the specific objects or details mentioned in the question?
2. **Completeness:** Are the objects complete, or are they meaningless background noise/fragments?
3. **Helpfulness:** Which set would better help a human answer the question without seeing the full image?
Output: 'Set A is better', 'Set B is better', or 'Tie'.

Table 6: GPT-4v preference rates on 500 samples from M3CoT and LLaVA-W. "Win" denotes that AIM-CoT provides better visual evidence than ICoT. "Tie" indicates equal quality.

Benchmark	**AIM-CoT Win**	Tie	ICoT Win
M3CoT (Reasoning)	**76.4%**	16.2%	7.4%
LLaVA-W (In-the-Wild)	**81.2%**	13.6%	5.2%

6 Conclusion

In this paper, we propose AIM-CoT, a novel I-MCoT framework that aims to frame the construction of interleaved-modal CoT as an active information-foraging process. Existing methods' static triggers fail to capture the dynamic needs of the VLM for visual information in complex reasoning, and their attention-based selectors depend heavily on the VLM's attention, which can be unreliable (for selection). In response to these challenges, AIM-CoT dynamically monitors the VLM's cognitive need for fine-grained visual evidence for subsequent generation, and accordingly, selects salient visual evidence in an information-driven manner. Extensive experiments demonstrate that AIM-CoT outperforms the leading methods across three benchmarks and four VLM backbones.

7 Acknowledgements

This work was partially supported by the Guangdong Basic and Applied Basic Research Foundation under Grant No. 2024A1515011949

and 2026A1515011672, the Shenzhen Science and Technology Program under Grant No. GXWD20231130110308001 and JCYJ20250604145617023.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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*.

Andrew An Bian, Joachim M Buhmann, Andreas Krause, and Sebastian Tschiatschek. 2017. Guarantees for greedy maximization of non-submodular functions with applications. In *International conference on machine learning*, pages 498–507. PMLR.

Donald Eric Broadbent. 2013. *Perception and communication*. Elsevier.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024a. M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8199–8221.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *CoRR*.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2024b. 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.

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.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1173–1203.

Abhimanyu Das and David Kempe. 2011. Submodular meets spectral: greedy algorithms for subset selection, sparse approximation and dictionary selection. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 1057–1064, Madison, WI, USA. Omnipress.

Shizhe Diao, Pengcheng Wang, Yong Lin, Rui Pan, Xiang Liu, and Tong Zhang. 2024. Active prompting with chain-of-thought for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1350.

Jennifer D'Souza, Hamed Babaei Giglou, and Quentin Münch. 2025. Yescieval: Robust llm-as-a-judge for scientific question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13749–13783.

Arduin Findeis, Floris Weers, Guoli Yin, Ke Ye, Ruoming Pang, and Tom Gunter. 2025. Can external validation tools improve annotation quality for llm-as-a-judge? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15997–16020.

Karl Friston. 2010. Friston, k.j.: The free-energy principle: a unified brain theory? nat. rev. neurosci. 11, 127-138. *Nature reviews. Neuroscience*, 11:127–38.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. 2025. Interleaved-modal chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19520–19529.

Ariel Gera, Odellia Boni, Yotam Perlitz, Roy Bar-Haim, Lilach Eden, and Asaf Yehudai. 2025. Justrank: Benchmarking llm judges for system ranking. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 682–712.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587.

Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. 2025. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge model is not a general substitute for gpt-4. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5880–5895.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Saeed Khorram, Tyler Lawson, and Li Fuxin. 2021. igos++ integrated gradient optimized saliency by bilateral perturbations. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 174–182.

Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, and 1 others. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026.

Andreas Krause, Ajit Singh, and Carlos Guestrin. 2008. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9:235–284.

Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. 2025. Scaffolding coordinates to promote vision-language coordination in large multimodal models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2886–2903.

Xiping Li, Aier Yang, Jianghong Ma, Kangzhe Liu, Shanshan Feng, Haijun Zhang, and Yi Zhao. 2025. Cpgrec+: A balance-oriented framework for personalized video game recommendations.

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.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. 1978. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294.

Mike Oaksford and Nick Chater. 1994. A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101:608–631.

Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological review*, 106(4):643.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and 1 others. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051.

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. 2013. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 2717–2739.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yanzhi Xu, Yueying Hua, Shichen Li, and Zhongqing Wang. 2024. Exploring chain-of-thought for multimodal metaphor detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 91–101.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2025. Improve vision language model chain-of-thought reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1662.

Zhuosheng Zhang, Aston Zhang, Mu Li, George Karypis, Alex Smola, and 1 others. a. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. b. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. 2023. Fast segment anything. *arXiv preprint arXiv:2306.12156*.

Cheng Zhen, Ervine Zheng, Jilong Kuang, and Geoffrey Jay Tso. 2025. Enhancing llm-as-a-judge through active-sampling-based prompt optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 960–970.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191.

A Experimental Setup and Reproducibility

A.1 Benchmarks

M3CoT (Chen et al., 2024a) is a novel multimodal CoT benchmark, which introduces complex, multi-step problems across science, mathematics, and commonsense domains, comprising 11,459 samples in total. M3CoT is characterized by succinct textual queries (<15 tokens on average) paired with intricate problems. This inherent text-vision imbalance makes it an ideal platform to validate the efficacy of our proposed CAG in mitigating this issue and the superiority of AVP in proactively selecting the salient visual regions.

ScienceQA (Saikh et al., 2022) is a popular benchmark for multiple-choice question answering with explanations on scholarly articles, comprising over 100,000 context-question-answer triples to address data scarcity in scientific machine reading comprehension.

LLaVA-Bench In-the-Wild (LLaVA-W) (Liu et al., 2024) is a challenging open-ended benchmark designed to evaluate the real-world capabilities of VLMs by mimicking the unpredictability of real-world scenarios. The answers generated by GPT-4v (Achiam et al., 2023) serve as the labels. LLaVA-W is exceptionally well-suited for evaluating the capability of our proposed framework to address complex, open-ended problems by generating a multimodal CoT, attending to salient regions within the image, and meticulously parsing the query.

A.2 Baselines

No-CoT prompts the VLM to answer questions directly based on the input query and image. In the 1-shot setting, an example containing the query, image, and corresponding answer is attached.

DDCoT (Zheng et al., 2023) deconstructs a multimodal problem into reasoning and recognition sub-questions, uses negative-space prompting to identify and fill visual information gaps with external models, and then integrates all information for a final joint reasoning step to generate rationales.

MMCoT (Zhang et al., a) first generates a rationale from fused language and vision inputs, and then uses this rationale along with the original multimodal data to infer the final answer.

CCoT (Mitra et al., 2024) first prompts the VLM to generate a scene graph from an image and then uses it as an intermediate reasoning step to produce the final response.

SCAFFOLD (Lei et al., 2025) promotes vision-language coordination in the VLM by overlaying a dot matrix with coordinates onto an image, which then serves as a visual anchor that can be explicitly referenced in the textual prompt.

ICoT (Gao et al., 2025) leverages the attention maps of the VLM to select relevant patches from the input image and insert them into the reasoning process, thereby generating sequential steps of paired visual and textual rationales.

A.3 Template of $\mathcal{P}_{CAG}$

Figure 3 provides an intuitive example showing the template of $\mathcal{P}_{CAG}$ and how it is used to prompt the VLM to carefully generate a guiding description for the input image. In particular, for multiple-choice questions, such as those in M3CoT and ScienceQA, we prepend the following brief explanation to $\mathcal{P}_{CAG}$ to aid the VLM in better understanding its designated task: "*This is a multiple-choice question. The question is based on the image provided.*"

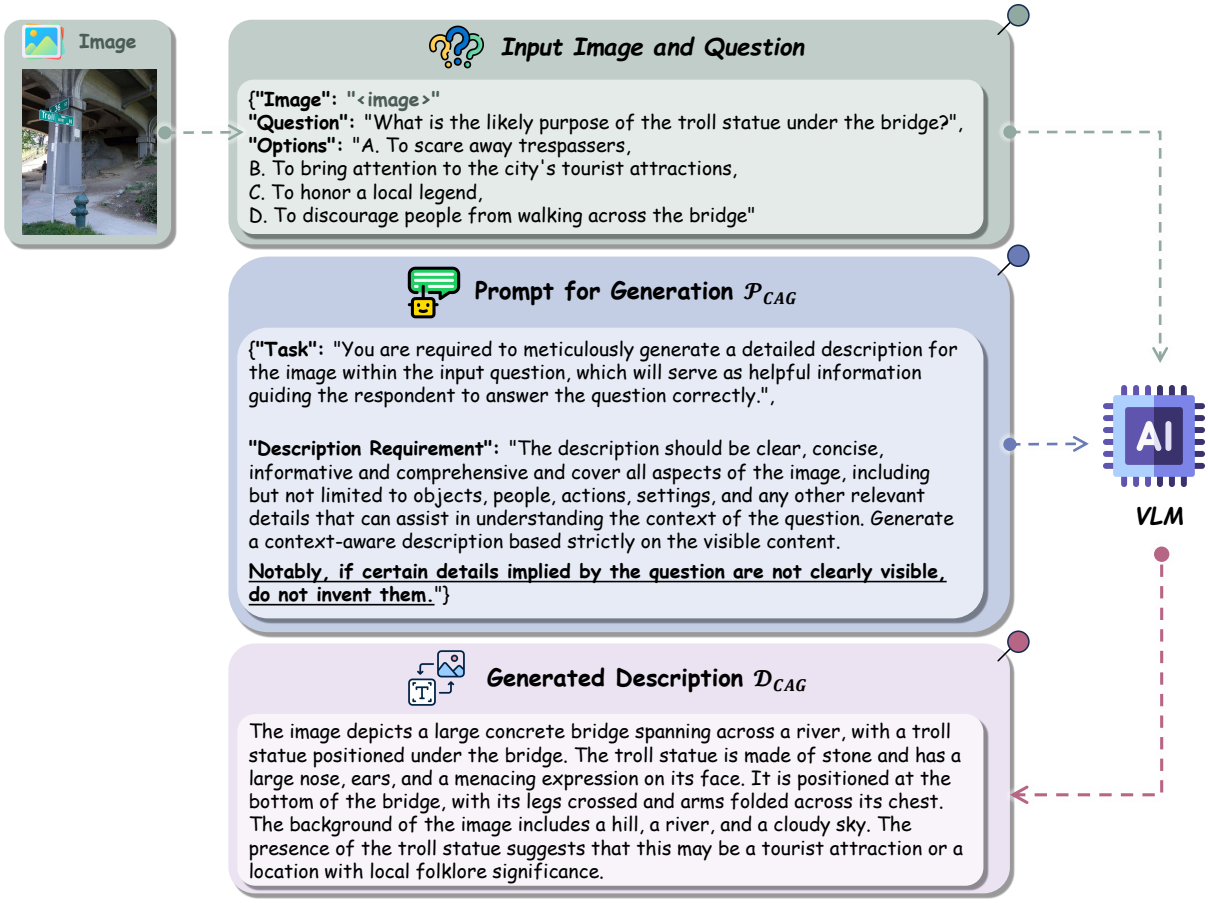


Figure 3: An illustration of the entire process of context enhancement by the CAG module, using problem physical-commonsense-1398 from the M3CoT benchmark as an example. This showcases both the template and usage of $\mathcal{P}_{CAG}$.

Then, the cross-attention weight matrix based on the enhanced context x' and I can be obtained as follows:

$$A' = \text{softmax}(\frac{(H_T W^Q)(H_V W^K)^T}{\sqrt{d_K}}), \quad (6)$$

where $H_T \in \mathbb{R}^{n_T \times d}, H_V \in \mathbb{R}^{n_V \times d}$ are the hidden states of the textual and visual input, respectively. W^Q, W^K are the weight matrices of the linear transformation layers for query and key, respectively.

A.4 Setting of Hyper-parameters

The hyper-parameter settings are listed in Table 7.

Table 7: Hyper-parameter settings across three datasets.

Parameter	M3CoT	ScienceQA	LLaVA-W
N_C	8	8	6
K	3	3	3
N	4	4	2
M	4	4	1
Region size for AVP s_r (grid)	1	1	1
Grid size for AVP s_g	4	4	4
δ	0.5	0.2	0.2
Temperature	0.7	0.7	0.7
Do sample	True	True	True
Top_p	0.9	0.9	0.9
Repetition_penalty	1.2	1.2	1.2
Min_new_tokens	32	32	32
Max_new_tokens	512	1024	1024

B Supplementary Important Definitions

Vision-Language Model. A VLM typically fuses a vision encoder for preprocessing visual input and a generative language model, which jointly enable it to respond in a human-like manner:

$$answer = \text{VLM}(I, x), \quad (7)$$

where I is the image and x the query.

Context Window of a VLM. An autoregressive generation of a VLM is conditioned on (1) the user's initial input and (2) its interaction with the user. A context window is a structured representation of this conditional information.

Specifically, a context window starts with the user's multimodal input question, including a textual query x and a paired image I. Built upon this, the VLM continuously extends the initial context window by iteratively incorporating each reasoning step and its corresponding visual evidence (if any) into this context, once they are acquired. This ensures that the prior part of the context window can consistently yet iteratively influence the subsequent generation and the final response.

In implementation, this multimodal information, whether textual or visual context, is incorporated into the context window in the form of tokens.

Visual Tokenization: Patches and Regions. In VLM's visual encoder preprocessing, the input image I is transformed into a sequence of patches and visual tokens. First, I is divided into a fixed number of non-overlapping and equal-sized **patches**. Each patch is a local fragment of I, which is then projected into a corresponding visual token, serving as the fundamental atomic unit for the model's attention mechanism.

A **region** is a spatial crop comprised of multiple spatially contiguous (i.e., neighboring) patches. Compared to a single patch, a region covers a larger area to capture higher-level, concrete semantic information that a single atomic unit (i.e., patch) lacks. Compared to the entire image I, a selected region enables the provision of finer-grained visual details.

Mapping from Patches to Regions. For a given input image, and given hyperparameters s_r and s_g, the mapping from patches to regions is deterministic. However, this relies on an intermediate concept, the "grid". Specifically, in the AVP module, the input image is first divided into $s_g \times s_g$ grids according to the set "Grid size" s_g. Each grid is comprised of multiple patches. For example, Chameleon (Team, 2024) divides the input image into $32 \times 32 = 1024$ patches. Therefore, each grid contains $(32/s_g) \times (32/s_g) = 1024/{s_g}^2$ patches. Furthermore, each region is a group of spatially continuous grids, consisting of $s_r \times s_r$ grids.

Interleaved-Modal CoT (I-MCoT). Built upon a VLM backbone, the I-MCoT paradigm fundamentally relies on two key components: selection and triggering. Compared to direct response (Equation 7), the trigger mechanism temporarily pauses the textual autoregressive process when specific conditions are met. During this suspension, the I-MCoT method selects salient visual evidence from the input image, which is then tokenized and concatenated into VLM's context to inform and support subsequent reasoning.

C Supplementary Details Regarding Motivation in Section 3.1

C.1 Experiment for Sufficiency Check

In the sufficiency check in Section 3.1, we assess the impact of the most attended regions on the baseline model (i.e., ICoT (Gao et al., 2025)) by masking them out.

Experimental Setup. We conduct experiments with four VLM backbones (Chameleon-7B (Team, 2024), Janus-Pro-7B (Chen et al., 2025), Qwen2-VL-7B (Wang et al., 2024), and Qwen2.5-VL-32B (Bai et al., 2025)) and report results averaged across all backbones. We adopt the zero-shot setting.

Experimental Results. The results are shown in Table 8. As K_{mask} increases, performance drops slightly, but the decline remains small. Even when $K_{\text{mask}} = 50$, the model does not exhibit a substantial degradation in performance.

Table 8: Performance difference of the baseline model (ICoT, 0-shot) when the Top K_{mask} regions on the attention map are masked.

K_{mask}	0	10	20	30	50
M3CoT	0%	+0.05%	-0.11%	-0.36%	-0.73%
ScienceQA	0%	-0.03%	-0.08%	-0.26%	-0.55%

C.2 Experiment for Text-vision Granularity Imbalance in Necessity Check

As mentioned in the necessity check in Section 3.1 and analyzed in Section 5.5, in the vast majority of cases (specifically, over 75%), the VLM fails to distribute its attention toward the truly crucial regions, i.e., those containing salient task-relevant information.

Table 9: Analysis of the impact of mitigating text-vision granularity imbalance on VLM's attention alignment. *Raw Query* denotes the baseline using the brief textual query, while *Enhanced Query* incorporates the fine-grained image description generated by CAG.

Input Context	Mean IoU (%)	Δ IoU	*p*-value
Raw Query (Baseline)	4.5	-	-
Enhanced Query (48 Tokens)	12.2	+7.7	< 0.001
Enhanced Query (96 Tokens)	18.6	+14.1	< 0.001
Enhanced Query (144 Tokens)	24.1	+19.6	< 0.001

In this section, we present experimental evidence demonstrating that mitigating the *text-vision granularity imbalance* inherent in VQA benchmarks significantly enhances the VLM's ability to accurately focus on salient visual content. Specifically, this is achieved by eliciting the VLM to extract the visual information from the informative image in textual form and transferring it to the otherwise brief raw query.

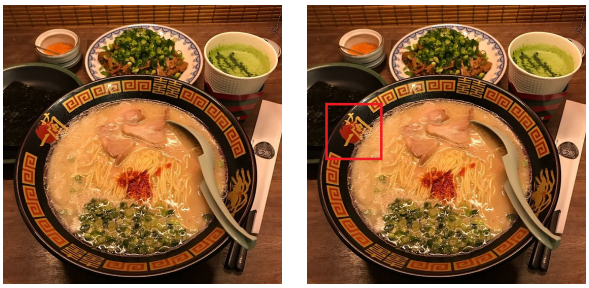
Experimental Setup. We adhere to the experimental setup detailed in Section 5.5. The mitigation of granularity imbalance is implemented via CAG component proposed in Section 4.2. To assess whether mitigating granularity imbalance significantly improves the VLM's attention distribution, we compare the Intersection over Union (IoU) between the most attended regions and the ground-truth crucial regions before and after the mitigation of imbalance. Furthermore, Wilcoxon signed-rank test is adopted to verify the statistical significance of the improvement.

Experimental Results. The results are reported in Table 9. Evidently, by enriching the originally brief textual query with additional semantic anchors, the VLM is able to more accurately localize the critical regions within the input image. Moreover, we observe that this improvement becomes increasingly pronounced as the length of the CAG-generated description increases. The Wilcoxon signed-rank test confirms that these improvements are highly statistically significant ($p < 0.001$).

D Explanation for Figures

D.1 Explanation for Figure 1

The example adopted in Section 3.2 is the 22nd question in the LLaVA-W benchmark. The query for this question is: *"What's the name of the restaurant serving these dishes?"* and the associated image is a close-up photo of a meal at ICHIRAN,



(a) Close-up meal photo (query: "What's the name of the restaurant serving these dishes?").

(b) Ground-truth cue is localized to a tiny area on the bowl rim (red box).

Figure 4: An intuitive example from LLaVA-W shows an ICHIRAN meal close-up. The left image is the raw image, while the small region containing the truly crucial information for answering the question is highlighted with the red box.

as shown in Figure 4a. This is a representative and challenging example because the query is extremely brief, while the image contains highly dense information: although it includes ramen, side dishes, and tableware, the truly task-relevant visual evidence occupies only a tiny portion of the image—a small region on the left side of the bowl's rim (Figure 4b).

Figure 1 compares the visual evidence selected by different methods. The left and right subfigures visualize the top-3 sets of attention-based regions and information-driven regions, respectively. The first, second, and third sets are colored red, purple, and blue, respectively; each set contains 72 patches.

As shown in Figure 1a, the attention-based regions are scattered, making it difficult to convey complete and concrete visual information. More importantly, they fail to cover the small area that contains the truly crucial evidence. In contrast, as shown in Figure 1b, the information gain-based selection guides the VLM to first focus on the inner rim of the bowl (red) accurately, where the critical information is contained. Although the VLM does not yield the final answer in this region, this indicates a correct line of reasoning, as the ground truth is situated in a highly similar area—a nearby location also on the inner wall of the bowl. Subsequently, the region ranked third (blue) precisely encompasses a large portion of the restaurant's name, which is just the answer to the question. This suggests that even for a challenging case where the text-vision granularity is highly disparate, information gain serves as a better foundation for region

selection.

D.2 Quick Guide to Framework Overview

Figure 2 illustrates AIM-CoT as a Trigger—Select—Insert loop over the VLM's multimodal context window during *Autoregressive Generation* (blue, bottom).

CAG (green, left) performs *Text Context Enhancement* by eliciting a fine-grained, query-conditioned image description and appending it to the original query. The enhanced query leads to *Refined Attention*, which both improves model reasoning and provides more reliable region candidates for subsequent selection.

During decoding, **DAT (yellow, top)** tracks the *Quantification of Text-to-vision Attention Shift*; when the shift is significant (i.e., $\Delta A_v(t) > \delta$), the *Triggering Condition* is met and a visual insertion is activated.

Once triggered, **AVP (purple, middle)** conducts *Candidate Region Set Construction* by merging attention-driven and exploratory regions, performs *Information Gain (IG) Quantification* (i.e., uncertainty reduction) for candidates, and runs a greedy *Sequential Selection Process* for K steps. In each step, AVP selects the region with the highest information gain and inserts it into the context window; the next step's selection is conditioned on the updated context.

Finally, the inserted regions are accompanied by a *Critical Integration Strategy* (safety instruction) that treats this visual evidence as supplementary and requires consistency checking in the VLM. Once included in the context window, they help refine subsequent next-token predictions.

E Statistical Significance Testing

To rigorously validate the superiority of AIM-CoT, we conduct statistical significance tests comparing our method against the runner-up performance. The one-sided approach is chosen because our hypothesis specifically posits that AIM-CoT outperforms the baseline.

M3CoT and ScienceQA. Since Accuracy (ACC) is derived from binary outcomes, we employ the one-sided McNemar's Test. This focuses on assessing whether the number of instances where AIM-CoT corrects ICoT's errors significantly exceeds the number of instances where it introduces new errors.

LLaVA-W. For the ROUGE-L metric, which involves continuous scores, we utilize the one-sided Wilcoxon Signed-Rank Test (alternative hypothesis: AIM-CoT outperforms the runner-up). This non-parametric paired test evaluates whether the distribution of improvement scores is significantly positive.

In Table 1, entries marked with † and ‡ denote statistical significance with $p < 0.05$ and $p < 0.01$, respectively. The detailed p-values and statistics are shown in Table 10.

F Validation of Motivation in Section 3.3: Attention Shifts as Dynamic Triggers

Experimental Setup. We take ICoT (Gao et al., 2025) as a baseline model, which is required to answer all questions from the LLaVA-W benchmark in a 0-shot setting, with ROUGE-L used as the evaluation metric. The hyper-parameters follow the default settings of the open-source implementation for ICoT, and all experiments are conducted with the Chameleon-7B backbone.

Formal Definition of Attention Shifts. To analyze attention shifts, we examine the averaged attention maps across all attention heads in the last three layers of the VLM during the prediction of each token t, following existing research (Jawahar et al., 2019; Tenney et al., 2019; Vig and Belinkov, 2019). The model's total attention scores allocated to the visual and text components of the input are respectively measured as follows:

$$A_{vision}(t) = \sum_{i \in \text{indices of } C_{vision}} \bar{a}_{t,i}, \quad (8)$$

$$A_{text}(t) = \sum_{j \in \text{indices of } C_{text}} \bar{a}_{t,j}, \quad (9)$$

where C_{vision}, C_{text} are the visual and text information within the context, respectively. Then, the shift in attention from the textual to the visual modality while generating token t is defined as follows:

$$\delta_t = A_{vision}(t) - A_{vision}(t-1). \quad (10)$$

$\Delta_k = [\delta_1, \delta_2, \cdots, \delta_{|\Delta_k|}]$ encompasses the model's attention shifts for each token when answering the arbitrary k-th question, where $|\Delta_k|$ is the number of tokens for answering the k-th question.

Formal Definition of Scores. For the predictions generated by the baseline model, the ROUGE-L scores are given by $List\_R$ =

Table 10: Statistical significance tests comparing AIM-CoT against the runner-up under the same backbone/setting. Entries report one-sided p-values with test statistics in parentheses.

Dataset	0/1-shot	
	0-shot	1-shot
Chameleon-7B		
M3CoT (ACC)	$0.0004^{\ddagger}$(11.5)	$0.0228^{\dagger}$(4)
ScienceQA (ACC)	$0.0000^{\ddagger}$(22.7)	$0.0246^{\dagger}$(3.87)
LLaVA-W (R-L)	$0.00001^{\ddagger}$(1502)	$0.00002^{\ddagger}$(1477)
Janus-Pro-7B		
M3CoT (ACC)	$0.0000^{\ddagger}$(15.6)	$0.0006^{\ddagger}$(10.4)
ScienceQA (ACC)	$0.0003^{\ddagger}$(11.7)	$0.0000^{\ddagger}$(23)
LLaVA-W (R-L)	$0.00012^{\ddagger}$(1414)	$0.00121^{\ddagger}$(1327)

Dataset	0/1-shot	
	0-shot	1-shot
Qwen2-VL-7B		
M3CoT (ACC)	$0.0224^{\dagger}$(4.02)	$0.0224^{\dagger}$(4.02)
ScienceQA (ACC)	$0.0223^{\dagger}$(4.03)	$0.0241^{\dagger}$(3.91)
LLaVA-W (R-L)	$0.02035^{\dagger}$(1193)	$0.03024^{\dagger}$(1170)
Qwen2.5-VL-32B		
M3CoT (ACC)	$0.0075^{\ddagger}$(5.92)	$0.0012^{\ddagger}$(9.18)
ScienceQA (ACC)	$0.0028^{\ddagger}$(7.67)	$0.0107^{\dagger}$(5.3)
LLaVA-W (R-L)	$0.00039^{\ddagger}$(1371)	$0.00000^{\ddagger}$(1538)

Notes: $^{\dagger}p < 0.05$, $^{\ddagger}p < 0.01$ (one-sided). For ACC. (M3CoT/ScienceQA) we use McNemar's test and report the χ^2 statistic in parentheses; for ROUGE-L on LLaVA-W we use the Wilcoxon signed-rank test and report the W statistic in parentheses.

$[R_1, R_2, \cdots, R_{|List\_R|}]$, where R_k is the score for the model's response to the k-th question, and $|List\_R|$ is the number of questions within the benchmark.

Based on these concepts, we design a two-part experiment:

Experiment 1: Correlation Analysis. We investigate the relationship between the proportion of visual insertions under significant attention shifts and the score of the corresponding generated prediction.

First, to identify whether a visual insertion is conducted during a significant attention shift, we define a high attention growth threshold, $\delta_k^{(h)}$ for the k-th response ($\delta_k^{(h)}$ is set to the 80% upper quantile of Δ_k by default). An insertion is considered to have been conducted under a significant shift and referred to as a *synchronized insertion* if and only if its corresponding attention shift value exceeds the threshold $\delta_k^{(h)}$.

Next, since the model can conduct multiple insertions per response for a question, we calculate P_k, the proportion of synchronized insertions out of the total number of insertions for the k-th question.

Finally, since the proportions of synchronized insertions $[P_1, P_2, \cdots, P_{|List\_R|}]$ and the ROUGE-L scores for all the questions $[R_1, R_2, \cdots, R_{|List\_R|}]$ are obtained, the Pearson Correlation coefficient can be computed. Specifically, the Pearson Correlation is 0.2166 with a p-value of 0.048, which suggests that the proportions of the synchronized insertions and the corresponding score are significantly positively related to each other.

Experiment 2: Group Analysis. We investigate the relationship between the proportion of synchronized insertions and the quality of the model's response.

To group the generated predictions according to response quality, we establish high- and low-scoring groups. All predictions are ranked in descending order by their ROUGE-L scores. The top 30% form the high-scoring group (high-quality responses) G_h, and the bottom 30% form the low-scoring group (low-quality responses) G_l.

Then, we calculate the mean proportion of synchronized insertions for groups G_h, G_l, which are denoted as $\bar{P}_h, \bar{P}_l$, respectively.

Finally, the means of the two groups $\bar{P}_h, \bar{P}_l$ are compared, and a t-test is performed to assess the statistical significance of the difference. Specifically, we find that $\bar{P}_h = 0.8889, \bar{P}_l = 0.5000$, which suggests that in the high-scoring group, approximately 89% of insertions are the synchronized insertions with significant attention shift from textual input to visual information; in contrast, in the low-scoring group, only about half of the insertions are synchronized insertions. Besides, the p-value of t-test is as low as 0.0019, which demonstrates that the result is highly statistically significant.

G Analysis of δ: Sensitivity and Adaptivity

The hyper-parameter δ within the DAT module serves as a crucial threshold to trigger the AVP module, which inserts salient visual regions to improve the construction of the multimodal CoT. In this section, we provide a comprehensive analysis of this parameter from two perspectives: (1) a sensitivity analysis to understand its impact on performance and triggering frequency, and (2) an exploration of an adaptive thresholding strategy to demonstrate the framework's robustness and generalizability without per-dataset tuning.

Sensitivity Analysis We detail a sensitivity analysis of δ by adjusting it across the range of $[0.1, 0.125, 0.15, 0.175, 0.2, 0.225]$ and examining not only the performance of AIM-CoT but also the number of times the AVP is triggered. The experiments are conducted under the 0-shot setting on the Chameleon-7B backbone and LLaVA-W benchmark. The experimental results are shown in Figure 5.

The left figure illustrates that AIM-CoT exhibits limited performance when the threshold, δ, is set too low. This underscores the importance of inserting visual information at critical moments: excessively frequent or inopportune visual insertions can disrupt the VLM's reasoning process, leading to suboptimal performance. As δ increases, the model's performance progressively improves, reaching its peak at $\delta = 0.2$ (our default setting), which corresponds to a ROUGE-L score of 0.2983. However, a further increase in δ results in a slight performance degradation. This highlights the criticality of visual information insertion for constructing an interleaved Chain of Thought: an overly stringent threshold excessively impedes the incorporation of visual data, preventing the AVP from supplying the model with necessary visual supplementation in a timely manner.

Conversely, the right figure demonstrates a consistent decrease in the number of times the AVP is triggered as δ is raised. This showcases the efficacy of δ as a threshold for modulating the activation frequency of the AVP.

Adaptive Thresholding Strategy While our main experiments utilize a fixed δ tuned for specific benchmarks (e.g., $\delta = 0.5$ for M3CoT, and $\delta = 0.2$ for ScienceQA and LLaVA-W), we acknowledge that the optimal threshold is intrinsically linked to the granularity of reasoning required by the task. M3CoT involves complex logic with sparse visual queries, leading to sharp, intense attention shifts (requiring a higher threshold to filter noise). In contrast, ScienceQA and LLaVA-W involve generating explanations or descriptions that require continuous visual grounding, resulting in smoother attention shifts (favoring a lower threshold).

To address this variance and enhance the deployability of AIM-CoT without requiring per-dataset tuning, we propose and evaluate an **Adaptive Z-Score Triggering** mechanism. Instead of a fixed absolute threshold, this method employs a dynamic relative threshold based on the statistical properties

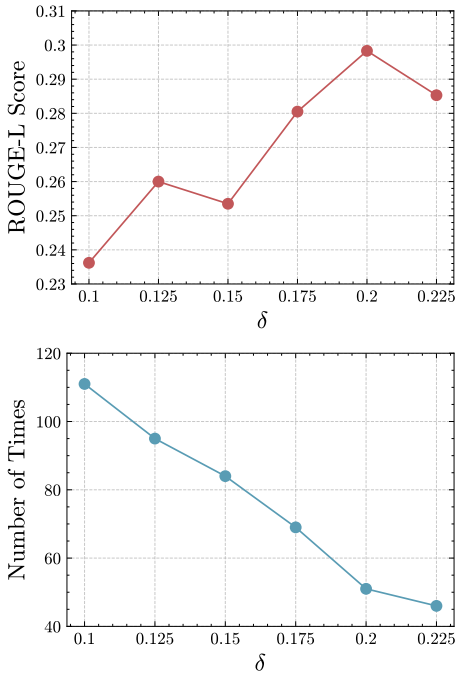


Figure 5: Experimental results of the sensitivity analysis of the hyper-parameter δ. The left figure illustrates the performance of AIM-CoT when δ takes different values, while the right one shows the number of times the AVP module within AIM-CoT is triggered.

of the attention shifts within the current generation context. Specifically, we calculate the Z-score of the current attention shift $\Delta A_{vision}(t)$ relative to the moving average of the previous W tokens:

$$Z_t = \frac{\Delta A_{vision}(t) - \mu_{t-W:t-1}}{\sigma_{t-W:t-1} + \epsilon}, \qquad (11)$$

where μ and σ are the mean and standard deviation of the attention shifts over a window W (set to 5 by default), and ϵ is a small constant for numerical stability. The trigger is activated when $Z_t > \lambda$, where λ is a universal sensitivity parameter (set to 3).

As presented in Table 11, the Adaptive Z-Score strategy demonstrates remarkable robustness. On the powerful Qwen2.5-VL-32B backbone, the adaptive method matches the performance achieved with the tuned threshold on ScienceQA (76.8) and shows negligible deviation on M3CoT (-0.1), suggesting that stronger models exhibit more distinct attention patterns that are easier to capture dynamically. Even on smaller models like Janus-Pro-7B, the performance gap remains minimal (e.g., only 0.1 drop on LLaVA-W). This confirms that while a tuned fixed threshold offers a slight edge, the adaptive strategy provides a highly competitive and deployment-friendly alternative.

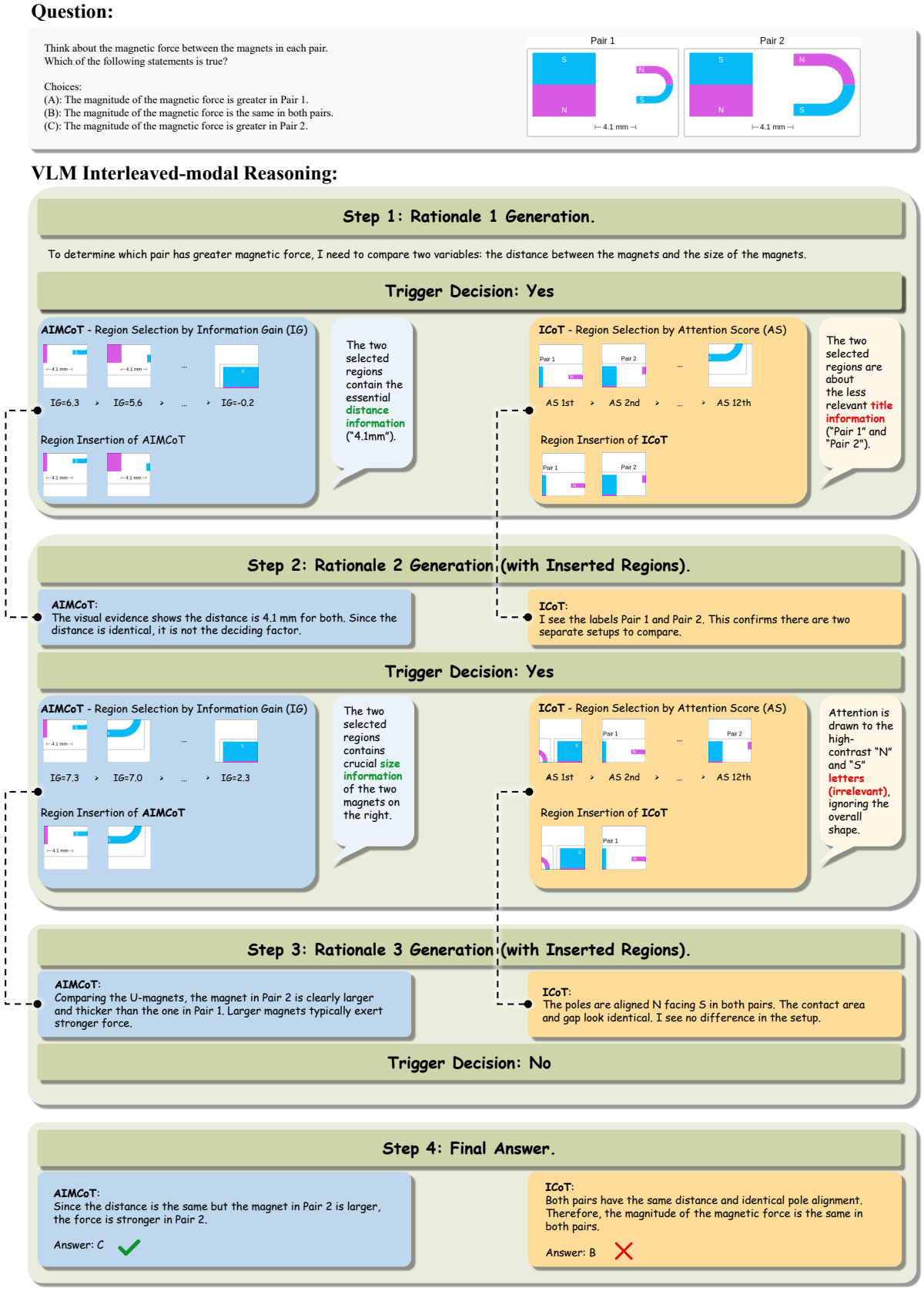


Figure 6: **Qualitative comparison on ScienceQA (Question 244) with Chameleon-7B. Left (Blue):** AIM-CoT actively selects regions based on Information Gain (IG). It correctly focuses on the *distance* (Step 1) and *magnet size* (Step 2), which are the actual visual variables required to solve the physics question. **Right (Yellow):** The baseline ICoT, relying on raw Attention Scores (AS), fails to filter out noise. It attends to irrelevant text ("Pair 1") and decorative letters ("N/S") rather than the physical attributes of the objects, resulting in reasoning failure.

Table 11: Performance comparison between the optimal Fixed Threshold (Tuned) and the Adaptive Z-Score Strategy across three backbones (0-shot setting).

Dataset	Method	Setting	Performance		
			Chameleon-7B	Janus-Pro-7B	Qwen2.5-VL-32B
M3CoT (ACC.)	Fixed Threshold	$\delta = 0.5$	31.4	39.7	58.7
	Adaptive Z-Score	$\lambda = 3$	31.1	39.5	58.6
ScienceQA (ACC.)	Fixed Threshold	$\delta = 0.2$	53.1	56.9	76.8
	Adaptive Z-Score	$\lambda = 3$	52.9	56.7	76.8
LLaVA-W (ROUGE-L)	Fixed Threshold	$\delta = 0.2$	29.8	35.5	46.5
	Adaptive Z-Score	$\lambda = 3$	29.5	35.4	46.3

H Qualitative Case Study: Information Gain vs. Attention

To complement the quantitative analyses in Sections 5.5 and 5.6, we provide a step-by-step visualization of the reasoning process to demonstrate the efficacy of our proposed AVP module.

Figure 6 presents a qualitative comparison between AIM-CoT (Ours) and the baseline ICoT on a challenging question from ScienceQA (Question ID: 244) using the Chameleon-7B (Team, 2024) backbone. As observed, the Information Gain-driven selection in AIM-CoT successfully identifies task-relevant visual evidence (e.g., *magnet size* and *distance*) that directly contributes to the correct reasoning chain. In contrast, the Attention-driven baseline is distracted by high-frequency but irrelevant features, such as the text labels ("Pair 1") or high-contrast letters ("N", "S"), leading to hallucinated reasoning and an incorrect answer. This visual evidence reinforces our hypothesis that high-attention regions do not necessarily equate to high-value information for reasoning.

I Comprehensive Analysis of AVP Component

In this appendix, we conduct a comprehensive analysis of the AVP component to validate its design choices and theoretical foundations. The detailed analyses are organized as follows:

- Appendix I.1 compares different strategies for constructing the candidate set to identify the most effective composition.
- Appendix I.2 quantifies the specific contribution of the exploratory set by analyzing the source distribution of the selected regions.
- Appendix I.3 evaluates the robustness of AVP against adversarial attention noise, demonstrating its intrinsic noise-rejection capability.
- Appendix I.4 provides a theoretical justification for (1) the approximate submodularity of the information gain function and (2) the optimality of our greedy selection algorithm.

I.1 Construction of Candidate Set C: Method Comparison

In this section, we investigate the influence of different compositions of the total set C on the performance of our proposed AIM-CoT. We specifically examine two primary configurations:

Constructing C using only C_{attn} or C_{exp}. For the latter, we evaluate three distinct construction methodologies for C_{exp}: (a) C_{rand}: uniform random sampling; (b) C_{ss}: the selective search algorithm (Uijlings et al., 2013), which is the seminal region proposal method utilized in R-CNN (Girshick et al., 2014); (c) C_{fsam}: FastSAM (Zhao et al., 2023), a computationally efficient variant of the foundational vision segmentation model, SAM (Kirillov et al., 2023).

Constructing C using both C_{attn} and C_{exp}. Similarly, as for C_{exp}, we also consider its diversified construction, including C_{rand}, C_{ss}, and C_{fsam}.

It is worth noting that although the construction is diverse, the size of C remains consistent. When C is composed of C_{attn} and C_{exp}, the two each account for half. The experimental results are shown in Table 12.

As observed, the combination of the two sets (i.e., $C = C_{attn} \cup C_{exp}$) invariably yields superior performance for AIM-CoT compared to configurations where either $C = C_{attn}$ or $C = C_{exp}$ is used exclusively. This highlights the importance of diversifying the sources of candidate visual regions.

When comparing the different construction methods for C_{exp}, the performance gap among models is marginal when used in conjunction with C_{attn}. Specifically, despite its simplicity, random sampling achieves highly competitive results, which motivates our choice to adopt it as the default method for constructing C_{exp}. Intuitively, the advantage of random sampling lies in its ability to provide regions across different parts of the image unbiasedly with maximal spatial diversity.

I.2 Source Distribution Analysis: Quantifying Contribution of Exploratory Set C_{exp}

In this section, we examine the distribution of sources for the visual regions selected by the AVP module of AIM-CoT. These regions are drawn from two sets, C_{attn} and C_{exp}, with their respective selection proportions denoted as P_{attn} and P_{exp}. Intuitively, P_{exp} reflects the significance of incorporating the exploratory set C_{exp} to construct a better multimodal CoT. A larger value of P_{exp} indicates that the exploratory set C_{exp} makes a greater contribution by providing informative salient regions to AIM-CoT, and vice versa.

Experimental Setup The experiments are conducted on the M3CoT and LLaVA-W benchmarks. Our proposed AIM-CoT is implemented with the Chameleon-7B backbone under a default 0-shot setting. To ensure the reliability of the results, we repeat each experiment three times on both benchmarks.

Results and Analysis As presented in Table 13, although the value of P_{exp} fluctuates across different experimental runs on the same benchmark, it remains consistently around 20% on M3CoT and 30% on LLaVA-W. This indicates that the influence of stochastic factors on the source distribution of the selected regions is limited, which validates our rationale of using this metric as a reflection of the relative importance of C_{attn} and C_{exp}. Furthermore, we observe that P_{exp} is significantly greater than zero. This demonstrates that the exploratory set C_{exp} consistently serves as a critical component of the total candidate set C, contributing a substantial portion of the informative regions for AIM-CoT.

I.3 Robustness of AVP Against Attention Noise

In Section 4.3, we posit that AVP possesses an intrinsic noise-rejection capability, formalized in Equation 4, allowing it to filter out high-attention regions that yield negligible information gain. To empirically validate this resilience against error propagation (e.g., when upstream modules produce misleading attention maps), we conduct an adversarial noise injection experiment.

Experimental Setup. We design an **Adversarial Noise Injection** protocol. For each image, we randomly select an irrelevant background patch (verified to have no overlap with ground-truth objects) and artificially override its attention score to be the global maximum in the attention map. We then compare the region selection behavior and downstream performance of two strategies using this corrupted attention map: (1) **Top-K Selection (Baseline):** Selects regions strictly based on attention scores. (2) **AVP (Ours):** Selects regions based on Information Gain (IG). Consistent with Appendix L.1, we evaluate on M3CoT, ScienceQA, and LLaVA-W using Chameleon-7B and Qwen2-VL-7B backbones.

Evaluation Metrics. (1) **Noise Rejection Rate (NRR):** The percentage of instances where the model successfully avoids selecting the injected noise patch, despite it having the highest attention score. (2) **Performance Drop (Δ):** The decline in task performance compared to the clean (non-adversarial) setting. Smaller Δ indicates higher robustness.

As shown in Table 14, the contrast between the two methods is stark:

First, Top-K is fundamentally vulnerable to attention errors. By definition, Top-K has an NRR of 0.0%, as it blindly trusts the manipulated attention scores. Consequently, it forces the inclusion of irrelevant visual noise, leading to significant performance degradation across all benchmarks (e.g., dropping 4.2% accuracy on M3CoT with Chameleon-7B).

Second, AVP demonstrates exceptional noise immunity. Even when the noise patch is forced to have the highest attention priority, AVP successfully rejects it in over 95% of cases (NRR $> 95\%$). This confirms that AVP correctly identifies that the noise patch, despite its high attention score, provides minimal reduction in uncertainty (low IG). As a result, the performance drop is negligible ($\Delta \approx 0$), proving that AVP effectively breaks the chain of error propagation, ensuring robust reasoning even when the guidance from the attention map is unreliable.

Table 12: Performance comparison of AIM-CoT variants on the basis of different constructions of the candidate set C.

Construction of C	M3CoT (ACC.)	LLaVA-W (ROUGE-L)
$C_{attn} \cup C_{rand} \quad (C_{exp} = C_{rand})$	31.4	29.8
$C_{attn} \cup C_{ss} \quad (C_{exp} = C_{ss})$	31.2	29.5
$C_{attn} \cup C_{fsam} \quad (C_{exp} = C_{fsam})$	31.0	29.6
C_{attn}	30.8	28.9
C_{rand}	30.4	28.6
C_{ss}	30.3	28.7
C_{fsam}	29.9	27.7

Table 13: Proportion of salient regions selected by the AVP module of our proposed AIM-CoT from the exploratory set C_{exp}.

Experiment Number	1	2	3
M3CoT	17.25%	20.44%	27.27%
LLaVA-W	31.33%	25.77%	26.67%

I.4 Theoretical Justification: Approximate Submodularity and Optimality of Greedy Selection

To offer a more comprehensive insight into the motivation for employing a greedy algorithm, this section provides a thorough analysis. We emphasize that for functions that are not theoretically submodular, a greedy approach remains one of the conventional methods for addressing their maximization, as established in recognized works (Bian et al., 2017; Sener and Savarese, 2018; Kim et al., 2016; Krause et al., 2008; Das and Kempe, 2011). In this part, we conduct meticulously designed experiments to investigate the extent to which the information gain function F approximates submodularity. The experimental results reveal that F empirically exhibits significant submodular characteristics. This finding motivates us to follow established works (Bian et al., 2017; Sener and Savarese, 2018; Kim et al., 2016; Krause et al., 2008; Das and Kempe, 2011) to propose a greedy algorithm to solve the maximization problem for F. The analysis is detailed as follows:

Firstly, we would like to introduce the definition of a submodular function. According to existing research (Nemhauser et al., 1978), a function f is a submodular function if it satisfies

$$f(A \cup \{R_i\}) - f(A) \geq f(B \cup \{R_i\}) - f(B) \tag{12}$$

for any sets $A \subseteq B \subset C$ and any element that satisfies $R_i \in C \backslash B$. In our scenario, the Inequality 12 is written as

$$F(A \cup \{R_i\}) - F(A) \geq F(B \cup \{R_i\}) - F(B) \tag{13}$$

for any $A \subset B \subset C$ and any $R_i \in C \backslash B$, which means that the information gain from incorporating a visual region exhibits a diminishing returns property.

To demonstrate this empirically, we design the experiment detailed as follows, aiming to show that for two sets of regions of different sizes, $S_{small} \subset S_{large} \subset C$, the information gain from incorporating a given visual region $R_{test} \in C \setminus S_{large}$ into the context of a VLM is greater when R_{test} is added to S_{small} than when it is added to S_{large}, ceteris paribus.

Experimental Setup. In our experimental design, each time the AVP process is triggered to select salient regions, we first execute it to select K_{small} regions from the total candidate pool C to form the set S_{small}. Subsequently, building upon S_{small}, we select an additional $K_{large} - K_{small}$ regions to construct the set S_{large}, where K_{small} and K_{large} are the respective set sizes. This construction inherently ensures that $S_{small} \subset S_{large}$.

Next, to compute the information gain contributed by a given region, we randomly sample a region R_{test} from $C \setminus S_{large}$. We then calculate the VLM's information contents, which are denoted as U_s, U_s^*, U_l, and U_l^*, when the context incorporates (1) S_{small}, (2) $S_{small} \cup \{R_{test}\}$, (3) S_{large}, and (4) $S_{large} \cup \{R_{test}\}$, respectively. We expect to observe in the majority of cases that:

$$U_s^* - U_s \geq U_l^* - U_l. \tag{14}$$

We conduct experiments on the M3CoT and LLaVA-W benchmarks, setting $K_{small} \in$

Table 14: Robustness analysis against adversarial attention noise. NRR denotes Noise Rejection Rate (↑ higher is better); Δ denotes performance drop (↓ lower is better, closer to 0).

Backbone	Method	M3CoT		ScienceQA		LLaVA-W	
		NRR (%) ↑	Δ Acc. ↓	NRR (%) ↑	Δ Acc. ↓	NRR (%) ↑	Δ ROUGE-L ↓
Chameleon-7B	Top-K	0.0	-4.2	0.0	-3.8	0.0	-5.1
	AVP (Ours)	**96.4**	**-0.2**	**98.1**	**-0.1**	**95.5**	**-0.4**
Qwen2-VL-7B	Top-K	0.0	-3.5	0.0	-2.9	0.0	-4.4
	AVP (Ours)	**97.8**	**-0.1**	**99.2**	**0.0**	**96.7**	**-0.2**

$\{2, 3, 4, 5\}$ and $K_{large} = K_{small} + 1$ for simplicity. In terms of evaluation, we record the proportion of instances for which the inequality $U_s^* - U_s \geq U_l^* - U_l$ holds, and further introduce a Binomial Test to rigorously examine the significance of the results.

Experimental Results. The experimental results are presented in Table 15. As we can see, the Inequality 12 holds in most instances across all settings and datasets. Furthermore, to confirm the significance of the obtained results, we introduce the Binomial Test, an exact statistical procedure for assessing the extent to which experimental outcomes with a binary structure are attributable to chance alone. The p-values, presented in Table 15, are all substantially below the 0.05 significance level. This demonstrates that the information gain function F behaves in a manner that is empirically near-submodular, which motivates us to follow existing research (Bian et al., 2017; Sener and Savarese, 2018; Kim et al., 2016; Krause et al., 2008; Das and Kempe, 2011) where greedy algorithms are proposed to solve the problem of maximizing approximately submodular functions.

Table 15: Proportions of instances on M3CoT and LLaVA-W benchmarks for which the approximate submodularity of information gain function F is manifested. The backbone is Chameleon-7B and the model is our proposed AIM-CoT. K_{large} is set to $K_{small} + 1$ for simplicity. The significance levels of these results are listed below them.

K_{small}	2	3	4	5
M3CoT (n=2318)	72.00%	62.99%	67.04%	61.09%
p-value	<1e-6	<1e-6	<1e-6	<1e-6
LLaVA-W (n=60)	61.67%	68.33%	61.67%	63.33%
p-value	0.0462	0.0031	0.0462	0.0249

J Extensive Ablation Studies on More Backbones

Section 5.3 presents an ablation study on the Chameleon-7B backbone. The results provide two key insights as follows:

- Each component (CAG, AVP, DAT) contributes substantially to the overall performance.
- Even without CAG, combining AVP and DAT is sufficient for AIM-CoT to outperform strong baseline models by a clear margin. This finding highlights the importance of addressing the "what to see" and "when to see it" questions.

To test whether these insights generalize beyond Chameleon-7B, in this section, we extend the same ablation study to two additional VLM backbones: Janus-Pro-7B and Qwen2.5-VL-32B.

As shown in Tables 16 and 17, the ablation trends are consistent with those in Section 5.3: removing any component leads to performance degradation across datasets and backbones, demonstrating the efficacy of our proposed components.

More importantly, the results highlight the central thesis of this study. **Even without CAG, the combination of AVP and DAT already yields strong performance and is sufficient to surpass strong baselines, including ICoT**. This demonstrates that explicitly and effectively addressing *what to see* (AVP) and *when to see it* (DAT) is the key driver of AIM-CoT's advantage.

Meanwhile, CAG remains a non-trivial contributor: although its absolute gains are comparatively smaller, it consistently improves results across all settings, suggesting that mitigating the text-vision granularity mismatch further complements AVP+DAT.

K In-depth Analysis of Performance Gain Variations Across Backbones

In Table 1, we observe that AIM-CoT yields varying degrees of improvement across different base

Table 16: Ablation study of AIM-CoT conducted on Janus-Pro-7B under 0-shot setting.

Dataset	AIM-CoT	w/o CAG	w/o AVP	w/o DAT
M3CoT (ACC.)	39.7	39.4 (-0.3)	38.6 (-1.1)	39.0 (-0.7)
ScienceQA (ACC.)	56.9	56.7 (-0.2)	55.8 (-1.1)	56.3 (-0.6)
LLaVA-W (ROUGE-L)	35.5	34.7 (-0.8)	33.5 (-2.0)	34.5 (-1.0)

Table 17: Ablation study of AIM-CoT conducted on Qwen2.5-VL-32B under 0-shot setting.

Dataset	AIM-CoT	w/o CAG	w/o AVP	w/o DAT
M3CoT (ACC.)	58.7	58.4 (-0.3)	57.6 (-1.1)	58.1 (-0.6)
ScienceQA (ACC.)	76.8	76.6 (-0.2)	75.9 (-0.9)	76.3 (-0.5)
LLaVA-W (ROUGE-L)	46.5	46.1 (-0.4)	44.2 (-2.3)	45.0 (-1.5)

models. Specifically, the gains on Chameleon-7B are substantially larger than those on Qwen2-VL-7B, while Qwen2.5-VL-32B exhibits higher improvements than its 7B counterpart. In this section, we analyze the underlying factors contributing to these phenomena from two perspectives: visual encoding mechanisms and model capability scaling.

K.1 Cross-Family Analysis: Architecture and Visual Encoding

The distinct performance gap between Chameleon-7B (up to 18.25% gain on LLaVA-W) and Qwen2-VL-7B (up to 6.14% gain) can be attributed to their fundamental differences in visual processing:

- Chameleon-7B (Early-Fusion & Fixed Resolution): As an early-fusion model utilizing a fixed tokenization strategy, Chameleon-7B often struggles with fine-grained visual details in high-resolution images, leading to the "myopic" attention behavior discussed in Section 3.1. AIM-CoT acts as a **perceptual correction** mechanism here. By actively cropping and zooming in on salient regions via the AVP module, AIM-CoT directly compensates for the model's native resolution limitations. The improvement is transformative as it solves a "can't see" problem, resulting in substantial gains.
- Qwen2-VL Series (NaViT & Dynamic Resolution): These models employ the NaViT architecture with dynamic resolution support, allowing them to process arbitrary aspect ratios and resolutions natively. Since Qwen2-VL can already perceive visual details relatively clearly, the marginal utility of "zooming in" is lower compared to Chameleon. For Qwen models, AIM-CoT serves more as an **attention optimization** tool rather than a vision repair tool, leading to more moderate base improvements.

K.2 Intra-Family Analysis: Capability Dependence and Scaling Laws

Within the Qwen family, we observe that the larger Qwen2.5-VL-32B benefits more from AIM-CoT (up to 9.84% gain) than the smaller Qwen2-VL-7B (up to 6.14% gain). This counter-intuitive trend (where a stronger baseline yields larger relative gains) highlights the **capability-dependent nature** of our framework:

- Quality of CAG Generation: AIM-CoT is a closed-loop system where the Context-enhanced Attention-map Generation (CAG) module relies on the model's own capability to describe the image. The 32B model generates more precise, hallucination-free descriptions than the 7B model. This high-quality context leads to more accurate attention maps, enabling the AVP module to select significantly more valuable image regions.
- Interleaved Reasoning Capability: The efficacy of AIM-CoT also depends on how well the model can leverage the inserted visual patches. Larger models (32B) possess stronger long-context understanding and multi-step reasoning capabilities. They can effectively synthesize the fragmented information provided by the inserted patches to deduce the correct answer, whereas smaller models might struggle to integrate these additional visual cues coherently.

In summary, AIM-CoT functions as a mechanism for Perceptual Correction on weaker architectures (Chameleon) and Reasoning Augmentation on stronger ones (Qwen 32B), demonstrating its adaptability across different model paradigms.

L Safety Mechanisms in AIM-CoT

To ensure the reliability of the reasoning chain and prevent error propagation, AIM-CoT incorporates explicit safety mechanisms at two critical stages. First, regarding **Source Quality (Input)**, we apply *Negative Constraints* within the CAG module to strictly ground the generated context and mitigate hallucinations. Second, regarding **Integration Safety (Process)**, we employ *Safety Instructions* during the visual insertion, guiding the model to verify relevance and filter out potential noise. In this section, we analyze the individual contributions of these safeguards to the framework's overall robustness.

L.1 Effectiveness of Negative Constraints in CAG Generation

In Section 4.2, we introduce negative constraints within the CAG prompt $\mathcal{P}_{CAG}$ to ensure the utmost precision of the generated descriptions. To rigorously assess the necessity of this design and the potential risk of error propagation, we conduct an ablation study focusing on the hallucination rate of the generated descriptions $\mathcal{D}_{CAG}$ and their impact on downstream performance.

Experimental Setup. We contrast the standard AIM-CoT against **AIM-CoT w/o Constraints**, a variant where the cautionary instructions are removed from $\mathcal{P}_{CAG}$. To demonstrate the generalizability of our design, we implement the framework on two representative backbones: Chameleon-7B and Qwen2-VL-7B. The experiments are conducted on M3CoT, ScienceQA, and LLaVA-W benchmarks under the 0-shot setting.

Evaluation Metrics. Two metrics are incorporated for evaluation: (1) **Hallucination Rate (HR):** We employ GPT-4v as an external evaluator to identify factual inconsistencies between the image content and the generated description $\mathcal{D}_{CAG}$. HR is reported as the percentage of descriptions containing details unsupported by the visual input. (2) **Downstream Performance:** We report Accuracy (ACC.) for M3CoT and ScienceQA, and ROUGE-L for LLaVA-W to measure the final reasoning outcome.

The results are summarized in Table 18, from which two key observations are derived:

First, the CAG module exhibits high intrinsic reliability across backbones. Even without explicit negative constraints, the hallucination rates remain relatively low (e.g., $< 6\%$ on M3CoT for both models). This suggests that the foundational strategy of using VLM-generated descriptions as context is robust and not inherently prone to generating misleading information, regardless of the underlying model architecture.

Second, negative constraints provide a critical layer of safety. Despite the solid baseline, the introduction of negative constraints consistently suppresses residual hallucinations across all datasets and backbones. Notably, for Chameleon-7B on M3CoT, the constraints reduce the hallucination rate by approximately 67% (from 5.8% to 1.9%). This demonstrates that while the models are generally capable, the constraints effectively filter out subtle noise and "over-interpretations." By driving the hallucination rate down to a negligible level, we ensure that the subsequent AVP module operates on a virtually noise-free foundation, thereby maximizing overall reasoning performance.

L.2 Effectiveness of Safety Instructions in DAT Integration

In Section 4.4, we introduce a robust integration strategy within the DAT module, utilizing specific Safety Instructions (e.g., instructing the model to treat inserted regions as "supplementary references" and verify semantic consistency) to mitigate the risk of integrating irrelevant or hallucinated visual cues. To validate the efficacy of this textual safeguard, we conduct a stress test involving Irrelevant Visual Injection.

Experimental Setup. We simulate a worst-case scenario where the upstream selection process fails completely. Whenever the DAT triggers a visual insertion, instead of inserting the semantically relevant region selected by AVP, we deliberately inject a random, irrelevant image patch from a different dataset sample. We then measure the model's resilience under this interference using two configurations: (1) **Naive Injection (w/o Safety):** The irrelevant patch is inserted directly into the reasoning chain without any accompanying safety prompts. (2) **AIM-CoT Integration (w/ Safety):** The irrelevant patch is inserted accompanied by our proposed safety instructions. We employ the same backbones (Chameleon-7B, Qwen2-VL-7B) and benchmarks (M3CoT, ScienceQA, LLaVA-W) as in previous sections.

Evaluation Metrics. We report the **Performance under Noise** and the **Performance Drop (Δ)** relative to the clean, standard AIM-CoT performance (where accurate regions are inserted). A smaller magnitude of Δ indicates greater robustness against misleading visual information.

The results in Table 19 provide compelling evidence for the necessity of safety instructions:

First, unprotected visual insertion is highly risky. When irrelevant patches are naively injected, the models suffer severe performance degradation (e.g., $\sim$7% drop on M3CoT for Chameleon-7B). This confirms that without guidance, VLMs tend to overtrust visual tokens, attempting to reason over noise and consequently hallucinating or diverging from the correct logic path.

Second, Safety Instructions act as an effective semantic firewall. By simply instructing the model to verify consistency, the performance drop is dras-

Table 18: Ablation study on the effectiveness of Negative Constraints across two backbones. HR denotes Hallucination Rate (↓ lower is better); Perf. denotes downstream performance (↑ higher is better).

Backbone	Method	M3CoT		ScienceQA		LLaVA-W	
		HR (%) ↓	Acc. ↑	HR (%) ↓	Acc. ↑	HR (%) ↓	ROUGE-L ↑
Chameleon-7B	w/o Constraints	5.8	30.1	4.5	52.5	8.4	27.9
	AIM-CoT (Ours)	**1.9**	**31.4**	**1.2**	**53.1**	**2.7**	**29.8**
Qwen2-VL-7B	w/o Constraints	3.5	43.9	2.8	56.9	5.1	35.1
	AIM-CoT (Ours)	**0.8**	**44.7**	**0.5**	**57.4**	**1.5**	**36.3**

Table 19: Ablation study on Safety Instructions under irrelevant visual injection. Δ indicates the performance drop compared to the standard AIM-CoT (↓ lower drop/magnitude is better).

Backbone	Method	M3CoT		ScienceQA		LLaVA-W	
		Acc.	Δ	Acc.	Δ	ROUGE-L	Δ
Chameleon-7B	Naive Injection (w/o Safety)	24.5	-6.9	46.2	-6.9	22.1	-7.7
	AIM-CoT (w/ Safety)	**29.9**	**-1.5**	**51.8**	**-1.3**	**28.4**	**-1.4**
Qwen2-VL-7B	Naive Injection (w/o Safety)	38.2	-6.5	50.9	-6.5	30.5	-5.8
	AIM-CoT (w/ Safety)	**43.5**	**-1.2**	**56.1**	**-1.3**	**35.1**	**-1.2**

tically reduced to a marginal level (e.g., only -1.5% on M3CoT). This demonstrates that the safety instruction successfully cues the model to exercise critical judgment: when the inserted visual evidence conflicts with the textual context (as in this random injection case), the model chooses to disregard the visual noise and rely on its internal knowledge, thereby maintaining robust performance even in the presence of upstream errors.