Mitigating Low-Quality Reasoning in MLLMs: Self-Driven Refined Multimodal CoT with Selective Thinking and Step-wise Visual Enhancement

Chongjun Tu[1*], **Peng Ye**[2,3*], **Dongzhan Zhou**[2], **Tao Chen**[1,4†], **Wanli Ouyang**[2,3]

[1]College of Future Information Technology, Fudan University
[2]Shanghai Artificial Intelligent Laboratory
[3]The Chinese University of Hong Kong
[4]Shanghai Innovation Institute

Abstract

Current Multimodal Chain-of-Thought (MCoT) methods suffer from *low-quality multimodal reasoning*, characterized by overthinking on simple queries and inefficient utilization of visual information, resulting in vast inefficient and ineffective computations. In this paper, we discover that Multimodal Large Language Models (MLLMs) possess inherent capabilities to distinguish between simple and difficult queries and enhance task-related visual information, which remain underutilized by existing approaches. Based on this insight, we propose Self-Driven Refined Multimodal CoT (SDR-MCoT), a training-free framework that mitigates these issues through two self-driven modules. First, our selective thinking module employs entropy-based confidence estimation to determine whether queries require detailed reasoning, preventing overthinking on simple questions. Second, our step-wise visual enhancement module strengthens attention to relevant visual regions at each reasoning step without inserting additional tokens, achieving fine-grained visual grounding and enhancement with minimal overhead. Moreover, SDR-MCoT can be seamlessly integrated into various MLLMs, offering a practical solution for improving multimodal reasoning. Comprehensive experiments across eight benchmarks from diverse domains (multimodal reasoning, visual understanding, hallucination, and mathematical reasoning) demonstrate that SDR-MCoT consistently outperforms existing MCoT methods on four different base models with reduced overhead. For instance, on Qwen2-VL-7B, our method improves average accuracy by over 6% while reducing token consumption by approximately 60% compared to zero-shot CoT.

1 Introduction

Chain-of-Thought (CoT) techniques have significantly enhanced the reasoning capabilities of Large Language Models (LLMs) by breaking down complex problems into intermediate steps (Kojima et al. 2022). As Multimodal Large Language Models (MLLMs) evolve, researchers have adapted CoT for vision-language tasks (Tu et al. 2025; Zhang et al. 2025a), introducing Multimodal Chain-of-Thought (MCoT) approaches. These methods aim to tackle complex tasks, such as scientific question answering (Zheng et al. 2025a; Yu et al. 2025), medical diagnosis (Liu et al. 2024b), and robotic navigation (Sun et al. 2024), which require both visual understanding and multi-step reasoning.

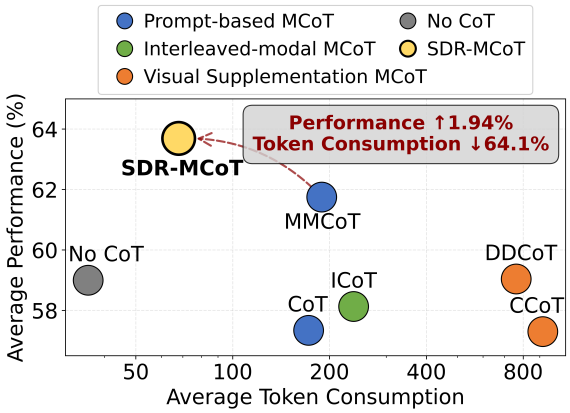

Figure 1: Performance and token consumption comparisons of different MCoT methods averaged across 8 benchmarks using Qwen2-VL-7B. Our proposed SDR-MCoT achieves superior accuracy while significantly reducing token consumption compared to existing MCoT approaches. Most existing MCoT methods show inconsistent performance across diverse task types, leading to limited average improvements.

Current MCoT approaches follow three main paradigms, as illustrated in Figure 2. Prompt-based methods apply text-based CoT prompting to MLLMs, generating text-only reasoning chains. For example, MM-CoT (Zhang et al. 2023) introduces a two-stage framework that generates rationales before answers. Visual supplementation methods convert visual information into textual context through scene graphs (Mitra et al. 2024), VQA-generated answers (Zheng et al. 2023), bounding boxes (Shao et al. 2024; Man et al. 2025), or knowledge graphs (Mondal et al. 2024). Recent interleaved-modal methods have taken a further step by directly inserting visual content at each reasoning step. Specifically, existing methods crop relevant patches from input images (Gao et al. 2025), use external tools for editing (Zheng et al. 2025b), or generate new images (Hu et al. 2024) to provide fine-grained visual information during reasoning.

Despite these advances, current MCoT methods still suf-

*These authors contributed equally.
†Corresponding author.

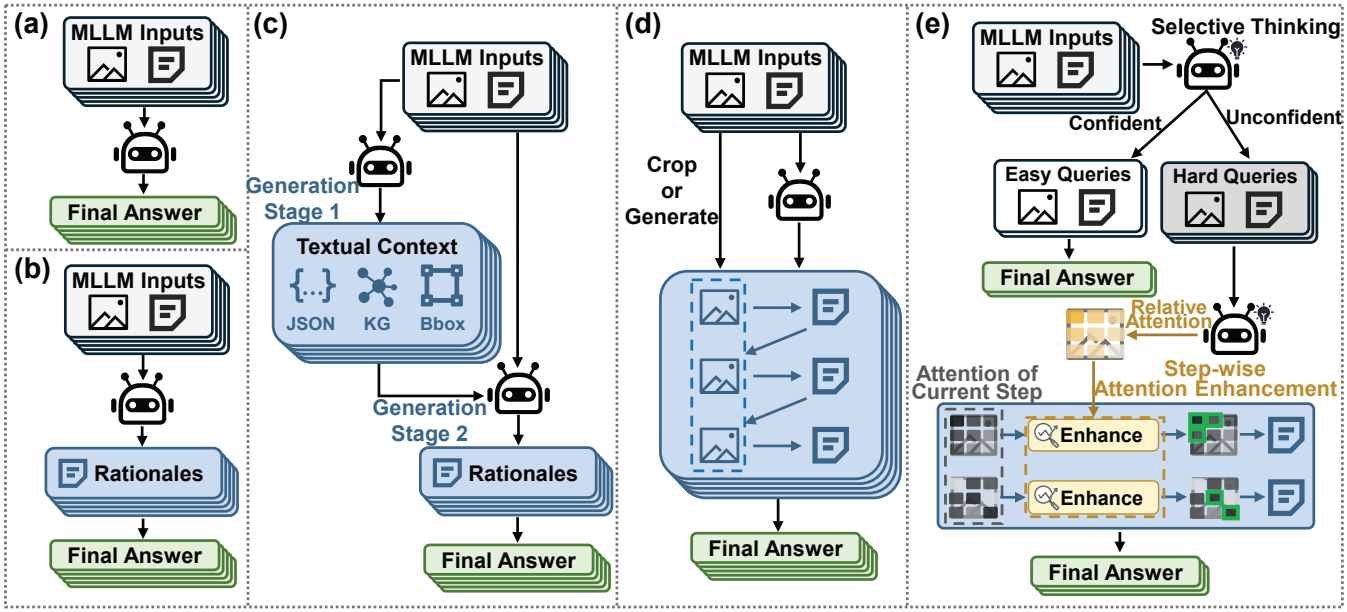

Figure 2: Illustration of different MCoT methods. (a) **No CoT**: MLLMs directly generate answers without intermediate reasoning steps. (b) **Prompt-based methods**: Generate structured reasoning chains through specific prompting instructions. (c) **Visual supplementation methods**: Convert visual evidence into textual context in the first generation stage (such as JSON file, knowledge graph and bounding box), and obtain answers in the second stage. (d) **Interleaved-modal methods**: Insert visual content (e.g., cropped patches, edited or generated images) at each reasoning step. (e) **Our proposed Self-Driven Refined MCoT**: Enables MLLMs to conduct selective thinking based on response certainty and enhance attention to relevant visual regions at each reasoning step without inserting tokens. Tokens with enhanced attention are marked with a green box.

fer from ***low-quality multimodal reasoning*** that limits their practical applications. First, most methods apply complex reasoning to all queries, resulting in the *overthinking* phenomenon, where unnecessary reasoning traces are generated for simple questions. While *overthinking* has been studied in Large Reasoning Models (LRMs) and LLMs (Chen et al. 2024c; Wu et al. 2025; Alomrani et al. 2025), it remains under-explored in MLLMs. Second, these methods utilize visual information inefficiently. Specifically, prompt-based approaches struggle to leverage visual evidence in reasoning steps. Visual supplementation methods primarily supplement at the input phase through extended prompts, resulting in substantial token overhead without fine-grained alignment with reasoning steps. Although interleaved-modal methods directly introduce visual information through image patches, they incur significant computational costs due to multiple visual processing and frequent external tool invocations, and mostly require specialized training or architectural modifications, limiting their applicability across different models.

In this paper, we discover that MLLMs possess inherent capabilities to distinguish between simple and difficult queries and enhance task-related visual information, which can effectively mitigate the *low-quality multimodal reasoning* problem but are underutilized by existing methods. Based on this insight, we propose Self-Driven Refined Multimodal CoT (**SDR-MCoT**), a framework with two key components: a self-driven selective thinking module and a self-driven step-wise visual enhancement module. Specifically, at the sample level, the selective thinking module employs an entropy-based confidence estimation mechanism via the model's predicted logits to autonomously determine whether each query requires detailed reasoning or can be answered directly. This prevents overthinking on simple questions while maintaining sufficient reasoning depth for complex problems. At the reasoning step level, the step-wise visual enhancement module strengthens attention to task-relevant visual regions at each reasoning step without inserting additional tokens. This is achieved through a relative attention mechanism that identifies and boosts attention to visual patches crucial for each reasoning step. Through the proposed SDR-MCoT, we achieve superior performance with significantly reduced token consumption. Moreover, our proposed method is completely training-free and seamlessly applicable to various MLLM architectures.

We evaluate the proposed SDR-MCoT on multiple benchmarks spanning diverse domains, including multimodal reasoning (M^3CoT (Chen et al. 2024b), CoMT (Cheng et al. 2025), ScienceQA (Lu et al. 2022)), general visual understanding (MMStar (Chen et al. 2024a), A-OKVQA (Schwenk et al. 2022), V* (Wu and Xie 2024)), hallucination evaluation (HallusionBench (Guan et al. 2024)), and mathematical reasoning (MathVista (Lu et al. 2024)). We compare SDR-MCoT with existing MCoT approaches across four base models: Qwen2-VL-7B (Wang et al. 2024), InternVL3-8B (Zhu et al. 2025), and LLaVA-1.5-7B/13B (Liu et al. 2024a). Experimental results demon-

strate that our method achieves state-of-the-art performance while utilizing significantly fewer tokens. For instance, as illustrated in Figure 1, SDR-MCoT achieves an average performance improvement of over 6% across all eight benchmarks while reducing token consumption by approximately 60% compared to the zero-shot CoT baseline on Qwen2-VL-7B. This validates that our approach effectively mitigates low-quality multimodal reasoning in MLLMs.

Our main contributions can be summarized as follows:

- We discover that MLLMs possess inherent abilities to distinguish between simple and difficult queries and enhance task-related visual information, based on which we propose a novel self-driven refined multimodal CoT (SDR-MCoT) framework that effectively improves the quality of multimodal reasoning, achieving superior performance with reduced token consumption.
- At the sample level, we introduce an entropy-based selective thinking module, enabling models to dynamically select reasoning strategies for each sample and avoid overthinking on simple queries. At the reasoning step level, we develop a step-wise visual enhancement module that strengthens attention to relevant visual regions for each reasoning step, thereby achieving fine-grained visual enhancement without requiring additional tokens.
- Comprehensive experiments on four base models and eight benchmarks from different domains demonstrate that our proposed method achieves superior performance while significantly reducing token consumption. Moreover, SDR-MCoT is totally training-free and can be seamlessly applied to various MLLM architectures.

2 Related Work

2.1 Multimodal Chain-of-Thought (MCoT)

Recent advances in MCoT aim to enhance the reasoning capabilities of MLLMs. These methods can be categorized into three main paradigms: prompt-based methods, visual supplementation methods, and interleaved-modal methods.

Prompt-based methods apply CoT prompting techniques from LLMs (Kojima et al. 2022) to MLLMs and generate structured textual reasoning steps. MM-CoT (Zhang et al. 2023) introduces a two-stage framework that generates rationales before answers. LLaVA-CoT (Xu et al. 2024) structures reasoning into four stages to maintain clarity. While these approaches demonstrate the feasibility of multimodal reasoning, they primarily rely on textual processes, potentially missing crucial visual details.

Visual supplementation methods enhance MCoT by transforming visual information into textual context. DDCoT (Zheng et al. 2023) decomposes problems and adopts VQA models to answer sub-questions for visual supplements. CCoT (Mitra et al. 2024) generates scene graphs as intermediate steps. Other methods train models to leverage bounding boxes (Shao et al. 2024; Man et al. 2025) and knowledge graphs (Mondal et al. 2024) to supplement visual evidence, thereby enhancing the reasoning process. These methods inject visual information at the input level through extended prompts, causing substantial token overhead without fine-grained alignment to different reasoning steps.

Interleaved-modal methods directly incorporate visual content into reasoning steps rather than relying solely on textual context. Several approaches utilize external tools to edit (Zhou et al. 2024; Hu et al. 2024) or generate (Xiao et al. 2024) images to reflect visual state changes during reasoning. ICoT (Gao et al. 2025) crops and inserts visual patches at each step based on attention mechanisms. Other methods use RL or SFT training to enable models to better master the paradigm of inserting (Jiang et al. 2025a; Zheng et al. 2025b) or generating (Li et al. 2025a) visual tokens during MCoT. While achieving stronger visual grounding and providing step-wise visual information, these methods incur significant overhead from multiple visual processing and tool invocations, and mostly require specialized training or architectural modifications.

In contrast, our proposed method introduces a novel step-wise visual enhancement mechanism that strengthens attention to task-relevant visual regions at each reasoning step without requiring additional tokens. This training-free approach achieves fine-grained visual grounding enhancement with minimal computational overhead and can be seamlessly integrated into various MLLM architectures, offering a practical alternative to existing MCoT paradigms.

2.2 Efficient Large Model Reasoning

With the increasing development of reasoning capabilities of LLMs and MLLMs, the simultaneously growing reasoning overhead has attracted attention. Currently, research on efficient reasoning mainly focuses on LLMs, while work on MLLMs remains relatively underexplored.

Efficient reasoning in LLMs aims to promote concise reasoning and mitigate unnecessary thinking processes to reduce inference costs. Several studies identify the overthinking problem (Chen et al. 2024c; Wu et al. 2025), where models often allocate excessive tokens to simple queries without performance gains or even causing performance decline. To address these inefficiencies, various strategies have been proposed. Token-budget-aware methods (Muennighoff et al. 2025; Han et al. 2024; Li et al. 2025b; Lin et al. 2025) estimate appropriate token budgets or problem difficulty (Huang et al. 2025; Shen et al. 2025) for each query. Other methods adaptively select reasoning depths (Zhang et al. 2025b; Xiang et al. 2025) or thinking modes (Jiang et al. 2025b; Liang et al. 2025) according to problem complexity.

Efficient reasoning in MLLMs remains less explored. Certainty-Based Adaptive Routing (Lu et al. 2025) uses perplexity-based confidence to route queries between short and long reasoning paths. Fast-Slow Thinking (Xiao et al. 2025) trains MLLMs via reinforcement learning to adaptively adjust reasoning depth based on visual complexity. Long or Short CoT (Zhang, Xiao, and Cao 2025) trains lightweight selectors to choose between reasoning strategies for multimodal tasks. However, these methods require either fine-tuning MLLMs or training external routing models, which limits their general applicability.

Compared to existing efficient MLLM reasoning methods, which require training or external components, our proposed method operates in a training-free manner by enabling

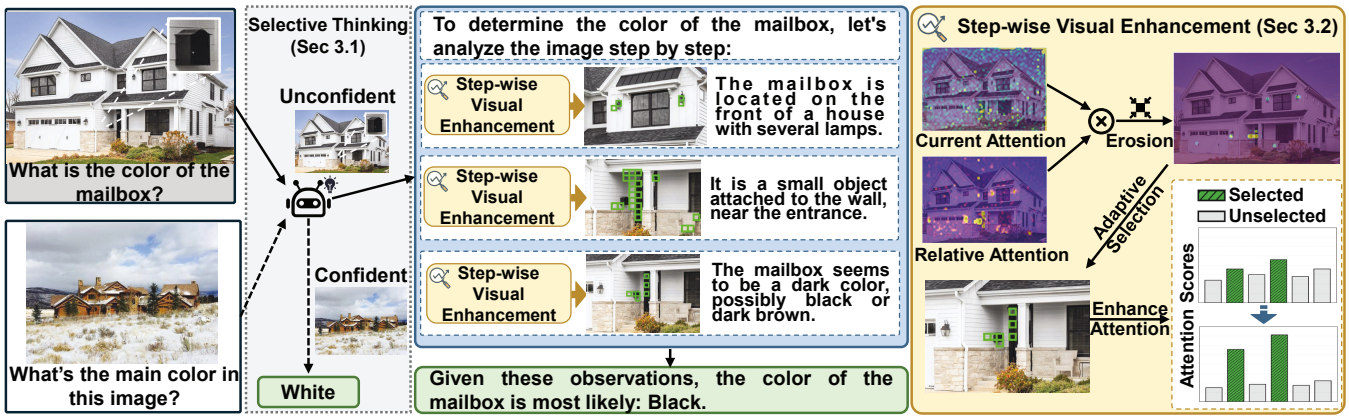

Figure 3: Overview of the proposed SDR-MCoT. **Selective Thinking:** For each query, MLLMs obtain answer confidence and determine whether to answer directly or conduct step-by-step reasoning with step-wise visual enhancement. **Step-wise Visual Enhancement:** At each reasoning step, the current visual attention map is multiplied by the relative attention and undergoes morphological erosion to obtain regions related to the question and the current step. Based on this, we adaptively select the tokens that need to be enhanced. Finally, we achieve visual enhancement for the current step by increasing the attention scores of selected tokens. Calculation of the relative attention is illustrated in Figure 5.

MLLMs to conduct selective thinking, thereby mitigating overthinking and reducing token consumption.

3 Method

In this section, we present the Self-Driven Refined Multimodal CoT (SDR-MCoT) framework, which mitigates low-quality multimodal reasoning in MLLMs. As illustrated in Figure 3, SDR-MCoT operates at both the sample level and reasoning step level. The selective thinking module assesses the query complexity at the sample level by estimating answer confidence through entropy-based analysis, determining whether a direct answer or detailed reasoning is required. For queries requiring detailed reasoning, the step-wise visual enhancement module strengthens attention to task-relevant visual regions at each reasoning step without inserting additional tokens, ensuring effective visual grounding throughout the reasoning process. This design enables SDR-MCoT to avoid overthinking on simple queries while enhancing visual understanding for complex problems.

3.1 Self-Driven Selective Thinking

Our selective thinking module is motivated by the observation that MLLMs possess inherent capabilities to assess their answer confidence. To investigate this capability, we analyze the correlation between answer accuracy and the entropy of predicted logits on the M^{3}CoT benchmark. Specifically, we compute the entropy of the first generated token when the model is prompted to provide a direct answer. Formally, given an input image I and question Q, we compute the entropy H of the predicted logits:

$$H = -\sum_{i=1}^{V} p_i \log p_i, \quad \text{where } p_i = \frac{\exp(z_i/T)}{\sum_{j=1}^{V} \exp(z_j/T)} \tag{1}$$

where V is the vocabulary size, z_i is the logit for the i-th token, and T is the generation temperature of the MLLM. We then sort samples by their entropy values and divide them into 10 groups with equal scales. For each group, we calculate the average entropy and accuracy. Figure 4 reveals a clear negative correlation: samples with lower entropy values (indicating higher confidence) achieve substantially higher accuracy, while those with higher entropy (lower confidence) show poor performance. This finding suggests that MLLMs can effectively estimate their answer certainty through the entropy of predicted logits, which forms the foundation of our selective thinking mechanism.

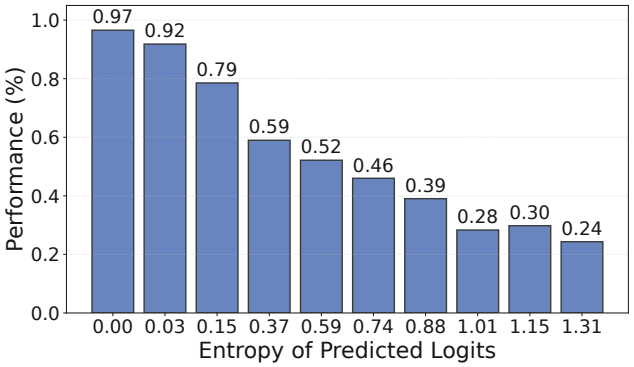

Figure 4: Correlation between performance and entropy of predicted logits on M^{3}CoT using Qwen2-VL-7B.

Based on this insight, we design an entropy-based confidence estimation mechanism. For each query, we prompt the model with the instruction *Answer the question using a single word or phrase.* This specific prompt serves multiple purposes: (1) it focuses the model's attention on answering the question directly, (2) it encourages the model to provide a conclusive response, and (3) it enables us to compute a reliable confidence estimate with just a single forward pass, avoiding the overhead of generating lengthy reasoning steps. We then use the obtained entropy value to determine the ap-

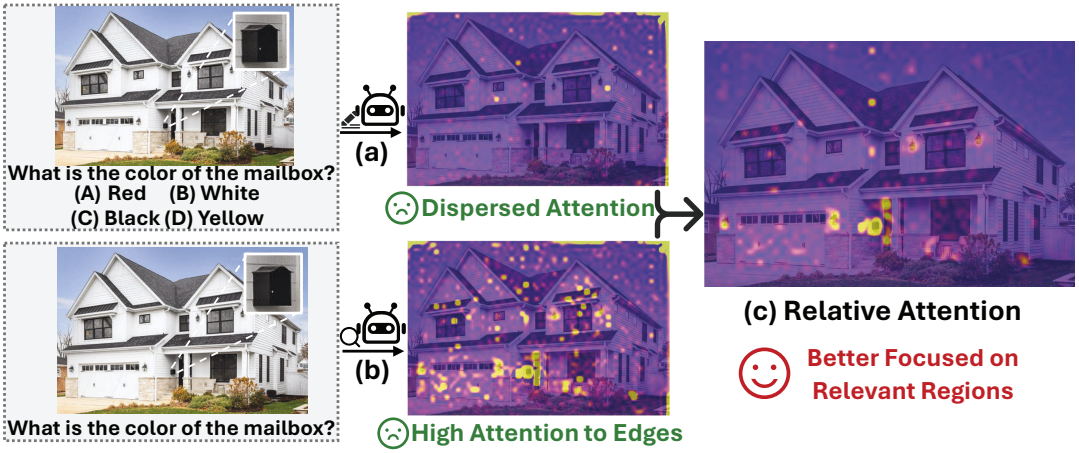

Figure 5: Illustration of the relative attention mechanism. **(a)** When provided with answer options, MLLMs focus on answering, with their visual attention dispersed. **(b)** Without options, models actively attend to question-relevant regions but still allocate attention to edges of the input image. **(c)** By contrasting these two attention patterns, we can obtain the relative attention, which eliminates edge attention while preserving the main focus on important visual regions.

propriate reasoning strategy:

$$\text{Strategy} = \begin{cases} \text{Direct Answer} & \text{if } H < \theta \\ \text{Step-by-Step Reasoning} & \text{if } H \geq \theta \end{cases} \quad (2)$$

where θ is an entropy threshold adaptively set by sampling a small subset and using the average entropy. Queries with low entropy are answered directly, while those with high entropy undergo detailed reasoning with visual enhancement.

This selective approach prevents overthinking on simple queries while ensuring sufficient reasoning depth for complex problems, effectively reducing token consumption without compromising performance.

3.2 Self-Driven Step-wise Visual Enhancement

Beyond assessing answer confidence, we discover that MLLMs also possess capabilities to identify task-relevant visual regions, which remain underutilized during multi-step reasoning. To address this, we develop a step-wise visual enhancement mechanism that strengthens attention to relevant regions throughout the reasoning process.

Relative Attention Mechanism. Figure 5 illustrates our key observation through three distinct attention patterns. When MLLMs are provided with answer options, they primarily focus on generating the answer, resulting in dispersed visual attention across the image. In contrast, when options are removed, models actively attend to question-relevant regions but also allocate substantial attention to image edges and corners that contain limited semantic information. By computing element-wise division of these two patterns, we obtain the relative attention that effectively eliminates edge attention while preserving high attention scores for important visual regions. The relative attention is calculated only once for each sample. For open-ended questions without answer options, we directly use the active attention pattern.

Step-wise Visual Enhancement. Leveraging the relative attention mechanism, we enhance visual grounding at each reasoning step through a multi-stage process that identifies and strengthens attention to relevant regions as follows:

(1) **Region Amplification.** First, we compute a region amplified attention map by combining the current attention with the relative attention:

$$A_{\text{amp}} = A_{\text{curr}} \odot A_{\text{rel}} \quad (3)$$

where A_{curr} represents the attention map at the current reasoning step, A_{rel} is the relative attention map, and $\odot$ denotes element-wise multiplication. Both A_{curr} and A_{rel} are computed by averaging attention weights across all layers. This multiplication highlights visual regions that are important for both the current reasoning context and the original question. (2) **Morphological Refinement.** We apply morphological erosion to the amplified attention map to filter out isolated high-attention points caused by the attention sink phenomenon (Kang et al. 2025), which can mislead our token selection process. This refinement ensures that we focus on coherent visual regions rather than scattered individual tokens. (3) **Adaptive Token Selection.** We adaptively determine the number of visual tokens to enhance. Since the distribution of high-attention tokens after erosion reflects the visual regions relevant to the current reasoning step, we select tokens to be enhanced based on the count of remaining high-attention patches. This adaptive approach can naturally adjust according to the resolution of the input image and the scale of the relevant region in the current step. (4) **Attention Enhancement.** Finally, we enhance the attention values of the selected tokens during the reasoning step. Specifically, we multiply their post-softmax attention values by an enhancement factor α and normalize attention values to maintain their sum to 1:

$$a'_k = \begin{cases} \alpha \cdot a_k \cdot \frac{1}{\sum_{i \notin S} a_i + \sum_{j \in S} \alpha \cdot a_j} & \text{if } k \in S \\ a_k \cdot \frac{1}{\sum_{i \notin S} a_i + \sum_{j \in S} \alpha \cdot a_j} & \text{if } k \notin S \end{cases} \quad (4)$$

where S is the set of selected tokens for enhancement, and the normalization ensures that $\sum_k a'_k = 1$.

Method	Token Consumption	Multi-modal Reasoning			General Visual Understanding			Hallucination	Math	Avg.
		M^3CoT	CoMT	ScienceQA	MMStar	A-OKVQA	V*	HalBench	MathVista	
Base Model Qwen2-VL-7B										
No CoT	35.54	46.81	28.08	76.15	51.22	83.12	68.06	65.34	53.2	59.00
CoT	172.08	45.90	28.37	72.19	53.62	80.02	67.54	63.76	47.3	57.34
DDCoT	760.41	53.36	31.35	76.25	51.52	81.04	62.75	65.54	50.5	59.04
MMCoT	189.15	54.01	31.56	78.83	55.08	85.21	70.16	66.14	53.0	61.75
CCoT	920.35	50.65	27.67	69.41	52.79	70.42	67.02	66.44	54.0	57.30
ICoT	237.54	46.50	29.30	72.46	53.47	79.72	68.59	65.87	49.1	58.13
SDR-MCoT	67.94	**56.17**	**32.39**	**81.01**	**56.39**	**86.67**	**70.68**	**69.12**	**57.1**	**63.69**
Base Model InternVL3-8B										
No CoT	65.40	56.54	45.95	92.58	66.58	87.95	74.73	**80.09**	57.8	70.28
CoT	224.46	47.90	44.39	89.24	65.07	83.67	70.37	78.97	47.6	65.90
DDCoT	1839.78	60.82	42.69	90.50	63.27	85.31	64.25	76.96	66.0	68.73
MMCoT	406.92	**61.52**	45.40	91.42	66.67	88.47	73.37	76.46	63.8	70.89
CCoT	1077.70	52.00	40.82	86.02	65.13	84.72	70.43	74.72	57.2	66.38
ICoT	511.57	48.60	39.55	85.55	44.00	73.97	72.61	75.00	32.2	58.94
SDR-MCoT	100.33	61.25	**46.21**	**95.38**	**67.00**	**89.09**	**75.81**	79.78	**68.4**	**72.87**
Base Model LLaVA-1.5-7B										
No CoT	11.76	39.01	26.24	59.87	33.06	77.09	43.98	45.96	23.1	43.54
CoT	104.77	37.75	23.06	61.62	31.91	73.99	40.57	49.27	24.5	42.83
DDCoT	832.25	34.95	24.79	53.41	30.33	68.99	42.41	49.55	23.5	40.99
MMCoT	169.64	35.29	24.89	55.47	32.61	73.70	37.37	42.51	23.2	40.63
CCoT	1112.41	37.85	25.25	59.65	32.15	76.24	39.47	51.23	23.5	43.17
ICoT	148.82	38.55	23.26	62.38	35.16	76.22	43.98	**52.22**	23.2	44.37
SDR-MCoT	59.40	**40.15**	**27.90**	**62.65**	**35.51**	**78.51**	**44.62**	51.93	**25.2**	**45.81**
Base Model LLaVA-1.5-13B										
No CoT	23.58	36.28	24.83	66.63	33.27	81.64	46.07	53.24	29.2	46.40
CoT	114.24	36.18	24.28	62.47	33.69	78.56	46.25	49.89	23.9	44.15
DDCoT	818.94	37.50	24.33	64.40	34.88	77.23	45.55	45.75	23.8	44.18
MMCoT	268.81	37.41	26.96	62.17	34.68	75.38	49.17	47.65	26.6	45.00
CCoT	1012.06	36.01	27.33	66.98	35.53	80.49	44.50	**57.96**	25.8	46.20
ICoT	197.34	37.80	26.11	62.38	34.96	78.61	43.83	52.34	23.9	44.99
SDR-MCoT	63.76	**39.31**	**27.37**	**67.56**	**36.01**	**84.23**	**49.44**	55.03	**29.6**	**48.57**

Table 1: Performance comparison of different MCoT methods across multiple benchmarks. Token consumption represents the average number of tokens used to obtain the final answer. Tokens are counted using tokenizers of the corresponding base models. Best results are **bolded** and sub-optimal results are underlined. HalBench and Avg. represent HallusionBench and Average Performance, respectively.

This step-wise visual enhancement module operates seamlessly within the reasonging process of MLLMs, strengthening visual grounding at each step without requiring additional token insertions or external tool invocations. By leveraging the model's inherent capabilities, we achieve fine-grained visual enhancement with minimal computational overhead.

4 Experiments

4.1 Baselines

We evaluate our method against several representative training-free MCoT approaches across four base models: LLaVA-1.5-7B/13B (Liu et al. 2024a), InternVL3-8B (Zhu et al. 2025) and Qwen2-VL-7B (Wang et al. 2024). The compared methods are introduced in brief as follows:

No CoT refers to vallina responses where MLLMs directly process the input image and question without additional prompt or explicit guidance on reasoning.

CoT (Kojima et al. 2022) encourages MLLMs to generate reasoning steps in a zero-shot manner by adding a simple instruction "Let's think step by step."

MM-CoT (Zhang et al. 2023) implements a two-stage reasoning framework. MLLMs generate textual rationales in the first stage and then derive the final answer based on the rationales as well as the original image and question.

DDCoT (Zheng et al. 2023) decomposes input questions into simpler sub-questions and utilizes a separate VQA model to generate corresponding sub-answers. Then all sub-questions/answers are provided to MLLMs as preliminary knowledge along with the original image and question.

CCoT (Mitra et al. 2024) first prompts MLLMs to generate JSON-formatted scene graphs that describe objects, attributes, and relationships in the image, then uses these scene

graphs to guide the generation of the final answer.

ICoT (Gao et al. 2025) constructs interleaved-modal CoT in a training-free way by selecting visual tokens with high attention scores and inserting them at each reasoning step.

4.2 Benchmarks

We conduct comprehensive experiments across 8 diverse benchmarks to evaluate different aspects of multimodal reasoning capabilities as follows. **(1) Multimodal Reasoning:** M^3CoT (Chen et al. 2024b), CoMT (Cheng et al. 2025), and ScienceQA (Lu et al. 2022); **(2) General Visual Understanding:** MMStar (Chen et al. 2024a), A-OKVQA (Schwenk et al. 2022), and V* (Wu and Xie 2024); **(3) Hallucination Evaluation:** HallusionBench (Guan et al. 2024); **(4) Mathematical Reasoning:** MathVista (Lu et al. 2024). Detailed descriptions of each benchmark are provided in the Appendix.

4.3 Main Results

Table 1 presents the comparative results across eight benchmarks. We also report the average token consumption, representing the number of tokens used to obtain the answer beyond the input question and image.

Our proposed SDR-MCoT demonstrates superior performance across diverse benchmarks while achieving significant token efficiency, validating its effectiveness in mitigating low-quality multimodal reasoning. In contrast, existing MCoT methods show inconsistent performance gains across different task types and base models. For instance, while MM-CoT achieves notable improvements on multimodal reasoning and general visual understanding tasks on Qwen2-VL-7B, it shows marginal or negative effects on other benchmarks and models. Similarly, other MCoT methods fail to deliver stable improvements. This inconsistency, aligning with recent findings (Gao et al. 2025), stems from two primary factors: First, extended reasoning processes introduce error accumulation and overthinking (Wu et al. 2025), particularly on simple queries. Second, as recent studies (Yang et al. 2025; Chu et al. 2025; Tu et al. 2025) have shown, MLLMs tend to neglect visual information during extended reasoning, causing them to make errors on questions they could answer correctly with direct responses. Our proposed SDR-MCoT mitigates these limitations by reducing overthinking on confident questions and enhancing attention to relevant visual evidence.

From the perspective of token consumption, visual supplementation methods incur significant overhead due to their inefficient provision of visual information. Taking Qwen2-VL-7B for example, DDCoT generates extensive sub-questions and sub-answers, consuming an average of 760.41 tokens. CCoT's JSON-formatted scene graphs result in even higher consumption (920.35 tokens), as they attempt to exhaustively describe visual elements regardless of task relevance. Besides, ICoT also results in higher overhead than zero-shot CoT due to multiple insertions of visual patches. In contrast, our step-wise visual enhancement module strengthens attention to relevant regions without inserting any additional tokens. Together with our self-driven selective thinking module, which effectively identifies queries that require detailed reasoning while avoiding overthinking on simple questions, SDR-MCoT uses only 67.94 tokens on average, reducing token consumption by approximately 60% compared to CoT.

ST	VE	Token Consumption	CoMT	MMStar	HallusionBench
✗	✗	212.91	28.37	53.62	63.76
✓	✗	74.07	29.25	54.47	67.39
✗	✓	216.57	31.24	54.91	67.76
✓	✓	69.84	**32.39**	**56.39**	**69.12**

Table 2: Ablation study on the two key components of SDR-MCoT. **ST** refers to *Selective Thinking*, while **VE** means *step-wise Visual Enhancement*. Token consumption represents the average tokens used to obtain the final answer.

4.4 Ablation Study

To understand the contribution of each component in our proposed SDR-MCoT framework, we conduct ablation experiments on Qwen2-VL-7B and three representative benchmarks by systematically removing the selective thinking module and step-wise visual enhancement module. Table 2 presents the results.

We first analyse the separate effect of selective thinking and step-wise visual enhancement. The selective thinking module significantly reduces the average token consumption from 212.91 to 74.07, achieving approximately a 65% reduction. Meanwhile, it also contributes to the performance due to the mitigation of error accumulation during overthinking. The step-wise visual enhancement module shows minimal impact on token consumption, while providing consistent performance gains across all benchmarks. This validates that our step-wise visual enhancement effectively strengthens attention to task-relevant regions during each reasoning step.

When combined, the two modules demonstrate complementary benefits. The SDR-MCoT framework not only reduces token consumption but also delivers the best performance across all benchmarks, which effectively addresses the low-quality multimodal reasoning problem from both efficiency and effectiveness perspectives.

5 Conclusion

In this paper, we addressed the low-quality multimodal reasoning problem in MLLMs and proposed Self-Driven Refined Multimodal CoT (SDR-MCoT), a training-free framework that leverages models' inherent capabilities for more efficient and effective reasoning. Our method introduces two key innovations: an entropy-based selective thinking module that dynamically determines whether detailed reasoning is needed for each query, and a step-wise visual enhancement module that strengthens attention to task-relevant regions without inserting additional tokens. Extensive experiments across eight benchmarks and four base models demonstrate that SDR-MCoT consistently outperforms existing MCoT methods while significantly reducing token consumption.

Acknowledgments

This work is supported by National Key Research and Development Program of China (No. 2022ZD0160101), Shanghai Natural Science Foundation (No. 23ZR1402900), Shanghai Science and Technology Commission Explorer Program Project (24TS1401300), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103). This work is also supported by the JC STEM Lab of AI for Science and Engineering, funded by The Hong Kong Jockey Club Charities Trust, the MTR Research Funding (MRF) Scheme (CHU-24003), the Research Grants Council of Hong Kong (Project No. CUHK14213224). The computations in this research were performed using the CFFF platform of Fudan University.

References

Alomrani, M. A.; Zhang, Y.; Li, D.; Sun, Q.; Pal, S.; Zhang, Z.; Hu, Y.; Ajwani, R. D.; Valkanas, A.; Karimi, R.; et al. 2025. Reasoning on a Budget: A Survey of Adaptive and Controllable Test-Time Compute in LLMs. *arXiv preprint arXiv:2507.02076*.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Wang, J.; Qiao, Y.; Lin, D.; et al. 2024a. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37: 27056–27087.

Chen, Q.; Qin, L.; Zhang, J.; Chen, Z.; Xu, X.; and Che, W. 2024b. 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)*, 8199–8221.

Chen, X.; Xu, J.; Liang, T.; He, Z.; Pang, J.; Yu, D.; Song, L.; Liu, Q.; Zhou, M.; Zhang, Z.; et al. 2024c. Do not think that much for 2+ 3=? on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*.

Cheng, Z.; Chen, Q.; Zhang, J.; Fei, H.; Feng, X.; Che, W.; Li, M.; and Qin, L. 2025. Comt: A novel benchmark for chain of multi-modal thought on large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 23678–23686.

Chu, X.; Chen, X.; Wang, G.; Tan, Z.; Huang, K.; Lv, W.; Mo, T.; and Li, W. 2025. Qwen Look Again: Guiding Vision-Language Reasoning Models to Re-attention Visual Information. *arXiv preprint arXiv:2505.23558*.

Gao, J.; Li, Y.; Cao, Z.; and Li, W. 2025. Interleaved-modal chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19520–19529.

Guan, T.; Liu, F.; Wu, X.; Xian, R.; Li, Z.; Liu, X.; Wang, X.; Chen, L.; Huang, F.; Yacoob, Y.; et al. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14375–14385.

Han, T.; Wang, Z.; Fang, C.; Zhao, S.; Ma, S.; and Chen, Z. 2024. Token-budget-aware llm reasoning. *arXiv preprint arXiv:2412.18547*.

Hu, Y.; Shi, W.; Fu, X.; Roth, D.; Ostendorf, M.; Zettlemoyer, L.; Smith, N. A.; and Krishna, R. 2024. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37: 139348–139379.

Huang, S.; Wang, H.; Zhong, W.; Su, Z.; Feng, J.; Cao, B.; and Fung, Y. R. 2025. AdaCtrl: Towards Adaptive and Controllable Reasoning via Difficulty-Aware Budgeting. *arXiv preprint arXiv:2505.18822*.

Jiang, C.; Heng, Y.; Ye, W.; Yang, H.; Xu, H.; Yan, M.; Zhang, J.; Huang, F.; and Zhang, S. 2025a. VLM-R^3: Region Recognition, Reasoning, and Refinement for Enhanced Multimodal Chain-of-Thought. *arXiv preprint arXiv:2505.16192*.

Jiang, L.; Wu, X.; Huang, S.; Dong, Q.; Chi, Z.; Dong, L.; Zhang, X.; Lv, T.; Cui, L.; and Wei, F. 2025b. Think only when you need with large hybrid-reasoning models. *arXiv preprint arXiv:2505.14631*.

Kang, S.; Kim, J.; Kim, J.; and Hwang, S. J. 2025. See What You Are Told: Visual Attention Sink in Large Multimodal Models. In *The Thirteenth International Conference on Learning Representations*.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Li, C.; Wu, W.; Zhang, H.; Xia, Y.; Mao, S.; Dong, L.; Vulić, I.; and Wei, F. 2025a. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*.

Li, Z.; Dong, Q.; Ma, J.; Zhang, D.; and Sui, Z. 2025b. Selfbudgeter: Adaptive token allocation for efficient llm reasoning. *arXiv preprint arXiv:2505.11274*.

Liang, G.; Zhong, L.; Yang, Z.; and Quan, X. 2025. Thinkswitcher: When to think hard, when to think fast. *arXiv preprint arXiv:2505.14183*.

Lin, J.; Zeng, X.; Zhu, J.; Wang, S.; Shun, J.; Wu, J.; and Zhou, D. 2025. Plan and Budget: Effective and Efficient Test-Time Scaling on Large Language Model Reasoning. *arXiv preprint arXiv:2505.16122*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, J.; Wang, Y.; Du, J.; Zhou, J.; and Liu, Z. 2024b. MedCoT: Medical Chain of Thought via Hierarchical Expert. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 17371–17389.

Lu, J.; Yu, H.; Xu, S.; Ran, S.; Tang, G.; Wang, S.; Shan, B.; Fu, T.; Feng, H.; Tang, J.; et al. 2025. Prolonged reasoning is not all you need: Certainty-based adaptive routing for efficient llm/mllm reasoning. *arXiv preprint arXiv:2505.15154*.

Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.-W.; Galley, M.; and Gao, J. 2024. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. In *The Twelfth International Conference on Learning Representations*.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521.

Man, Y.; Huang, D.-A.; Liu, G.; Sheng, S.; Liu, S.; Gui, L.-Y.; Kautz, J.; Wang, Y.-X.; and Yu, Z. 2025. Argus: Vision-Centric Reasoning with Grounded Chain-of-Thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14268–14280.

Mitra, C.; Huang, B.; Darrell, T.; and Herzig, R. 2024. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14420–14431.

Mondal, D.; Modi, S.; Panda, S.; Singh, R.; and Rao, G. S. 2024. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 18798–18806.

Muennighoff, N.; Yang, Z.; Shi, W.; Li, X. L.; Fei-Fei, L.; Hajishirzi, H.; Zettlemoyer, L.; Liang, P.; Candes, E.; and Hashimoto, T. 2025. s1: Simple test-time scaling. In *Workshop on Reasoning and Planning for Large Language Models*.

Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; and Mottaghi, R. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, 146–162. Springer.

Shao, H.; Qian, S.; Xiao, H.; Song, G.; Zong, Z.; Wang, L.; Liu, Y.; and Li, H. 2024. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37: 8612–8642.

Shen, Y.; Zhang, J.; Huang, J.; Shi, S.; Zhang, W.; Yan, J.; Wang, N.; Wang, K.; Liu, Z.; and Lian, S. 2025. Dast: Difficulty-adaptive slow-thinking for large reasoning models. *arXiv preprint arXiv:2503.04472*.

Sun, Q.; Hong, P.; Pala, T. D.; Toh, V.; Tan, U.; Ghosal, D.; Poria, S.; et al. 2024. Emma-x: An embodied multimodal action model with grounded chain of thought and look-ahead spatial reasoning. *arXiv preprint arXiv:2412.11974*.

Tu, C.; Ye, P.; Zhou, D.; Bai, L.; Yu, G.; Chen, T.; and Ouyang, W. 2025. Attention reallocation: Towards zero-cost and controllable hallucination mitigation of mllms. *arXiv preprint arXiv:2503.08342*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wu, P.; and Xie, S. 2024. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13084–13094.

Wu, Y.; Wang, Y.; Ye, Z.; Du, T.; Jegelka, S.; and Wang, Y. 2025. When more is less: Understanding chain-of-thought length in llms. *arXiv preprint arXiv:2502.07266*.

Xiang, V.; Blagden, C.; Rafailov, R.; Lile, N.; Truong, S.; Finn, C.; and Haber, N. 2025. Just Enough Thinking: Efficient Reasoning with Adaptive Length Penalties Reinforcement Learning. *arXiv preprint arXiv:2506.05256*.

Xiao, W.; Gan, L.; Dai, W.; He, W.; Huang, Z.; Li, H.; Shu, F.; Yu, Z.; Zhang, P.; Jiang, H.; et al. 2025. Fast-slow thinking for large vision-language model reasoning. *arXiv preprint arXiv:2504.18458*.

Xiao, Z.; Zhang, D.; Han, X.; Fu, X.; Yu, W. Y.; Zhong, T.; Wu, S.; Wang, Y.; Yin, J.; and Chen, G. 2024. Enhancing llm reasoning via vision-augmented prompting. *Advances in Neural Information Processing Systems*, 37: 28772–28797.

Xu, G.; Liu, P.; Zhu, K.; Zhang, W.; Xu, C.; and Huang, Z. 2024. LLaVA-CoT: Let Vision Language Models Reason Step-by-Step. *arXiv preprint arXiv:2411.10440*.

Yang, S.; Niu, Y.; Liu, Y.; Ye, Y.; Lin, B.; and Yuan, L. 2025. Look-Back: Implicit Visual Re-focusing in MLLM Reasoning. *arXiv preprint arXiv:2507.03019*.

Yu, F.; Wan, H.; Cheng, Q.; Zhang, Y.; Chen, J.; Han, F.; Wu, Y.; Yao, J.; Hu, R.; Ding, N.; et al. 2025. HiPhO: How Far Are (M) LLMs from Humans in the Latest High School Physics Olympiad Benchmark? *arXiv preprint arXiv:2509.07894*.

Zhang, D.; Lei, J.; Li, J.; Wang, X.; Liu, Y.; Yang, Z.; Li, J.; Wang, W.; Yang, S.; Wu, J.; et al. 2025a. Critic-v: Vlm critics help catch vlm errors in multimodal reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9050–9061.

Zhang, J.; Lin, N.; Hou, L.; Feng, L.; and Li, J. 2025b. Adaptthink: Reasoning models can learn when to think. *arXiv preprint arXiv:2505.13417*.

Zhang, R.; Xiao, C.; and Cao, Y. 2025. Long or short CoT? Investigating Instance-level Switch of Large Reasoning Models. *arXiv preprint arXiv:2506.04182*.

Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; and Smola, A. 2023. Multimodal Chain-of-Thought Reasoning in Language Models. In *TMLR*.

Zheng, G.; Yang, B.; Tang, J.; Zhou, H.-Y.; and Yang, S. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36: 5168–5191.

Zheng, S.; Cheng, Q.; Yao, J.; Wu, M.; He, H.; Ding, N.; Cheng, Y.; Hu, S.; Bai, L.; Zhou, D.; et al. 2025a. Scaling physical reasoning with the physics dataset. *arXiv preprint arXiv:2506.00022*.

Zheng, Z.; Yang, M.; Hong, J.; Zhao, C.; Xu, G.; Yang, L.; Shen, C.; and Yu, X. 2025b. DeepEyes: Incentivizing" Thinking with Images" via Reinforcement Learning. *arXiv preprint arXiv:2505.14362*.

Zhou, Q.; et al. 2024. Image-of-Thought Prompting for Visual Reasoning Refinement in Multimodal Large Language Models. *arXiv preprint arXiv:2405.13872*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.