CoFFT: Chain of Foresight-Focus Thought for Visual Language Models

Xinyu Zhang[1,2] **Yuxuan Dong**[1,2] **Lingling Zhang**[1,2] * **Chengyou Jia** [1,2]
ZhuoHang Dang [1,2] **Basura Fernando** [4,6] **Jun Liu**[1,3] **Mike Zheng Shou** [5] *
[1]School of Computer Science and Technology, Xi'an Jiaotong University
[2]Ministry of Education Key Laboratory of Intelligent Networks and Network Security, China
[3]Shaanxi Province Key Laboratory of Big Data Knowledge Engineering, China
[4]IHPC, Agency for Science, Technology and Research, Singapore
[5]Show Lab, National University of Singapore
[6]College of Computing and Data Science, Nanyang Technological University, Singapore
zhang13938697l6@stu.xjtu.edu.cn, zhanglling@xjtu.edu.cn, mikeshou@nus.edu.sg

Abstract

Despite significant advances in Vision Language Models (VLMs), they remain constrained by the complexity and redundancy of visual input. When images contain large amounts of irrelevant information, VLMs are susceptible to interference, thus generating excessive task-irrelevant reasoning processes or even hallucinations. This limitation stems from their inability to discover and process the required regions during reasoning precisely. To address this limitation, we present the Chain of Foresight-Focus Thought (CoFFT), a novel training-free approach that enhances VLMs' visual reasoning by emulating human visual cognition. Each Foresight-Focus Thought consists of three stages: (1) Diverse Sample Generation: generates diverse reasoning samples to explore potential reasoning paths, where each sample contains several reasoning steps; (2) Dual Foresight Decoding: rigorously evaluates these samples based on both visual focus and reasoning progression, adding the first step of optimal sample to the reasoning process; (3) Visual Focus Adjustment: precisely adjust visual focus toward regions most beneficial for future reasoning, before returning to stage (1) to generate subsequent reasoning samples until reaching the final answer. These stages function iteratively, creating an interdependent cycle where reasoning guides visual focus and visual focus informs subsequent reasoning. Empirical results across multiple benchmarks using Qwen2.5-VL, InternVL-2.5, and Llava-Next demonstrate consistent performance improvements of 3.1-5.8% with controllable increasing computational overhead.

1 Introduction

Vision Language Models (VLMs) have demonstrated remarkable progress across numerous domains [1, 2, 3], particularly in visual reasoning [4, 5, 6, 7]. However, their performance remains significantly constrained by the inherent complexity and redundancy of visual inputs [8, 9, 10]. Visual Language Models demonstrate high sensitivity to large, salient elements in images, often struggling to effectively mitigate the impact of visually dominant but semantically irrelevant information, which leads to deteriorated performance and flawed reasoning [11]. These limitations are especially pronounced in complex reasoning tasks that require fine-grained image understanding such as mathematics problem-solving [12, 13], chart understanding [14, 15], and geolocation inference [16].

*Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

 https://doi.org/10.52202/085713-3544

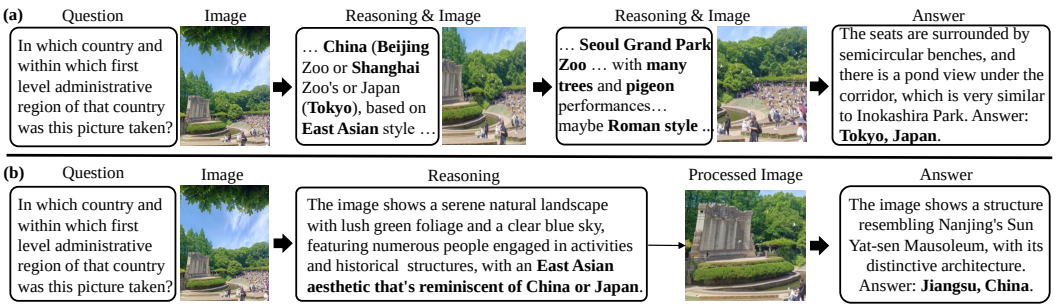

Figure 1: An example from the SeekWorld [16]. (a) is the reasoning process of o3, and (b) is the reasoning process of o3 after human visual cognition. The correct answer is Jiangsu, China.

To address this limitation, researchers have proposed Multi-modal Chain of Thought approaches [17, 18, 19], which explore the integration of visual operations during reasoning and leverage image information across multiple stages for understanding the image comprehensively. Taking the recently prominent OpenAI-O3 [20] as an example, we illustrate its reasoning process as shown in Figure 1 (a). Despite multiple attempts, O3 ultimately arrives at an incorrect result. Based on the redundant and interference-laden reasoning process, we attribute this failure to its tendency to continuously explore different image regions without evaluating their contribution to the question, thus introducing substantial interference from irrelevant visual information.

However, when humans view this image, they evaluate different regions based on their potential contribution to question-solving. This leads them to reduce attention to the commonplace elements such as crowds, pigeons, and trees while prioritizing attention toward distinctive historical buildings. We send the region of historical buildings to guide O3, as shown in Figure 1 (b), which illustrates the necessity of learning from human visual cognition. This performance stems from two human visual cognition capabilities when analyzing complex visual scenes: **(1) Foresight** - the foresight evaluation of which visual regions will be most valuable for future reasoning [21], and **(2) dynamic visual focus** - precisely shift attention toward the most future reasoning-relevant regions [22, 23].

Inspired by these observations, we propose the Chain of Foresight-Focus Thought (CoFFT), a training-free approach that enhances VLMs' visual reasoning capabilities, as illustrated in Figure 2. Each **Foresight-Focus Thought** iteration consists of three stages: **(1) Diverse Samples Generation (DSG)**: Based on the current reasoning process, the VLM generates diverse reasoning samples under different temperature parameters, where each sample contains multiple subsequent reasoning steps. **(2) Dual Foresight Decoding (DFD)**: Evaluates samples by considering both visual focus and reasoning progression to select the optimal reasoning sample, incorporating its first step into the reasoning process. **(3) Visual Focus Adjustment (VFA)**: First, the image is evaluated based on two criteria: relevance to the question and correlation with future reasoning steps in the optimal sample. Then, a sliding window is used to select, crop, and magnify the best region as the next visual focus image. Finally, this image obtained from VFA serves as visual input for the next iteration, cycling back to stage (1) to generate new reasoning samples with the updated reasoning from stage (2) until reaching the final answer. This creates a cycle where reasoning guides visual focus, and visual focus informs subsequent reasoning steps, enhancing VLMs' visual reasoning capabilities.

We conduct extensive experiments across several benchmarks using different VLMs, including Qwen2.5-VL [1], InternVL-2.5 [2], and Llava-Next [24]. CoFFT demonstrates consistent performance gains of 3.1-5.8% on average across these benchmarks. Furthermore, analysis across models with varying parameter counts reveals that CoFFT's effectiveness scales positively with model size, yielding greater improvements for larger VLMs. Our computational overhead analysis demonstrates that while CoFFT requires more computation than direct VLM inference, it remains more efficient than Monte Carlo Tree Search approaches, highlighting the practical efficiency of CoFFT.

1. We propose CoFFT, a novel training-free approach that improves VLM's performance on complex visual reasoning tasks without requiring model modifications or retraining.

2. We propose Dual-Foresight Decoding, which evaluates reasoning samples by optimizing visual focus and reasoning progression jointly, along with Visual Focus Adjustment that directs visual focus to regions relevant to the question and essential for subsequent reasoning.

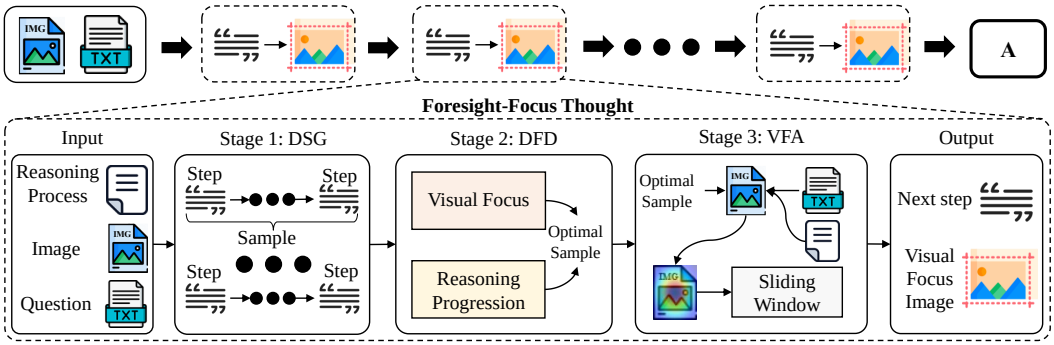

Figure 2: The overall approach of CoFFT, where Dual Foresight Decoding and Visual Focus Adjustment will be introduced in detail later.

3. We demonstrate through extensive experiments across multiple benchmarks that CoFFT significantly improves performance on tasks requiring fine-grained visual understanding and complex reasoning, without excessively increasing computational overhead.

2 Related work

2.1 Visual Language Model

Large Language Models' (LLMs) success has fundamentally transformed Visual-Language Models (VLMs) development, advancing from basic dispatcher architectures (Visual ChatGPT [25], HuggingGPT [26], MM-REACT [27]) to more sophisticated vision-language approaches. Key architectural innovations include LLaVA's [28] learned image-token projectors, BLIP-2's [29, 30] question transformers, and MoVA's [31] innovative task-specific adaptive routing. Recent advances feature Qwen2.5-VL's [1] novel window attention optimizations and InternVL2.5's [2] enhanced data processing combining Random JPEG Compression with Square Averaging techniques.

2.2 Visual Search

Early research emulated human visual search through Bayesian models with saliency maps[32, 33], deep similarity mapping networks[34], and inverse reinforcement learning[35]. These approaches primarily replicated gaze samples but lacked precise target localization capabilities and employed fixed-size attention windows unsuitable for complex scenarios. Recent methods such as SEAL[36] incorporate localization modules and visual memory systems to connect visual search with large multimodal models, though requiring additional training. DyFo[37] achieves training-free dynamic focus through visual expertise integration, based on language segment-anything modal. Both remain limited by their one-time image processing, lacking iterative analytical capabilities during reasoning. Our approach, by contrast, enables dynamic image manipulation throughout the reasoning process while maintaining a training-free methodology, offering a more robust visual understanding approach.

2.3 Multi-modal Chain-of-Thought

Research on multi-modal chain-of-thought reasoning integration follows two main approaches: (1) specialized visual expert models [17, 18], such as Sketchpad and VoT/MVoT, which equip VLMs with drawing tools and enhanced spatial reasoning, and (2) improved training processes [38, 19] featuring autonomous region selection and built-in visual operations for preserving reasoning evidence. Both approaches face limitations - expert models lack generalizability and require adaptation costs, while enhanced training demands extensive resources. Additionally, Multimodal Chain-of-Thought Prompting is an alternative strategy to enhance VLMs' visual reasoning capabilities through prompt engineering [39, 40, 41]. However, this method generates only text-based rationales for answers, unlike the aforementioned approaches which can produce interleaved visual-textual rationales.

https://doi.org/10.52202/085713-3544

3 Method

3.1 Overview

We introduce CoFFT (Chain of Foresight-Focus Thought), a training-free approach designed to address VLM's limitations in complex visual reasoning, as shown in Figure 2. CoFFT is an iterative execution of Foresight-Focus Thought to obtain the final result. To illustrate the workflow in Foresight-Focus Thought, we take the current $t+1$ iteration of Foresight-Focus Thought as an example, given original image V, question Q, current visual focus image V_t, and existing reasoning process $R_t = \{r_1, \ldots, r_t\}$, to introduce the following three stages, as shown in Algorithm 1.

1. **Diverse Sample Generation**: The VLM generates k candidate reasoning samples $S_{t+1} = \{s_1, \ldots, s_k\}$ based on the current reasoning process R_t, visual focus image V_t, and original question Q. Different temperature parameters are used to get the samples, ensuring sample diversity, where each sample s retains a maximum of l (foresight length) steps.
2. **Dual Foresight Decoding**: The samples are evaluated using a combination of visual focus score E_{att} and reasoning progression score E_{prob} to ensure a comprehensive assessment. E_{att} evaluates the relevance between the reasoning process and image to suppress image-irrelevant hallucination, while E_{prob} assesses reasoning progression by measuring probability improvements across reasoning steps. The first step of the optimal sample s_i is integrated into the evolving reasoning process (R_{t+1}), for continuous iterations.
3. **Visual Focus Adjustment**: First, a scoring mechanism is used to evaluate images based on two criteria: relevance to the question Q and correlation with future reasoning steps in optimal samples s_i. Then, a sliding window selects, crops, and magnifies the highest-scoring region as a visual focus image. Through this stage, the approach switches between global views and local details, avoiding the oversight of critical local information.

These three stages constitute a complete **Foresight-Focus Thought** iteration as illustrated in Figure 3. They create a synergistic cycle where the reasoning path directs visual focus, and optimized visual focus subsequently improves reasoning quality. This cognitive-inspired process continues iteratively until the final answer is derived. By simulating human visual cognition processes, particularly Dual Foresight Decoding and Visual Focus Adjustment stages, CoFFT reduces VLMs' hallucination tendencies while improving reasoning accuracy. Next, we will first introduce the basics of CoFFT, the relative attention mechanism, and then explain the DFD and VFA stages in detail.

3.2 Relative Attention Mechanism

Due to the influence of redundancy information in images, relying solely on the original text-image paired attention maps tends to contain noise and makes it challenging to directly distinguish semantically relevant regions [42, 11]. To address this limitation, we introduce a relative attention mechanism that normalizes any text input attention against a baseline descriptive attention distribution.

Formally, taking question Q and image V as an example, we define relative attention $A^{rel}(V, Q)$ as:

$$A^{rel}(V, Q) = Softmax(\frac{A(V, Q)}{A(V, D) + \epsilon}), \quad A \in \mathbb{R}^{H \times W} \tag{1}$$

where $A(V, Q)$ and $A(V, D)$ represent the attention maps for the question and descriptive prompts (e.g., "Describe the image in detail") respectively, the $Softmax$ function is used to normalize, with ϵ (10^{-10}) ensuring numerical stability. The division operation is performed element-wise. It is worth noting that the attention map A between the input text and the image is provided by VLM itself and does not require additional modification during reasoning process.

This relative attention mechanism is effective for two reasons: (1) it suppresses attention to generally salient but input text-irrelevant regions by normalizing against the descriptive attention, and (2) it amplifies attention to regions that are specifically relevant to the input text, thereby encouraging the model to focus on regions that are relevant to the input text rather than salient elements in the image.

3.3 Dual-Foresight Decoding

Our Dual-Foresight Decoding mechanism evaluates reasoning samples by jointly combining visual focus and reasoning progression. Given a set of candidate samples S_{t+1} generated at reasoning step

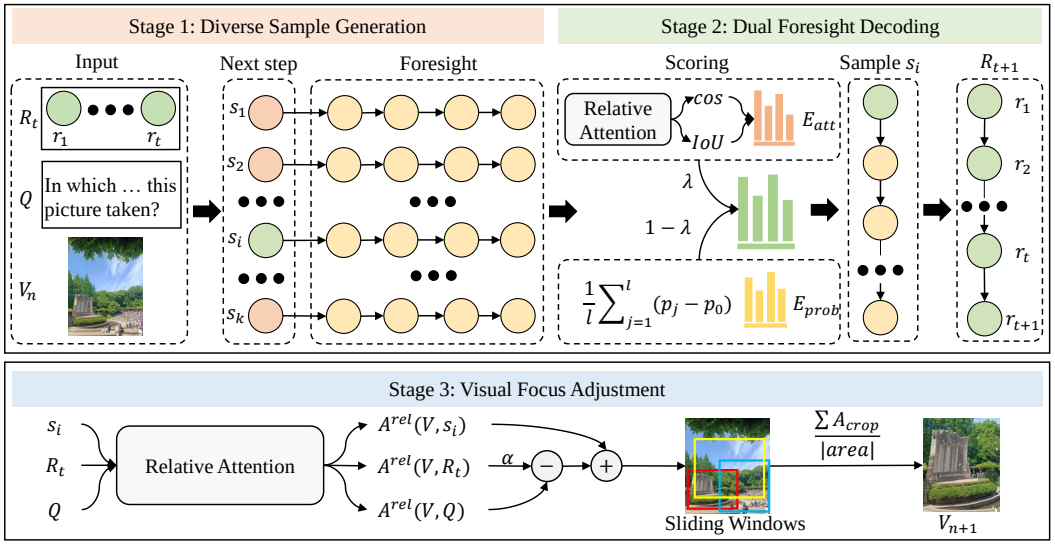

Figure 3: The two primary components of CoFFT: (a) Dual Foresight Decoding, which evaluates different reasoning samples and selects the best sample from both visual focus and reasoning progression to enhance decision robustness, and (b) Visual Focus Adjustment, which adaptively modulates visual focus adjustment to reasoning-relevant regions for optimized information understanding.

$t+1$, we evaluate each sample $s \in S_{t+1}$ (containing up to l steps) through a combined score:

$$E_{t+1} = \lambda \cdot Softmax([E_{att}(s)]\ \forall s \in S_{t+1}) + (1-\lambda) \cdot Softmax([E_{prob}(s)]\ \forall s \in S_{t+1}), \quad (2)$$

where $\lambda \in (0, 1]$ (empirically set to 0.3) balances visual focus and reasoning progression. The $Softmax$ function normalizes scores across all samples to obtain comparable distributions.

For the visual focus score, we design E_{att} that measures the focus between each sample and the image, which comprehensively integrates cosine similarity and IoU metrics for robust evaluation::

$$E_{att}(s) = 0.5 \cdot \cos(A^{rel}(V, Q), A^{rel}(V, s)) + 0.5 \cdot IoU^{30\%}(A^{rel}(V, Q), A^{rel}(V, s)), \quad (3)$$

where $A^{rel}(V, s)$ computes the relative attention following Equation 1, and $IoU^{30\%}$ calculates the intersection over union of high-attention regions (top 30%). This dual-metric formulation captures both global attention distribution and local focus similarity, effectively reducing hallucination by enforcing visual coherence through comprehensive spatial analysis.

Inspired by [43, 44], we observe that scores based on dynamic improvement tend to be more reliable than those simply using confidence values. Therefore, we propose the reasoning progression score E_{prob} that captures the stepwise improvement in reasoning quality using new reasoning steps of different lengths to compare with the current reasoning process:

$$E_{prob}(s) = \frac{1}{l}\sum_{j=1}^{l}(p_j - p_0), \quad (4)$$

where p_j represents the mean log probability of tokens including R_t and first j-th step of sample s, and l denotes the number of evaluated steps. This formulation identifies samples demonstrating maximum average improvement in reasoning confidence, leading to more reliable sample selection.

3.4 Visual Focus Adjustment

This stage dynamically adjusts visual focus by evaluating and selecting informative image regions based on a dual-criteria scoring mechanism: question relevance and future reasoning relevance. This enables effective switching between global views and local details throughout the reasoning process.

For question relevance, we encourage VLM to focus on regions that have not been explored by the current reasoning process R_t. Therefore, we define $C^{rel}(V, Q, R_t)$ as following:

$$C^{rel}(V, Q, R_t) = max(A^{rel}(V, Q) - \alpha \cdot A^{rel}(V, R_t), 0). \quad (5)$$

https://doi.org/10.52202/085713-3544

where $A^{rel}(V, R_t)$ represents the relative attention from n completed reasoning steps (following Equation 1), and $\alpha \in (0, 1]$ (empirically set to 0.3) controls the influence of processed regions.

For future reasoning relevance, we leverage the optimal sample s_i selected by the Dual Foresight Decoding stage to obtain $A^{rel}(V, s_i)$ in the same way as Equation 1.

Based on those, the final attention map evaluation combines both criteria as following:

$$A_{crop} = 0.5 \cdot C^{rel}(V, Q, R_t) + 0.5 \cdot A^{rel}(V, s_i). \tag{6}$$

Then, a dynamic sliding window algorithm determines the optimal visual region V_{t+1}:

$$V_{t+1} = \begin{cases} \arg\max_{B \in \Omega} \frac{1}{|B|} \sum_{(x,y) \in B} A_{crop}(x, y), & \text{if } \mu_{B^*} > \mu_{V_0} + \beta \\ V_0, & \text{otherwise} \end{cases} \tag{7}$$

where $\beta = \sigma_{V_0} \cdot (1 - \cos(C^{rel}(V, Q, R_t), A^{rel}(V, s_i)))$, and B^* is the optimal region. Here, σ_{V_0} denotes the standard deviation of A_{crop}, Ω contains regions spanning 40%-90% (interval is 10%) of original dimensions, and $\mu_{B^*} = \frac{1}{|B^*|} \sum_{(x,y) \in B^*} A_{crop}(x, y)$ and $\mu_{V_0} = \frac{1}{|V_0|} \sum_{(x,y) \in V_0} A_{crop}(x, y)$ represent the maximum region and global mean scores. The selected region is scaled maintaining an aspect ratio within the original dimensions for subsequent reasoning iterations.

The mechanism employs adaptive thresholding based on the two attention maps: requiring minimal improvement over the global mean when attention centers converge ($\beta \approx 0$) and stricter thresholds when they diverge ($\beta \approx \sigma_{V_0}$). This strategy prevents fixation on misleading local information while ensuring comprehensive visual understanding through balanced global-local focus transitions.

4 Experiment

4.1 Settings

Benchmarks We conduct comprehensive evaluations across multiple complementary benchmarks to assess various aspects of visual reasoning capabilities. For mathematical and geometric reasoning, we employ MathVista [45] and MathVision [46]. To evaluate cross-domain visual reasoning abilities, we utilize the multi-subject benchmarks M3CoT [47] and MMStar [48]. For chart comprehension assessment, we leverage Charxiv [14]. Additionally, we contribute to the geographical domain by introducing two novel datasets in SeekWorld [16]: SeekWorld-Global, which utilizes Google Maps panoramic imagery, and SeekWorld-China, which incorporates data from the Xiaohongshu App.

Comparison methods and experimental setup Our experimental approach incorporates state-of-the-art Vision Language Models (VLMs): Qwen2.5-VL-Instruct (7B, 32B) [1], InternVL2.5-Instruct (8B) [2], and Llava-Next (7B) [24], selected for their architectural capabilities and superior performance in visual reasoning. We establish comprehensive baseline comparisons using search-based method (MCTS [49]), foresight reasoning method (Predictive Decoding [43]), visual search methodology (DyFo [37]), and multi-modal chain-of-thought prompting method (ICoT [41]). This diverse selection enables systematic evaluation against both text-based reasoning, visual search and multi-modal chain-of-thought methods. All experiments are run on four NVIDIA A100 GPUs with parallel processing. To ensure sample diversity, the temperature parameter ranges from 0.4 to 1 with an interval of 0.1, and a sample is randomly selected each time it is generated. To prevent repeated selections, the probability weight of each chosen parameter is reduced by half in subsequent sampling processes. The weights are reset to their initial values once all parameters have been selected, ensuring a balanced exploration of different temperature values. For Predictive Decoding and CoFFT, inference is considered complete when the model outputs 'REASONING_COMPLETE'.

Performance Metrics We adopt Pass@1 accuracy (Acc.) as our primary performance metric across all benchmarks. To quantify the computational efficiency-performance trade-off, we calculate floating point operations (FLOPS) following the methodology in [50], where FLOPS $\approx 6nP$ (P represents model parameters, n denotes generated tokens). By computing the average number of tokens generated per example, we provide a standardized measure of computational cost across different methods based on Qwen2.5-VL-7B-Instruct to enable direct efficiency comparisons.

Table 1: Main results, where Claude-3.5, Gemini-2, Pred. Dec. are Claude-3.5-sonnet, Gemini-2.0-Flash and Predictive decoding. The optimal results are highlighted in bold, whereas suboptimal results are underlined. The Avg. column indicates the averaged results across the six benchmarks.

Models	Math		Multi-subjects		Chart	Geography		Avg.
	MathVista	MathVision	MMStar	M3CoT	Charxiv	S.W-China	S.W-Global	
Closed source VLMs								
GPT-4o	63.8	18.75	64.7	65.75	50.5	31.90	56.50	50.27
Claude-3.5	65.4	26.21	65.1	66.05	60.2	29.22	52.50	52.10
Gemini-2	73.1	31.83	69.4	67.73	53.2	30.83	55.31	54.49
Qwen2.5VL-7B-Instruct								
Baseline	68.2	18.09	63.9	59.62	42.5	21.45	25.31	42.72
MCTS	69.6	18.75	66.2	60.87	44.5	26.27	26.56	44.68
Pred. Dec.	69.9	19.73	65.7	61.34	45.3	26.54	26.86	45.05
ICoT	47.5	10.53	42.1	61.17	27.5	29.22	26.37	34.91
DyFo	68.4	16.78	64.5	61.26	43.7	32.44	27.81	44.98
CoFFT	**70.4**	**23.36**	**69.4**	**62.47**	**47.2**	**35.12**	**29.37**	**48.19**
LLaVA-NeXT-7B								
Baseline	34.6	9.87	34.2	38.27	13.9	10.72	15.31	22.41
MCTS	35.1	10.53	36.7	39.52	14.6	12.33	16.56	23.62
Pred. Dec.	34.8	11.36	35.1	40.16	15.3	11.80	17.19	23.67
ICoT	27.3	11.18	31.2	39.34	9.5	12.87	16.56	21.14
DyFo	34.8	8.22	36.1	39.86	15.7	13.67	17.50	23.69
CoFFT	**35.6**	**12.17**	**38.3**	**40.68**	**16.8**	**15.55**	**19.69**	**25.54**
InternVL2.5-8B-Instruct								
Baseline	64.4	22.00	60.5	57.16	32.9	23.32	25.63	40.84
MCTS	65.0	24.67	62.0	58.24	34.3	25.20	26.56	42.28
Pred. Dec.	65.4	25.00	62.4	58.76	35.1	26.81	27.19	42.95
ICoT	42.9	16.12	39.4	58.50	16.4	27.35	26.88	32.51
DyFo	64.7	20.39	61.5	58.58	34.2	29.22	28.13	42.39
CoFFT	**66.5**	**28.29**	**64.5**	**59.19**	**36.6**	**31.37**	**30.63**	**45.30**
Qwen2.5VL-32B-Instruct								
Baseline	74.7	25.33	69.5	62.81	44.5	24.13	28.41	47.05
MCTS	76.2	27.31	70.6	64.15	47.6	28.69	29.06	49.09
Pred. Dec.	76.6	27.96	71.1	64.62	48.2	29.22	30.31	49.72
ICoT	58.7	21.05	54.6	63.93	47.3	32.17	31.56	44.19
DyFo	75.6	24.67	70.1	64.32	47.7	35.38	32.19	49.99
CoFFT	**77.5**	**29.93**	**72.7**	**66.08**	**50.9**	**38.61**	**34.38**	**52.96**

Table 2: FLOPS denotes the calculated computational cost, with lower values indicating lower costs.

Models	Baseline	MCTS	Predictive decoding	ICoT	DyFo	CoFFT
FLOPS	8.35e+12	4.05e+14	1.85e+14	1.88e+13	1.98e+13	2.38e+14

4.2 Results

As shown in Tables 1 and 2, we compare our method against current state-of-the-art approaches across seven datasets and several VLMs. Our analysis reveals several significant findings:

Comprehensive Performance Improvements CoFFT consistently outperforms all approaches across all VLMs and seven benchmarks, as demonstrated in Table 1. Performance gains are substantial across diverse model scales, from 7B to 32B parameters, highlighting CoFFT's versatility and effectiveness in enhancing both visual perception and reasoning capabilities of foundation models.

Balancing Visual Understanding and Reasoning VLMs must balance image comprehension with reasoning capabilities. Language-based methods like MCTS and Predictive Decoding show broad improvements but struggle with fine-grained visual analysis tasks in the SeekWorld datasets, indicating single-pass image understanding is often insufficient. Similarly, DyFo and ICoT fail to significantly improve performance on reasoning-intensive tasks, primarily due to the fragility of the reasoning process. Our proposed CoFFT addresses these limitations by simultaneously enhancing both visual comprehension and reasoning capabilities, achieving superior performance.

https://doi.org/10.52202/085713-3544

Table 3: Ablation Studies on Qwen2.5VL-7B-Instruct. *w/o DFD* refers to sample evaluation using reasoning progression score only. *w/o VFA* indicates the absence of adaptive image cropping, relying instead on original images throughout the reasoning process.

Models	Math		Multi-subjects		Chart	Geography	
	MathVista	MathVision	MMStar	M3CoT	Charxiv	S.W-China	S.W-Global
Our	**70.4**	**23.36**	**69.4**	**62.47**	**47.2**	**35.12**	**29.37**
w/o DFD	68.5	20.42	66.5	61.39	44.8	28.42	27.19
w/o VFA	69.3	21.71	67.4	61.09	44.7	27.08	26.25

Table 4: Performance comparison of method combinations on Qwen2.5-VL-7B. We evaluate various combinations of existing approaches (DyFo, Predictive Decoding) and our proposed components (VFA, DFD) to demonstrate the effectiveness of our integrated CoFFT framework.

Method/Combination	MathVista	SeekWorld-China	Average
Qwen2.5-VL-7B (Baseline)	68.2	21.45	44.83
DyFo + Predictive Decoding	69.0	33.24	51.02
VFA + Predictive Decoding	68.5	31.37	50.24
DyFo + Dual Foresight Decoding	69.3	34.05	51.73
CoFFT (VFA + DFD)	**70.4**	**35.12**	**52.76**

Fine-grained Detail Extraction CoFFT excels particularly in extracting easily overlooked fine-grained details from images, evidenced by substantial improvements on Charxiv, SeekWorld-China, and SeekWorld-Global benchmarks. Nevertheless, on SeekWorld-Global, our method still underperforms compared to advanced closed-source models like GPT-4o and Gemini-2.0-Flash. We attribute this gap to differences in knowledge capabilities across regional contexts, where these closed-source models likely benefit from more extensive pre-training on globally diverse datasets.

Scaling Properties Notably, the benefits of CoFFT increase with model size in a systematic manner. The absolute performance improvements are more pronounced with larger models, suggesting that CoFFT effectively leverages the enhanced capabilities of more powerful base models. For instance, comparing improvements on Qwen2.5VL-7B-Instruct versus Qwen2.5VL-32B-Instruct reveals larger absolute gains across most benchmarks for the 32B model. This scaling trend suggests CoFFT could yield even greater benefits when applied to larger foundation models.

Computational Efficiency Analysis As shown in Table 2, while CoFFT introduces additional computational costs compared to the baseline, it remains more computationally efficient than MCTS while delivering superior performance across all benchmarks. The computational overhead primarily stems from the iterative reasoning process. Given the significant performance improvements observed, particularly on challenging reasoning tasks, this computational trade-off is well justified.

4.3 Ablation Studies

To validate the effectiveness of our proposed components, we conduct ablation experiments across all benchmarks using Qwen2.5-VL-7B-Instruct as the backbone model. We specifically examine two critical components: *w/o DFD* removes the visual focus score E_{att} and relies solely on the reasoning progression score E_{prob}, while *w/o VFA* eliminates the dynamic image update module and uses only the original image for inference during the reasoning process. Results in Table 3 demonstrate performance degradation across all benchmarks when either component is removed, confirming the integral contribution of each mechanism to the overall effectiveness of CoFFT.

Additionally, we explore combinations of existing methods such as DyFo for visual search and Predictive Decoding for reasoning enhancement, with results shown in Table 4. Our evaluation demonstrates that CoFFT achieves superior performance with an overall accuracy of 48.19% compared to the baseline accuracy of 42.72%. Experiments reveal that simply combining existing visual search and reasoning methods often yields suboptimal results due to fundamental incompatibilities between their design principles. DyFo's static region determination conflicts with Predictive Decoding's reasoning requirements, while VFA's dynamic adjustments cannot be effectively utilized by static reasoning

Table 5: Analysis of the effects of Foresight parameter l and Sampling number K on experimental outcomes. During k parameter studies, l is maintained constant at 5, while l parameter studies are conducted with k fixed at 4.

Foresight	MathVista	S.E-China	FLOPS	Sampling	MathVista	S.E-China	FLOPS
$l = 3$	68.8	33.51	1.41e+14	$k = 2$	68.8	34.32	1.19e+14
$l = 4$	69.3	34.58	1.91e+14	$k = 4$	70.4	35.12	2.38e+14
$l = 5$	70.4	35.12	2.38e+14	$k = 6$	70.8	35.65	3.56e+14
$l = 6$	70.5	35.65	2.86e+14	$k = 8$	71.4	36.19	4.75e+14
$l = 7$	70.7	35.92	3.33e+14	$k = 10$	72.2	37.27	5.93e+14

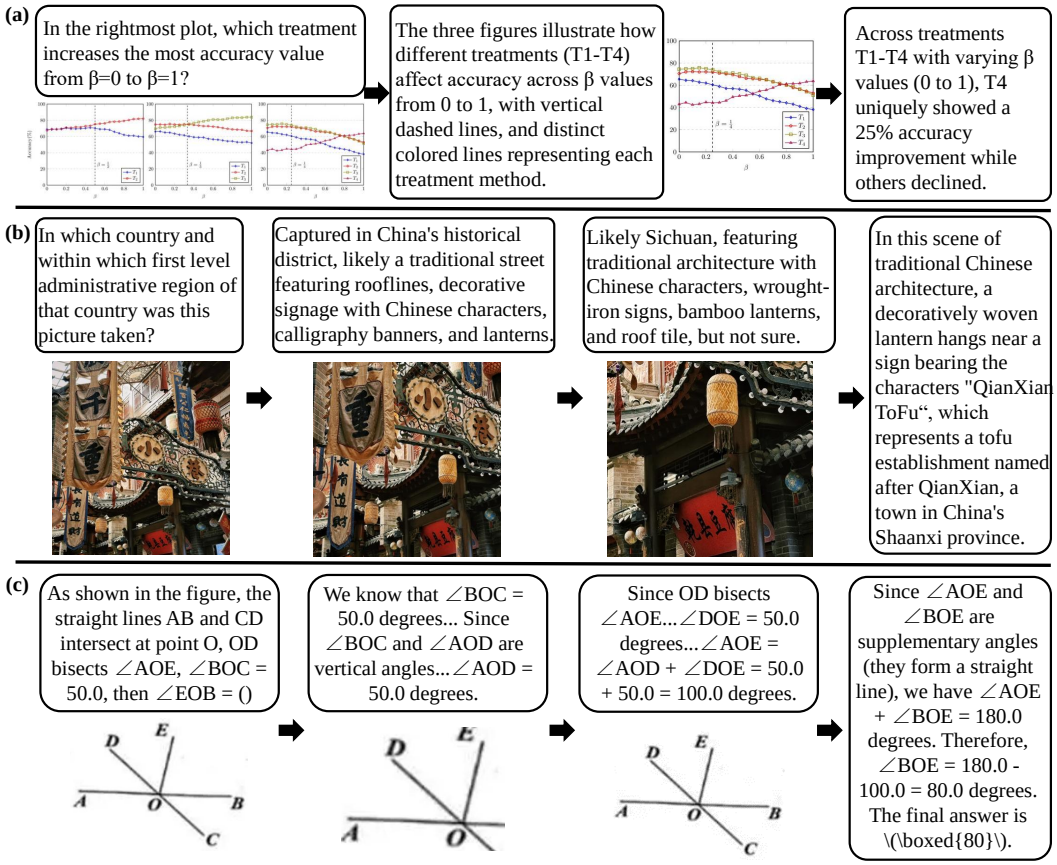

Figure 4: Illustrative cases demonstrating the reasoning process of CoFFT. (a), (b), and (c) show an example from Charxiv, SeekWorld-China, and MathVista benchmarks, respectively.

approaches. CoFFT's breakthrough lies in its synergistic design where DFD evaluates reasoning paths based on both logical progress and visual relevance to guide VFA's region selection, while VFA provides high-quality visual inputs that enable DFD to make more reliable judgments. This mutual reinforcement between focus and reasoning represents CoFFT's core contribution and explains its significant performance improvements over alternative combinations.

4.4 Parameter Analysis

Our CoFFT approach incorporates two critical hyperparameters: the number of foresight statements (l) and the number of samples generated per inference step (k). For our primary experiments, we select $l = 5$ and $k = 4$ to balance computational efficiency and performance. To investigate parametric sensitivity, we systematically varied l from 3 to 7 (interval is 1) and k from 2 to 10 (interval is 2) using Qwen2.5-VL-Instruct as our backbone model. Experiments on MathVista and SeekWorld-China, which represent different visual reasoning aspects, show consistent performance gains with increasing

https://doi.org/10.52202/085713-3544

l and k (Table 5). This scalability suggests potential for further improvements with larger parameters, though practical deployment requires balancing performance with computational costs.

4.5 Case Study

Figure 4 illustrates the performance of CoFFT through three different cases. In case (a), CoFFT successfully identifies relevant chart to correctly answer this question. In case (b), despite the presence of substantially irrelevant information in the complete image, CoFFT exhibits remarkable discrimination by precisely identifying and focusing on critical regions, thereby extracting essential information that leads to accurate results. In case (c), CoFFT successfully identifies relevant regions while demonstrating the ability to dynamically return to the original image when necessary for continued reasoning. These cases demonstrate the approach's adaptability across varying visual complexity levels and information densities, highlighting its robustness in real-world applications.

5 Conclusion

We have introduced CoFFT, a training-free approach that enhances visual reasoning capabilities in VLMs by emulating human visual cognition through three stages: Diverse Sample Generation, Dual-Foresight Decoding, and Visual Focus Adjustment. CoFFT effectively bridges the gap between static visual processing and dynamic reasoning by implementing a sophisticated visual cognition mechanism that continuously refines visual focus based on comprehensive iterative reasoning outputs. Extensive experiments demonstrate a significant 3.1%-5.8% average improvement across challenging visual reasoning tasks without requiring specialized visual expert models or system modifications. However, while CoFFT successfully solves previously unsolvable problems for VLMs, it may introduce unexpected errors in cases where VLMs alone perform well, indicating potential interference with their original well-established capabilities. This suggests that although CoFFT represents a promising approach for VLM enhancement, its robustness requires further improvement, making the systematic optimization of the reasoning process an important direction for future research.

6 Acknowledgements

This work was supported by the National Key Research and Development Program of China (2022YFC3303600), National Natural Science Foundation of China (No. 62137002, 62293550, 62293553, 62293554, 62450005, 62437002, 62477036, 62477037, 62176209, 62192781, 62306229), 'LENOVO-XJTU' Intelligent Industry Joint Laboratory Project, the Shaanxi Provincial Social Science Foundation Project (No. 2024P041), the Natural Science Basic Research Program of Shaanxi (No. 2023-JC-YB-593), the National Research Foundation, Singapore, under its NRF Fellowship (Award# NRF-NRFF14-2022-0001), the Youth Innovation Team of Shaanxi Universities 'Multi-modal Data Mining and Fusion', Project of China Knowledge Center for Engineering Science and Technology, and the Youth AI Talents Fund of China Association of Automation (Grant No._HBRC-JKYZD-2024-311). Mike Shou does not receive any funding for this work.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We show the main claim in the abstract and introduction. And, we have concluded three contributions in Section 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 explicitly points out that VLMs using CoFFT can correctly solve many problems that were previously unsolvable, but there are also some problems that can be solved correctly using VLM directly but CoFFT causes errors.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

https://doi.org/10.52202/085713-3544

Answer: [NA] .

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have introduced the detailed settings such as implementation details and baselines in Section 4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, all codes will be open-sourced after the review process.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have described the experimental details in Section 4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All experimental results as shown in Section 4.2 are the averages obtained after running three trials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

 https://doi.org/10.52202/085713-3544

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All information about computer resources has been described in Section 4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research complies with NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes] .

    Justification: This paper primarily focuses on technical performance impacts, showing positive improvements of 3.1%-5.8% on visual reasoning tasks, which can improve the practicality and reliability of current VLMs in complex visual reasoning scenarios, as shown in Section 4.2.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: This paper does not release any risky data.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All benchmark and VLMs are the license and terms of use explicitly mentioned and properly respected in Section 4.1.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

 https://doi.org/10.52202/085713-3544

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes] .

    Justification: Yes, all assets will be publicly available after the review process.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We use VLM to perform inference tests on multiple public benchmarks to observe the performance improvement of our approach, as shown in Section 4.1.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

References

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

[2] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. Internvideo2. 5: Empowering video mllms with long and rich context modeling. arXiv preprint arXiv:2501.12386, 2025.

[3] Chengyou Jia, Changliang Xia, Zhuohang Dang, Weijia Wu, Hangwei Qian, and Minnan Luo. Chatgen: Automatic text-to-image generation from freestyle chatting. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 13284–13293, 2025.

[4] Xinyu Zhang, Yuxuan Dong, Yanrui Wu, Jiaxing Huang, Chengyou Jia, Basura Fernando, Mike Zheng Shou, Lingling Zhang, and Jun Liu. Physreason: A comprehensive benchmark towards physics-based reasoning. arXiv preprint arXiv:2502.12054, 2025.

[5] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14455–14465, 2024.

[6] Bowen Ping, Minnan Luo, Zhuohang Dang, Chenxi Wang, and Chengyou Jia. Autogps: Automated geometry problem solving via multimodal formalization and deductive reasoning. arXiv preprint arXiv:2505.23381, 2025.

[7] Muye Huang, Lingling Zhang, Han Lai, Wenjun Wu, Xinyu Zhang, and Jun Liu. Vprochart: Answering chart question through visual perception alignment agent and programmatic solution reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pages 3689–3696, 2025.

[8] Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. arXiv preprint arXiv:2410.07149, 2024.

[9] Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. arXiv preprint arXiv:2412.04467, 2024.

 https://doi.org/10.52202/085713-3544

[10] Xinyu Zhang, Lingling Zhang, Xin Hu, Jun Liu, Shaowei Wang, and Qianying Wang. Alignment relation is what you need for diagram parsing. IEEE Transactions on Image Processing, 33:2131–2144, 2024.

[11] Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to look: Training-free perception of small visual details with multimodal llms. 2025.

[12] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? arXiv preprint arXiv:2407.01284, 2024.

[13] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3828–3850, 2024.

[14] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. Advances in Neural Information Processing Systems, 37:113569–113697, 2024.

[15] Muye Huang, Han Lai, Xinyu Zhang, Wenjun Wu, Jie Ma, Lingling Zhang, and Jun Liu. Evochart: A benchmark and a self-training approach towards real-world chart understanding. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pages 3680–3688, 2025.

[16] Kaibin Tian, Zijie Xin, and Jiazhen Liu. SeekWorld: Geolocation is a natural RL task for o3-like visual clue-tracking. `https://github.com/TheEighthDay/SeekWorld`, 2025. GitHub repository.

[17] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.

[18] Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind's eye of llms: Visualization-of-thought elicits spatial reasoning in large language models. In The Thirty-eighth Annual Conference on Neural Information Processing Systems.

[19] Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, et al. Cogcom: A visual language model with chain-of-manipulations reasoning. 2025.

[20] OpenAI. Introducing openai o3 and o4-mini. `https://openai.com/index/introducing-o3-and-o4-mini/`, 2025.

[21] Roy M Pritchard, Woodburn Heron, and DO Hebb. Visual perception approached by the method of stabilized images. Canadian Journal of Psychology/Revue canadienne de psychologie, 14(2):67, 1960.

[22] Keith Rayner. Eye movements and cognitive processes in reading, visual search, and scene perception. In Studies in visual information processing, volume 6, pages 3–22. Elsevier, 1995.

[23] Michele Rucci and Martina Poletti. Control and functions of fixational eye movements. Annual review of vision science, 1(1):499–518, 2015.

[24] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. CoRR, 2024.

[25] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671, 2023.

[26] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. Advances in Neural Information Processing Systems, 36:38154–38180, 2023.

[27] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. arXiv preprint arXiv:2303.11381, 2023.

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

[29] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International conference on machine learning, pages 12888–12900. PMLR, 2022.

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International conference on machine learning, pages 19730–19742. PMLR, 2023.

[31] Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context. arXiv preprint arXiv:2404.13046, 2024.

[32] Melanie Sclar, Gaston Bujia, Sebastian Vita, Guillermo Solovey, and Juan Esteban Kamienkowski. Modeling human visual search: A combined bayesian searcher and saliency map approach for eye movement guidance in natural scenes. In NeurIPS 2020 Workshop SVRHM.

[33] Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychological review, 113(4):766, 2006.

[34] Mengmi Zhang, Jiashi Feng, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Gabriel Kreiman. Finding any waldo with zero-shot invariant and efficient visual search. Nature communications, 9(1):3730, 2018.

[35] Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Predicting goal-directed human attention using inverse reinforcement learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 193–202, 2020.

[36] Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13084–13094, 2024.

[37] Geng Li, Jinglin Xu, Yunzhen Zhao, and Yuxin Peng. Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding. 2025.

[38] Runpeng Yu, Xinyin Ma, and Xinchao Wang. Introducing visual perception token into multimodal large language model. arXiv preprint arXiv:2502.17425, 2025.

[39] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 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, 2024.

[40] Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 19162–19170, 2024.

[41] Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. 2025.

[42] Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. Mllm can see? dynamic correction decoding for hallucination mitigation. 2025.

 https://doi.org/10.52202/085713-3544

[43] Chang Ma, Haiteng Zhao, Junlei Zhang, Junxian He, and Lingpeng Kong. Non-myopic generation of language models for reasoning and planning. 2025.

[44] Fangzhi Xu, Hang Yan, Chang Ma, Haiteng Zhao, Jun Liu, Qika Lin, and Zhiyong Wu. ϕ-decoding: Adaptive foresight sampling for balanced inference-time exploration and exploitation. arXiv preprint arXiv:2503.13288, 2025.

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

[46] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2024.

[47] Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 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, 2024.

[48] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? In The Thirty-eighth Annual Conference on Neural Information Processing Systems.

[49] Rémi Coulom. Efficient selectivity and backup operators in monte-carlo tree search. In International conference on computers and games, pages 72–83. Springer, 2006.

[50] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.