Fast or Slow? Integrating Fast Intuition and Deliberate Thinking for Enhancing Visual Question Answering

Songtao Jiang[1*], Chenyi Zhou[1*] Yan Zhang[2], Yeying Jin[3], Zuozhu Liu[1,4†]

[1]Zhejiang University, [2]Byte Dance, [3]National University of Singapore
[4]Zhejiang Key Laboratory of Medical Imaging Artificial Intelligence
Correspondence[†]: zuozhuliu@intl.zju.edu.cn

Abstract

Multimodal large language models (MLLMs) still struggle with complex reasoning tasks in Visual Question Answering (VQA). While current methods have advanced by incorporating visual prompts, our study uncovers critical limitations: these approaches indiscriminately annotate all detected objects for every visual question, generating excessive visual markers that degrade task performance. This issue stems primarily from a lack of focus on key visual elements, raising two important questions: *Are all objects equally important, and do all questions require visual prompts?* Motivated by Dual Process Theory, which distinguishes between instinctive and deliberate cognitive modes in human reasoning, we propose FOCUS, a plug-and-play approach that dynamically adapts to the complexity of questions, combining fast intuitive judgments with deliberate analytical reasoning to enhance the vision-language reasoning capability of the MLLM. For straightforward questions, FOCUS supports efficient zero-shot reasoning. For more complex tasks, it employs the conceptualizing before observation strategy to highlight critical elements. Extensive experiments on four benchmarks—ScienceQA, TextQA, VizWiz, and MME—demonstrate that FOCUS consistently improves the performance of both open-source and black-box MLLMs, achieving significant gains across all datasets. Ablation studies further validate the importance of combining diverse cognitive strategies with refined visual information for superior performance. Code will be released.

1 Introduction

Multimodal large language models (MLLMs) have demonstrated promising capabilities in visual question answering (VQA) tasks (OpenAI, 2023b; Liu et al., 2023b; Chen et al., 2023). However, MLLMs

---

* Equal contribution.

Figure 1: Comparison of methods: baseline (original image), FOCUS, and SoM (Yang et al., 2023) (previous SoTA method). See Appendix A.1 for more discussion.

still face challenges in addressing complex scenarios, particularly those requiring fine-grained visual perception and the ability to effectively leverage visual information for reasoning tasks (Chen et al., 2024; Jiang et al., 2024b).

Current approaches (Yang et al., 2023; Cai et al., 2024) employ segmentation models to generate visual prompts for all questions by annotating all objects in VQA images as visual markers, aiming to enhance MLLMs' attention to visual information for improving accuracy. While effective, as shown in Figure 1, this coarse-grained guidance of adding visual markers to all objects fails to emphasize key elements. Consequently, MLLMs' attention is often distracted by redundant information, while also introducing significant computational overhead.

In this paper, we introduce FOCUS, a novel approach inspired by *Dual Process Theory* (Kahneman, 2011), which delineates two cognitive modes: a fast, instinctive mode (System 1) and a slower, deliberate mode (System 2). FOCUS enables MLLMs to dynamically alternate between rapid intuitive judgments and thorough analytical reasoning, adapting seamlessly to questions of varying complexity. Specifically, for each visual question, FOCUS first assesses its difficulty level. For simple questions, the MLLM performs efficient zero-shot reasoning. For more complex questions, FOCUS employs a *conceptualizing before observation* strat-

egy, which identifies and highlights critical visual elements using a segmentation model, enabling the MLLM to reason based on refined visual inputs. Prior research has demonstrated that MLLMs struggle when visual elements in the input are ambiguous. As illustrated in Figure 1, FOCUS addresses this limitation by guiding MLLMs to focus on key visual elements, thereby unlocking their full reasoning potential. Furthermore, FOCUS not only enhances the accuracy of responses to complex questions but also optimizes computational efficiency, improving the overall performance.

We applied FOCUS to five MLLMs with diverse architectures and parameter scales, including three open-source models: LLaVA-1.5 (Liu et al., 2023a), MiniGPT4-V2 (Chen et al., 2023), and InstructBLIP (Dai et al., 2023), as well as two black-box models: GPT-4V (OpenAI, 2023b) and Gemini Pro (Gemini Team, Google, 2023). We evaluated their performance on four popular benchmarks, demonstrating that FOCUS consistently enhances model capabilities. Notably, FOCUS combined with LLaVA-1.5-13B achieves state-of-the-art (SoTA) performance across all four benchmarks, while FOCUS with LLaVA-1.5-7B delivers performance comparable to LLaVA-1.5-13B. Compared to the previous SoTA method, our approach not only surpasses SoM on all datasets but also reduces inference time by nearly 44%. Additionally, extensive analysis and ablation studies validate the effectiveness of our dual-thinking strategy and demonstrate its capability to achieve attention calibration.

2 Methods

2.1 Problem Definition

In this paper, we denote the dataset as $\mathcal{D} = \{(I_m, Q_m, A_m)\}_{m=1}^{M}$ where M is the number of data samples. The goal of our tasks is to develop a mapping function $\mathcal{F}(\cdot)$ that can generate answers in response to the questions, represented as:

$$A = \mathcal{F}(I, Q), \tag{1}$$

Here, I denotes the visual input. Q represents the natural language question pertaining to the visual input I. A is the output of the MLLM $\mathcal{F}(\cdot)$. FOCUS begins by evaluating question complexity to determine whether to apply the conceptualizing before observation strategy in the inference process. The overview of FOCUS is demonstrated in Figure 2. The following sections will delve into these components in detail.

2.2 Question Complexity Evaluation

Past research (Xiong et al., 2023) has shown that large language models effectively express their confidence in answers by mimicking human behavior, though they are prone to overconfidence. Building on this insight, we guide the MLLM to indicate their confidence before answering the question. If the MLLM demonstrates high confidence, we classify the question as simple. Otherwise, it is considered complex. Additionally, to mitigate the issue of overconfidence, we employ the self-consistency (Wang et al., 2022) method along with strict confidence evaluation criteria.

Specifically, we first set a high temperature for the MLLM and utilized the prompt strategy from VizWiz (Gurari et al., 2018) to activate the MLLM, generating N responses $\{r_i\}_{i=1}^{N}$, where r_i is either "Answerable" or "Unanswerable". A question is deemed simple if all responses are "Answerable." Otherwise, it is considered complex. In this paper, we set N to three. We also discuss the effectiveness of the selected prompt and chosen N in Appendix B.1.

2.3 Conceptualizing before Observation

This strategy is specifically designed to address the complex question where the original visual information is confusing or insufficient. By conditioning the segmentation process on key extracted elements, it ensures that MLLMs can concentrate on the most crucial parts of the image corresponding to the question. To be specific, the first step involves extracting K key elements $\{k_i\}_{i=1}^{K}$ from the question Q using the language model. Then the key elements $\{k_i\}_{i=1}^{K}$ are combined with bounding boxes generated by the segmentation model $\mathcal{S}$, which serves as an open-set object detector. The segmented regions $\{s_i\}_{i=1}^{K}$ are defined as:

$$s_i = \mathcal{S}(I, k_i) \tag{2}$$

The final processed image I' is produced by aggregating the relevant segments:

$$I' = \sum_{i \in K} s_i, \tag{3}$$

Finally, the processed image I', along with the original question Q, is then fed into the MLLM to generate a more accurate response. In this paper, we used GPT-3.5-turbo (OpenAI, 2023a) as the language model and Grounded-SAM (Ren
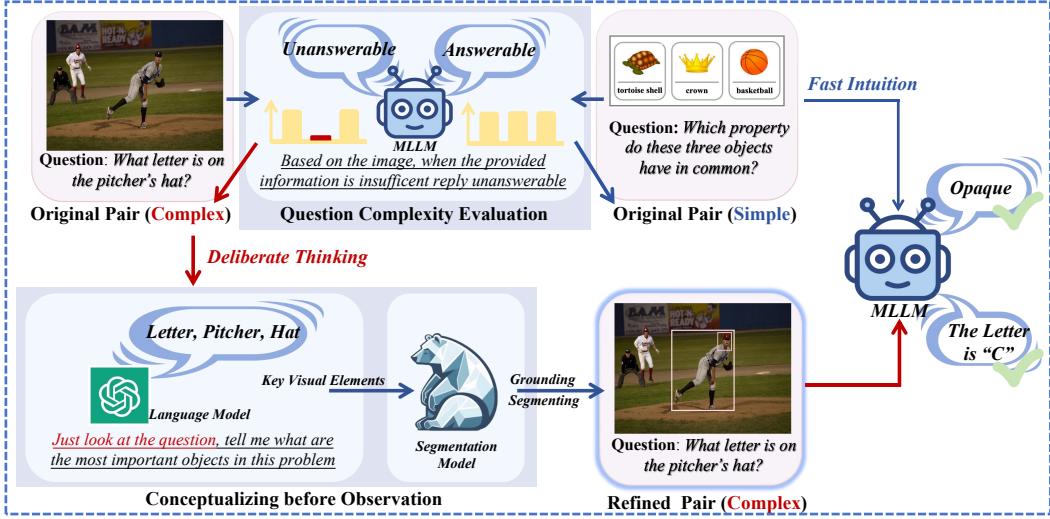
Figure 2: Overview of our model pipeline. (1) Question complexity evaluation for MLLM-based behavior to determine which questions require deliberate thinking. (2) Conceptualizing before observation, helping the model identify the most important visual information in the image.

et al., 2024) as the segmentation model. The effectiveness of Grounded-SAM is discussed in Appendix B.2, while the effectiveness of GPT-3.5-turbo is discussed in Appendix B.3.

3 Experiement

Datasets. We evaluate FOCUS using three popular open-source MLLMs and two black-box MLLMs. The evaluation is conducted on four datasets specifically chosen to assess distinct capabilities: ScienceQA (Lu et al., 2022), TextVQA (Singh et al., 2019), VizWiz (Gurari et al., 2018), and MME (Fu et al., 2023). For ScienceQA, TextVQA, and VizWiz, we measure accuracy, while for MME, we utilize the total score to evaluate overall performance. These datasets collectively address challenges in *logical reasoning, text recognition, real-world visual understanding, and generalization*, offering a comprehensive assessment of FOCUS's capabilities. Additional details are in Appendix B.4.

Baselines. We evaluate two types of baselines for comparison. The first category comprises training-based approaches (Zhang et al., 2023; Li et al., 2023), which leverage additional data or larger models. The second category includes inference-based methods, which utilize techniques such as cot reasoning or visual prompts (Yang et al., 2023; Zheng et al., 2023; Mitra et al., 2024).

3.1 Main Results

Comparison with Previous Methods: As shown in Table 1, our method significantly outperforms previous approaches. Compared to the best-performing method, SoM, our approach not only achieves better performance but also offers a cost advantage, as it does not require adding unnecessary visual markers to the entire image.

Applicability to Different MLLMs: We evaluated FOCUS on multiple MLLMs, as detailed in Table 2 and Table 3. The results show that FOCUS consistently improves performance across diverse model architectures and parameter scales. For black-box MLLMs, we tested on a randomly selected subset of 250 samples from ScienceQA due to the cost.

3.2 Ablation Study and Analysis

Ablation of Thinking Strategies. As shown in Table 4, we assess the impact of Fast Intuition and Deliberate Thinking in FOCUS. Results reveal that uniformly applying a single strategy across all questions reduces performance. While Deliberate Thinking offers limited gains, it incurs significant computational costs. Conversely, relying solely on Fast Intuition hampers the MLLM's ability to address complex questions. Thus, adapting strategies based on question complexity is essential for optimal performance.

How does Deliberate Thinking outperform SoM? We compare the superiority of our Deliberate Thinking approach, which emphasizes Conceptualizing before Observation, with SoM's coarse-grained strategy of globally adding visual markers. As shown in Table 5, our Deliberate Thinking method consistently outperforms SoM across all

Method	LLM	ScienceQA(%)	TextVQA(%)	VizWiz (%)	MME
BLIP-2 (Li et al., 2023)	Vicuna-13B	61.0	42.5	19.6	1293.8
IDEFICS-9B (Laurençon et al., 2024)	LLaMA-7B	—	25.9	35.5	—
IDEFICS-80B (Laurençon et al., 2024)	LLaMA-65B	—	30.9	36.0	—
Qwen-VL-Chat (Bai et al., 2023)	Qwen-7B	68.2	61.5	38.9	1487.5
LVIS-INSTRUCT4V (Wang et al., 2023)	Vicuna-13B	69.0	62.1	51.4	1572.0
CCoT (Mitra et al., 2024)	Vicuna-7B	68.7	58.9	51.1	—
DDCoT (Zheng et al., 2023)	Vicuna-7B	67.6	57.9	50.5	—
DCoT (Jia et al., 2024)	Vicuna-7B	69.3	59.1	51.9	—
SoM (Yang et al., 2023)	Vicuna-13B	71.3	61.5	54.0	1540.1
FOCUS + LlaVA-1.5-7B	Vicuna-7B	70.2	60.1	54.5	1528.4
FOCUS + LlaVA-1.5-13B	Vicuna-13B	**74.4 (+3.1)**	**63.6 (+2.1)**	**58.5 (+4.5)**	**1551.0 (+10.9)**

Table 1: Comparison of methods across ScienceQA, TextVQA, VizWiz, and MME. Rows highlighted in yellow represent training-based methods, green represent inference-based methods, which are our primary comparison targets, and pink highlight FOCUS.

Method	LLM	ScienceQA(%)	TextVQA(%)	VizWiz (%)	MME
LLaVA-1.5	Vicuna-7B	66.8	58.2	50.0	1510.7
LLaVA-1.5	Vicuna-13B	71.6	61.3	53.6	1531.3
LLaVA-1.5 + FOCUS	Vicuna-7B	**70.2 (+3.4)**	**60.1 (+1.9)**	**54.5 (+4.5)**	**1528.4 (+17.7)**
LLaVA-1.5 + FOCUS	Vicuna-13B	**74.4 (+2.8)**	**63.6 (+2.3)**	**58.5 (+4.9)**	**1551.0 (+19.7)**
MiniGPT4-V2	LLaMA-7B	56.7	34.1	44.1	1316.5
MiniGPT4-V2 + FOCUS	LLaMA-7B	**59.6 (+2.9)**	**36.2 (+2.1)**	**47.4 (+3.3)**	**1334.3 (+17.8)**
InstructBLIP	Vicuna-7B	60.5	50.1	34.5	1174.5
InstructBLIP + FOCUS	Vicuna-7B	**62.3 (+1.8)**	**52.3 (+2.2)**	**36.8 (+2.3)**	**1189.8 (+15.3)**

Table 2: Comparative performance of open-source MLLMs across ScienceQA, TextVQA, VizWiz and MME. Both LLaVA-1.5 and MiniGPT-4-V2 use MLP as the connecting component between the ViT (Dosovitskiy, 2020) and LLM backbones, whereas InstructBLIP employs Q-Former (Li et al., 2023)

datasets, demonstrating its effectness in VQA.
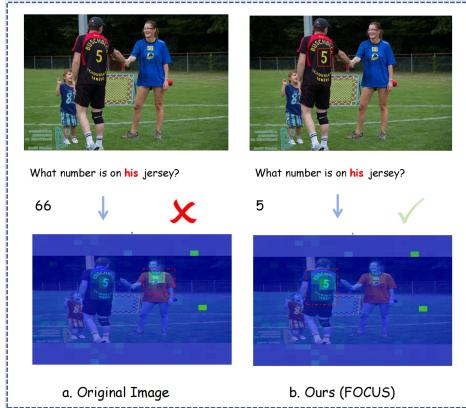
Figure 3: Visual attention visualization in LLaVA-1.5. The green areas indicate higher attention weights.

Visual Attention Calibration. We assessed whether FOCUS's fine-grained visual markers enhance MLLMs' ability to filter distractions by analyzing the attention patterns of LLaVA-1.5. As shown in Figure 3, in the original image, the MLLM mistakenly focuses on the woman's jersey number, resulting in an incorrect answer. In con-

trast, FOCUS guides the MLLM's attention to the relevant image regions. This highlights FOCUS's ability to help MLLMs identify critical visual information, enhancing their perceptual accuracy.
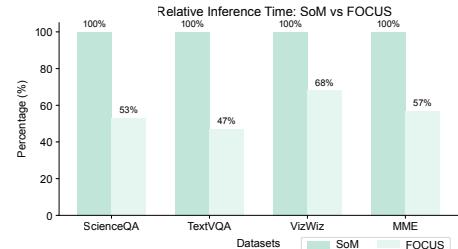

Figure 4: Relative inference time comparasion

Cost Analysis. As illustrated in Figure 4, our method achieves faster inference while maintaining superior performance compared to SoM. This efficiency stems from our approach's ability to skip the unnecessary addition of visual prompts for certain questions, as well as its avoidance of annotating all objects in the image—a process that is both time-consuming and detrimental to performance. These advantages demonstrate that our method is a

cost-effective solution for VQA tasks.

Analysis of Answerable and Unanswerable Proportions. We perform an analysis of the proportions of answerable versus unanswerable responses across different VLMs and benchmarks. Specifically, we compute average unanswerable proportions for three base VLMs, obtaining 39.00% for LLaVA-1.6, 42.40% for LLaVA-1.5, and 54.83% for InstructBlip. These results indicate a clear trend: stronger models show lower unanswerable proportions, implying that more capable models have greater confidence and consequently prefer FI strategy. Additionally, we analyze average unanswerable proportions across four benchmarks—VizWiz, MME, TextVQA, and ScienceQA—and observe a ranking by decreasing uncertainty: VizWiz (58.33%), MME (48.97%), TextVQA (44.20%), and ScienceQA (30.13%). Our findings suggest that benchmarks demanding more detailed perceptual abilities elicit higher uncertainty, whereas simpler visual tasks, such as ScienceQA with cartoon-like images, enable models to respond with higher confidence.

More Analysis of Unanswerable Questions. We further carefully analyze the unanswerable questions along with their associated images, identifying three prevalent scenarios where the model demonstrates increased uncertainty and thus necessitates DT strategy. First, images containing excessive elements, especially those with more than two primary objects, consistently confuse the models, irrespective of whether the style is realistic or cartoon-like (e.g., in ScienceQA). Second, images with unclear or ambiguous regions, notably common in the VizWiz benchmark, directly contribute to heightened uncertainty. Third, dense textual information in images, predominantly observed in TextVQA, visually distracts and overwhelms the VLMs, frequently causing confusion and errors. Additionally, we notice that some questions require specialized external knowledge, particularly within the ScienceQA benchmark, which involves physics or chemistry concepts. Although our FOCUS effectively directs model attention to relevant image regions, the absence of domain-specific knowledge occasionally leads to incorrect predictions. This limitation suggests a promising direction for future work, such as integrating Retrieval-Augmented Generation (RAG) techniques to enrich the model's knowledge base.

Analysis of Confidence Evaluation Strategies. We conducted experiments to further investigate the impact of different confidence evaluation strategies. Specifically, we sampled 100 challenging questions from LLaVA-1.5, consisting of 50 questions incorrectly answered using DT but correctly with FI, and another 50 incorrectly answered using FI without employing DT. We compared three confidence measurement strategies: (1) our original approach (prompt with self-consistency), (2) semantic entropy computed from logits for binary classification of "answerable/unanswerable," and (3) self-consistency combined with semantic entropy. The results are summarized in Table 6. Our analysis reveals several key observations: First, semantic entropy slightly improves performance by addressing the overly stringent requirement of consistent answerable predictions in the original method, reducing unnecessary DT usage on simpler questions. Second, incorporating self-consistency into semantic entropy further enhances performance by averaging entropy scores across multiple predictions, thereby reducing noise and improving robustness.

4 Conclusion

We propose FOCUS, a novel approach that enhances MLLMs in VQA. By adapting thinking strategies to question complexity, FOCUS improves performance and resource efficiency. Experiments show it surpasses existing methods, offering a scalable solution for multimodal reasoning.

5 Limitations

While our FOCUS shows promise in enhancing MLLMs for vision-language tasks, it has several limitations. The method relies on external tools to provide fine-grained visual markers, which could introduce dependencies. Additionally, addressing the fundamental issues of visual hallucinations in MLLMs likely requires modifications at the model architecture and training levels, rather than solely relying on post-processing techniques. Therefore, exploring how to integrate this method more deeply with training processes presents an interesting direction for future research.

Acknowledgements

References

Shengnan An, Yifei Li, Zeqi Lin, Qian Liu, Bei Chen, Qiang Fu, Weizhu Chen, Nanning Zheng, and Jian-Guang Lou. 2022. Input-tuning: Adapting unfamiliar inputs to frozen pretrained models. *arXiv preprint arXiv:2203.03131*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. 2024. Vip-llava: Making large multi-modal models understand arbitrary visual prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12914–12923.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *arXiv preprint arXiv:2402.04788*.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P Xing, and Zhiting Hu. 2022. Rlprompt: Optimizing discrete text prompts with reinforcement learning. *arXiv preprint arXiv:2205.12548*.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Gemini Team, Google. 2023. Gemini 1 report. Google.

Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2021. Ppt: Pre-trained prompt tuning for few-shot learning. *arXiv preprint arXiv:2109.04332*.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Zixi Jia, Jiqiang Liu, Hexiao Li, Qinghua Liu, and Hongbin Gao. 2024. Dcot: Dual chain-of-thought prompting for large multimodal models. In *The 16th Asian Conference on Machine Learning (Conference Track)*.

Songtao Jiang, Yuan Wang, Sibo Song, Yan Zhang, Zijie Meng, Bohan Lei, Jian Wu, Jimeng Sun, and Zuozhu Liu. 2025. Omniv-med: Scaling medical vision-language model for universal visual understanding. *arXiv preprint arXiv:2504.14692*.

Songtao Jiang, Yan Zhang, Ruizhe Chen, Yeying Jin, and Zuozhu Liu. 2024a. Modality-fair preference optimization for trustworthy mllm alignment. *arXiv preprint arXiv:2410.15334*.

Songtao Jiang, Yan Zhang, Chenyi Zhou, Yeying Jin, Yang Feng, Jian Wu, and Zuozhu Liu. 2024b. Joint visual and text prompting for improved object-centric perception with multimodal large language models. *arXiv preprint arXiv:2404.04514*.

Songtao Jiang, Tuo Zheng, Yan Zhang, Yeying Jin, Li Yuan, and Zuozhu Liu. 2024c. Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3843–3860.

Daniel Kahneman. 2011. *Thinking, fast and slow*. Farrar, Straus and Giroux.

S KAZEMZADE, V Ordonez, M MATTENV, et al. 2014. Referring to objects in photographs of natural scenes. In *Empirical Methods in Natural Language Processing*, volume 28, pages 787–789.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. 2024. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023c. Gpt understands, too. *AI Open*.

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

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

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

OpenAI. 2023a. Gpt-3.5 turbo fine-tuning and api updates. https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates.

OpenAI. 2023b. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf. 2023.

R OpenAI. 2023c. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2.

Jing Qian, Li Dong, Yelong Shen, Furu Wei, and Weizhu Chen. 2022. Controllable natural language generation with contrastive prefixes. *arXiv preprint arXiv:2202.13257*.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Hao Sun. 2023. Offline prompt evaluation and optimization with inverse reinforcement learning. *arXiv preprint arXiv:2309.06553*.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Cer. 2021. Spot: Better frozen model adaptation through soft prompt transfer. *arXiv preprint arXiv:2110.07904*.

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. 2023. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Qinwei Xu, Xingkun Xu, Chenyi Zhou, Zuozhu Liu, Feiyue Huang, Shaoxin Li, Lifeng Zhu, Zhian Bai, Yuchen Xu, and Weiguo Hu. 2025. Towards normalized clinical information extraction in chinese radiology report with large language models. *Expert Systems with Applications*, 271:126585.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, G. Karypis, and Alexander J. Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

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

A Appendix

A.1 Comparison with SoM

Although SoM (Yang et al., 2023) also employs visual prompts to enhance the performance of VLMs, our approach differs significantly in several key aspects:

1. **Scope and Focus**: SoM primarily aims to unleash the visual grounding abilities of VLMs, with its experiments centered on tasks such as Open-vocabulary Image Segmentation, Referring Segmentation, Phrase Grounding, and

Video Object Segmentation. These tasks are inherently focused on visual grounding rather than addressing challenges in the VQA domain. In contrast, we explore the potential of SoM in VQA and propose targeted improvements to adapt it for this domain.

2. **Annotation Strategy**: SoM annotates all content in an image using a SAM-based model (Kirillov et al., 2023). However, our study reveals that this approach is less suitable for VQA tasks, as it fails to highlight critical information and may introduce coarse-grained visual markers that potentially disrupt the semantic integrity of the original image, thereby degrading VQA performance.

3. **Prompting Efficiency**: Current visual or text prompting methods for VQA typically apply prompts uniformly to all questions. Our research demonstrates that not all questions require prompting, and we propose a more cost-efficient solution. This insight provides a new direction for the future development of visual prompting techniques.

B More Analysis and Discussion

B.1 Effect of Chosen Prompt

To evaluate whether our *question complexity evaluation* strategy effectively identifies images that require further fine-grained processing, we used GPT-4V (OpenAI, 2023b). We input a sample of 100 images from VizWiz (Gurari et al., 2018) along with their corresponding questions, using the prompt: "Is this VQA question a complex task requiring fine-grained perception, or a simpler task that only requires coarse-grained perception? Please provide a detailed explanation." After manually reviewing GPT-4V's reasoning, we labeled the images, determining that 59 required fine-grained visual information extraction, while 41 only needed coarse-grained perception. We then explored the accuracy of question complexity evaluation with 1, 3, and 5 response iterations to assess the need for further fine-grained visual information extraction. The accuracies for each iteration were 59%, 81%, and 83%, respectively. These results demonstrate that question complexity evaluation is highly effective in identifying which visual question answering (VQA) questions might confuse the model. This enables us to selectively apply fine-grained

information marking to specific images, achieving satisfactory results with just three iterations.

B.2 Effect of Segmentation Model

We also evaluated the segmentation model's ability to mark the extracted keywords. Given that Grounded-SAM (Ren et al., 2024) was trained on a diverse set of open-world objects, and most VQA images come from the COCO dataset (Lin et al., 2015), the marking accuracy is high. We randomly selected 100 samples from VizWiz, achieving an accuracy of over 85%, confirming the effectiveness of our method. However, we set the threshold to 0.7 for the segmentation model. When the model is uncertain about the segmentation, it refrains from making further modifications to the image to avoid disrupting the original visual content. This ensures that the segmentation model does not negatively impact the MLLMs in cases of failure.

B.3 Effect of Keywords Extraction

We randomly selected 100 samples from VizWiz to assess GPT-3.5-Turbo's (OpenAI, 2023a) ability to extract key concepts from VQA questions. Due to GPT-3.5-Turbo's strong information extraction capabilities (Wei et al., 2023) and the relatively simple nature of the VQA questions, 87% of the samples allowed for accurate keyword extraction. However, in the remaining 13%, the relevant objects were not explicitly mentioned in the questions, preventing extraction. In such cases, the segmentation model refrains from adding markers, preserving the image's original semantic content.

Method	ScienceQA(%)
GPT-4V	79.2
Gemini Pro	77.0
GPT-4V + FOCUS	82.4 (+3.2)
Gemini Pro + FOCUS	79.4 (+2.4)

Table 3: Comparative performance of black-box MLLMs on a sample of 250 ScienceQA examples.

Method	ScienceQA (%)	TextVQA (%)	VizWiz (%)	MME
FOCUS	74.4	63.6	58.5	1551.0
Only FI.	71.6	61.3	53.6	1531.3
Only DT.	72.2	61.6	54.9	1538.6

Table 4: Ablation study on Fast Intuition (FI) and DT (Deliberate Thinking). In this experiment, we used LLaVA-1.5 as the MLLM.

Method	SciQA. (%)	TVQA. (%)	VizWiz (%)	MME
SoM	71.3	61.5	54.0	1540.1
FI. + SoM	72.6	61.8	55.2	1544.9
FOCUS (FI. + DT.)	**74.4**	**63.6**	**58.5**	**1551.0**

Table 5: Comparison of Deliberate Thinking (DT) and SoM. SciQA denotes ScienceQA, and TVQA denotes TextVQA.

B.4 Details of Datasets

We evaluate FOCUS using three popular open-source MLLMs and two black-box MLLMs with diverse architectures: LLaVA-1.5 (Liu et al., 2023a), MiniGPT4-V2 (Chen et al., 2023), Instruct-BLIP (Dai et al., 2023), GPT-4V (OpenAI, 2023b), and Gemini Pro (Gemini Team, Google, 2023). The evaluation is conducted on four datasets specifically chosen to assess distinct capabilities:

ScienceQA (Lu et al., 2022) primarily features images not from the real world, focusing on testing the logical reasoning capabilities of the models. We explore whether finer-grained visual information could enhance the models' reasoning abilities.

TextVQA (Singh et al., 2019) contains images rich in challenging textual information, primarily evaluating the models' text recognition capabilities. We explore our model's ability to extract challenging fine-grained text information in complex visual scenes.

VizWiz (Gurari et al., 2018) consists mostly of real-life images, providing a comprehensive assessment of the MLLMs' proficiency in processing real-world imagery. This dataset includes images that contain complex visual elements, some of which are not very clear, making it difficult for models to interpret the content.

MME (Fu et al., 2023) is a benchmark designed to evaluate MLLMs, addressing the lack of comprehensive assessment tools by measuring both perception and cognition across 14 diverse subtasks. We explore whether our approach could enhance the overall performance of MLLMs on this dataset.

For ScienceQA, TextVQA, and VizWiz, we assess accuracy, while for MME, we use the total score to evaluate the models' overall ability to understand and process different types of data.

C Related Work

Prompt Optimization in VQA Multimodal large language models (MLLMs) have shown promising capabilities in performing visual question answering (VQA) tasks (OpenAI, 2023b; Xu et al., 2025; Liu et al., 2023b; Chen et al., 2023; Jiang et al., 2024c, 2025, 2024a). Prior research aimed at improving MLLMs for Visual Question Answering (VQA) has primarily focused on gradient-based methods (Li and Liang, 2021; Vu et al., 2021; Gu et al., 2021; Liu et al., 2023c; Mokady et al., 2021; Qian et al., 2022; An et al., 2022) and prompt optimization techniques (Deng et al., 2022; Sun, 2023; Jiang et al., 2024b; Zhang et al., 2023). Among these, the Multimodal Chain of Thought (MM-CoT) method (Zhang et al., 2023) is particularly notable for its ability to integrate visual and textual information within LLMs, achieving superior performance in reasoning tasks. However, this approach incurs higher training costs.

Recently, visual prompt strategies have gained prominence in MLLMs. The Set-of-Mark (SoM) method (Yang et al., 2023) is the first to experiment with applying visual marks as prompts on image inputs for MLLMs like GPT-4V, aiming to enhance their grounding capabilities. SoM has demonstrated state-of-the-art results on datasets such as RefCOCOg and DAVIS2017 (KAZEMZADE et al., 2014; OpenAI, 2023c). However, its application to VQA tasks remains underexplored. Moreover, we observe that SoM often segments the entire image when applied to VQA tasks, which may overlook critical information necessary for accurate answers. In contrast, our approach introduces more precise segmentation in MLLMs, focusing on the most relevant parts of the image to significantly improve VQA performance.

C.1 Case Study

There are more case studies on GPT-4V:

Strategy	DT (Count)	FI (Count)	Corrected (DT/FI)
Original Strategy (Prompt + Self-Consistency)	50	50	0/0
Semantic Entropy	41	59	3/9
Self-Consistency Entropy	43	57	4/7

Table 6: Comparison of different confidence evaluation strategies on a subset of 100 challenging cases.
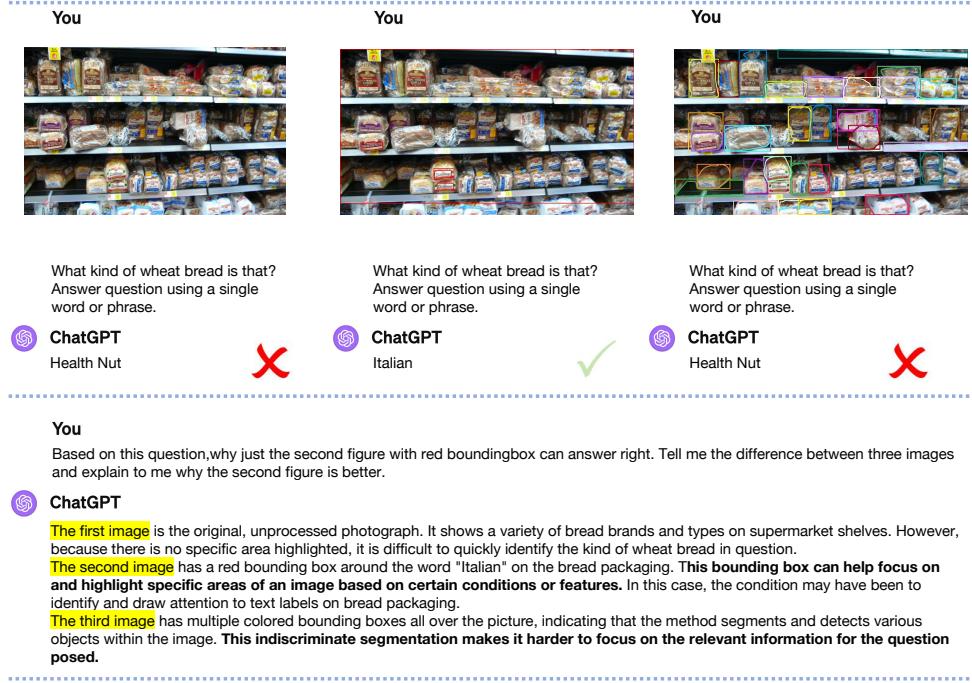
Figure 5: More cases on GPT-4V
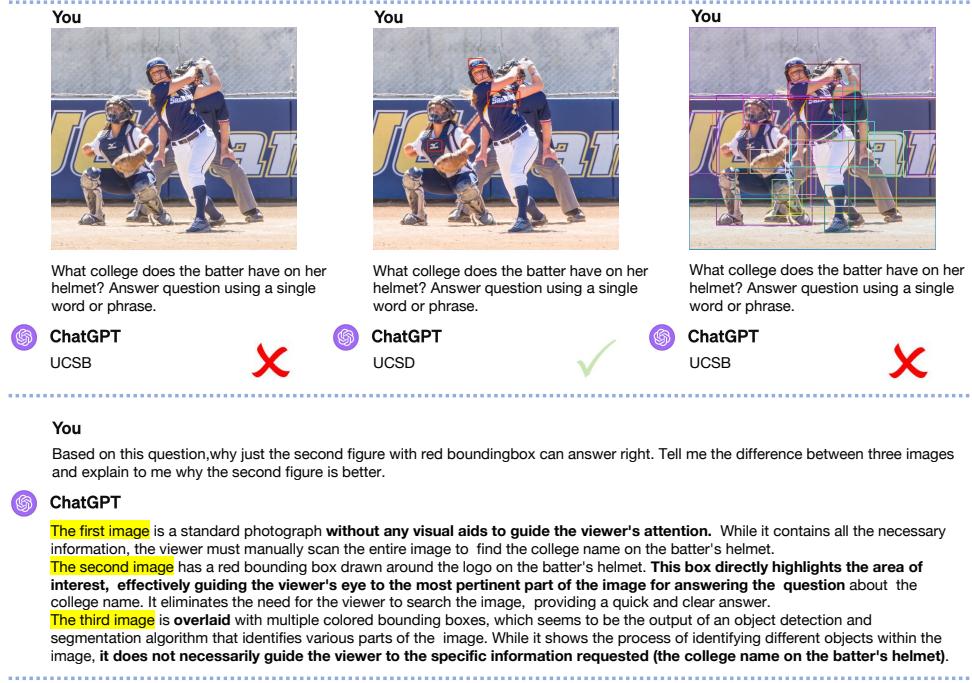
Figure 6: More cases on GPT-4V