Enhancing Spatial Understanding in MLLMs via Fine-Grained Image-Text Dual Prompt Learning

Anonymous ACL submission

Abstract

Multimodal large language models (MLLMs) perform excellently in cross-modal tasks, but their spatial understanding capabilities are still far from human-level performance, and existing prompt learning methods have not fully unlocked their potential. Therefore, we propose a fine-grained image-text dual prompt learning framework aimed at enhancing the spatial understanding ability of MLLMs. Our method utilizes three mechanisms—target detection, image segmentation, and attention visualization—to provide fine-grained prompts for the input image from different angles, and employs an LLM-based refined Chain of Thought method to transform the text into fine-grained prompts. This approach strengthens the interaction between the image and text prompts, facilitating a deeper semantic analysis by MLLMs. We evaluate our proposed method using the BLINK dataset, with two tasks—counting and relative depth judgment—that effectively assess spatial understanding capabilities. Experimental results show that MLLMs prompted by our method demonstrate significant improvement in both tasks, which strongly validates the effectiveness of our approach.

1 Introduction

In recent years, the rise of MLLMs such as LLaVA (Liu et al., 2024b) and GPT-4V (Achiam et al., 2023) has not only expanded the boundaries of natural language processing and computer vision but also revealed emergent behaviors exhibited by these models on tasks they were not explicitly trained for. For example, they can integrate image and text information to ask and answer questions about the content of images, demonstrating their potential in multimodal interaction. This phenomenon is partly attributed to the vastness of the model's training corpus— the internet— which inherently contains rich cross-modal communication patterns, including image-text pairs, image descriptions, visual question answering, and other diverse data forms.

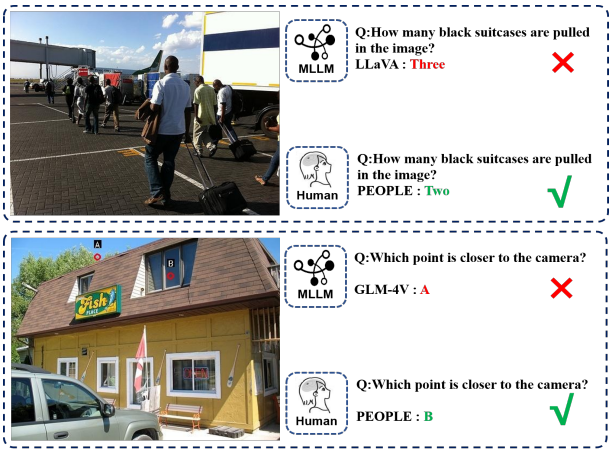

Figure 1: In images containing multiple objects and complex backgrounds, humans can quickly identify the main objects, understand their relationships, and accurately answer related questions. However, MLLMs may provide incorrect answers due to their failure to capture spatially critical details in the image.

However, when faced with common spatial reasoning tasks, MLLMs often exhibit discrepancies in understanding images compared to humans, as shown in Figure 1. This difference arises because humans not only rely on visual information when interpreting images but also integrate rich background knowledge, life experiences, and intuitive judgments from the real physical world. As a result, humans can quickly simulate a 3D scene from a 2D image and make inferences about patterns and causal relationships. In contrast, MLLMs primarily rely on statistical patterns in training data, lacking genuine "understanding" abilities(Fu et al., 2025). Particularly, current MLLMs are pre-trained on 2D images and text, without explicit modeling of spatial location information. Furthermore, the coarse-grained interaction between image and text during training makes it difficult for the model to

fully grasp the image details, especially the relationships between them. Therefore, when encountering complex scenes that require spatial reasoning, the models struggle to achieve satisfactory results.

Although MLLMs lack explicit 3D spatial modeling, the abundant spatial location information embedded in 2D images gives them the potential to understand spatial relationships. In the exploration of unimodal large models, researchers have focused on using carefully designed text prompts to trigger specific behaviors in the model, a strategy that has been successful in many downstream tasks. However, in the task of spatial understanding for MLLMs image information plays a more crucial role, and relying solely on text prompts is insufficient to enable the model to deeply understand the posed questions.

To address the above issues, we propose a fine-grained image-text dual prompt learning framework aimed at enhancing the spatial understanding of MLLMs. In terms of image prompts, our framework constructs fine-grained object boundary information through image segmentation, fine-grained object location information through target detection, and highlights the important information understood by external multimodal pretrained models through image attention heatmaps generated from image-text interactions, thus bridging the gap between image and text prompts. In terms of text prompts, we use a fine-grained chain of thought from unimodal large models to decompose the question into step-by-step fine-grained information, thereby enhancing the model's understanding of the image content. We evaluate the proposed method on the BLINK dataset(Fu et al., 2025) using object counting and relative depth estimation tasks, which effectively measure spatial understanding capabilities. Experimental results demonstrate that the performance of MLLMs prompted by our method shows significant improvement in both tasks.In summary, our contributions are as follows:

- As far as we know, we are the first to systematically explore the role of fine-grained image prompt learning in enhancing the spatial understanding capabilities of MLLMs.

- We propose a fine-grained image prompt learning strategy that effectively integrates image segmentation, target recognition, and attention heatmaps to enhance the spatial understanding capabilities of MLLMs.

- We propose a method that uses a unimodal large model to decompose textual questions into fine-grained chain of thought and combine it with the fine-grained image prompt learning strategy, forming a fine-grained image-text dual prompt learning framework. This combined approach further enhances the MLLMs' ability to understand spatial location information.

- We conducted experiments on multiple MLLMs, and the results show that the proposed method performs excellently in spatial understanding-related tasks, providing direction for the design and optimization of future MLLMs.

2 Related Work

2.1 Multimodal Large Language Models

Large language models (LLMs) have achieved widespread success in the field of natural language processing (NLP). From early models like BERT(Kenton and Toutanova, 2019) and GPT-2(Radford et al., 2019), to more recent ones like GPT-3(Brown et al., 2020), instructGPT(Ouyang et al., 2022), and various other large-scale open-source language models such as LLaMA(Touvron et al., 2023a) and LLaMA2(Touvron et al., 2023b), significant advancements have been made in NLP, especially in natural language understanding and generation.

In multimodal research, how to apply these powerful LLMs to multimodal tasks has also gradually gained widespread attention. Early studies, such as Frozen(Tsimpoukelli et al., 2021), achieved impressive performance by training a visual encoder to encode image inputs as prefixes to a pre-trained language model. BLIP(Li et al., 2022) pre-trained a multimodal encoder-decoder hybrid model to further enhance performance on vision-language tasks. BLIP2(Li et al., 2023) introduced a Q-former to efficiently align visual features with LLMs. Additionally, other studies such as MiniGPT4(Zhu et al., 2023), LLaVA(Liu et al., 2024b), and Qwen-VL(Bai et al., 2023) use adapters (e.g., linear layers or multi-layer perceptrons) to further align image features extracted from visual encoders.

2.2 Visual Prompts

In multimodal large models, a common strategy is to insert a set of learnable tokens before the text input, visual input, or both, as visual prompts,

to guide a frozen model (i.e., one whose parameters are not updated) to perform a specific task. A method has been proposed to enable a frozen model to perform new tasks through a single image perturbation(Bahng et al., 2022). Additionally, some researchers use red circles on images to prompt the CLIP model, improving its performance(Shtedritski et al., 2023). Other studies have enhanced the performance of large models by segmenting images and adding labels such as numbers on them (Yang et al., 2023).

2.3 Fine-grained Segmentation

The YOLO algorithm(Jiang et al., 2022) treats target detection as a regression problem, predicting bounding boxes and class probabilities directly through a single forward pass, enabling fast and accurate object localization. Object detection provides an overall segmentation of objects, which can serve as an initial division of image content when handling complex scene images. In the field of image segmentation, the Set-of-Mark (SoM)(Yang et al., 2023) technique offers a more fine-grained segmentation method, dividing the image into different regions and using interactive segmentation models to identify these regions. SoM not only focuses on the overall objects in an image but also strives to recognize regions with different granularities, allowing for a deeper understanding of these areas, thus achieving more precise image segmentation.

2.4 Multimodal Attention Interaction

Multimodal interaction considers four types of attention interactions between text and images(Chefer et al., 2021). For each type, a relevance map is constructed and calculated on the attention layers through forward propagation. Before performing attention operations, self-attention interactions are initialized as an identity matrix, while cross-modal interactions are initialized to zero. As the attention layers contextualize the tokens, the relevance map is updated using attention maps and gradients, accounting for the importance and relevance of heads in multi-head attention. The update rules differ between self-attention and multimodal attention. Finally, by examining the row corresponding to the [CLS] token in the relevance map, the relevance of each token can be extracted for the final classification task of the Transformer model.

3 Method

In this section, we outline our approach. First, we describe the tasks used to evaluate the spatial understanding capabilities of MLLMs. Then, we present our proposed prompt learning framework based on image-text dual prompts. Finally, we provide a detailed implementation of each module within the framework.

3.1 Task Description

The BLINK(Fu et al., 2025) evaluation framework, widely recognized in the academic community, is effective for assessing the capabilities of MLLMs. The two core tasks in BLINK—counting and relative depth judgment(Depth)—are particularly adept at testing a model's spatial and positional awareness. Therefore, we chose these two tasks to evaluate the effectiveness of our framework in enhancing the spatial understanding abilities of MLLMs.

Counting: The model is given an image-text pair as input, where the text requires the MLLM to answer the number of objects of a particular category in the image. Each sample includes an image, a question, and a numerical answer. In addition to the correct answer, three other numbers are randomly selected as distractors. This task effectively reflects the spatial understanding ability of MLLMs by requiring them to count the objects in the image. This involves logical reasoning about spatial relationships between images at different locations, especially in complex scenes where objects may overlap, be occluded, or vary in size and appearance. The questions are selected from the TallyQA dataset(Acharya et al., 2019).

Depth: The model is given an image-text pair, where the text asks the MLLM to determine which labeled point is closer to the camera. Each question includes an image and two specified points. The task is to determine which point is closer to the observer. This task serves as an alternative metric for validating whether the geometric understanding abilities of current MLLMs are close to human-level. It effectively reflects the spatial understanding of objects in images. Samples for testing are constructed using manually annotated data from the Depth in the Wild dataset(Chen et al., 2016), ensuring the authenticity and challenge of the task.

3.2 Prompt Framework

The MLLM M takes an image $I \in R^{H\times W\times 3}$ and a text sequence of length l_i ,denoted as $T^i =$

$[t_1^i, ..., t_{l_i}^i]$,as input, and the model outputs a text sequence of length l_0 , denoted as $T^O = [t_1^0, ..., t_{l_0}^0]$, formulated as:

$$T^O = M(I, T^i) \tag{1}$$

Building on this, in this section, we propose a framework that significantly enhances the spatial understanding ability of MLLMs. The framework integrates four prompt learning modules to form a comprehensive prompt learning template, aimed at processing the input image and text more precisely. Specifically, we introduce fine-grained image prompts to process the image I,generating the processed image C,and simultaneously apply a text-level fine-grained chain of thought (CoT) to optimize the text sequence T^i,resulting in the processed text T^Q.Through our method, the large model outputs T^O,fully exploiting the spatial understanding potential of MLLMs while ensuring deep interaction between the image and text, as shown in the following formula. The overall framework is illustrated in Figure 2.

$$T^O = M(C, T^Q) \tag{2}$$

3.3 Fine-grained Image Prompts

In this section, we describe the specific process of three types of image prompts.First, we divide the input image of size $H \times W$ into K distinct regions S, with the output represented by S , consisting of K binary masks. The MLLM, based on the prompt text, may not treat the object of interest in the image as a complete whole, but rather as a part of it. This makes it difficult to capture the boundaries of the region. Therefore, by performing a finer-grained segmentation of the image, the boundaries of objects can be more easily distinguished.

$$S = [s_1, ..., s_K] \in \{0, 1\}^{K \times H \times W} \tag{3}$$

Next, we apply an target detection algorithm to perform fast target detection on the image I, generating a set of bounding boxes Y. Since the attention capability of MLLMs is limited when it comes to the content of the image, we use an target detection algorithm because it can quickly identify and accurately locate the main elements in the image. With the bounding boxes generated by target detection, the model can rapidly focus on key objects. Thus, the use of the target detection algorithm allows the large model to better attend to the main parts of the image as indicated by the text prompt.

$$Y = F_y\left(I\right) \tag{4}$$

Finally, we generate an attention map A for the input image I and text T^i after processing. This prompting method primarily targets the multimodal attributes of large models, aiming to fully leverage the role of the text modality. It highlights the weak supervision of the text modality over the image modality, focusing on the interaction between images and text, and reflecting the attention of the text within the image. By generating attention maps, the MLLM can more accurately understand the key issues in the text prompts.

$$A = softmax\left(\frac{Q_{T^i} \cdot K_I^T}{\sqrt{d_h}}\right) \tag{5}$$

In the formula, Q_{T^i} is the query matrix for the text,K_I is the key matrix for the image, and d_h is the dimension of the attention head. The $softmax$ operation ensures that the attention of each text token to the image is correctly normalized at each layer.

In summary, we propose a comprehensive image processing method that integrates target detection algorithms, image segmentation algorithms, and text attention mechanisms in images. Specifically, we first use an object detection algorithm to quickly locate the main elements in the image, then combine it with image segmentation algorithms for fine-grained partitioning. At the same time, we leverage the attention areas of the text in the image to achieve fine-grained segmentation driven by the text. This method not only retains the advantages of image segmentation techniques but also compensates for the limitations of previous work through object detection and text attention mechanisms, helping large models to more comprehensively and deeply understand the spatial relationships in images. The weighted results of the three modules are integrated to generate the final fused image C.The parameters λ_1,λ_2,λ_3 represent the weight parameters for the segmentation region set S,bounding box set Y, and attention map A, respectively.

$$C = fusion(\lambda_1 \times S, \lambda_2 \times A, \lambda_3 \times Y) \tag{6}$$

3.4 Fine-Grained CoT

In this section, we describe the specific process of text prompting. While exploring how to enhance the understanding of the interaction between

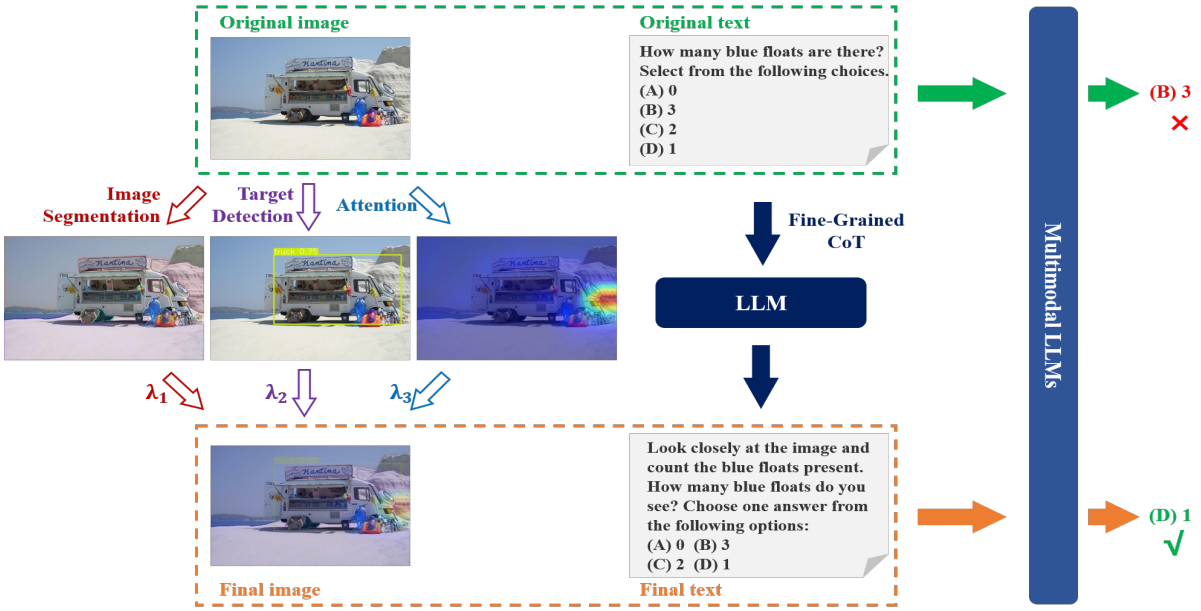

Figure 2: Our method consists of two main components: image prompts and text prompts. In the image prompt phase, we apply three different processing techniques to the raw input image and combine them based on their respective importance, assigning different weights for fusion. In the text prompt phase, we use fine-grained Chain of Thought techniques and design a question decomposition template to structure complex problems into sub-questions, which are then decomposed by the large language model following our approach.

images and text in multimodal large models, we not only delved into image processing techniques but also proposed an innovative fine-grained chain of thought (CoT) for text processing. Since large models have limited understanding of text, we decompose the text by fine-grained segmentation and guide it step by step. This significantly improves the model's focus on key elements of the problem, thus enhancing its ability to solve complex problems.

For example, in a counting task, suppose the scene is an image containing several dogs, only some of which are blue. If the model is directly asked, "How many blue dogs are there in the image?", it might need to scan the entire image and identify all the dogs. In a complex or unclear background, this approach could increase the error rate and lower accuracy. To solve this problem, we propose a decomposition strategy based on a fine-grained CoT at the text level.

We decompose the original question into two more specific and manageable sub-questions: "What are the dogs in the image?" and "How many of these dogs are blue?" This decomposition helps the model first focus on identifying dog features in the image, and then the second question further guides the model to identify and count the blue dogs. In this way, the model can more efficiently use the information in the text prompt to guide its search and recognition process in the image, reducing unnecessary computational overhead and errors. The following formula demonstrates this process:

$$T^Q = Decompose\left(T^i\right) = \{Q_1, Q_2\} \quad (7)$$

We input this case along with the text prompt T^i into the large language model, allowing it to output the decomposed text prompt T^Q,where T^Q is the combination of the two sub-questions Q_1 and Q_2,The primary reason for choosing a large language model (LLM) for question decomposition is that the training corpora for both MLLMs and unimodal large language models are largely derived from the internet. This means that they are already adapted to the language patterns and expressions commonly found online. Therefore, compared to manual question decomposition by humans, LLMs are better at generating structured queries that align with the model's expectations, thereby improving the accuracy and efficiency of the parsing process.

By introducing a fine-grained CoT at the text level, we transform complex problems into a series of organized and specific queries. This not only simplifies the problem-solving process but also significantly enhances the accuracy and efficiency of MLLMs in understanding and answering questions.

Specifically, the step-by-step guidance enables the model to focus on key areas in stages, avoiding the processing of too much information at once, thereby improving processing speed and accuracy. Additionally, each sub-question directs the model's attention to specific image features, such as object categories, colors, etc., allowing the model to capture details in the image more precisely and reducing the likelihood of background interference and misjudgments.

By combining the fine-grained CoT at the text level with the three image prompting methods described above, we can better enhance the spatial understanding ability of MLLMs.

4 Experiment

In this experimental section, we first provide a detailed description of the experimental setup in Section 4.1. This section covers the selection of datasets, the choice of MLLMs, and the baseline standards established. In Section 4.2, we perform an in-depth performance evaluation of the MLLMs we selected. We further explore the performance of each model on different tasks. Additionally, we conducted ablation studies to analyze the contribution of each individual method to specific tasks and the potential interference it may cause. Finally, in Section 4.3, we present a detailed analysis of specific cases, demonstrating the powerful effectiveness of the proposed framework in practical applications.

4.1 Experimental Setup

Dataset. In this experiment, for the sake of fairness and rationality, we used the original text and images from the BLINK dataset(Fu et al., 2025), which includes 120 image-text pairs in the Counting task. The questions for the Counting task were selected from the TallyQA dataset(Acharya et al., 2019), known for its challenging human-written counting problems. Each sample consists of an image, a question, and a numerical answer. In addition to the correct answer, we randomly selected three other numbers as distractor options. The Depth task contains 124 image-text pairs, with human annotations from the Depth in the Wild dataset(Chen et al., 2016) used to organize our samples. Each question contains an image and two specified points. The task is to determine which point is closer to the observer.

MLLMs. We selected four MLLMs for the experiment: LLaVA-v1.5-7B(Liu et al., 2023), Qwen-VL-Chat(Bai et al., 2023), GLM-4V(Wang et al., 2023), and LLaVA-v1.6-vicuna-7B(Liu et al., 2024a).

Baselines. The accuracy of human recognition and random choice is referenced against the benchmark data published in the BLINK(Fu et al., 2025) paper. All experimental data were processed using locally deployed MLLMs, including the question and answer stages. When evaluating the accuracy of the model's responses, if the answer fails to provide specific and correct information, it is considered incorrect. This setting is designed to rigorously test the model's ability to understand and respond to questions, ensuring the reliability and validity of the experimental results.

Processing Method. For the image segmentation task, we chose the SoM algorithm(Yang et al., 2023) and removed the identifiers. For target detection, we applied YOLO v5(Jocher, 2020), which is known for its fast and accurate detection capabilities, enabling efficient identification and localization of multiple objects in an image. To generate attention heatmaps, we selected CLIP(Radford et al., 2021), a powerful multimodal model that establishes a deep connection between text and images, allowing it to generate high-quality heatmaps that highlight important regions of the image. The weight parameters λ_1,λ_2, and λ_3 are selected using the Monte Carlo method(Rubinstein and Kroese, 2016).For text decomposition, we used the LLM ChatGLM2-6B(GLM et al., 2024).

4.2 Main Results and Ablation Study

In this section, Counting refers to the counting task, and Depth refers to the relative depth estimation task. It is important to note that the Depth task is not suitable for using the YOLO method alone, as the bounding boxes in the image may occlude the markers. The best results are highlighted in bold. The numbers in the table represent accuracy, calculated as the correct count divided by the total, and are rounded to two decimal places.

As shown in Table 1 , for the Counting and Depth tasks, the accuracy of the MLLMs is significantly lower than human performance, indicating a clear deficiency in the models' ability to understand spatial relationships in images. Specifically, for the Counting task, the accuracy of the LLaVA-v1.5-7B and Qwen-VL-Chat is only about 40%, while the best-performing GLM-4V has an accuracy of only 67.5%. For the Depth task, except for the

Table 1: Main Experimental Results. The table presents the experimental results for single methods, as well as the effects of combining two different methods, including the results of our proposed approach. Additionally, we compare the accuracy of random selection versus human selection probabilities. The numbers in the table represent accuracy, calculated as the number of correct answers divided by the total number of answers, rounded to two decimal places. In the table, S stands for SoM, Y stands for YOLO v5, and A stands for attention.Baseline is the result obtained on a dataset without any operations.

	Counting(Acc%)				Depth(Acc%)			
Random Choice	25				50			
Human	93.75				99.19			
Open-source multimodal LLMs								
LLaVA -v1.5-7B	Baseline		40.83		Baseline		51.61	
	Y	43.33	Y+S	43.33	Y	-	Y+S	57.26
	S	46.67	S+A	44.17	S	45.97	S+A	56.45
	A	38.33	A+Y	39.17	A	49.19	A+Y	52.42
	Ours		**47.50**		Ours		**58.06**	
Qwen-VL-Chat	Baseline		42.50		Baseline		57.26	
	Y	44.17	Y+S	44.17	Y	-	Y+S	53.23
	S	44.17	S+A	47.50	S	54.84	S+A	58.06
	A	40.83	A+Y	45.00	A	49.19	A+Y	54.84
	Ours		**48.33**		Ours		**59.68**	
GLM-4V	Baseline		67.50		Baseline		65.32	
	Y	59.17	Y+S	61.67	Y	-	Y+S	62.10
	S	61.67	S+A	67.50	S	61.29	S+A	65.32
	A	56.67	A+Y	60.83	A	66.13	A+Y	60.48
	Ours		**68.33**		Ours		**70.97**	
LLaVA -v1.6-vicuna-7B	Baseline		49.17		Baseline		52.42	
	Y	**50.00**	Y+S	48.33	Y	51.61	Y+S	53.23
	S	42.50	S+A	46.67	S	51.61	S+A	53.23
	A	45.83	A+Y	47.50	A	**54.03**	A+Y	53.23
	Ours		**50.00**		Ours		**54.03**	

GLM-4V, the other three models have accuracies just above 50%, which is close to the probability of random selection. This suggests that during the initial training process, MLLMs may not have adequately focused on datasets and training strategies related to spatial understanding, and these models might lack fine-grained segmentation capabilities.

Further observation reveals that the performance of the LLaVA-v1.6-vicuna-7B on the Counting task improved by 8.34% compared to LLaVA-v1.5-7B, but the accuracy on the Depth task did not show significant improvement. This indicates that while certain improvements can enhance performance on specific tasks, the overall spatial understanding ability remains limited.

In addition, compared to the baseline methods, our proposed method significantly improved the accuracy of multimodal large models on both tasks, demonstrating its effectiveness. Specifically, on the Counting task, our method improved the accuracy by an average of 3.54% over the baseline, with an improvement of 6.67% for LLaVA-v1.5-7B. On the Depth task, our method improved accuracy by an average of 4.03% over the baseline, with the most notable improvement of 6.42% for LLaVA-v1.5-7B. These results not only demonstrate the effectiveness of our method but also suggest that targeted optimization can help mitigate the spatial understanding deficiencies in MLLMs.

Since no prior researchers have conducted similar experiments, we analyzed the main experimental results alongside the ablation experiments. We present the results of single methods and combinations of two methods. A notable finding in the experiments is that, in most cases, the combination of two methods actually reduced the accuracy of MLLM responses. This phenomenon suggests that each method has inherent limitations, and simply

combining two methods may lead to interference, affecting the model's judgment ability. This further proves that our proposed method effectively compensates for the shortcomings of various methods, helping MLLMs better understand spatial knowledge in images.

Moreover, we observed in the experiments that different MLLMs show varying sensitivity to different prompt methods when handling different types of tasks. For example, LLaVA-v1.5-7B is more sensitive to SoM-based prompts when facing the Counting task, with an accuracy improvement of 5.84%. However, when dealing with the Depth task, the SoM prompt method actually led to a performance decline. This phenomenon reveals that different tasks have significantly different requirements for prompt methods, and the selection of prompt strategies should be optimized according to the specific characteristics of each task.

4.3 Case Study

Case 1:

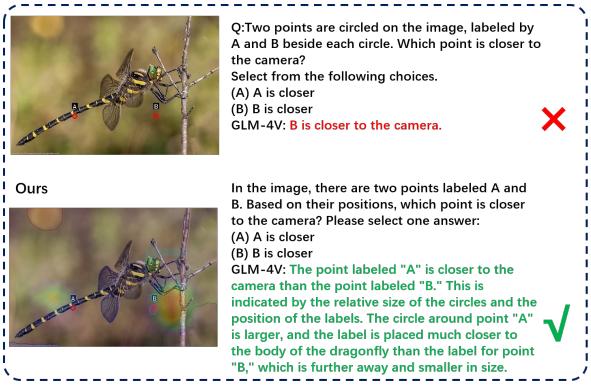

Figure 3: An example of the Depth task.

In Figure 3, when only the upper image and the original text are input into the GLM-4V, the model fails to generate an accurate answer. In contrast, for the lower image, after applying our prompt framework, GLM-4V not only provides the correct answer but also explains its reasoning process and the logic behind the conclusion. This case demonstrates that in the Depth task, our method can significantly enhance the spatial understanding ability of MLLMs.

Case 2:

In Figure 4, when only the top image and the original text were input into the LLaVA-v1.5-7B, the model failed to generate an accurate answer. However, when processing the lower image, image segmentation techniques were applied to divide the

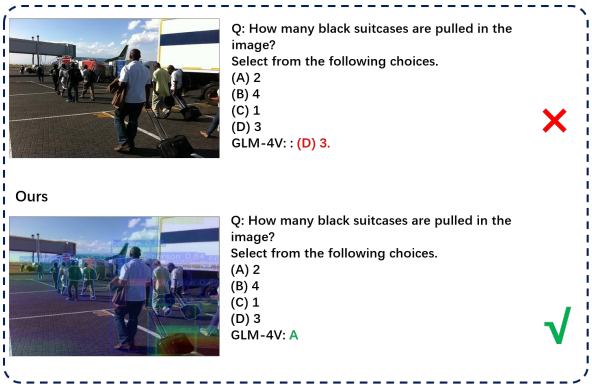

Figure 4: A counting task example.

image into multiple semantic regions, and the target detection algorithm accurately identified the specific locations of different objects. Additionally, with the help of the attention visualization mechanism, the model highlighted the attention region for key items (e.g., the suitcase), significantly enhancing its understanding and localization of important elements. As a result, the model was able to provide a more accurate answer. This case demonstrates that in the Counting task, our method significantly improves the spatial understanding ability of MLLMs.

5 Conclusion

This paper proposes a novel large model prompting framework aimed at enhancing the spatial understanding capabilities of MLLMs. The framework develops a comprehensive image processing approach that cleverly integrates target detection, image segmentation, and attention visualization mechanisms. This approach not only retains the inherent advantages of image segmentation technology but also effectively addresses the limitations of previous works by incorporating object detection and attention visualization. Furthermore, we innovatively introduce a fine-grained CoT decomposition strategy at the text level, which improves the accuracy and efficiency of the model in understanding and answering complex questions.

6 Limitations

In the LLaVA-v1.6-vicuna-7B, it was observed that the model is less sensitive to prompts compared to other MLLMs, and the exact reasons for this remain to be further investigated. Additionally, there are differences in sensitivity to weight distribution across different MLLMs, and the underlying causes of this phenomenon are also not yet clear.

References

Manoj Acharya, Kushal Kafle, and Christopher Kanan. 2019. Tallyqa: Answering complex counting questions. In *AAAI*.

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

Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. 2022. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*.

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*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406.

Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. 2016. Single-image depth perception in the wild. *Advances in neural information processing systems*, 29.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2025. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Peiyuan Jiang, Daji Ergu, Fangyao Liu, Ying Cai, and Bo Ma. 2022. A review of yolo algorithm developments. *Procedia computer science*, 199:1066–1073.

Glenn Jocher. 2020. Yolov5 by ultralytics.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota.

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. In *International conference on machine learning*, pages 19730–19742. PMLR.

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

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Reuven Y Rubinstein and Dirk P Kroese. 2016. *Simulation and the Monte Carlo method*. John Wiley & Sons.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. 2023. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11987–11997.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogvlm: Visual expert for pretrained language models. *Preprint*, arXiv:2311.03079.

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*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Appendix