Guiding Medical Vision-Language Models with Explicit Visual Prompts: Framework Design and Comprehensive Exploration of Prompt Variations

Anonymous ACL submission

Abstract

With the recent advancements in vision-language models (VLMs) driven by large language models (LLMs), many researchers have focused on models that comprised of an image encoder, an image-to-language projection layer, and a text decoder architectures, leading to the emergence of works like LLava-Med. However, these works primarily operate at the whole-image level, aligning general information from 2D medical images without attending to finer details. As a result, these models often provide irrelevant or non-clinically valuable information while missing critical details.

Medical vision-language tasks differs significantly from general images, particularly in its focus on fine-grained details, while excluding irrelevant content. General domain VLMs tend to prioritize global information due to their design, which compresses the entire image into a multi-token representation that is passed into the LLM decoder. Therefore, current VLMs all lack the capability to constrain their attention to particular areas.

To address this critical issue in the medical domain, we introduce MedVP, an visual prompt generation and fine-tuning framework, which involves extract medical entities, generate visual prompts, and adapt datasets for visual prompt guided fine-tuning. To the best of our knowledge, this is the first work to explicitly introduce visual prompt into medical VLMs, and we successfully outperform recent state-of-the-art large models across multiple medical VQA datasets. Extensive experiments are conducted to analyze the impact of different visual prompt forms and how they contribute to performance improvement. The results demonstrate both the effectiveness and clinical significance of our approach.

1 Introduction

As human beings, we inherently rely on visual cues or prompts to understand complex visual content in greater detail. These visual prompts may take various forms, yet they serve a fundamentally similar purpose: to direct our attention toward critical areas that contain rich information or are of particular interest.

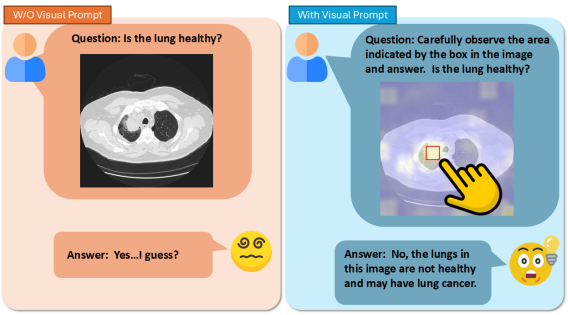

Figure 1: The figure illustrates the difference in the reasoning process of the Vision-Language Model (VLM) with and without visual prompts. By emulating the way humans approach vision-language tasks, we use an auxiliary visual prompt, akin to pointing at a specific region, to help the model more easily focus on the relevant details and generate accurate responses.

Current mainstream pre-trained vision-language models often lack mechanisms to direct their attention to specific areas of interest. Instead, their attention is typically distributed according to priors learned during the pre-training stage. However, these pre-trained attention patterns are not universally applicable, particularly in specialized domains such as biomedical images. Medical images often exhibit large regions that appear visually similar, with only small, detailed areas—such as regions of abnormality, small nodules, or lesions—providing critical information. Therefore, inspired by the human recognition process, we propose a framework to adjust the prior attention distribution of vision-language models to a posterior distribution given the language context by embedding explicit visual prompts within the images. To comprehensively explain our approach, it is important to first provide a brief review of the current de-

sign and mechanisms underlying vision-language models.

Vision-Language Models (VLMs) are a class of models designed to process both image and text inputs for cross-modal tasks. Depending on the specific task, the architecture and pre-training strategies of these models can vary significantly. For the purposes of this discussion, we will focus specifically on Large Vision-Language Models (LVLMs), a subset of VLMs that are particularly designed for image-to-language tasks, such as Visual Question Answering (VQA) and Image Captioning. In this line of research, foundation models such as BLIP-2 (Li et al., 2023) and LLaVA (Liu et al., 2024) have significantly influenced the design of current mainstream LVLMs. Taking LLaVA as an example, these models are typically composed of an image encoder, a projection layer, and a Large Language Model (LLM)-driven decoder. The image inputs are first encoded into image embeddings using pre-trained models like ViT (Dosovitskiy, 2020) or CLIP (Radford et al., 2021), which represent the image in a latent space. The projection layer then maps these image embeddings into the language representation space, and concated the M length visual embedding vectors with the language embedding vectors to be processed by the a transformer-based decoder. Due to the Multi-Head Self-Attention(MHSA) mechanism in transformer blocks, image embeddings will be attended by other language tokens to capture the image-level information. In the majority of Vision-Language Models (VLMs) (Li et al., 2023; Liu et al., 2024), the pre-training task involves generating descriptions of image content in an evenly distributed manner. This approach often results in the attention of language tokens being uniformly allocated across all visual embedding tokens, without any explicit constraints or guidance. Although this training method facilitates the VLM's understanding of overall visual information, it prevents the model from paying special attention to critical details, and it often leads to the inclusion of irrelevant information. As previously mentioned, such an attention mechanism is not conducive to solving medical image understanding tasks, as most medical images maintain consistency in anatomical structures (e.g., all individuals have the same organs). However, it is the details, such as tumors, that distinguish one medical image from another. To mitigate this problem, recent research (Yang et al., 2023; Rasheed et al., 2024; Cai et al., 2024; Zhang et al., 2024; Kirillov et al., 2023) has increasingly recognized the value of visual prompts and their application in vision-language tasks. Visual prompts can be categorized as either implicit or explicit in relation to the visual inputs. In the case of implicit visual prompts, no direct markers are applied to the images; rather, they are introduced at the level of visual features within hidden layers. In this study, we will focus on explicit visual prompts, which are directly added to the input images, allowing for more direct interaction with the visual model.

Inspired by ViP-LLaVA (Cai et al., 2024) and SAM (Kirillov et al., 2023), we decide to add visual markers such as bounding box, scribbles, and circles into the input images to highlight the Region of Interest (ROI) for the visual-prompt guided VLMs to focus on important features. However, current works(Zhang et al., 2024; Cai et al., 2024; Kirillov et al., 2023) usually assume the visual markers are provided interactively by human annotators. Given the high cost of acquiring human annotations in medical images, we propose a novel method, MedVP(**Med**ical **V**isual **P**rompting) that will automatically generate visual prompts for the interested region given the context. Specifically, our framework first extracts entities or relevant keywords from the queries, as our focus is on medical Visual Question Answering (VQA) tasks. Next, we fine-tune an open-set grounding model that generates coordinates for the target area based on the identified medical entities. Finally, these coordinates are used to generate visual prompts in various formats—such as scribbles, bounding boxes, or circles—on the input images. Along with the visual-prompted images, we adapt the text Question-Answer pairs from SLAKE, VQA-RAD, and PMC-VQA datasets to ensure that the language decoder becomes aware of the presence of our visual prompt markers within the images. For example, we incorporate text such as "Carefully observe the area in the red box..." to guide the LLM-driven decoder to focus on the visually prompted area, ensuring that the model pays attention to the regions highlighted by the visual prompts. The visual-prompted images are fine-tuned with Vision-Language Models (VLMs) to develop our model, **MedVP-LLaVA**—the first medical VLM guided by visual prompts. We find that our **MedVP-LLaVA** performs effectively in medical VQA tasks, significantly enhancing performance across multiple datasets.

As conclusion, our MedVP method has at least

following contribution:

1. We are the first to introduce explicit visual prompts into medical-specialized Vision-Language Models, and we validate the effectiveness of visual prompts in enhancing performance for medical VQA tasks.

2. We design a whole framework from extracting keywords from VQA queries, to visual grounding given medicle entities, and, generate the visual prompts to the images, and adapted three VQA datasets to include the information of our visual prompts.

3. We then fine-tune the **MedVP-LLaVA**, a visual-prompt-aware Vision-Language Model (VLM) tailored for the medical imaging domain, and validate its superiority across multiple medical VQA datasets.

4. We will release the model weights for both **MedVP-LLaVA** and the medical entity grounding model, which has been trained on a large dataset. Additionally, we will make our modified VQA datasets publicly available to accelerate research in the development of medical VLMs.

2 Preliminaries

2.1 Medical Vision Language Model

Medical Large Vision-Language Models (Med-VLMs) combine large language models (LLMs) with medical-specific visual modules, allowing them to process medical images alongside clinical queries. When presented with a medical image, the vision encoder extracts visual features, which are then transformed by an adapter module into a format interpretable by the LLM. Utilizing this multimodal input, the model predicts the next token in the sequence autoregressively. A critical capability of these models is Visual Question Answering (VQA), where the model is tasked with answering questions accurately based on the content of the given image.

2.2 Region-Level Image Understanding

In most VLMs, the input typically consists of both an image and text, with the model generating responses by integrating the global information from the image and the accompanying text input. However, a common issue arises when the model fails to focus adequately on local regions and fine details, which often leads to incorrect answers. In the domain of natural images, there are three primary approaches to enhancing large models' attention to local image information, one is incorporating learnable soft tokens into visual inputs for parameter-efficient finetuning (Bahng et al., 2022; Khattak et al., 2023). The second approach is concatenating an image sequence to demonstrate new tasks (Bai et al., 2024; Bar et al., 2022), or using Region of Interest (ROI) features [42] to align language with specific image regions, and (iii) grounding regions by overlaying visual markers (such as masks, boxes, or circles) onto the visual inputs (Yao et al., 2024; Zellers et al., 2021). In medical imaging, the focus on local regions is particularly crucial, as accurate diagnosis and assessment of abnormalities typically require a combination of both local and global information. However, research on incorporating local region attention in Med-VLMs is still relatively underexplored.

3 Method

In this section, we outline the proposed method, detailing the training pipeline and objectives for Med-VLMs, including the visual prompt extraction process and fine-tuning on downstream medical visual question answering tasks.

3.1 Region Level Medical Knowledge Alignment

Med-VLMs primarily perform alignment training at the image level, as seen in LLaVA-Med (Li et al., 2024) using visual instruction tuning (Liu et al., 2024) to align medical context and vision knowledge. However, in medical imaging, diagnostic reasoning often hinges on specific details or localized regions, such as lesions or organs. Failing to adequately focus on these critical areas can lead to incorrect model predictions. Therefore, our training approach ensures that while the model captures global information, it also strengthens the model's focus and understanding of key regions. Specifically, during instruction tuning for Med-VLMs, we incorporate annotations of local regions in the images, and the corresponding descriptions contain text that explicitly references these annotated areas. Therefore, we call these annotations visual prompts for the VLMs, since they are prompting the models to attend on specific areas. Each training sample thus consists of three components: the medical

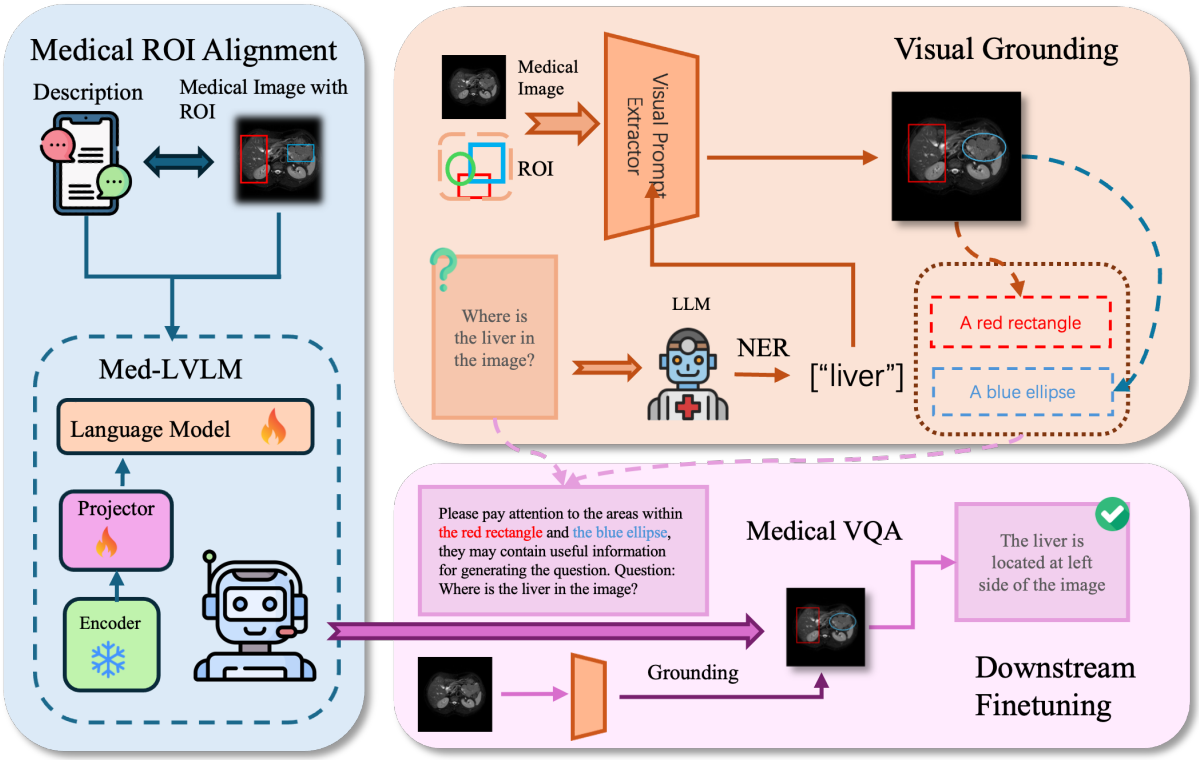

Figure 2: The framework of MedVP-LLaVA

image ($\mathbf{I}$), the coordinates of the visual prompts referring to the region of interest within the image ($\mathbf{P}$), and the associated text description ($\mathbf{T}$). The description includes both whole-image-level information and specific details about the region of interest. Following the setting in ViP-LLaVA (Cai et al., 2024), for the vision part, we merge the image $\mathbf{I}$ and its visual prompts using alpha blending, generating a composite representation that emphasizes the relevant local areas.

$$\hat{\mathbf{I}} = \alpha \cdot \mathbf{P} + (1 - \alpha) \cdot \mathbf{I}, \tag{1}$$

where $\alpha \in [0, 1]$ denotes the transparency level of the visual prompt.

Using the blended image with the highlighted regions and the accompanying detailed text description, we perform autoregressive language modeling to maximize the likelihood of generating the tokens of the ground-truth answer T_x:

$$P(\mathbf{T}_x \mid \hat{\mathbf{I}}, \mathbf{T}_{\text{instruct}}) = \prod_{i=1}^{L} P_\theta(x_i \mid \hat{\mathbf{I}}, \mathbf{T}_{\text{instruct}}, \mathbf{T}_{x,<i}) \tag{2}$$

where θ represents the trainable parameters, $X_{instruct}$ is the text instruction for generating the target description, L is the sequence length of the answer T_x, and $T_{x,<i}$ denotes all the answer tokens before the current prediction token x_i, where i denotes the steps during text token generation.

3.2 Entity Recognition

Since the Med-VLM requires regions of interest (ROI) to establish region-level cognition and understanding, it is crucial to obtain effective visual prompts for medical images. A key challenge we face is how to generate accurate region-level annotations for medical images. We adopted a two-step approach: first, extracting region-level entities, and then feeding these entities into a visual prompt extractor to generate the visual prompt. In the medical visual question answering task, we start by analyzing the question to identify potential region-level entities that could aid in answering it. To generate these entities, we employ an LLM, which offers greater flexibility and can be easily guided with prompts to produce helpful entities compared to other entity recognition models. For each question Q in the visual question answering task, we prompt the LLM to identify a set of entities:

$$\mathcal{E} = \text{LLM}(Q, T_{\text{prompt}}) \tag{3}$$

where $\mathcal{E}$ represents the set of entities the LLM captures from question Q, and T_{prompt} represents the prompt for entity extraction. The extracted entities

may include specific organs or diseases mentioned in Q, or more general terms for potential relevant organs and diseases. They are the preliminary for genearting visual prompts using the visual prompt extractor.

3.3 Visual Prompt Position Extraction

In the previous step, we identified region-level entities that could potentially assist in answering the question. The next step involves using a visual prompt extractor to obtain the visual prompt coordinates, utilizing these entities along with the corresponding image. Specifically, we employed the Grounding DINO model (Liu et al., 2023), which is a detection model that supports (image,text) input for open-vocabulary detection of entities described by the provided text within the image. Grounding DINO has shown strong performance in open-vocabulary detection on natural images. However, due to the extensive use of specialized terminology for organs and lesions in the medical field, the existing Grounding DINO model struggles to accurately detect region-level entities in medical images domain. To address this, we first fine-tune Grounding DINO for the medical imaging domain.

Using a medical dataset consisting of (image,text,coordinates), we adapt Grounding DINO to the specific requirements of medical image detection. Fine-tuning involves applying contrastive loss (Radford et al., 2021) between predicted objects and language tokens for classification, alongside L1 loss and GIoU loss (Rezatofighi et al., 2019) for bounding box regression.

Following fine-tuning, we obtain a Grounding DINO model capable of accurately detecting medical entities. By inputting the region-level entity names into the model, we extract the visual prompt coordinates corresponding to the relevant regions in the medical images:

$$\mathbf{P} = \text{DINO}(\mathbf{I}, \mathcal{E}) \tag{4}$$

3.4 Visual Prompt Generation

We incorporate the visual prompts detected by Grounding DINO into medical images through the alpha blending technique outlined in Equation 1. Since manual annotation of medical images is a common practice in clinical settings, it is essential for our model to be capable of recognizing and interpreting the diverse types of visual prompts frequently employed in these contexts. To address this, we introduce a set of clinically prevalent visual prompts (`scribble`, `rectangle`, `ellipse`) during training to enhance the model's capability in recognizing different annotation types. For each set of coordinates, one of these shapes is randomly selected and applied through alpha blending, resulting in images that contain diverse forms of visual prompt annotations.

By integrating these region-specific visual prompts into the image, we steer the model's attention towards the relevant areas indicated in the question Q. To facilitate this, we extract two key attributes (`color`, `category`) of the visual prompt to describe the visual annotation and guide the model's focus to the appropriate region.

4 Experiment

4.1 Experiment Setup

Base Model. We utilize ViP-LLaVA 7B (Cai et al., 2024) as our base model. ViP-LLaVA (Cai et al., 2024) was initially pre-trained on natural image datasets and allows for user interaction by incorporating visual prompts into the images, enhancing the model's ability to interpret and respond to inputs. We use it as our base model becasue it has a better capability of identifying region in marked regions

Knowledge Pre-training of Med-VLM. To inject medical knowledge into our model, we pre-train ViP-LLaVA 7B using a subset of the MedTrinity-20M dataset (Xie et al., 2024), which consists of triplets in the format image, ROI, description. Each ROI (region of interest) corresponds to an abnormality annotated with a bounding box. The descriptions provide a multi-granular textual explanation, covering disease or lesion type, imaging modality, region-specific details, and inter-regional relationships. This pre-training on medical images equips ViP-LLaVA with domain-specific knowledge, enabling the model to focus on region-specific information in medical contexts.

Finetuning of Visual Prompts Extractor. We use Grounding DINO (Liu et al., 2023) as our visual prompt extractor, which takes text input and identifies the corresponding regions in images. To adapt Grounding DINO for annotation tasks in the medical domain, we fine-tune it using the combination of Slake training dataset and SAMed-2D dataset (Ye et al., 2023). The SLAKE (Liu et al., 2021) dataset has region level annotations of different organs and disease on the x-ray image. The SAMed-2D dataset (Ye et al., 2023) includes over

Table 1: Performance of MedVP-LLaVA on three medical visual question answering datasets. Following the routing in LLaVA-Med (Li et al., 2024), for open-set questions, we report the recall value in column Open. For closed-set questions, we report the accuracy in column Closed. The best results are **bold**. The asterisk (*) indicates results from our own tests.

Models	VQA-RAD		SLAKE		PMC-VQA
	Open	Closed	Open	Closed	Closed
Zero Shot					
BLIP-2 (Li et al., 2023)	-	-	-	-	24.3
Open-Flamingo (Awadalla et al., 2023)	15.4	61.4	13.8	29.5	26.4
LLaVA-Med (Li et al., 2024)	28.2	61.4	39.2	52.2	32.9
Finetuned on the Respective Dataset					
PMC-CLIP (Lin et al., 2023)	52.0	75.4	72.7	80.0	37.1
MedVInT-TE (Zhang et al., 2023)	69.3	84.2	88.2	87.7	39.2
MedVInT-TD (Zhang et al., 2023)	73.7	86.8	84.5	86.3	40.3
LLaVA-Med++ (Xie et al., 2024)	77.1	86.0	86.2	89.3	30.0*
MedVP-LLaVA (ours)	**89.3**	**97.3**	**91.6**	**92.9**	**40.6**

20,000 images from various modalities, such as MRI, CT, ultrasound, PET, X-ray, fundus, and endoscopy. The annotated masks are converted into bounding box coordinates, covering a wide range of organs and disease types. Here is a detailed data distribution of the selected finetuning dataset for Grounding DINO.

Choice of Visual Prompts. We utilize three types of visual prompts for image annotation: scribble, rectangle, and ellipse, chosen for their frequent application in medical imaging. During fine-tuning, different visual prompt shapes are randomly applied to the images, allowing the model to learn how to interpret and respond to a variety of prompt shapes, thereby improving its robustness and flexibility. The examples of different types of visual prompts are shown in appendix.

4.2 Results

In this section, we assess the performance of MedVP-LLaVA across three distinct medical VQA datasets, comparing our approach against several state-of-the-art methods in the field.

4.2.1 Evaluation of the Benchmarks

We evaluated our method on three medical VQA datasets: SLAKE (Liu et al., 2021), VQA-RAD (Lau et al., 2018), and PMC-VQA (Zhang et al., 2023). Both SLAKE and VQA-RAD contain two types of questions: Open and Closed. PMC-VQA is a multi-modal, multiple-choice VQA dataset. As shown in Table 1, our approach consistently outperforms the current state-of-the-art methods. On the SLAKE and VQA-RAD datasets, our method achieves significant improvements. On the VQA-RAD dataset, our method demonstrates an improvement of over 10% on both the open and closed metrics. On the SLAKE dataset, our approach also achieves a significant average improvement of approximately 4% compared to the best-performing models. The fine-tuned MedVP-LLaVA shows particularly strong gains on the closed-set questions in VQA-RAD dataset. For the PMC-VQA dataset, where each question requires selecting the correct answer from four options, our method also achieves the best performance, surpassing a range of advanced models in the medical VQA setting.

4.3 Analysis

In this section, we primarily analyze the role of visual prompts in the medical VQA task and evaluate the performance of MedVP-LLaVA under different visual prompt shape settings. We also analyze why the inclusion of visual prompts leads to performance improvements in the medical VQA task. We will also examine the clinical value of our approach, discuss its applicability, and provide examples of specific cases.

4.3.1 Ablation Studies on Visual Prompts

During the fine-tuning phase, the visual prompt plays a crucial role in improving the performance on Medical VQA tasks. As shown in Table 2, incorporating visual prompts during fine-tuning significantly enhances the model's test results compared to the absence of visual prompts. This highlights the importance of Med-VLM's ability to observe and analyze local information in medical VQA tasks.

Figure 3: Visualization results of our cross-attention map. The red boxes are the visual prompts generated by our grounding model. As illustrated, the yellow area indicates high attention values and is largely overlapped with the visual prompted area.

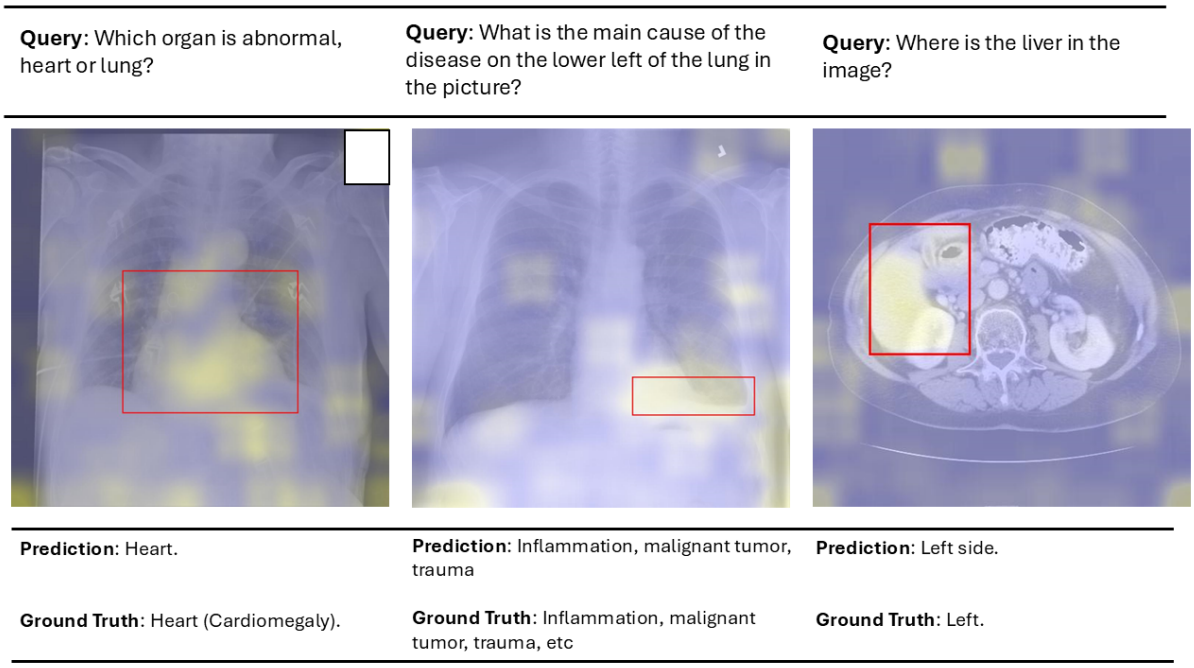

Table 2: Ablation study on the impact of using visual prompts during fine-tuning on downstream medical VQA datasets. 'VP' in the first column refers to visual prompts. Both the open and close set accuracy are reported.

MedVP-LLaVA	VQA-RAD		SLAKE		PMC-VQA
	Open	Closed	Open	Closed	Closed
w/o VP	79.81	92.92	72.56	87.23	28.09
w VP	**80.41**	**97.27**	**79.22**	**92.88**	**40.60**

Table 3: Performance of different visual prompt categories on three medical VQA datasets during inference. The 'Mix' row represents the use of a combination of the three visual prompt types. We report both the accuracy for the open set and close set questions.

Category	VQA-RAD		SLAKE		PMC-VQA
	Open	Closed	Open	Closed	Closed
Scribble	79.38	96.10	80.62	91.82	40.11
Rectangle	81.90	96.40	78.29	91.35	40.43
Ellipse	**82.47**	96.80	79.06	92.54	40.47
Mix	80.41	**97.27**	**79.22**	**92.88**	**40.60**

4.3.2 Performance on Various Visual Prompts

We compared the performance of our model on medical VQA under different visual prompt categories, as shown in Table 3. The results indicate that the ellipse visual prompt slightly outperforms the Rectangle prompt, which in turn performs better than the scribble prompt. This outcome can be attributed to the varying visual saliency and coverage of different prompts—rectangle and ellipse prompts generally cover a broader area and are more visually prominent compared to the scribble prompt. As observed in the table, using a mix of all three prompt types enhances overall performance. Incorporating various prompts during both training and testing improves the model's robustness. As a result, even when tested with a single type of visual prompt, the model's performance does not exhibit significant degradation, remaining relatively consistent across different prompt types.

4.3.3 Analysis of Model's Reliance on the Visual Prompt.

In this section, we design an experiment to evaluate the impact of reducing the annotation level of Grounding DINO during inference, specifically analyzing the effect of different visual prompt ratios when the annotation quantity is decreased to 80%, 60%, 40%, and 20% of the original dataset. As shown in table 4, reducing the labeling ratio

results in a noticeable performance degradation compared to full annotation. For instance, in the VQA-RAD dataset, when the labeling ratio is reduced to 20%, Open accuracy drops by approximately 1.5%, while Close accuracy declines by about 0.5%. This highlights the model's dependence on visual prompts to assist in answering questions. Moreover, we observe that as the visual prompt labeling ratio decreases, the model's performance during inference remains relatively stable. This suggests that the use of visual prompts during training has significantly enhanced the model's decision-making capabilities. Visual prompts appear to guide the model in attending to the relevant areas of the image, allowing it to learn more effectively which details and regions to prioritize. As a result, even when visual prompts are limited during inference, the model is still able to generate accurate responses by focusing on the key areas.

Table 4: Performance comparison across three medical VQA datasets with varying visual prompt annotation ratios in the test set images. We report both the accuracy for the open set and close set questions.

Ratio	SLAKE		VQA-RAD		PMC-VQA
	Open	Close	Open	Close	Ratio
20%	78.22	89.45	78.86	96.88	40.47
40%	78.37	90.35	78.35	96.88	40.40
60%	78.69	89.84	78.86	97.27	40.36
80%	78.69	91.73	78.35	96.10	40.48
100%	**79.22**	**92.88**	**80.41**	**97.27**	**40.60**

4.3.4 Visualization of Model's Attention with Visual Prompts

To further investigate whether the model has correctly established region-level attention and understanding, we visualized MedVP-LLaVA's attention on the medical images, as shown in Figure 5. Following the setting in this work (Yu et al., 2024), we visualize the cross-attention map between the generated tokens and the visual embedding tokens. The regions colored yellow shows the region that affect more on the model's response. In the first example, the heatmap shows that our model accurately focuses on the region of interest. Additionally, in the rightmost figure, we can see that the model correctly attends to the liver. This attention map illustrates the regions of the image that the model focuses on while generating responses to the queries. As illustrated in the figure, we can conclude that after our visual-prompt-guiding fine-tuning, our model can constraint its attention to our visual prompts area.

4.3.5 Clinical Significance of Applying Visual Prompts

Our approach incorporates visual prompts to assist in the medical VQA task, which not only enhances the model's performance but also holds clinical significance for medical image analysis. During training, visual prompts guide the model to focus on the most pertinent regions, improving both the efficiency and effectiveness of the learning process. In the inference stage, visual prompts provide valuable reference points for the model when answering questions, thereby improving the interpretability of its responses and increasing its safety. These factors are critical in the context of medical imaging applications.

5 Limitations

Also, during knowledge pre-training stage, the dataset we utilize contain AI-generated Question-Answering contents. These contents might not be factually accurate, considering the LLMs are suffering Hallucination problems for challenging tasks, especially in medical domain. However, considering the high cost of acquiring human annotated data for medical images, we have to choose such data to augment our training corpus. In our future work, we will dedicate to more rigorous data cleaning methods and implement RAG algorithms to our model to avoid factual errors.

6 Conclusion

In this work, we propose the MedVP method, which introduces explicit visual prompts into medical images to guide Med-VLMs in focusing on specific Regions of Interest (ROIs) for medical VQA tasks. Our approach includes a medical visual prompt extraction process. First, we use LLM to identify medical entities from the questions. Next, we trains a grounding model to locate the ROIs in the images based on the extracted entity names. Finally, we blend the visual prompts with the medical images. Our MedVP-LLaVA method significantly outperforms recent state-of-the-art large models across three medical VQA datasets by a substantial margin, demonstrating the effectiveness of utilizing visual prompts to guide medical VLMs in the medical VQA task.

References

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

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

Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. 2024. Sequential modeling enables scalable learning for large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22861–22872.

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. 2022. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35:25005–25017.

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

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

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. *Preprint*, arXiv:2304.02643.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *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. In *International conference on machine learning*, pages 19730–19742. PMLR.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE.

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

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

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.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji Mullappilly, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S. Khan. 2024. Glamm: Pixel grounding large multimodal model. *Preprint*, arXiv:2311.03356.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666.

Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, et al. 2024. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*.

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. *Preprint*, arXiv:2310.11441.

Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2024. Cpt: Colorful prompt tuning for pre-trained vision-language models. *AI Open*, 5:30–38.

Jin Ye, Junlong Cheng, Jianpin Chen, Zhongying Deng, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan

Huang, Jilong Chen, Lei Jiang, et al. 2023. Samed2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks. *arXiv preprint arXiv:2311.11969*.

Runpeng Yu, Weihao Yu, and Xinchao Wang. 2024. Attention prompting on image for large vision-language models. *Preprint*, arXiv:2409.17143.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. *Advances in neural information processing systems*, 34:23634–23651.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. 2024. Gpt4roi: Instruction tuning large language model on region-of-interest. *Preprint*, arXiv:2307.03601.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

A Data

A.1 Data Statistics

The quantities of all the datasets have been shown in Table 5. During knowledge pre-training stage, we extract a subset of Med-Trinity datasets to fintune our model. This subset include images from SLAKE, VQA-RAD, and PATH-VQA.

During finetuning, we only use the training set data of SLAKE, VQA-RAD, and PMC-VQA datasets.

A.2 Involved Datasets

We leveraged three publicly available medical VQA datasets: VQA-RAD, SLAKE, and PMC-VQA.

- **SLAKE**: SLAKE is a bilingual dataset in English and Chinese, comprising 642 images and 14,028 question-answer pairs, designed for training and evaluating Med-VQA systems.
- **VQA-RAD**: A radiology-focused VQA dataset that includes radiological images and corresponding questions, aimed at assessing the model's performance in answering domain-specific questions about various types of radiological findings.
- **PMC-VQA**: A large-scale dataset constructed from PubMed Central (PMC) articles, focusing on medical visual question answering with figures and charts extracted from research papers, testing models on understanding complex visual and textual medical data.

B Visualization of Visual Prompts

We present the visualization results of visual prompts integrated into the images, demonstrating three types of visual prompts with varying sizes and transparency levels.

C Implementation Details

All experiments were conducted on RTX 4090 and H100 GPUs. For region-level medical domain alignment, we used a learning rate of 2e-5 and a batch size of 64. When detecting bounding boxes with Grounding DINO, we set the prediction score threshold to 0.2 and the batch size to 24.

For downstream fine-tuning, we applied a learning rate of 2e-5, a batch size of 128, and a warm-up ratio of 0.03. The fine-tuning process on 4 RTX 4090 GPUs took approximately 6 hours for SLAKE, 2 hours for VQA-RAD, and 20 hours for PMC-VQA. Grounding DINO fine-tuning, also on 4 RTX 4090 GPUs, took around 9 hours.

D Instructions and Prompts

We show the prompt used for the entity recognition task and the instruction to the Med-VLM when doing downstream finetuning and inference in Figure 5

Table 5: Data statics for various datasets in different training stage

Dataset	SLAKE		VQA-RAD		PMC-VQA		PATH-VQA	
	QA-pair	Image	QA-pair	Image	QA-pair	Image	QA-pair	Image
Pre-train(Med-Trinity)	642	642	1754	1754	✗	✗	13371	13371
Fine-tuning	9834	642	1793	313	91.35	✗	✗	
Test	2094	642	451	203	33430	164360	✗	✗

Figure 4: Visualization of the visual prompts blended into medical images.

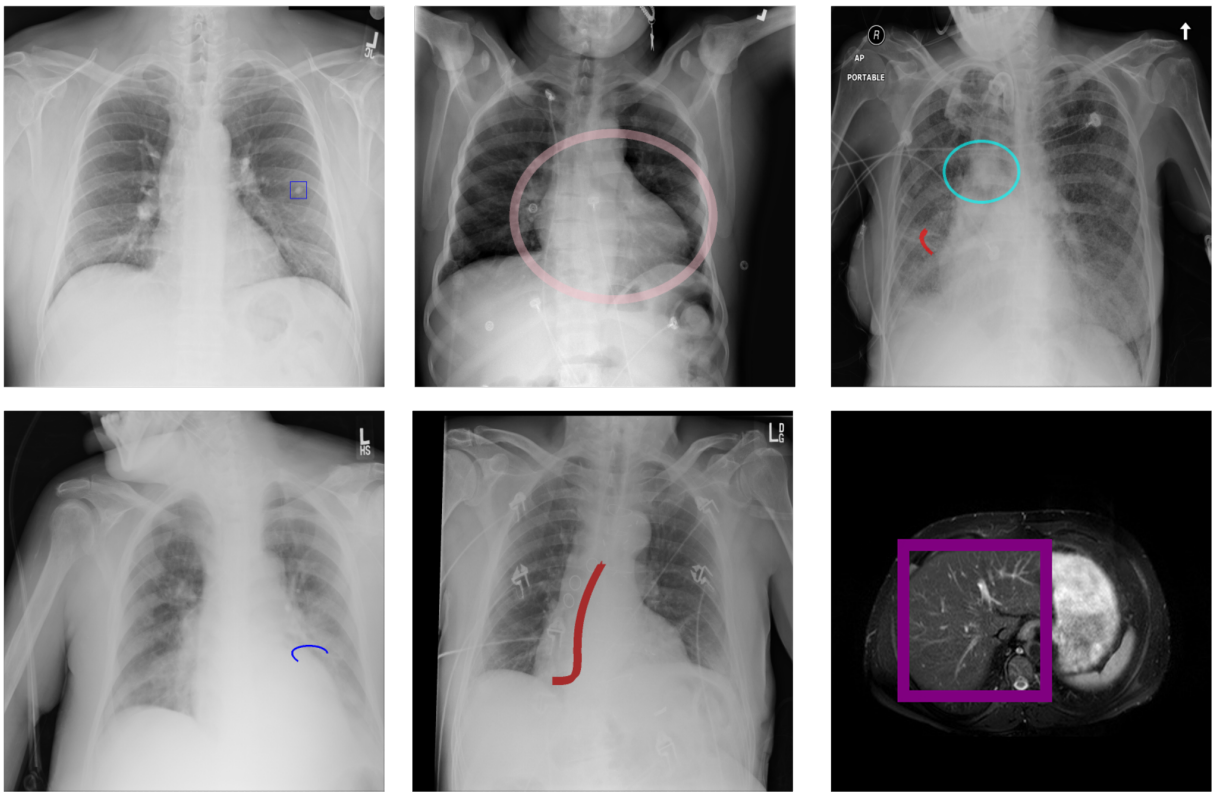

Figure 5: Prompts and Instructions for different procedure.

Prompts for the LLM in the entity recognition task:

```
You are a medical entity identifier. Please help me find the
medical entity that could help visual question answering if the
entity is cropped by bounding box in the corresponding image.
Basically you need to identify the organ and disease.If any
specific organ is mentioned in the question, include the unmodified
noun in your response. Otherwise, add exact 'organ' in your
response if that labeling can be helpful. If you think labeling a
disease would help answering the question, add exact 'disease' in
your response. Here are some example sentences and answers:
Sentence: Does the kidneys look abnormal? Answer:
['kidney','disease']. Sentence: Is there swelling of the grey
matter? Answer: ['disease']. Sentence: What is the largest organ in
the picture? Answer: ['organ']. Sentence: Which is the biggest in
this image,lung,liver or heart? Answer:['lung','liver','heart'].
Sentence: Which organs appear in pairs? Answer: ['organ'].
Following the example, answer the question: Sentence: [SENTENCE]
```

Prompt for inference and finetuning prompt with visual prompts:

```
Please pay attention to the areas in <color> <category>. It/They
may contain useful information for answering the question. Question:
<image>\n<QUESTION>
```