Fine-grained Adaptive Visual Prompt for Generative Medical Visual Question Answering

Ting Yu[1], **Zixuan Tong**[1], **Jun Yu**[2], **Ke Zhang**[3,*]

[1]School of Information Science and Technology, Hangzhou Normal University, Hangzhou, China
[2]School of Intelligence Science and Engineering, Harbin Institute of Technology (Shenzhen), China
[3]Key Laboratory of Complex Systems Modeling and Simulation, Hangzhou Dianzi University, Hangzhou, China
yut@hznu.edu.cn, tongzixuan@stu.hznu.edu.cn, yujun@hit.edu.cn, ke.zhang@hdu.edu.cn

Abstract

Medical Visual Question Answering (MedVQA) serves as an automated medical assistant, capable of answering patient queries and aiding physician diagnoses based on medical images and questions. Recent advancements have shown that incorporating Large Language Models (LLMs) into MedVQA tasks significantly enhances the capability for answer generation. However, for tasks requiring fine-grained organ-level precise localization, relying solely on language prompts struggles to accurately locate relevant regions within medical images due to substantial background noise. To address this challenge, we explore the use of visual prompts in MedVQA tasks for the first time and propose fine-grained adaptive visual prompts to enhance generative MedVQA. Specifically, we introduce an Adaptive Visual Prompt Creator that adaptively generates region-level visual prompts based on image characteristics of various organs, providing fine-grained references for LLMs during answer retrieval and generation from the medical domain, thereby improving the model's precise cross-modal localization capabilities on original images. Furthermore, we incorporate a Hierarchical Answer Generator with Parameter-Efficient Fine-Tuning (PEFT) techniques, significantly enhancing the model's understanding of spatial and contextual information with minimal parameter increase, promoting the alignment of representation learning with the medical space. Extensive experiments on VQA-RAD, SLAKE, and DME datasets validate the effectiveness of our proposed method, demonstrating its potential in generative MedVQA.

Code — https://github.com/OpenMICG/FAVP

Introduction

The convergence of breakthroughs in computer vision (CV) and natural language processing (NLP) has catalyzed significant interest in multimodal tasks, such as visual captioning (Zhang et al. 2023a,b, 2024; Yu et al. 2024d), visual grounding (Yuan et al. 2024), and visual question answering (VQA) (Yu et al. 2019, 2020, 2024a,b). Among these, Medical Visual Question Answering (MedVQA) (Yu et al. 2024c) stands out as a crucial extension in the medical domain. MedVQA necessitates a deep understanding of medical images at various granular levels and establishing reliable cross-modal associations, which are essential for assisting physicians in diagnosis, preventing misdiagnoses, and enhancing patient care efficiency and experience. Consequently, MedVQA has emerged as a prominent focus in computer-aided diagnosis. In addition to traditional discriminative MedVQA models, the success of large language models (LLMs) in general domains has inspired their adaptation to the medical field (Liu et al. 2024; Wu et al. 2023; Li et al. 2024). These models leverage their robust generative capabilities to address MedVQA tasks better. Despite medical LLMs' strong zero-shot transfer abilities, there remain two challenges. (i) For fine-grained, instance-level precise localization required by MedVQA, language prompts alone in medical LLMs struggle to accurately identify relevant information amid substantial background noise, as shown in Figure 1. (ii) Typically, medical LLMs are featured by large parameter sizes and extensive training times. Full parameter fine-tuning across diverse downstream tasks is parameter-inefficient and compromises model generalization.

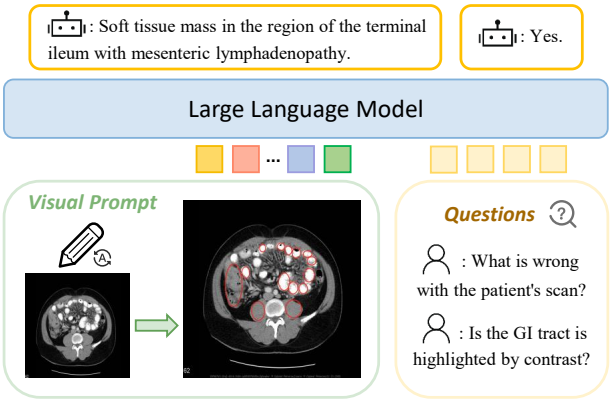


Figure 1: **Main idea of FAVP.** We introduce eight types of visual prompts (e.g., circle, contour, mask, box) for accurate localization assistance. Here we show an example of using the circle.

To address these challenges, we treat MedVQA as a generative task and introduce Fine-grained Adaptive Visual

*The corresponding author.

Prompts (FAVP) to enhance generative MedVQA. Our proposed framework integrates an Adaptive Visual Prompt Creator, which adaptively generates region-level visual prompts based on the characteristics of different organs within medical images. This mechanism provides fine-grained visual references for LLMs, thereby improving the accuracy of cross-modal localization when generating answers from the medical domain. To our knowledge, we are the first to explore the use of visual prompts in MedVQA, demonstrating its significant effectiveness in open-set questions. We further explore different types of instance-level visual prompts, and ultimately select the most effective visual prompt based on the characteristics of different datasets. Furthermore, our model incorporates a Hierarchical Answer Generator, designed to extract hierarchical high-semantic representations from fine-grained visual features and map these representations into the language space to generate medical answers. Additionally, we integrate parameter-efficient fine-tuning (PEFT) techniques, which significantly enhance the model's spatial and contextual comprehension with minimal parameter augmentation, thereby promoting the representation learning process within the medical domain.

Overall, the main contributions of this paper can be summarized as follows:

- We propose Fine-grained Adaptive Visual Prompts (FAVP) to enhance the performance of generative MedVQA. The Adaptive Visual Prompt Creator (AVPC) adaptively generates region-level visual prompts based on image characteristics of various organs, providing fine-grained references for medical LLMs during answer generation, thereby improving the model's precise cross-modal localization capabilities on original images.
- We are the first to explore different types of organ-level visual prompts in MedVQA tasks and ultimately select the most effective visual prompt for eliciting intrinsic knowledge from LLMs based on the distinct characteristics of different datasets.
- We introduce a Hierarchical Answer Generator (HAG) to extract hierarchical high-semantic visual representations and fully retrieve and generate accurate answers in the language space from LLMs. Additionally, we integrate parameter-efficient Fine-Tuning (PEFT) techniques within HAG, significantly enhancing the model's spatial and contextual understanding with minimal parameters.
- Extensive experiments on VQA-RAD, SLAKE, and DME datasets validate the effectiveness of our proposed method, demonstrating its potential in open medical question answering.

Related Works

Medical Visual Question Answering (MedVQA) aims to automatically provide answers to questions based on given medical images. Current mainstream approaches are mainly divided into discriminative and generative methods. The discriminative methods select answers from a predefined answer set via classification. PubMedCLIP (Eslami, Meinel, and De Melo 2023) validated the effectiveness of transferring the CLIP architecture to the Med-VQA task. M3AE (Chen et al. 2022) proposed a multimodal self-supervised pretraining paradigm based on masked autoencoders to learn domain knowledge in the medical field. PMC-CLIP (Lin et al. 2023) constructed a larger dataset, PMC-OA, and pretrained a CLIP-style model. (Zhang et al. 2023c) aligned the pre-trained vision encoder and LLM via visual instruction tuning and constructed the PMC-VQA dataset. Although these approaches perform well on closed-set problems, they completely restrict the model's ability to handle open-set questions. Consequently, another generative method has emerged, generally combined with LLMs. Generative methods are not limited by a candidate answer set and can provide detailed answers to diverse open-set questions in the real world, significantly improving generalization. Due to the success of LLMs such as GPT-4 (Achiam et al. 2023) and LLaMA-2 (Touvron et al. 2023), numerous LLMs tailored for the medical field have emerged, becoming the mainstream in generative methods. Notable works include ChatDoctor (Yunxiang et al. 2023), PMC LLaMA (Wu et al. 2023), and Huatuo (Wang et al. 2023), etc. These models are fine-tuned on specific medical datasets based on open-source LLMs and can provide accurate and detailed guidance to patients in need. Visual Med Alpaca (Shu et al. 2023) is currently the earliest known attempt to incorporate medical images as input in a multimodal medical model. It converts images into intermediate representations and combines them with text input for the LLM. However, this approach may be limited by the pre-trained image captioning models. LLaVA-Med (Li et al. 2024) proposed a new curriculum learning method using a biomedical multimodal dataset constructed with GPT-4, adapting LLaVA (Liu et al. 2024) to the biomedical domain. Unlike the existing methods, our approach utilizes fine-grained visual prompts to facilitate precise localization of image regions by large multimodal models, thereby providing more accurate answers.

Method

Overview

Traditional methods treat MedVQA as a classification task where the goal is to select the most likely answer $\hat{a}$ from the set $\mathcal{A}$ containing all candidate answers, conditioned on the question $\mathcal{Q}$ and the image $\mathcal{I}$. The task can be formulated as:

$$\hat{a} = \underset{a \in \mathcal{A}}{\operatorname{argmax}} P(a \mid \mathcal{I}, \mathcal{Q}). \tag{1}$$

However, candidate answer sets are not provided in advance in the actual diagnosis, which hinders the openness of the system. In this paper, we leverage the powerful generative capabilities of the large language model to address the open MedVQA task in a generative manner.

$$p(\tilde{a}|\mathcal{I}, \mathcal{Q}) = \prod_{t=1}^{T} p(\tilde{a}_t|\hat{a}_{1:t-1}, \mathcal{I}, \mathcal{Q}), \tag{2}$$

where $\tilde{a}$ is the ground truth answer while T denotes the length of $\tilde{a}$. As illustrated in Figure 2, our framework mainly consists of an Adaptive Visual Prompt Creator (AVPC) and

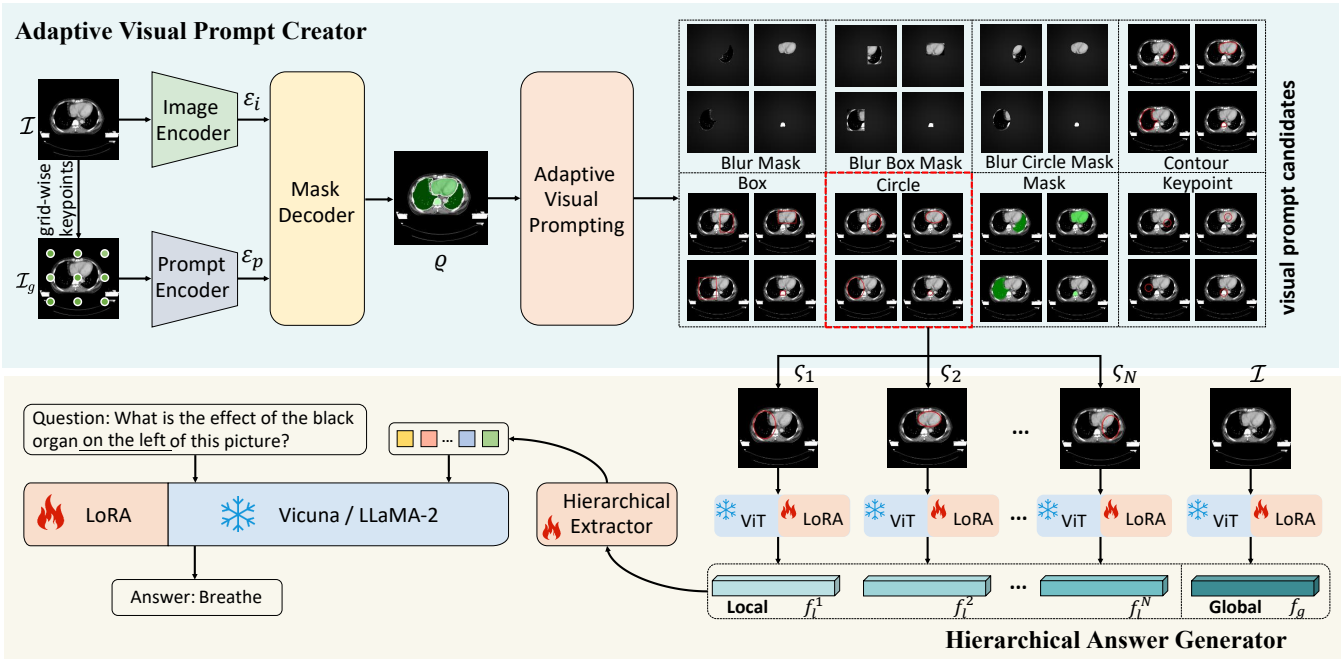


Figure 2: **FAVP framework** consists of two main components: the Adaptive Visual Prompt Creator (AVPC) at the top and the Hierarchical Answer Generator (HAG) at the bottom. The AVPC generates organ-level visual prompts to enhance the model's precise localization capabilities. Meanwhile, the HAG utilizes fine-grained visual prompts to learn hierarchical visual representations and to retrieve implicit knowledge from LLMs through cross-modal retrieval. Magnification is recommended to ensure detailed visualization.

a Hierarchical Answer Generator (HAG). The AVPC adaptively generates organ-level visual prompt candidates for images using grid-wise keypoints. The HAG aims to leverage these visual prompt candidates along with higher-level semantic feature representations extracted from the global image to query the intrinsic knowledge within the LLM in the language feature space, which enables the generation of more comprehensive open responses. The HAG includes a shared Vision Transformer (ViT), a Hierarchical Extractor, and the LLM.

Adaptive Visual Prompt Creator

Visual prompts significantly enhance model capabilities in interpreting visual information by offering essential guidance and context, thereby improving the accuracy of image content analysis. Recent studies have utilized CLIP's capabilities to encode images and superimpose visual cues, yet this approach did not translate effectively to MedVQA datasets. Unlike datasets used for visual grounding tasks, which include annotations specifying object locations for easy application of visual prompts, MedVQA datasets lack such annotations for organs or specific objects. To enable the model to focus more on fine-grained image details, it is natural to apply visual prompts to enhance the model's understanding of region-level visual features. In scenarios where ground truth masks are unavailable, existing methods typically rely on proposals generated by pre-trained detectors. However, detectors are not commonly utilized in the MedVQA domain. To generate fine-grained visual prompts, we propose the Adaptive Visual Prompt Creator (AVPC), which can adaptively produce various forms of visual prompts based on automatically generated keypoints, such as boxes, circles, masks, etc. Prior to generating the visual prompts, for the original image $\mathcal{I} \in \mathbb{R}^{H\times W\times C}$, where H, W, and C represent height, width, and channel, respectively, we generate a uniform grid of points according to the predefined number of keypoints. By overlaying this keypoint grid onto the original image $\mathcal{I}$, we obtain the augmented image $\mathcal{I}_g$. Subsequently, $\mathcal{I}$ and $\mathcal{I}_g$ are fed into the Image Encoder and Prompt Encoder, respectively, to obtain the Image embedding $\mathcal{E}_i$ and Prompt embedding $\mathcal{E}_p$.

$$\mathcal{E}_i = \mathbf{IE}(\mathcal{I}),\ \mathcal{E}_p = \mathbf{PE}(\mathcal{I}_g), \tag{3}$$

where $\mathbf{IE}$, $\mathbf{PE}$ denote the Image Encoder and the Prompt Encoder.

The process of generating visual prompts using the AVPC module is primarily divided into two steps. In the first step, we use SAM-Med2D (Cheng et al. 2023) to obtain a global mask ϱ that encompasses all relevant organs as follows:

$$\varrho = \mathbf{SAMMed}(\mathcal{E}_i, \mathcal{E}_p, \tau), \tag{4}$$

where τ represents the Non-maximum Suppression (NMS) threshold, a critical hyperparameter that influences the final outcome of the global mask. Using the global mask directly as a visual prompt may lead to coupling between local details. Therefore, to further enhance the focus on fine-grained

local details, in the second step, we employ adaptive visual prompting to segment the global mask ϱ into instance-level local masks $\varsigma_i \in \varrho = \{\varsigma_1, \ldots, \varsigma_N\}$, where N is the total number of instances. These single-organ local masks can be extended into various forms of local visual prompts, such as boxes, circles, contours, etc. Various forms of visual prompt candidates, encapsulating precise regional details, are illustrated in Figure 2. Ultimately, only one type of local visual prompt from the prompt candidates will be selected as the final input for the Hierarchical Answer Generator.

Hierarchical Answer Generator

To avoid being limited by a predefined set of answers, we treat MedVQA as a generative task and propose the Hierarchical Answer Generator. By using fine-grained visual prompts generated by the AVPC to query the intrinsic knowledge of LLM, we leverage its powerful generative capabilities to achieve true open question answering. Specifically, the Hierarchical Answer Generator comprises a shared ViT, a Hierarchical Extractor, and a LLM.

First, we input the set of local masks marked with visual prompts, $\{\varsigma_1, \ldots, \varsigma_N\}$, along with the original image $\mathcal{I}$ into the shared ViT, extracting hierarchical visual features and obtaining both instance-level local representations of organs and the global representation of the image.

$$f_l, f_g = \mathbf{ViT}(\varsigma_1, \ldots, \varsigma_N, \mathcal{I}), \tag{5}$$

where $f_l = \{f_l^1, \ldots, f_l^N\} \in \mathbb{R}^{B\times(N\times P)\times D}$ and $f_g \in \mathbb{R}^{B\times P\times D}$, with B, P, and D representing batch size, number of patches, and dimension, respectively.

Subsequently, we concatenate the local and global visual features along the second dimension to obtain the final visual features $f_v \in \mathbb{R}^{B\times((N+1)\times P)\times D}$. To shift the extraction of visual representations from the natural image space to the medical image space while introducing minimal parameters, we use Parameter-Efficient Fine-Tuning (PEFT) techniques by integrating LoRA (Hu et al. 2021) on the shared ViT. Specifically, during training, the shared ViT parameters are frozen, and only the low-rank adaptation layer is updated.

Following this, we employ our proposed hierarchical extractor $\Psi(\cdot)$ to learn higher-semantical visual representations from f_v and map f_v into the language feature space to obtain hierarchical visual tokens $\mathcal{V}$. The hierarchical extractor includes a Q-Former (Li et al. 2023) and a linear layer.

$$\mathcal{V} = \Psi(f_v) = \mathbf{Linear}(\mathbf{Qformer}(f_v)). \tag{6}$$

Finally, the LLM takes $\mathcal{V}$ and the question $\mathcal{Q}$ as inputs, producing the final answer $\hat{a}$.

$$\hat{a} = \mathbf{LLM}(\mathcal{V}, \mathcal{Q}). \tag{7}$$

Training LLM from scratch for MedVQA demands substantial computational resources and extensive annotated datasets. Therefore, we also employ the PEFT techniques on LLM to reduce computational cost. Notably, we only apply PEFT to the LLM during the fine-tuning phase.

The final objective is to minimize the cross-entropy loss using Teacher Forcing strategy with ground truth answer $\tilde{a}$.

$$\begin{aligned} &\min_{\theta} \ \mathcal{L}_{ce}(\theta), \\ &s.t.\ \mathcal{L}_{ce} = -\sum_{t=1}^{T} \log p_\theta\left(\tilde{a}_t \mid \mathcal{I}, \mathcal{Q}, \tilde{a}_{1:t-1}\right). \end{aligned} \tag{8}$$

Here, T is the length of the ground-truth answer, and $\tilde{a}_{1:t-1}$ denotes the preceding tokens in the ground-truth answer sequence. The symbol θ signifies the trainable parameters.

Training Stages

Our overall framework training is carried out in three stages, progressively adapting the general-domain multimodal LLM model to the biomedical domain.

Stage 1. To achieve cross-modal alignment between medical images and text, we utilize the radiology part of the ROCO dataset (Pelka et al. 2018). For each sample, given the image input, the model is required to predict the original caption. In training, we only update the linear layer in the hierarchical extractor and the low-rank adaptation layer in the shared ViT while keeping other weights frozen.

Stage 2. We keep the weights of the LLM and shared ViT frozen, and update the whole Hierarchical Extractor and the low-rank adaptation layer of ViT. To train our proprietary VQA model, we utilize PMC-VQA (Zhang et al. 2023c) in stage 2, a large-scale dataset that encompasses a broad range of modalities and diseases.

Stage 3. We train and evaluate our model on three downstream MedVQA datasets, i.e., VQA-RAD (Lau et al. 2018), SLAKE (Liu et al. 2021), and DME (Tascon-Morales, Márquez-Neila, and Sznitman 2022). During training, we freeze the original weights of the LLM while updating the low-rank adaptation layer. Concurrently, the Hierarchical Extractor and the low-rank adaptation layer of the shared ViT are updated. For evaluation, we set the model weights to a frozen state, prohibiting any updates.

Experiments

Experiments Setup

Datasets and Evaluation Metrics. We fine-tune and evaluate FAVP on three Medical VQA datasets, i.e., VQA-RAD, SLAKE, and DME. Following LLaVA-Med (Li et al. 2024), for closed-set questions, we report the accuracy based on the presence of ground-truth tokens in the generated sequences. For open-set questions, we use recall to assess the proportion of ground-truth tokens appearing in the generated sequences. Compared to discriminative methods that predict directly from a set of candidates, our approach more closely aligns with the nature of open-set settings and presents a greater challenge. For the DME dataset, we report overall accuracy and consistency metrics.

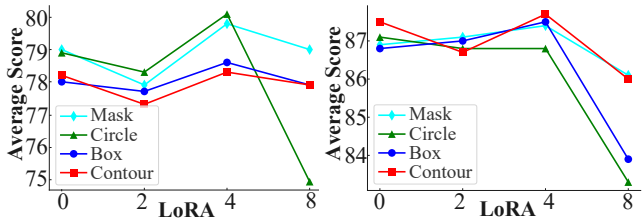
Implementation Details. We conduct all experiments on GeForce RTX 4090 GPUs. For the HAG module, we utilize ViT-G/14 (Zhai et al. 2022) as the shared ViT, Q-Former and linear layer as the Hierarchical Extractor while Vicuna 7B (Chiang et al. 2023) as the LLM. The Hierarchical Extractor consists of Q-Former and a linear layer. The Q-Former

Method	VQA-RAD			SLAKE			Parameter
	Ref	Open	Closed	Ref	Open	Closed	
Generative Methods							
Med-MoE (Jiang et al. 2024)	-	52.6	84.6	-	85.3	86.8	3.6B
LLaVA-Med (Li et al. 2024)	-	64.8	83.1	-	87.1	86.8	7B
FAVP (From Vicuna)	-	**71.9**	**88.2**	-	87.2	**88.1**	0.1B
FAVP (From LLaMA-2)	-	68.1	**89.0**	-	85.6	87.9	0.1B
Discriminative Methods							
Prefix T. Medical LM (Van Sonsbeek et al. 2023)	-	-	-	84.3	-	82.0	60M
PubMedCLIP (Eslami, Meinel, and De Melo 2023)	60.1	-	80.0	78.4	-	82.5	0.1B
M3AE (Chen et al. 2022)	67.2	-	83.5	80.3	-	87.8	0.4B
PMC-CLIP (Lin et al. 2023)	67.0	-	84.0	81.9	-	88.0	0.2B
MedVInT-TE (Zhang et al. 2023c)	69.3	-	84.2	88.2	-	87.7	0.2B

Table 1: Performance comparison with prior state-of-the-art methods on VQA-RAD and SLAKE datasets. For open-set questions, recall is reported for the free-form generative method under the column *Open*, while accuracy for discriminative methods is listed under the column *Ref*. For closed-set questions, accuracy is documented in the column *Closed*. **Bold** indicates FAVP achieves new SoTA.

Method	Accu.	Cons.
SQuINT (Selvaraju et al. 2020)	80.58	88.17
MVQA (Tascon-Morales et al. 2022)	81.15	89.95
MVQA-CPQA	83.49	94.20
LIMOD (Tascon-Morales et al. 2023a)	83.59	95.78
LQ (Tascon-Morales et al. 2023b)	84.20	-
FAVP	**84.73**	**97.82**

Table 2: Performance comparison of FAVP with SoTA approaches on DME dataset. Accu. and Cons. are the abbreviations for overall accuracy and consistency criteria. MVQA-CPQA and MVQA are different variants in the same paper.



(a) VQA-RAD (b) SLAKE

Figure 3: Ablation study on different LoRA ranks of the shared ViT across VQA-RAD and SLAKE datasets.

learns $K = 32$ query tokens with a hidden dimension of size 768. Based on preliminary experiments, we establish the LoRA rank of ViT at 4 and that of LLM at 8, and we also discuss them in ablation studies. The trainable components of the FAVP consist of a Hierarchical Extractor with 108M parameters and LoRA layers with 5M parameters, resulting in a total of 113M activation parameters.

During training, we employ the AdamW optimizer with a learning rate of 1e-4, following a cosine learning rate schedule. The values of β_1 and β_2 are set to 0.9 and 0.999, respectively. To enhance model generalization and mitigate overfitting, we apply a weight decay of 0.05. In HAG, images are resized to 224×224 to align with the encoder.

Quantitative Evaluations

Table 1 presents the comparison of FAVP with existing generative and discriminative methods on the VQA-RAD and SLAKE datasets. Firstly, on closed-set questions across both datasets, FAVP significantly outperforms the current leading generative and discriminative approaches, demonstrating its robust capability to perform precise localization and closed-set categorization under fine-grained visual prompts. Secondly, for open-set questions, FAVP achieves SoTA results among generative methods on both datasets, notably surpassing the advanced generative method LLaVA-Med by 11.0% on the VQA-RAD dataset. Remarkably, our open-set generation performance even exceeds the SoTA discriminative method MedVInT-TE, which uses a predefined set of candidate answers. Although the metrics for generative methods differ from discriminative methods, this indirectly shows that FAVP can generate answers that are comparable to, or even more accurate than, those from discriminative methods without restricting answer choices. Thirdly, with the aid of lightweight technology, FAVP requires only 0.1B trainable parameters, equivalent to just 1.4% of the parameter size used by LLaVA-Med, yet surpasses SoTA methods. Meanwhile, the performance variance across different LLM initializations on FAVP is minor, with Vicuna generally proving more adept at the MedQA task.

Table 2 demonstrates that FAVP achieves SoTA performance in both accuracy and consistency metrics on the DME ocular dataset, indicating its excellent generalizability across different medical imaging modalities. Furthermore, a 2.04% higher consistency metric compared to LIMOD suggests that FAVP's proficiency in ocular lesion assessment benefits from precise regional localization, substantially reducing contradictions in answers generated for different questions within the same image.

LLM	LoRA	VQA-RAD			SLAKE		
		Open	Closed	Avge.	Open	Closed	Avge.
Vicuna	0	67.8	85.7	76.8	83.0	85.8	84.4
	2	67.4	88.2	77.8	84.8	86.4	85.6
	4	69.6	85.7	77.7	83.3	86.5	84.9
	8	**71.9**	88.2	**80.1**	**87.2**	**88.1**	**87.7**
	16	67.8	87.1	77.5	83.2	86.3	84.8
LLaMA-2	0	63.8	85.3	74.6	82.4	84.1	83.3
	2	66.7	**89.3**	78.0	83.9	86.3	85.1
	4	69.9	84.9	77.4	85.3	87.0	86.2
	8	68.1	89.0	78.6	85.6	87.9	86.8
	16	59.6	84.6	72.1	84.4	86.5	85.5

Table 3: Ablation study on different LLMs and the setting of LLM's LoRA ranks. *Avge.* denotes the average value of accuracy across *open* and *closed* column. **Bold** indicates the best performance, while underline denotes the second-best performance.

Prompt	VQA-RAD		SLAKE	
	Open	Closed	Open	Closed
Mask	70.9	**88.6**	86.2	**88.6**
Blur Mask	70.7	86.8	86.7	88.4
Circle	**71.9**	88.2	85.6	87.9
Blur Circle Mask	65.8	87.1	86.0	87.4
Box	71.1	86.0	86.4	**88.6**
Blur Box Mask	66.9	87.1	85.9	87.4
Contour	68.3	88.2	**87.2**	88.1
Keypoint	69.0	87.1	86.2	87.0
w/o Prompt	67.5	86.4	85.1	87.2

Table 4: Ablation study on different types of prompt. **Bold** indicates the best performance, while underline denotes the second-best performance.

Stage 1	Stage 2	VQA-RAD		SLAKE	
		Open	Closed	Open	Closed
✗	✗	57.7	78.3	83.1	82.4
✓	✗	48.7	76.8	79.8	82.0
✗	✓	55.8	83.1	82.3	83.7
✓	✓	**71.9**	**88.2**	**87.2**	**88.1**

Table 5: Ablation study on different training stages. Stage 3 is enabled across all settings as the downstream fine-tuning phase, with details omitted in the table for conciseness.

Ablation Studies

LoRA rank settings and different LLMs. To determine the impact of different LoRA ranks of the shared ViT on the enhancement of Answer generation performance through visual prompts, we manually select four representative visual prompts and conduct ablation studies with LoRA ranks set at 0, 2, 4, and 8. For conciseness, the reported metrics are the average scores for open-set and closed-set questions. It can be observed from Figure 3 that optimal performance for most visual prompts is achieved when the rank is set to 4. Both excessively high and low LoRA ranks tend to degrade answer quality. We hypothesize that too low LoRA rank may hinder the model's transition from general to medical domains, while too high rank might introduce excessive trainable parameters, potentially diluting the domain knowledge already acquired in the frozen layers. Therefore, in subsequent experiments, we set the LoRA rank of ViT to 4. Simultaneously, as shown in Table 3, we conduct ablation experiments on the LoRA rank setting of LLM and various LLM architectures. The results indicate that both Vicuna and LLaMA-2 achieve optimal performance when the LoRA rank is set to 8, as measured by average QA performance. In particular, Vicuna slightly outperforms LLaMA-2, which may be attributed to Vicuna's specific optimization and fine-tuning for medical-related data and tasks, while LLaMA-2's general design and optimization strategy can result in comparatively lower performance for specialized domain tasks.

Different types of visual prompts. To evaluate the effectiveness of different types of visual prompts in querying latent medical knowledge from Large Language Models (LLMs), we conduct ablation studies on eight visual prompts, including Mask, Circle, Box, and Contour. As observed in Table 4, most visual prompts contribute to performance gains in the MedVQA task, and the optimal visual prompt varies across different datasets. While focusing on performance in open-set problems and also considering the effects on closed-set questions, Circle and Contour emerged as the most effective visual prompts for the VQA-RAD and SLAKE datasets, respectively.

Impact of different training stages. In addition to fine-tuning on the downstream datasets in stage 3, separate ablation studies are conducted on stages 1 and 2. As observed in Table 5, using either stage 1 or stage 2 in isolation does not yield optimal results. Notably, employing only stage 1 results in performance that is even slightly lower than the non-pretrained baseline. This can be attributed to the use of image captioning as the pretraining task in stage 1, where the general medical knowledge acquired does not generalize well across the large-scale MedVQA dataset in stage 2,

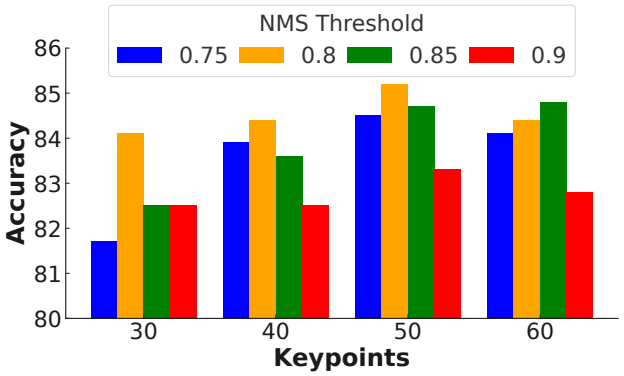


Figure 4: Ablation study on the NMS threshold τ and the number of keypoints.

resulting in increased noise during fine-tuning in stage 3. The substantial improvement in accuracy when both stages 1 and 2 are active further confirms that only their collaboration can effectively enhance the model's capabilities in multimodal knowledge mining and feature representation learning, thereby benefiting subsequent fine-tuning.

Hyperparameters analysis. Figure 4 investigates the impact of the hyperparameters, specifically the number of keypoints and the NMS threshold τ on the accuracy of answer generation. Empirically, we select keypoints spanning from 30 to 60, while τ ranges cover [0.75, 0.80, 0.85, 0.90]. It can be observed that, firstly, when τ increases to 0.90, there is a significant performance decline, likely due to the high threshold causing the omission of masks corresponding to crucial organs, thereby degrading the quality of visual prompts input during the answer generation phase. Secondly, as τ increases, the accuracy associated with each set of keypoints initially rises then falls, with most peaks occurring at $\tau = 0.80$. Furthermore, when $\tau = 0.80$, the expansion of keypoints also shows a pattern of initial increase followed by a decrease in accuracy. This reveals the complex interplay and potential trade-offs between these two hyperparameters in generating visual prompts.

Qualitative Evaluations

Figure 5 presents representative cases and the corresponding visual prompts generated by FAVP across three datasets. The first three cases demonstrate FAVP's capability to accurately localize lesions and organs across different imaging modalities using fine-grained generated visual prompts. Notably, the first-row second column shows that FAVP, due to its strong localization accuracy, can provide more precise location descriptions than GT, although the corresponding Recall value decreases significantly. This suggests potential limitations in current generative MedVQA evaluation metrics. Additionally, the final example from the DME dataset reveals that both FAVP and LLaVA-Med encounter difficulties in inferring diabetic macular edema grades, indicating that general large models still face limitations in scenarios

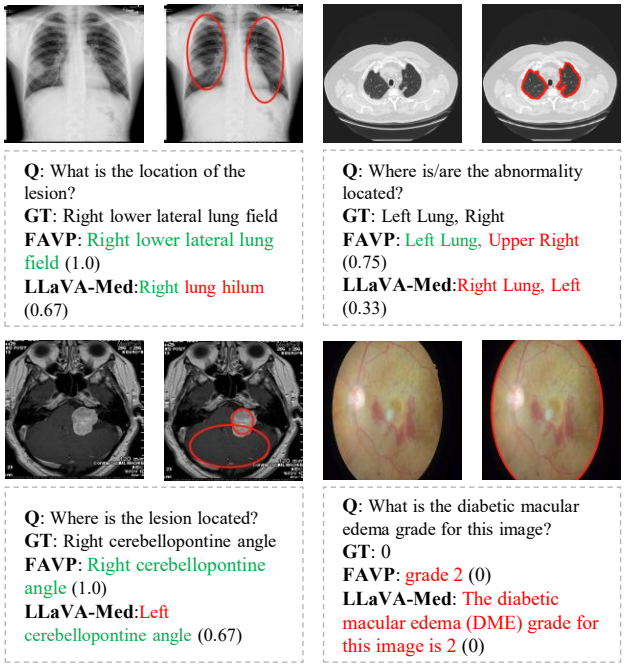


Figure 5: Qualitative examples: incorrect in red, correct in green. Left column: VQA-RAD dataset. Right column: SLAKE and DME datasets. The values in parentheses indicate the *recall* of the current answer compared to GT.

requiring complex medical knowledge reasoning. This multifaceted exploration provides insights into the capabilities and potential improvements for FAVP.

Conclusion

This paper treats MedVQA as a generative task and introduces fine-grained adaptive visual prompts to enhance generative MedVQA. Our framework integrates an Adaptive Visual Prompt Creator, which adaptively generates region-level visual prompts based on the characteristics of different organs within medical images. We further explore different types of instance-level visual prompts, and ultimately select the most effective visual prompt based on the characteristics of different datasets. Furthermore, FAVP incorporates a Hierarchical Answer Generator, designed to extract hierarchical high-semantic representations from fine-grained visual features and map them into the language space. We believe that FAVP represents a significant step towards enhancing the precise localization capabilities in open MedVQA tasks. Extensive experiments showcasing FAVP's prowess. Nevertheless, as FAVP has only been validated on common medical modalities (e.g., X-Ray, CT, MRI), further validation across additional modalities, such as PET scans, mammography, and histopathological images, is necessary to assess its generalizability. Future work aims to enhance the capability of FAVP in the reliability of zero-shot generation. We hope that FAVP can inspire further exploration and application of visual prompts in various medical verticals.

Acknowledgments

This work was supported by National Natural Science Foundation of China No. 62125201, 62002314, and Zhejiang Provincial Natural Science Foundation of China under Grant No. LY23F020005.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Chen, Z.; Du, Y.; Hu, J.; Liu, Y.; Li, G.; Wan, X.; and Chang, T.-H. 2022. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 679–689. Springer.

Cheng, J.; Ye, J.; Deng, Z.; Chen, J.; Li, T.; Wang, H.; Su, Y.; Huang, Z.; Chen, J.; Jiang, L.; et al. 2023. Sam-med2d. *arXiv preprint arXiv:2308.16184*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3): 6.

Decenciere, E.; Cazuguel, G.; Zhang, X.; Thibault, G.; Klein, J.-C.; Meyer, F.; Marcotegui, B.; Quellec, G.; Lamard, M.; Danno, R.; et al. 2013. TeleOphta: Machine learning and image processing methods for teleophthalmology. *Irbm*, 34(2): 196–203.

Eslami, S.; Meinel, C.; and De Melo, G. 2023. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, 1181–1193.

Gu, T.; Yang, K.; Liu, D.; and Cai, W. 2024. LaPA: Latent Prompt Assist Model For Medical Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4971–4980.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jiang, S.; Zheng, T.; Zhang, Y.; Jin, Y.; and Liu, Z. 2024. MoE-TinyMed: Mixture of Experts for Tiny Medical Large Vision-Language Models. *arXiv preprint arXiv:2404.10237*.

Khan, Z.; and Fu, Y. 2024. Consistency and Uncertainty: Identifying Unreliable Responses From Black-Box Vision-Language Models for Selective Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10854–10863.

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

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Lin, W.; Zhao, Z.; Zhang, X.; Wu, C.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 525–536. Springer.

Liu, B.; Zhan, L.-M.; Xu, L.; Ma, L.; Yang, Y.; and Wu, X.-M. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 1650–1654. IEEE.

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

Pelka, O.; Koitka, S.; Rückert, J.; Nensa, F.; and Friedrich, C. M. 2018. Radiology Objects in COntext (ROCO): A Multimodal Image Dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, 180–189. Cham: Springer International Publishing.

Porwal, P.; Pachade, S.; Kamble, R.; Kokare, M.; Deshmukh, G.; Sahasrabuddhe, V.; and Meriaudeau, F. 2018. Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. *Data*, 3(3): 25.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Selvaraju, R. R.; Tendulkar, P.; Parikh, D.; Horvitz, E.; Ribeiro, M. T.; Nushi, B.; and Kamar, E. 2020. Squinting at vqa models: Introspecting vqa models with sub-questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10003–10011.

Shu, C.; Chen, B.; Liu, F.; Fu, Z.; Shareghi, E.; and Collier, N. 2023. Visual med-alpaca: A parameter-efficient biomedical llm with visual capabilities.

Tascon-Morales, S.; Márquez-Neila, P.; and Sznitman, R. 2022. Consistency-preserving visual question answering in medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 386–395. Springer.

Tascon-Morales, S.; Márquez-Neila, P.; and Sznitman, R. 2023a. Logical Implications for Visual Question Answering Consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6725–6735.

Tascon-Morales, S.; Márquez-Neila, P.; and Sznitman, R. 2023b. Localized questions in medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 361–370. Springer.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Van Sonsbeek, T.; Derakhshani, M. M.; Najdenkoska, I.; Snoek, C. G.; and Worring, M. 2023. Open-ended medical visual question answering through prefix tuning of language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 726–736. Springer.

Wada, Y.; Kaneda, K.; Saito, D.; and Sugiura, K. 2024. Polos: Multimodal Metric Learning from Human Feedback for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13559–13568.

Wang, H.; Liu, C.; Xi, N.; Qiang, Z.; Zhao, S.; Qin, B.; and Liu, T. 2023. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*.

Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*, 2(5): 6.

Yu, T.; Fu, K.; Wang, S.; Huang, Q.; and Yu, J. 2024a. Prompting Video-Language Foundation Models with Domain-specific Fine-grained Heuristics for Video Question Answering. *IEEE Transactions on Circuits and Systems for Video Technology*.

Yu, T.; Fu, K.; Zhang, J.; Huang, Q.; and Yu, J. 2024b. Multi-Granularity Contrastive Cross-Modal Collaborative Generation for End-to-End Long-Term Video Question Answering. *IEEE Transactions on Image Processing*.

Yu, T.; Ge, B.; Wang, S.; Yang, Y.; Huang, Q.; and Yu, J. 2024c. Consistency Conditioned Memory Augmented Dynamic Diagnosis Model for Medical Visual Question Answering. *IEEE Journal of Biomedical and Health Informatics*.

Yu, T.; Lin, X.; Wang, S.; Sheng, W.; Huang, Q.; and Yu, J. 2024d. A Comprehensive Survey of 3D Dense Captioning: Localizing and Describing Objects in 3D Scenes. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(3): 1322–1338.

Yu, T.; Yu, J.; Yu, Z.; Huang, Q.; and Tian, Q. 2020. Long-term video question answering via multimodal hierarchical memory attentive networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(3): 931–944.

Yu, T.; Yu, J.; Yu, Z.; and Tao, D. 2019. Compositional attention networks with two-stream fusion for video question answering. *IEEE Transactions on Image Processing*, 29: 1204–1218.

Yuan, Z.; Ren, J.; Feng, C.-M.; Zhao, H.; Cui, S.; and Li, Z. 2024. Visual programming for zero-shot open-vocabulary 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20623–20633.

Yunxiang, L.; Zihan, L.; Kai, Z.; Ruilong, D.; and You, Z. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*, 2(5): 6.

Zhai, X.; Kolesnikov, A.; Houlsby, N.; and Beyer, L. 2022. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12104–12113.

Zhang, K.; Jiang, H.; Zhang, J.; Huang, Q.; Fan, J.; Yu, J.; and Han, W. 2023a. Semi-supervised medical report generation via graph-guided hybrid feature consistency. *IEEE Transactions on Multimedia*, 26: 904–915.

Zhang, K.; Yang, Y.; Yu, J.; Fan, J.; Jiang, H.; Huang, Q.; and Han, W. 2024. Attribute Prototype-guided Iterative Scene Graph for Explainable Radiology Report Generation. *IEEE Transactions on Medical Imaging*.

Zhang, K.; Yang, Y.; Yu, J.; Jiang, H.; Fan, J.; Huang, Q.; and Han, W. 2023b. Multi-task paired masking with alignment modeling for medical vision-language pre-training. *IEEE Transactions on Multimedia*.

Zhang, X.; Wu, C.; Zhao, Z.; Lin, W.; Zhang, Y.; Wang, Y.; and Xie, W. 2023c. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.