LaPA: Latent Prompt Assist Model For Medical Visual Question Answering

Tiancheng Gu
University of Sydney
Sydney, NSW, Australia
tigu8498@uni.sydney.edu.au

Kaicheng Yang
DeepGlint
Beijing, China
kaichengyang@deepglint.com

Dongnan Liu
University of Sydney
Sydney, NSW, Australia
dongnan.liu@sydney.edu.au

Weidong Cai
University of Sydney
Sydney, NSW, Australia
tom.cai@sydney.edu.au

Abstract

*Medical visual question answering (Med-VQA) aims to automate the prediction of correct answers for medical images and questions, thereby assisting physicians in reducing repetitive tasks and alleviating their workload. Existing approaches primarily focus on pre-training models using additional and comprehensive datasets, followed by fine-tuning to enhance performance in downstream tasks. However, there is also significant value in exploring existing models to extract clinically relevant information. In this paper, we propose the **La**tent **P**rompt **A**ssist model (LaPA) for medical visual question answering. Firstly, we design a latent prompt generation module to generate the latent prompt with the constraint of the target answer. Subsequently, we propose a multi-modal fusion block with latent prompt fusion module that utilizes the latent prompt to extract clinical-relevant information from uni-modal and multimodal features. Additionally, we introduce a prior knowledge fusion module to integrate the relationship between diseases and organs with the clinical-relevant information. Finally, we combine the final integrated information with image-language cross-modal information to predict the final answers. Experimental results on three publicly available Med-VQA datasets demonstrate that LaPA outperforms the state-of-the-art model ARL, achieving improvements of 1.83%, 0.63%, and 1.80% on VQA-RAD, SLAKE, and VQA-2019, respectively. The code is publicly available at https : //github.com/GaryGuTC/LaPA_model.*

1. Introduction

Medical visual question answering (Med-VQA) plays a critical role in disease detection and diagnosis. In clinical practice, the review of numerous medical images and their corresponding questions by physicians is both costly and error-prone [16]. To address this challenge, there has been a growing interest in the development of automatic Med-VQA techniques [2, 6, 10, 21, 24, 33]. While deep learning models have achieved remarkable success in predicting accurate answers in standard visual-question answering (VQA) tasks by given images and questions [11, 32], Med-VQA poses unique challenges [2]. The size of Med-VQA datasets is relatively small, and medical images are complex and challenging due to the small region of interest related to the disease that physicians need to focus on [8, 29]. Consequently, extracting clinically relevant information from medical images becomes a difficult task for the model [28].

Numerous Med-VQA methods [2, 10, 21, 33] have been proposed to address the aforementioned challenges and have demonstrated impressive performance. For instance, methods such as MEVF model [24], MMQ model [4], and CPCR [19] have proposed pretraining the model using external complementary datasets to enhance the model's analytical capabilities, followed by fine-tuning for downstream tasks. Similarly, M2I2 model [15] and m3ae model [2] have utilized self-supervised learning to enable the model to autonomously learn clinical features from both image and language modalities. Notably, despite their remarkable achievements, none of these approaches consider the latent prompt. However, the latent prompt is a crucial aspect that warrants research attention due to its enhanced flexibility in information extraction, as evidenced by its widespread utilization in the field of natural language processing [9, 26, 35].

This study presents the LaPA (Latent Prompt Assist) model for medical visual question answering (Med-VQA), as illustrated in Fig. 1. The LaPA model incorporates the latent prompt to filter different modal information and extract
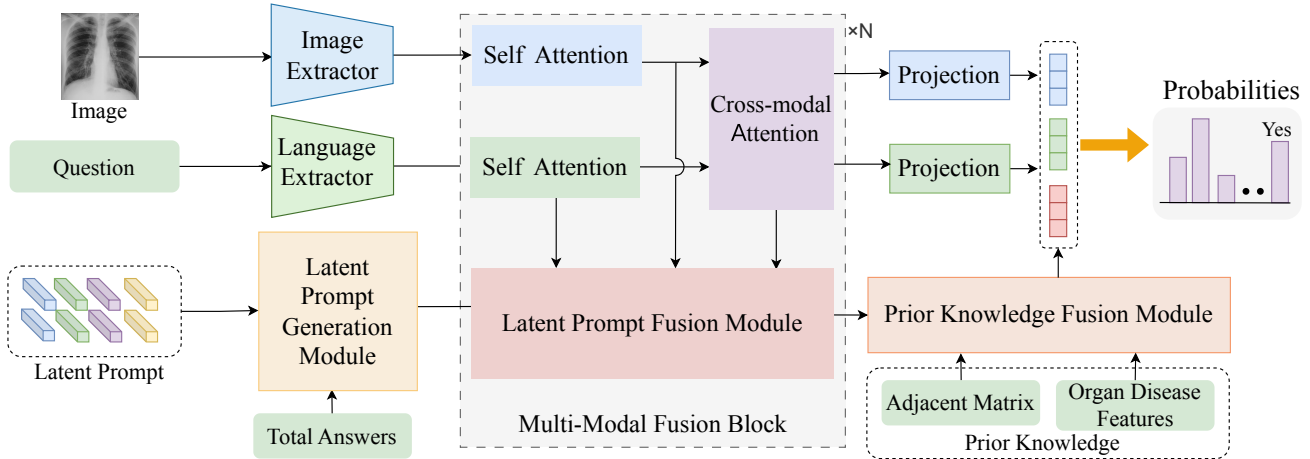
Figure 1. The overall structure of our proposed LaPA model. The input feature is denoted by a block with rounded corners, while the square-angled structure represents a module. The language and image pipelines are represented by green and blue modules, respectively. The final tokens in blue, green, and red correspond to the cross-modal image, language, and integrated information, respectively. For optimal viewing, it is recommended to zoom in for detailed examination.

clinic-relevant information, aiding in the prediction of the final answer. Firstly, we introduce the latent prompt generation module, which generates the latent prompt. The latent prompt interacts with the total answer tokens and is constrained by the target answer tokens to focus on the relevant tokens associated with the target answer. Subsequently, the latent prompt is fed into the multi-modal fusion block to fuse with uni- and multi-modal information, enabling the filtering of different modal information and extraction of clinic-relevant details. Additionally, the latent prompt interacts with the prior knowledge derived from the relationship between organs and diseases, obtained from a knowledge graph [18], resulting in the generation of the final interacted information to further assist in the prediction of the final answer. Lastly, the latent prompt combines with the image-language cross-modal information to produce the final answer.

The main contributions of our work can be summarized as follows:

- We propose the latent prompt generation model that generates a latent prompt and utilize a multi-modal fusion block to filter different modal information and extract clinic-relevant information.

- We leverage prior knowledge regarding the relationship between organs and diseases by employing a graph neural network to interact with the latent prompt, ultimately assisting in answer prediction.

- Our proposed LaPA model demonstrates its effectiveness by achieving exceptional performance on VQA-RAD [12], SLAKE [18], and VQA-2019 [1] datasets.

2. Related Works

Prompt Learning. Prompt learning is a research focus aimed at leveraging prompts to enhance various aspects of a model's performance, such as efficiency, flexibility, and knowledge transfer [9, 37, 38]. Recent studies [26, 27] have explored the utilization of prompts to extract relevant information from pre-trained models for downstream tasks, yielding promising results. Notably, the ChatExtract method proposed by [27] employs engineered prompts to aid in sentence differentiation and data extraction, thereby improving answer accuracy. In contrast, [35] focuses on using latent prompts, encompassing controlled and uncontrolled signals, to extract valuable and highly relevant information, thereby enhancing text summarization quality. Building upon these studies, we introduce the concept of latent prompts to the domain of Med-VQA.

Medical Visual Question Answering. The field of automatic prediction of answers for medical visual questions based on medical images has been extensively studied, yielding numerous notable works [10, 21, 24, 33]. Notably, some approaches have been proposed to train models based on external knowledge, such as MEVF model [24] and MMQ model [4]. These methods initialize the weights of specific modules (e.g., visual encoder or decoders) using pre-trained large language models (LLMs) and subsequently fine-tune the overall frameworks for downstream Med-VQA tasks. Q2ATransformer [21] introduces a novel approach that combines the advantages of both classification and generation techniques, achieving a unified treat-

(a) Latent Prompt Generation Module **(b) Latent Prompt Fusion Module** **(c) Prior Knowledge Fusion Module**
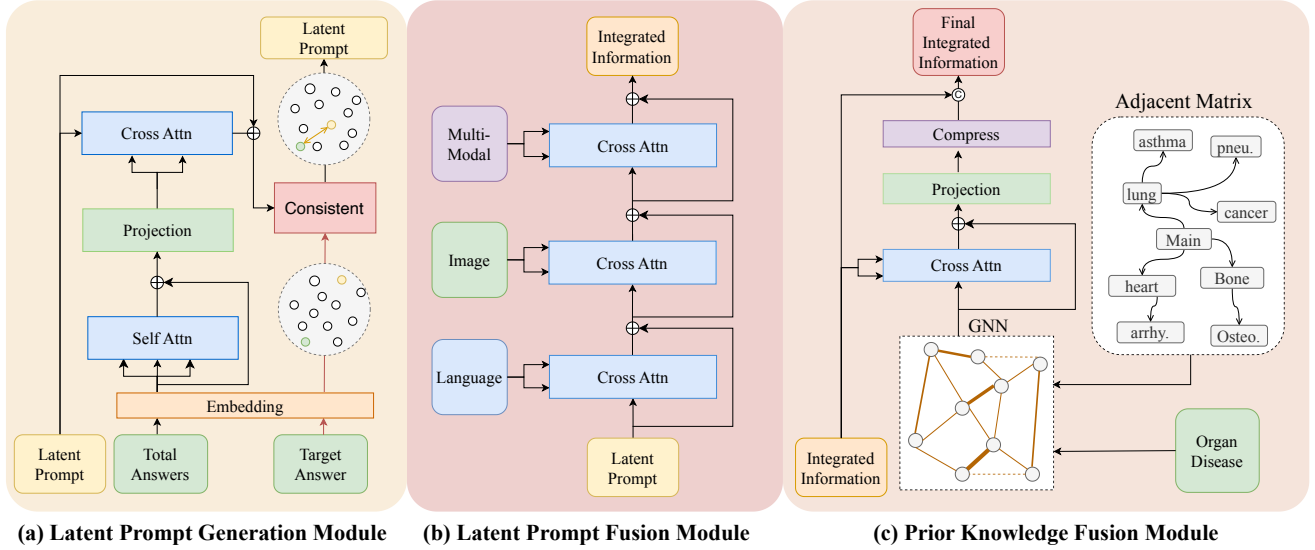
Figure 2. The structure of the main modules in LaPA is illustrated as follows: (a), (b), and (c) represent the latent prompt generation module (Sec. 3.1), the latent prompt fusion module (Sec. 3.2), and the prior knowledge fusion module (Sec. 3.3), respectively. For optimal visualization, it is recommended to zoom in for detailed examination.

ment for closed-end and open-end questions. By employing learnable candidate answer embeddings, Q2ATransformer queries the presence of each answer class for a given image-question pair. Additionally, MeDVInt [34] and LLaVA-Med [13] are generative models for Med-VQA understanding that align visual information from a pre-trained vision encoder with a large language model (LLM) or large vision language model such as ChatGPT and LLaVA. In contrast to these existing works, our proposed approach utilizes latent prompts to filter uni- and multi-modal information and extract clinic-relevant information, thereby enhancing the final answer prediction process.

3. LaPA Model

The architectural overview of our proposed LaPA (Latent Prompt Assist) model for medical visual question answering is presented in Fig. 1. The model comprises three key components: the latent prompt generation module (Sec. 3.1), the multi-modal fusion block (Sec. 3.2), and the prior knowledge fusion module (Sec. 3.3). Further insights into the training process can be found in Sec. 3.4.

3.1. Latent Prompt Generation Module

We first propose a latent prompt generation module (Fig. 2 (a)) to generate the learnable latent prompt, which is initialized using the normal distribution. To improve training efficiency and performance, we interact the generated latent prompt with total answer tokens. Under the constraint of answer tokens, the latent prompt can fo-

cus on the tokens associated with the answer. To this end, we treat all the answer tokens in the downstream datasets as prior knowledge, embedding them as features X_{TA} using RoBERTa [20]. Subsequently, the total answer tokens undergo self-attention followed by a projection layer to obtain the total token features F_{TA} as follows:

$$F_{TA} = \text{Proj}(\text{SA}(X_{TA})), \qquad (1)$$

where $\text{SA}(\cdot)$ and $\text{Proj}(\cdot)$ represent self-attention mechanism and projection layer respectively. After that, we employ cross-attention to integrate the total answer tokens with the latent prompt:

$$\hat{X}_{LP} = \text{CA}(X_{LP}, F_{TA}, F_{TA}), \qquad (2)$$

where $\text{CA}(\cdot)$ represents the cross-attention mechanism [30] with the query, key and value as input. To focus on answers-related tokens, we introduce a consistent loss $\mathcal{L}_{CS}$ to constrain the latent prompt with the target answer, thereby bringing it closer to the target answer in the semantic space. The process is defined as:

$$\mathcal{L}_{CS}(\hat{X}_{LP}, X_A) = 1 - \frac{\hat{X}_{LP}^{\top} X_A}{||\hat{X}_{LP}|| \, ||X_A||}, \qquad (3)$$

where X_A is the token embeddings of the target answer.

3.2. Multi-modal Fusion Block

To make the latent prompt fully extract clinic-relevant information from uni-modal and multi-modal information,

we introduce a multi-modal feature fusion block. As shown in Fig. 1, the image features and language features are extracted by the Swin Transformer [22] and the RoBERTa [20], and the uni-modal features F_I and F_L can be obtained through self-attention as follows:

$$F_I = SA(E_I(X_I)),\qquad(4)$$

$$F_L = SA(E_L(X_L)).\qquad(5)$$

After that, the image and language features are fused through the cross-attention to get the multi-modal features F_{MM}:

$$F_{MM} = [Proj(CA(F_I, F_L, F_L)); Proj(CA(F_L, F_I, F_I))],\qquad(6)$$

where $Proj(\cdot)$ represents the projection layer.

After getting the uni-modal and multi-modal features F_I, F_L, and F_{MM}, we design the latent prompt fission module (Fig. 2 (b)) to make the latent prompt to integrate clinic-relevant information through cross-attention:

$$X_{II} = CA(\hat{X}_{LP}, F_I, F_I),\qquad(7)$$

$$\check{X}_{II} = CA(X_{II}, F_L, F_L),\qquad(8)$$

$$\tilde{X}_{II} = CA(\check{X}_{II}, F_{MM}, F_{MM}),\qquad(9)$$

where $CA(\cdot)$ is the cross-attention mechanism. The X_{II} represents the integrated information obtained by combining latent prompts with image features. Similarly, the $\check{X}_{II}$ and the $\tilde{X}_{II}$ denote the integrated information resulting from the fusion of language features and multi-modal features, respectively. The fusion process follows a sequential order, where language features are integrated first, followed by images, and finally multi-modal features. We have conducted experiments to explore various approaches for information fusion and extraction, and the current form yields the most optimal results.

In the multi-modal fusion module, the latent prompt is utilized to integrate with language features to extract clinically relevant information within the textual semantic space. Subsequently, it is combined with image features to extract clinically relevant information within the image semantic space. Finally, the integrated information undergoes fusion with the combined language-image cross-modal features to filter out diverse modal information and consolidate the uni-modal features of both language and image, along with their multi-modal combination features, resulting in the generation of the final clinically relevant information.

3.3. Prior Knowledge Fusion Module

Following the previous works [31, 36], we incorporate a prior knowledge graph [18] that captures the relationships between organs and diseases to enhance the accuracy of answer prediction in Med-VQA. We employ a graph neural network ($GNN(\cdot)$) to analyze the organ-disease relationships and improve the performance of answer prediction. Additionally, we propose a prior knowledge fusion module that integrates the prior knowledge with the integrated information to facilitate the final answer prediction.

As depicted in Fig. 2 (c), the adjacent matrix X_{adj} is derived from the aforementioned prior knowledge [18], representing the relationship between organs and diseases using binary values (0 and 1). The organ-disease feature F_{OD} is tokenized and embedded using RoBERTa [20]. Subsequently, it is fed into the GNN module to extract valuable information regarding the organ-disease relationships denoted as F_G, which can be summarized as follows:

$$F_G = GNN(F_{OD}, X_{adj}).\qquad(10)$$

Then, the extracted information is combined with the previous integrated information $\tilde{x}_{LP}$ to get the final integrated information ($\hat{X}_{II}$), and the process is indicated below:

$$\hat{X}_{II} = [\tilde{X}_{II}; Proj(CA(F_G, X_{LP}, X_{LP}))],\qquad(11)$$

where $CA(\cdot)$ is the cross attention mechanism and $Proj(\cdot)$ is the projection layer. Finally, the interacted relationship-based features will concatenate ($[;]$) with latent prompt as the final integrated information to assist the final answer predicted for Med-VQA.

3.4. Training Details

After the processes mentioned above, we add the cross-modal information F_{FI} and F_{FL} of the cross-modal attention in the last multi-modal fusion block with the final integrated information $\hat{X}_{II}$ to predict the answer:

$$X_F = \alpha\hat{X}_{II} + \theta F_{FI} + \beta F_{FL},\qquad(12)$$

where α, θ, and β are weight to balance different types of information. This final total loss ($\mathcal{L}_T$) is shown below:

$$\mathcal{L}_T = \mathcal{L}_{BCE}(X_F, F_T) + \eta\mathcal{L}_{CS},\qquad(13)$$

where the $\mathcal{L}_{BCE}$ is the binary cross-entropy loss [7] and $\mathcal{L}_{CS}$ is the consistent loss used to minimize the semantic distance between the latent prompt and the target answer. η is a loss weight to adjust the influence of different losses.

4. Experiments and Results

4.1. Implementation Details

For our model, we adopted the Swin-Transformer [22] as the image extractor model, RoBERTa [20] as the language extractor model, the graph attention network [31]

Method	Venue	VQA-RAD			SLAKE			VQA-2019
		Open	Closed	Overall	Open	Closed	Overall	Overall
BAN [11]	NeurIPS[18]	37.40	72.10	58.30	74.60	79.10	76.30	-
CPRD-BAN [17]	MICCAI[21]	52.50	77.90	67.80	79.50	83.40	80.10	-
MMBERT [10]	ISBI[21]	63.10	77.90	72.00	-	-	-	67.20
M3AE*† [2]	MICCAI[22]	64.80	82.72	75.61	79.22	85.10	81.53	78.40
M2I2 [15]	ISBI[22]	61.80	81.60	73.70	74.70	**91.10**	81.20	-
ARL* [3]	MM[22]	65.10	85.96	77.55	79.70	89.30	84.10	79.80
PubMedCLIP [5]	EACL[23]	60.10	80.00	72.10	78.40	82.50	80.10	-
CPCR [19]	TMI[23]	60.50	80.40	72.50	80.50	84.10	81.90	-
LaPA	Ours	**68.72**	**86.40**	**79.38**	**82.17**	88.70	**84.73**	**81.60**

Table 1. The results of the LaPA model and other tested models in VAR-RAD, SLAKE and VQA-2019. * indicates that we tested the results ourselves, which may differ from those reported in the models' original papers. † denotes the baseline model. The results for other models were obtained from their original papers. The highest-performing result in each category is highlighted in bold for clarity.

#	Method	VQA-RAD			SLAKE			VQA-2019
		Open	Closed	Overall	Open	Closed	Overall	Overall
1	BL.	64.80	82.72	75.61	79.22	85.10	81.53	78.40
2	+GM.$_{w/o\,cs}$ & LF.	68.16	84.93	78.27	80.93	87.74	83.60	80.80
3	+GM. & LF.	**69.27**	85.29	78.94	81.24	87.50	83.70	81.30
4	+GM. & LF.+PF.	68.72	**86.40**	**79.38**	**82.17**	**88.70**	**84.73**	**81.60**
-	Δ	↑ 3.92	↑ 3.68	↑ 3.77	↑ 2.95	↑ 3.60	↑ 3.20	↑ 3.20

Table 2. The ablation study for the LaPA model was conducted on the VQA-RAD, SLAKE, and VQA-2019 datasets to ascertain the contribution of individual components to the overall performance. In this context, GM., LF., and PF. represent the latent prompt generation module, latent prompt fusion module, and prior knowledge fusion module, respectively. The term $w/o\,cs$ denotes the exclusion of the consistency method from the model configuration. The final row delineates the performance enhancement achieved by the LaPA model relative to the established baseline model.

with eight heads as the GNN model, and utilized six multi-modal fusion blocks. Training was conducted on a single NVIDIA GeForce RTX3090 GPU with 24GB memory, employing half-precision training. Following the approach in M3AE [2], we utilized the AdamW optimizer [23] with a learning rate of 5e-6 for optimization. The input images were resized to 384×384, and the feature dimension was set to 768. Furthermore, we utilized the pre-training weights from the M3AE model, which were pre-trained on the ROCO [25] and MedICaT [28] datasets. For evaluation purposes, we report the matching accuracy for both closed-set and open-set questions. The overall metrics are calculated by combining the results from open-set and closed-set questions using coefficients, as outlined in M3AE [2].

4.2. Datasets

In order to comprehensively evaluate the effectiveness of our proposed method, we conducted experiments on three widely-used Med-VQA benchmarks: VQA-RAD [12], SLAKE [18], and VQA-2019 [1]. The dataset splits provided by existing works, such as M3AE [2], were used in our experiments. The questions in VQA-RAD and SLAKE are categorized into two types: open-ended (free-form) and closed-ended (YES/NO) forms. VQA-RAD dataset consists of 315 radiology images with 3064 question-answer pairs, and a subset of 451 pairs was used for testing purposes. SLAKE dataset is composed of 642 radiology images, with 14028 question-answer (QA) pairs. The dataset was divided into a ratio of 70:15:15 for training, validation, and testing, respectively. It it worth noting that we only evaluated the English subset of SLAKE. VQA-2019 dataset comprises 3200 medical images, with 12792 QA pairs for training, 500 images with 2000 QA pairs for validation, and 500 images with 500 QA pairs for testing.

4.3. Comparison Experiments

Our proposed LaPA model was benchmarked against eight contemporary state-of-the-art (SOTA) Med-VQA methodologies: BAN [11], CPRD [17], MMBERT [10], M3AE [2], M2I2 [15], ARL [3], PubMedCLIP [5] and CPCR [19]. As delineated in Tab. 1, LaPA consistently surpassed the aforementioned models on all three datasets in the majority of evaluative metrics. Notably, for the VQA-RAD dataset, our model demonstrated a consider-
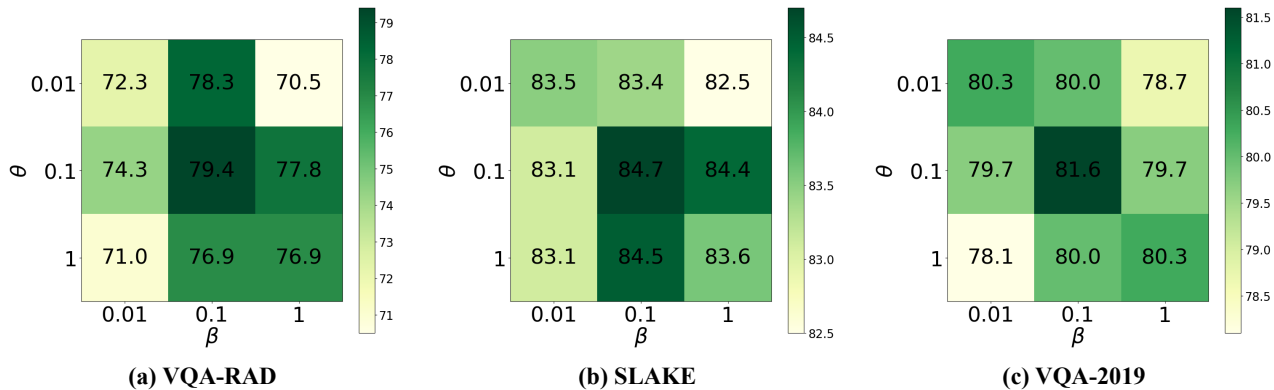
(a) VQA-RAD   **(b) SLAKE**   **(c) VQA-2019**

Figure 3. Ablation on the θ and β.

Interact Order	VQA-RAD			SLAKE			VQA-2019
	Open	Closed	Overall	Open	Closed	Overall	Overall
I.⇒L.⇒MM.	55.31	84.56	72.95	81.40	87.74	83.88	78.93
L.⇒I.⇒MM.	**68.72**	**86.40**	**79.38**	**82.17**	**88.70**	**84.73**	**81.60**

Table 3. The results of the change in the fusion direction by latent prompt in the latent prompt fusion module. The I., L., and MM. are the abbreviations of image, language, and multi-modal.

Latent Prompt size	4	8	16	32	64	128	256
VQA-RAD	77.16	78.27	78.49	**79.38**	76.94	76.49	75.61
SLAKE	84.17	84.35	83.88	**84.73**	84.26	84.26	84.45
VQA-2019	80.00	80.53	79.47	**81.60**	79.47	78.93	80.00

Table 4. Ablation on the latent prompt size.

able enhancement in performance across all question types, achieving an overall accuracy of 79.38%, an improvement of 1.83 percentage points over the second-best model. In the SLAKE dataset, LaPA achieved an overall accuracy of 84.73%, outperforming the runner-up by approximately 0.63 percentage points. For VQA-2019, our model registered a significant overall accuracy of 81.6%, which represents a 1.8 percentage point augmentation compared to the second-best performing model. The M2I2 model exhibited proficiency in answering closed-ended questions but showed limitations with open-ended question types, potentially attributable to disparities in pre-training datasets. The Q2ATransformer [21] and MUMC [14] models were precluded from our comparison due to the unavailability of their source code, checkpoints, and pre-training datasets, which hindered reproducibility of their results. Moreover, the MeDVInT [34] and LLaVA-Med [13] models possess a parameter count exceeding 7 billion, nearly 17 times that of our LaPA model (0.405B). Despite some superior results from these models, we posit that the comparison would not be equitable due to the vast difference in model size and

complexity. Consequently, these models were also excluded from our comparative analysis.

4.4. Ablation Study

In this section, we present an ablation study designed to evaluate the impact of each module within our proposed methodology. The results are summarized in Tab. 2, encompassing three benchmark datasets. We utilize the following abbreviations: BL. for baseline, GM. for the latent prompt generation module (detailed in Section 3.1), LF. for the latent prompt fusion module (described in Section 3.2), and PF. for the prior knowledge fusion module (elucidated in Section 3.3). The notation $w/o\ cs$ specifies configurations that omit the consistency method, which allows for the assessment of its effectiveness. The concluding line quantifies the enhancement our LaPA model offers over the baseline.

Due to the indirect interaction of the latent prompt generation module with image and language modalities, we investigate its influence by conducting an ablation study in conjunction with the GM. and LF. modules. The comparison between conditions #1 and #2 in Tab. 2 demonstrates

η	0.01	0.05	0.1	0.5	1
VQA-RAD	71.84	72.28	**79.38**	72.28	72.95
SLAKE	83.69	83.60	**84.73**	83.22	83.60
VQA-2019	80.27	78.40	**81.60**	78.40	80.80

Table 5. Ablation on the η.

that the integration of the latent prompt markedly enhances the model's capability in addressing Med-VQA tasks. Further, we examine the efficacy of the consistency method; the comparative improvement of condition #3 over #2 underscores its utility. The incorporation of the prior knowledge fusion module further augments model performance (Comparison #4 and #3). Ultimately, the amalgamation of all enhancements into the baseline model culminates in a substantial performance leap, as evidenced in condition #5. The aggregate improvement across all three benchmarks is nearly 3% relative to the baseline, as detailed in the concluding line of our ablation analysis.

Ablation on the θ and β. The hyperparameters θ and β are pivotal in modulating the interaction of cross-modal information, subsequently influencing the accuracy of the final predictive responses. Fig. 3 employs a triad of heatmaps to elucidate the effects of various θ and β coefficients on the fusion of cross-modal image and language features within three benchmark datasets. With the coefficient for the latent prompt (α) held constant at 1 and the latent prompt size fixed at 32, we systematically vary θ and β from 0.01, through 0.1, to 1 to assess their impact on model performance. The visual representation in Fig. 3 indicates that the combination of $\beta = 0.1$ and $\theta = 0.1$ is optimal across all three evaluated datasets.

Ablation on the interaction order. The sequence of interactions within the latent prompt fusion module exerts a direct influence on the efficacy of information extraction via the latent prompts. Tab. 3 delineates the impact of various fusion sequences on the accuracy of the resultant outputs. It is observed that the optimal fusion sequence commences with language, subsequently incorporates image modality, and concludes with a multi-modal fusion, thereby yielding the most favorable outcomes.

Ablation on the latent prompt size. The size of the latent prompt critically determines the parameter count within the latent prompt framework. Tab. 4 presents an analysis of how varying the latent prompt size from 4 to 256 influences performance across the three benchmark datasets. Initially, an increase in latent prompt size correlates with enhanced performance across benchmarks. However, a decline in model accuracy is observed when the latent prompt exceeds a size of 32. The optimal performance, as evidenced by accuracy

metrics, is achieved with a latent prompt size of 32 across all evaluated datasets. We hypothesize that excessively large latent prompt dimensions may introduce superfluous and potentially disruptive noise into the information extraction process, thereby detrimentally impacting the precision of the final answer prediction in the Med-VQA context.

Ablation on the η. The hyperparameter η exerts a direct influence on the weighting of the consistency loss within the aggregate loss function. Tab. 5 illustrates the impact of varying η from 0.01 to 1 on the overall performance across three benchmark datasets. The empirical results indicate that setting η to 0.01 yields the most favorable outcomes on all three benchmarks.

4.5. Qualitative Analysis

To further elucidate the efficacy of our Latent Prompt Assist (LaPA) model, a qualitative analysis was conducted on six Medical visual question answering (Med-VQA) instances, specifically three from the VQA-RAD dataset (cases a, b, c) and three from the SLAKE dataset (cases d, e, f), as depicted in Fig. 4. Examination of cases a, b, c, d, and f reveals that the incorporation of latent prompts facilitates the model in accurately responding to both closed-ended and open-ended queries across the two benchmarks. However, in case e, the model's integration of solely the latent prompt proved insufficient for distinguishing between two highly similar responses. The addition of the prior Knowledge fusion module (PF.) was instrumental in rectifying the model's response. These six cases collectively demonstrate that our proposed enhancements substantively bolster the model's performance in resolving both closed-ended and open-ended VQA challenges.

5. Conclusion

This study introduces a novel Latent Prompt Assist (LaPA) model designed to enhance the accuracy of responses in the domain of medical visual question answering (Med-VQA). It employs the latent prompt to filter different modal information and extract clinic-relevant information to assist in predicting the final answer. Our innovative framework entails a latent prompt generation module that synthesizes latent prompts under the constraint of target answer tokens. These prompts are then integrated with both uni-modal and multi-modal information streams to isolate clinical insights. Further, the model incorporates prior knowledge encapsulated in a knowledge graph, detailing disease-organ relationships, to interact with the latent prompt and refine the final answer prediction. Empirical validation of our approach across three well-established benchmarks demonstrates its superiority in generating accurate answers within the Med-VQA context. Looking forward, we aim to deploy the latent prompt mechanism within

(a) **Question**: Has the midline of the mediastinum shifted?
Baseline	: yes
+ GM. & LF.	: no
+ GM. & LF. + PF.	: no

(b) **Question**: Does this look like a healthy liver?
Baseline	: no
+ GM. & LF.	: yes
+ GM. & LF. + PF.	: yes

(c) **Question**: What structures are involved?
Baseline	: brain
+ GM. & LF.	: caudate putemen left parietal
+ GM. & LF. + PF.	: caudate putemen left parietal

(d) **Question**: Which part of the body does this image belong to?
Baseline	: abdomen
+ GM. & LF.	: chest
+ GM. & LF. + PF.	: chest

(e) **Question**: Where is/are the abnormality located?
Baseline	: left lung right
+ GM. & LF.	: left lung right
+ GM. & LF. + PF.	: right lung right

(f) **Question**: Is the brain healthy?
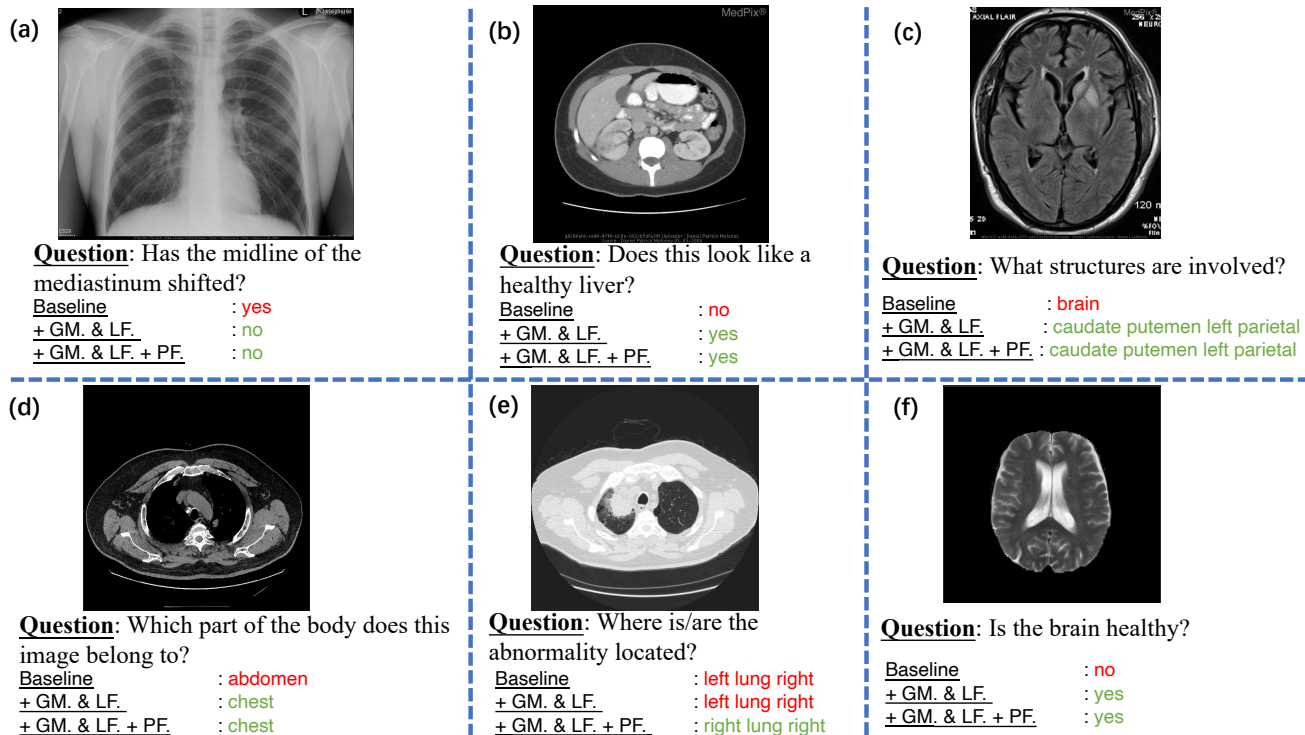|	
Baseline	: no
+ GM. & LF.	: yes
+ GM. & LF. + PF.	: yes

Figure 4. Six examples of the LaPA model that use different modules to do the ablation study. Instances a, b, and c are extracted from the VQA-RAD dataset, whereas instances d, e, and f originate from the SLAKE dataset. Within the provided illustrations, responses are annotated with green to denote correctness and with red to signify erroneous predictions by the model. The GM., LF., and PF. are the abbreviations of the latent prompt generation module, latent prompt fusion module, and prior knowledge fusion module.

a large-scale, highly-parameterized model to fully explore the potential of latent prompts in complex inference tasks.

References

[1] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. *CLEF (working notes)*, 2(6), 2019. 2, 5

[2] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 679–689. Springer, 2022. 1, 5

[3] Zhihong Chen, Guanbin Li, and Xiang Wan. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5152–5161, 2022. 5

[4] Tuong Do, Binh X Nguyen, Erman Tjiputra, Minh Tran, Quang D Tran, and Anh Nguyen. Multiple meta-model quantifying for medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg,*

France, September 27–October 1, 2021, Proceedings, Part V 24, pages 64–74. Springer, 2021. 1, 2

[5] Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv preprint arXiv:2112.13906*, 2021. 5

[6] Haifan Gong, Guanqi Chen, Mingzhi Mao, Zhen Li, and Guanbin Li. Vqamix: Conditional triplet mixup for medical visual question answering. *IEEE Transactions on Medical Imaging*, 41(11):3332–3343, 2022. 1

[7] Irving John Good. Rational decisions. *Journal of the Royal Statistical Society: Series B*, 14(1):107–114, 1952. 4

[8] Tiancheng Gu, Dongnan Liu, Zhiyuan Li, and Weidong Cai. Complex organ mask guided radiology report generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7995–8004, 2024. 1

[9] Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. PPT: Pre-trained prompt tuning for few-shot learning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022. 1, 2

[10] Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021*

IEEE 18th International Symposium on Biomedical Imaging, pages 1033–1036. IEEE, 2021. 1, 2, 5

[11] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear Attention Networks. In *Advances in Neural Information Processing Systems 31*, pages 1571–1581, 2018. 1, 5

[12] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. 2, 5

[13] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 6

[14] Pengfei Li, Gang Liu, Jinlong He, Zixu Zhao, and Shenjun Zhong. Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 374–383. Springer, 2023. 6

[15] Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong. Self-supervised vision-language pretraining for medial visual question answering. In *2023 IEEE 20th International Symposium on Biomedical Imaging*, pages 1–5. IEEE, 2023. 1, 5

[16] Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, page 102611, 2023. 1

[17] Bo Liu, Li-Ming Zhan, and Xiao-Ming Wu. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 210–220. Springer, 2021. 5

[18] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging*, pages 1650–1654. IEEE, 2021. 2, 4, 5

[19] Bo Liu, Li-Ming Zhan, Li Xu, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning and contrastive learning. *IEEE transactions on medical imaging*, 42(5):1532–1545, 2022. 1, 5

[20] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3, 4

[21] Yunyi Liu, Zhanyu Wang, Dong Xu, and Luping Zhou. Q2atransformer: Improving medical vqa via an answer querying decoder. In *International Conference on Information Processing in Medical Imaging*, pages 445–456. Springer, 2023. 1, 2, 6

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 4

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018. 5

[24] Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. Overcoming data limitation in medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 522–530. Springer, 2019. 1, 2

[25] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. 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: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 180–189. Springer, 2018. 5

[26] Jiaren Peng, Wenzhong Yang, Fuyuan Wei, and Liang He. Prompt for extraction: Multiple templates choice model for event extraction. *Knowledge-Based Systems*, page 111544, 2024. 1, 2

[27] Morgan D. Polak, M.P. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nat Commun 15, 1569 (2024).*, 2024. 2

[28] Sanjay Subramanian, Lucy Lu Wang, Sachin Mehta, Ben Bogin, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. Medicat: A dataset of medical images, captions, and textual references. *arXiv preprint arXiv:2010.06000*, 2020. 1, 5

[29] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442, 2023. 1

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[31] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. 4

[32] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016. 1

[33] Anda Zhang, Wei Tao, Ziyan Li, Haofen Wang, and Wenqiang Zhang. Type-aware medical visual question answering. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4838–4842. IEEE, 2022. 1, 2

[34] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Vi-

sual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. 3, 6

[35] Yubo Zhang, Xingxing Zhang, Xun Wang, Si qing Chen, and Furu Wei. Latent prompt tuning for text summarization, 2022. 1, 2

[36] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI open*, 1:57–81, 2020. 4

[37] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 2

[38] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2