

## A. Dataset Details

Dataset	Image	Description	QA-Pairs	Type	Usage
VQA-RAD [22]	315	315	3,515	Radiology	Downstream Fine-tuning & Evaluation & RAG
SLAKE [29]	642	642	7,000	Radiology	Downstream Fine-tuning & Evaluation & RAG
PathVQA [14]	4,998	4,998	32,799	Pathology	Downstream Fine-tuning & Evaluation & RAG
ROCO [38]	70,306	70,306	—	Radiology	Retrieval Augmentation
PEIR [40]	33,572	33,572	—	Pathology	Retrieval Augmentation
PubMedVision [5]	1,009,700	647,031	647,031	Comprehensive	Alignment & Instruct Tuning

Table 4. **Dataset statistics**, including the number of images, image descriptions, QA-pairs, and data types are detailed. For MedVQA data, only the **training set** is used for knowledge retrieval.

## B. Prompting GPT-4 to Convert Question and Answer into Caption

Considering that BioMedCLIP [55] is trained on paired medical images and textual captions, we reformulated the question-answer pairs in the MedVQA dataset into concise captions to ensure semantic consistency. Specifically, questions with answers like “no” were excluded, as such negative responses provide no meaningful information for generating captions. Figure 7 illustrates the GPT-4 prompt used for converting question-answer pairs into coherent and clinically relevant captions.

GPT-4 Prompt:
<b>Instruction:</b> Combine the following pairs of questions and answers into concise sentences while maintaining clarity and relevance. Use proper grammar and punctuation. <b>Example Input:</b> Are regions of the brain infarcted? yes Where is the pathology in this image? vasculature What type of imaging is this? diffusion weighted MR What organ system is pictured? the brain <b>Expeted Ouput:</b> The brain shows infarcted regions in the vasculature, as seen in diffusion-weighted MRI.

Figure 7. **GPT-4 prompt for converting question-answer pairs into concise medical captions.**

## C. Prompting GPT-4 to Generate Expert Annotations

Expert-CFG requires experts to predefine semantically coherent and query-relevant keywords or keyword groups as highlights. However, relying entirely on experts for this task is clearly impractical. To address this, we propose a semi-automated pipeline that combines expert guidance with automated keyword extraction techniques to streamline the process. To further optimize token usage, we divide the process into two steps: 1) extracting keywords from each caption in the knowledge base (see Figure 8); 2) matching these extracted keywords with the query to identify the most relevant highlights (see Figure 9). Since there is no keyword annotation in the existing MedVQA, we only preliminarily evaluated the performance of this method on the VQA-RAD dataset with the smallest amount of data, as shown in Table . A recall rate of over 98% indicates that GPT-4o can extract keywords from captions well.

Method	Keyword Extraction on VQA-RAD		
	Recall	Precision	F1
GPT4o	89.57	98.36	93.76

Table 5. Performance of GPT-4o in extracting keyword from captions.

<b>GPT-4 Prompt:</b>
<b>Instruction:</b> As a medical expert, extract and highlight key medical terms or phrases from the given question and caption. Focus on identifying specific diseases, locations, and relevant contextual information. Separate multiple terms with a semicolon (;).
<b>Example Input:</b> This MRI image was taken using T2-weighted MRI technology and shows a transverse plane view of the head. The image primarily shows the head organ system, with the left brain lobe revealing abnormalities while the right lobe appears normal.
<b>Expeted Ouput:</b> T2-weighted MRI; Transverse Plane; Head; The Left Brain Lobe; Abnormalities; The Right Brain Lobe

Figure 8. GPT-4 prompt for extracting medical terms from captions.

<b>GPT-4 Prompt:</b>
<b>Instruction:</b> Given a query, select the most relevant keywords from the provided list and highlight them. If multiple keywords are relevant, separate them with a semicolon (;).
<b>Example Input:</b> <b>Query:</b> Is this an MRI image with T2 weighted? yes <b>Keywords:</b> T2-weighted MRI; Transverse Plane; Head; The Left Brain Lobe; Abnormalities; The Right Brain Lobe. <b>Expeted Ouput:</b> T2-weighted MRI

Figure 9. GPT-4 prompt for highlighting relevant key medical terms based on a query.

## D. Discussion on Uncertainty Estimation Metric

For uncertainty estimation, we selected the entropy of the generated answer. This choice was motivated by a combination of effectiveness, efficiency, and system compatibility. Specifically, entropy proved to be a reliable uncertainty indicator for our model and integrates seamlessly with our Expert-CFG framework. A key advantage is its computational efficiency: the entropy can be calculated from a single forward pass of our VLM without requiring any architectural modifications or additional inference costs. We considered alternative methods, such as those based on multiple forward passes (e.g., Monte Carlo dropout). However, these were deemed unsuitable as they introduce significant computational overhead, conflicting with our design goals. Our approach aligns with recent findings that favor direct, single-pass uncertainty measures for large language models [44].

	Method	VQA-RAD(Closed)				SLAKE(Closed)				Implementation
		ECE↓	ECE-t↓	BS-t↓	AUC↑	ECE↓	ECE-t↓	BS-t↓	AUC↑	Details & Explanation
Phi3.5V-Med Zero-Shot	Label prob.	0.333	0.042	0.230	0.636	0.117	0.039	0.196	0.691	do_sample=10, temp=1
	Is True prob.	0.174	0.095	0.201	0.727	0.294	0.094	0.212	0.617	$\mathcal{Q} + \mathcal{A}$ “is True?”
	Entropy	0.108	0.037	0.145	0.879	0.119	0.066	0.148	0.827	$\mathcal{A}$ token prob (greedy)
Phi3.5V-Med Fine-Tune	Label prob.	0.089	0.019	0.077	0.889	0.082	0.026	0.071	0.895	This model can not follow “is True” prompt
	Entropy	0.068	0.024	0.066	0.958	0.073	0.024	0.063	0.949	
HuatuoV-32B Zero-Shot	Label prob.	0.266	0.200	0.231	0.811	0.190	0.056	0.154	0.768	Stronger instruct following. Fixed outputs(e.g., 0.95).
	Is True prob.	0.114	0.116	0.172	0.828	0.082	0.074	0.151	0.851	
	Verb1s top-1	0.166	0.001	0.166	0.754	0.141	0.053	0.145	0.808	
	Entropy	0.041	0.044	0.141	0.886	0.055	0.048	0.126	0.885	

Table 6. Uncertainty metrics on VQA-RAD / SLAKE (“yes/no” QA) under zero-shot / fine-tuned settings.

## E. Entropy and Retrieval Hit Rate of Answers

The relationship between entropy and MedVLM outputs was explored by analyzing the high-entropy answers across three MedVQA datasets, as illustrated in Figure 10-12 (a). It can be observed that questions involving pronouns like “this” and “their,” attributes such as “small” and “large,” numerical or size-related types, and positional references like “left” and “right” exhibit higher uncertainty. Specifically, for the PathVQA dataset, answers like “foot,” “face,” “breast,” and “blood” often correspond to ambiguous questions such as “What is present?” Similarly, for the VQA-RAD dataset, inconsistencies between abbreviations like “pa” and their full forms “posterior-anterior” also contribute to high uncertainty.

To further investigate whether entropy correlates with the information present in the dataset, we analyzed the low-hit-rate answers, as illustrated in Figure 10-12 (b). A notable correlation between high-entropy answers and low hit rates is observed. For numerical and attribute-related answers, most are not directly retrieved. Moreover, expressions not appearing in the training data, such as “PA,” also result in lower hit rates.

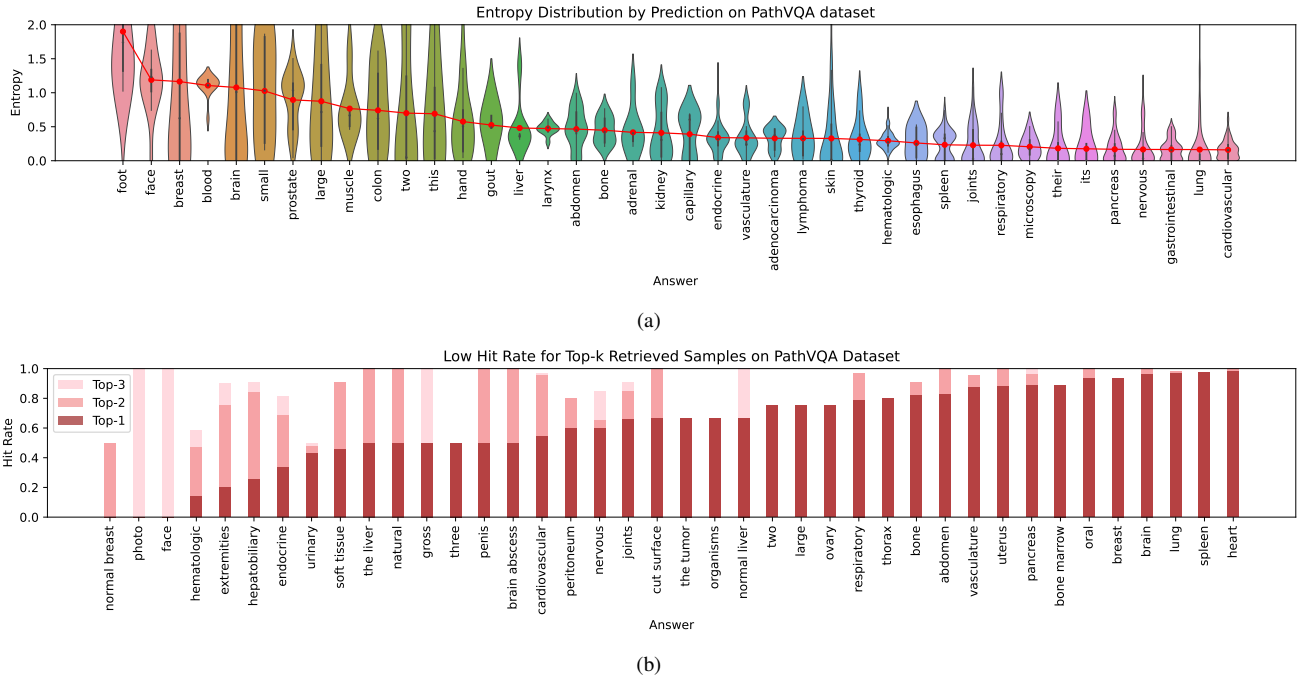


Figure 10. **PathVQA Dataset:** (a) Entropy Distribution and (b) Overlapping Answer Hit Rate for Top-k.

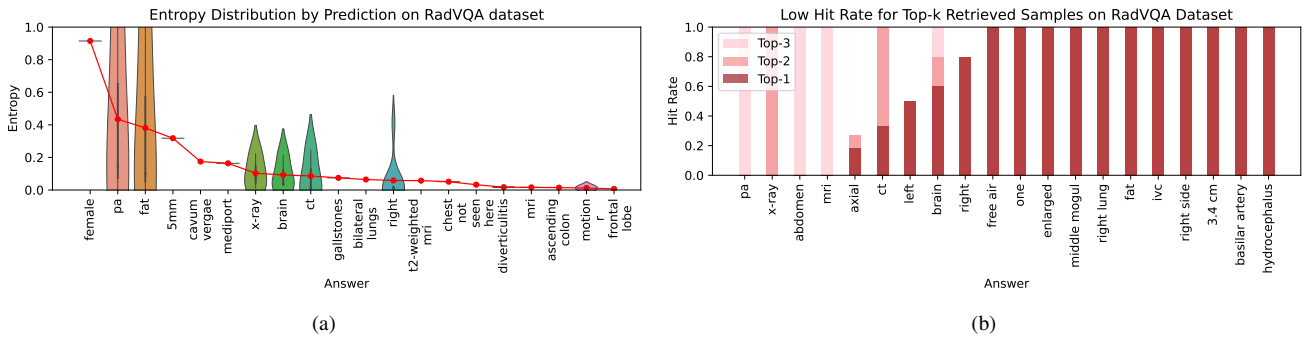


Figure 11. **VQA-RAD Dataset:** (a) Entropy Distribution and (b) Overlapping Answer Hit Rate for Top-k.

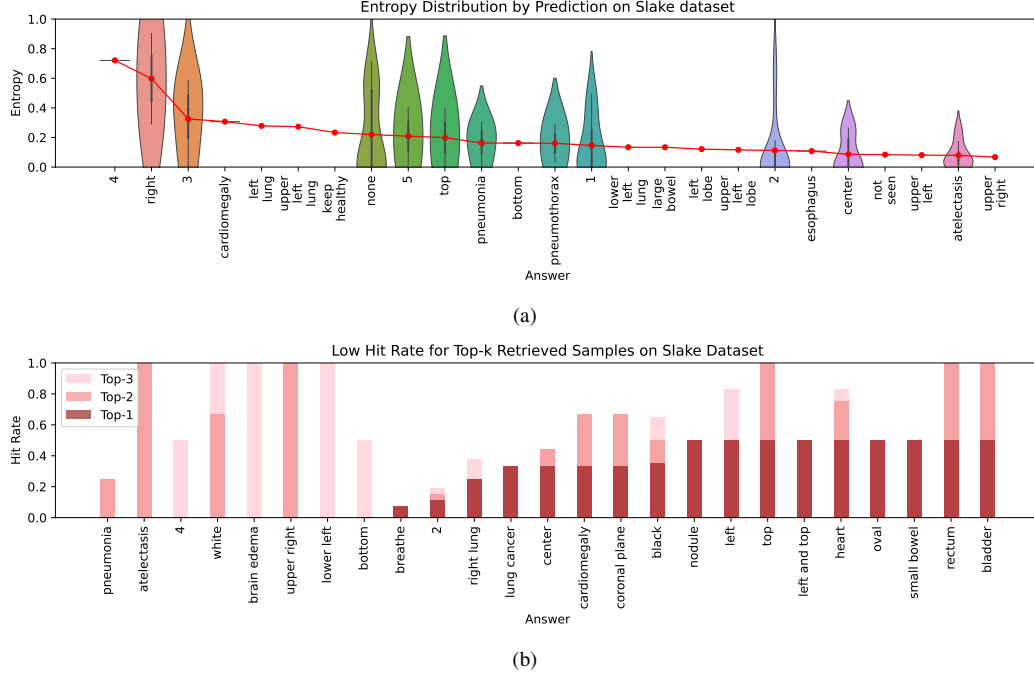


Figure 12. **Slake dataset**: (a) Entropy Distribution and (b) Overlapping Answer Hit Rate for Top-k.

## F. Knowledge Retrieval

Our knowledge retrieval process utilizes CLIP scores for a preliminary relevance assessment. We rank reference texts by their CLIP scores and select the Top-K most relevant ones. This Top-K strategy is employed instead of applying a fixed score threshold, as any such threshold is highly sensitive to sample variations and lacks generalizability. The selected texts are then merged and provided as input to the GPT-4 API, whose sole function is to extract relevant keywords **k** for highlighting purposes. Figure 13 illustrates the significant overlap in CLIP scores between documents that contain an explicit answer (“Hits”) and those that do not (“Misses”), demonstrating the unreliability of a simple threshold. It is noteworthy that a ‘Miss’ does not necessarily imply irrelevance. Our empirical findings on the current datasets suggest that documents with CLIP scores below 0.6 are generally not informative.

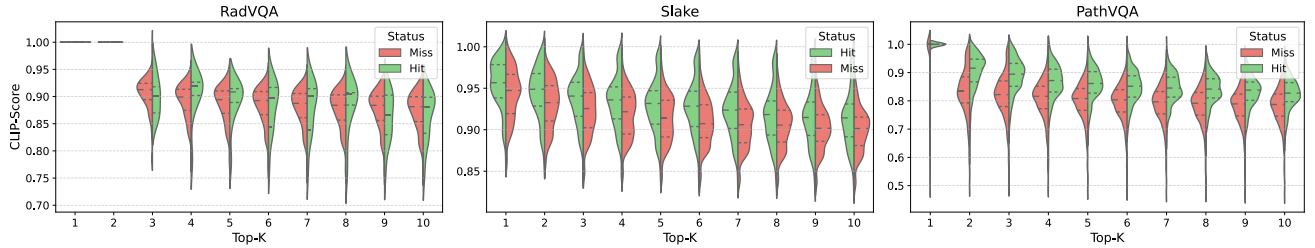


Figure 13. CLIP score distributions for open-ended QA hits and misses.

## G. Hyperparameter Ablation of Classifier-Free Guidance

The hyperparameters  $\alpha$ ,  $\beta$ , and  $\gamma$  represent the weights applied to different components in the Expert-CFG framework. Specifically,  $\alpha$  adjusts the weight of the token embeddings in the normal branch,  $\beta$  modulates the attention probability weights in the normal branch, and  $\gamma$  scales the weight applied to the logits. The hyperparameter settings aim to balance the strength of alignment between the original textual output and expert annotations. Table 7 compares the overall accuracy on the VQA-RAD, SLAKE, and PathVQA datasets for various hyperparameter configurations, highlighting the impact of each parameter on performance. When  $\alpha = 0$ , Eq (7) degenerates to fully masking the conditional input, leading to incomplete context and performance degradation.

VQA-RAD Overall Accuracy				Slake Overall Accuracy				PathVQA Overall Accuracy			
Phi3.5-V 4.2B	82.04			Phi3.5-V 4.2B	85.20			Phi3.5-V 4.2B	65.24		
w/ 100% RAG	81.15			w/ 100% RAG	85.58			w/ 100% RAG	72.90		
w/ 100% Expert-RAG	95.34			w/ 100% Expert-RAG	97.36			w/ 100% Expert-RAG	88.46		
w/ 100% Expert-CFG	<b>Hyper Parameters</b>			w/ 100% Expert-CFG	<b>Hyper Parameters</b>			w/ 100% Expert-CFG	<b>Hyper Parameters</b>		
$\alpha$ (in Eq. (7))	0	0.01	0.1	$\alpha$ (in Eq. (7))	0	0.01	0.1	$\alpha$ (in Eq. (7))	0	0.01	0.1
( $\alpha$ , 3.0, 1.3)	96.23	<b>98.44</b>	97.11	( $\alpha$ , 3.0, 1.3)	98.20	<b>99.62</b>	98.58	( $\alpha$ , 3.0, 1.3)	91.36	<b>94.76</b>	89.94
$\beta$ (in Eq. (9))	1	3	5	$\beta$ (in Eq. (9))	1	3	5	$\beta$ (in Eq. (9))	1	3	5
(0.01, $\beta$ , 1.3)	98.22	<b>98.44</b>	98.00	(0.01, $\beta$ , 1.3)	<b>99.62</b>	<b>99.62</b>	99.43	(0.01, $\beta$ , 1.3)	93.64	<b>94.76</b>	92.63
$\gamma$ (in Eq. (8))	1	1.3	1.5	$\gamma$ (in Eq. (8))	1	1.3	1.5	$\gamma$ (in Eq. (8))	1	1.3	1.5
(0.01, 3.0, $\gamma$ )	98.00	<b>98.44</b>	97.11	(0.01, 3.0, $\gamma$ )	99.24	<b>99.62</b>	98.86	(0.01, 3.0, $\gamma$ )	93.09	<b>94.76</b>	92.03

Table 7. Comparison of Overall Accuracy with Different Hyperparameter Settings for VQA-RAD, Slake, and PathVQA.

## H. Zero-Shot Cases of Baseline MedVLM and Expert-CFG

We conducted additional zero-shot case studies on images and questions from three datasets: VQA-RAD, SLAKE, and PathVQA. While the baseline MedVLM’s responses are relevant to the input images, they exhibit certain errors due to a lack of knowledge or fine-grained understanding. For instance, the baseline model fails to distinguish between “fat” and “fat stranding,” recognize fine-grained details such as EKG electrodes, or differentiate between “T1-weighted MRI” and “T2-weighted MRI,” among other issues. These limitations highlight the importance of incorporating external knowledge and expert guidance for accurate and clinically reliable outputs.



Expert highlights the reference text



Expert manually adds additional text


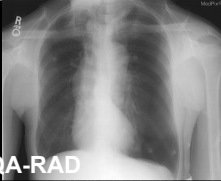
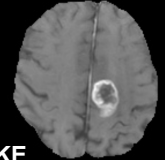
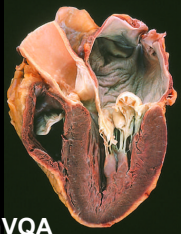
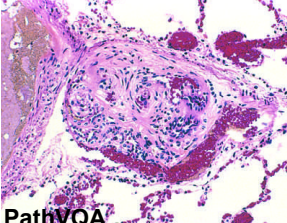
 <p>VQA-RAD</p>	<p><b>Question:</b> Is there fat stranding? <b>Ground Truth:</b> No</p> <p><b>Knowledge Reference (VQA-RAD):</b> The hypoattenuated tissue near or in the skin is fat. A diagnosis of appendicitis cannot be made from this image. <b>No fat stranding.</b></p> <p><b>Original Answer:</b> <b>Yes</b>, there is fat stranding visible in the image, indicating the presence of inflammation or edema in the surrounding fat tissue.</p> <p><b>Expert-CFG:</b> <b>No</b>, there is no fat stranding visible in the ct scan image.</p>
 <p>VQA-RAD</p>	<p><b>Question:</b> Are there EKG leads on this patient? <b>Ground Truth:</b> Yes</p> <p><b>Knowledge Reference (VQA-RAD):</b> The descending aortic silhouette has a normal contour and size. There is tortuosity of the descending aorta. <b>EKG electrodes are visible.</b></p> <p><b>Original Answer:</b> <b>No</b>, there are no EKG leads visible on this patient's radiographic image</p> <p><b>Expert-CFG:</b> <b>Yes</b>, the EKG leads are visible on the patient in the image. They are located on the right side of the chest, indicating the presence of a pacemaker.</p>
 <p>SLAKE</p>	<p><b>Question:</b> What is the mr weighting in this image? <b>Ground Truth:</b> T1</p> <p><b>Knowledge Reference (SLAKE):</b> This MRI image was taken using <b>T1-weighted MRI</b> technology and shows a transverse plane view of the head.</p> <p><b>Original Answer:</b> The MR weighting in this image is <b>T2-weighted MRI</b> technology.</p> <p><b>Expert-CFG:</b> The image is weighted T1 MR, which means it is a <b>T1-weighted</b> magnetic resonance imaging (MRI) scan.</p>
 <p>PathVQA</p>	<p><b>Question:</b> What is shown on the right in this four-chamber view? <b>Ground Truth:</b> The left ventricle</p> <p><b>Knowledge Reference (PathVQA):</b> The shape and volume of <b>the left ventricle</b> have been distorted by the enlarged right ventricle.</p> <p><b>Original Answer:</b> The image displays a four-chamber view of a human heart, where the right atrium and right ventricle are visible. The right atrium is positioned on the left side of the image, and <b>the right ventricle</b> is located on the right side.</p> <p><b>Expert-CFG:</b> The image shows a four-chamber view of the heart, with the right ventricle appearing significantly enlarged. This enlargement has caused a distortion in the shape and volume of the left ventricle, indicating a pathological condition affecting the heart's structure and function.</p>
 <p>PathVQA</p>	<p><b>Question:</b> What is seen in small arteries? <b>Ground Truth:</b> Plexiform lesion</p> <p><b>Knowledge Reference (PEIR):</b> In the image, there is a histological view of the respiratory system specifically showing the lung tissue. At lower magnification, typical grade 3 lesions are visible, including a <b>plexiform lesion</b> and surrounding dilation lesions.</p> <p><b>Original Answer:</b> In the small arteries, there is a proliferation of smooth muscle cells and a thickened intima, which is the innermost layer of the artery.</p> <p><b>Expert-CFG:</b> In the small arteries, there is a <b>plexiform lesion</b>, which is a type of arteriosclerosis characterized by a complex, tangled appearance of the arterial wall.</p>

Figure 14. Zero-shot case study examples from three datasets: VQA-RAD, SLAKE, and PathVQA.

# I. Additional Zero-Shot Chinese Capability of Baseline MedVLM and Expert-CFG

We were surprised to discover that Phi3V-Med demonstrated zero-shot capability in Chinese as shown in Figure 15, although the baseline outputs in Chinese and English for the same question were inconsistent. For example, the Chinese response was able to correctly recognize “fat stranding,” whereas the English response failed to do so. Expert-CFG also proved effective in incorporating expert annotations in Chinese.

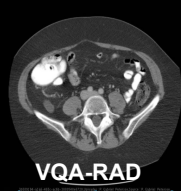
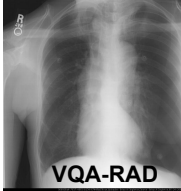

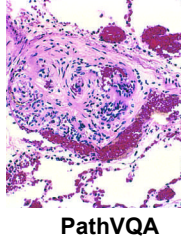
 VQA-RAD	<p>Question: 图中存在脂肪束吗?</p> <p>Ground Truth: 否</p> <p>Original Answer: 在这张图像中, 脂肪组织的密度与周围组织相似, 没有明显的增强区域, 表明<b>没有显著的脂肪束</b>。</p>	 VQA-RAD	<p>Question: 该患者有心电电极导联吗?</p> <p>Ground Truth: 是</p> <p>Original Answer: 根据提供的X光片, <b>没有显示心电电极导联的迹象</b>。</p> <p>Expert-CFG: 根据提供的X光片, 患者胸部右侧<b>存在心电电极导联</b>。</p>
 PathVQA	<p>Question: 此四腔视图右侧显示的是什么?</p> <p>Ground Truth: 左心室</p> <p>Original Answer: 这张四腔视图显示的是<b>右心室</b>的内部结构。可以看到<b>右心室的壁层厚</b>, 呈现出不同的颜色和质地, 这可能是由于.....</p> <p>Expert-CFG: 这张四腔视图显示<b>左心室</b>的形状和体积已受到<b>扩大的右心室</b>的影响而发生了扭曲。右心室的增大压迫了左心室, 导致.....</p>	 PathVQA	<p>Question: 在小动脉中可以看到什么?</p> <p>Ground Truth: 丛状病变</p> <p>Original Answer: 在这张显微镜下的小动脉中, 可以清楚地看到一个异常的细胞团块, 这种团块被称为<b>肉芽肿</b>。肉芽肿是由于炎症.....</p> <p>Expert-CFG: 在这张显微镜下的小动脉中, 可以清楚地看到一个异常的细胞团块, 这种团块被称为<b>丛状病变</b>。这个病变由多个异型的.....</p>

Figure 15. Additional zero-shot chinese case study examples from three datasets: VQA-RAD, SLAKE, and PathVQA.

## J. Low-Entropy Error Analysis and Mitigation Pathways

While our uncertainty estimation achieves  $AUC > 0.8$  across benchmarks, persistent low-entropy errors reveal critical directions for improvement. Excluding ambiguous cases such as questions like “What does this image show?” and answers like “CT” or its full name, errors predominantly stem from anatomical positional confusion. For example, the inability to distinguish left from right due to the distinction between PA/AP views accounts for 21.73% and 42.02% in the VQA-RAD and SLAKE datasets respectively. In addition, there are several other issues. There is a lack of the concept of degree (such as “larger”, “mild”), and there is an inability to accurately handle numbers, which is an inherent problem of VLM. Disease misinterpretation accounts for 17.39% and 14.49%, and unlearned medical concepts (such as words that do not appear in the training set) account for 26.08% and 7.24% in the relevant datasets respectively. We attribute these to sparse visual-textual grounding in training data. For instance, insufficient annotations for radiographic projection labels hinder left-right discrimination. Addressing such issues requires fine-grained synthetic data augmentation (such as view-specific anatomical templates) and knowledge-anchored multi-turn dialogue to reinforce spatial reasoning.

Type	Examples of low entropy (<0.35) but incorrect results			Proportion (%)	
	Question	GT	Prediction (Entropy)	VQA-RAD (23)	SLAKE (69)
Question ambiguity	What is under the right hemidiaphragm?	free air	stomach bubble (0.011)	4.34	7.24
Synonymous answers	What modality is used to take this image?	xr	x-ray (0.013)	26.08	2.89
Position	Which side is more clearly visualized?	left	right (0.032)	21.73	42.02
Degree	Is the heart size in this image smaller or larger than if the image was taken AP?	smaller	larger (0.031)	4.34	5.79
Disease	What is abnormal about the pancreas?	enlarged	fatty infiltration (0.078)	17.39	14.49
Numbers	How many lungs have existed in this image?	2	1 (0.2524)	0.00	18.84
Shapes	What is the shape of larynx in this image?	oval	irregular(0.0046)	0.00	1.44
Not appeared in the training set	What organ system is the pathology?	lymphatic	cardiovascular (0.329)	26.08	7.24

Table 8. Examples of Low-entropy (<0.35) incorrect results.

## K. Limitations and Future Works

Despite the fact that our framework has achieved state-of-the-art performance on three MedVQA datasets, due to the significant gaps in different sub-fields within the medical domain, it is necessary to further expand the experimental evaluation. Besides, the greedy decoding strategy may not be globally optimal for long sequences. However, currently, there is a lack of long-sequence benchmarks in the medical field for further research on the estimation of uncertainty in and knowledge representation of MedVLMs. Meanwhile, the framework we proposed provides the medical community with a cost-effective method to align MedVLMs with expert knowledge. It can be combined with high-accuracy classifiers or detectors to synthesize data to further improve the performance and generalization ability of the model.