



# MedFG-VQA: Low-Frequency Memory and Graph Attention for Lightweight Medical VQA

## Supplementary Material

This supplementary material provides details on the construction of SynMedVQA, including the prompting strategy for VQA generation and the formatting of both open-ended and close-ended questions. We also analyze the dataset’s linguistic features, including word frequency and question length. Finally, additional visual comparisons are included to support the quantitative results in the main paper.

### 1. SynMedVQA Construction Details

**Prompt for VQA Generation.** As mentioned in §4.2, we provide the detailed prompts used to generate VQA data with GPT-4o [1]. For each image, we create five multiple-choice and five open-ended questions. The multiple-choice prompts cover four key perspectives: image features, anatomical structures, pathological manifestations, and clinical significance. Each question lists ten options, consisting of five correct answers and five distractors. This structure supports systematic generation and future expansion. Figures 4 and 5 illustrate the prompts used. To ensure the output is easy to parse, we enforce a strict JSON format using example templates.

**Q&A Reformation.** Following data generation, we format the questions to match the strategy used in NanoVLM [9]. Open-ended questions use the generated Q&A pairs directly with the template “<Question> Answer:”. For close-ended questions, we apply a specific mix: 75% contain one correct answer and three distractors, while the remaining 25% consist of four distractors only, requiring a “no answer” response. This design helps the model handle uncertainty and improves its reasoning. The template for close-ended questions is shown in Figure 1.

### 2. Benchmarks Details

**Word Frequency Analysis of Question.** Following [5], we analyze the word frequency of the questions in Figure 2. Frequent terms like *features*, *retinal*, *OCT*, *image*, *layer*, and *visual* reflect the dataset’s focus on anatomy and imaging, especially in the ophthalmic field. Meanwhile, words such as *fluid*, *lesion*, and *abnormal* highlight the emphasis on pathology and diagnosis. This distribution confirms that the questions cover diverse clinical aspects, including structural interpretation and disease assessment.

**Question Length Analysis.** Figure 3 shows the distribution of question lengths. Most questions fall between 10 and 20 words, indicating that the generated text is concise and con-

```
<Question>
A. <option1>
B. <option2>
C. <option3>
D. <option4>
Answer with the option’s letter from the given choices directly.
Answer:
```

Figure 1. Question template for close-ended question.

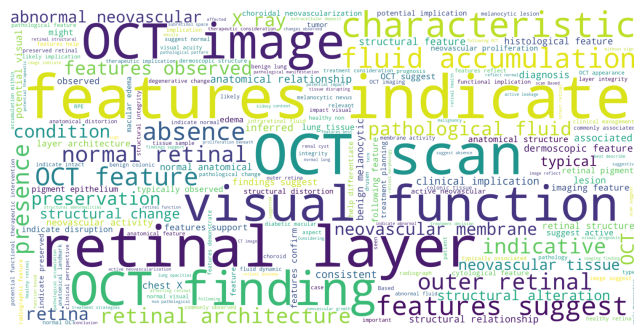


Figure 2. Word cloud of questions.

sistent. Some datasets, such as LungColon [3] and COVID-19 [6], show a wider range due to more complex descriptions. For clarity, we truncate the x-axis at the 95th percentile. These patterns confirm that our dataset maintains a balanced complexity across different medical modalities, keeping the questions natural and informative.

### 3. Fine-Tuned Baseline Results

The initial comparison evaluated existing models on our new benchmark. To ensure a fair architectural comparison, we fine-tuned the best baseline, *Qwen3-VL-4B* [2], on SynMedVQA using identical hyperparameters. As shown in the Table 1, our model **outperforms the fine-tuned baseline across all 11 categories**, with a **3.87% higher average accuracy**, despite having significantly fewer parameters. We will update our paper with these results.

### 4. Additional Evaluation Metrics

To provide a more comprehensive evaluation, we report the per-class Precision, Recall, and F1-score in Table 2, complementing the Accuracy metrics in the main paper. The no answer category shows high Precision (0.835), indicating conservative and clinically reliable refusal behavior, whereas its lower Recall (0.349) reflects the intrinsic difficulty of medical negative samples. The averaged Precision

Table 1. Comparison with the fine-tuned Qwen3-VL baseline on SynMedVQA. Dataset names are abbreviated due to space constraints, following the same order as Table 2.

Method	OCTMNIST	CHMNIST	LungColon	CTKidney	COVID-19	BTMRI	KneeXray	RETINA	Kvasir	DermaMNIST	BUSI	Average
Qwen3-VL (FT)	0.6498	<b>0.5533</b>	0.6109	0.5994	0.6214	0.5758	0.6114	<b>0.5475</b>	0.5252	0.6650	<b>0.6288</b>	0.6054
<b>Ours</b>	<b>0.7056</b>	0.5372	<b>0.6493</b>	<b>0.6239</b>	<b>0.6733</b>	<b>0.6155</b>	<b>0.6609</b>	0.5372	<b>0.5440</b>	<b>0.7225</b>	0.6284	<b>0.6441</b>

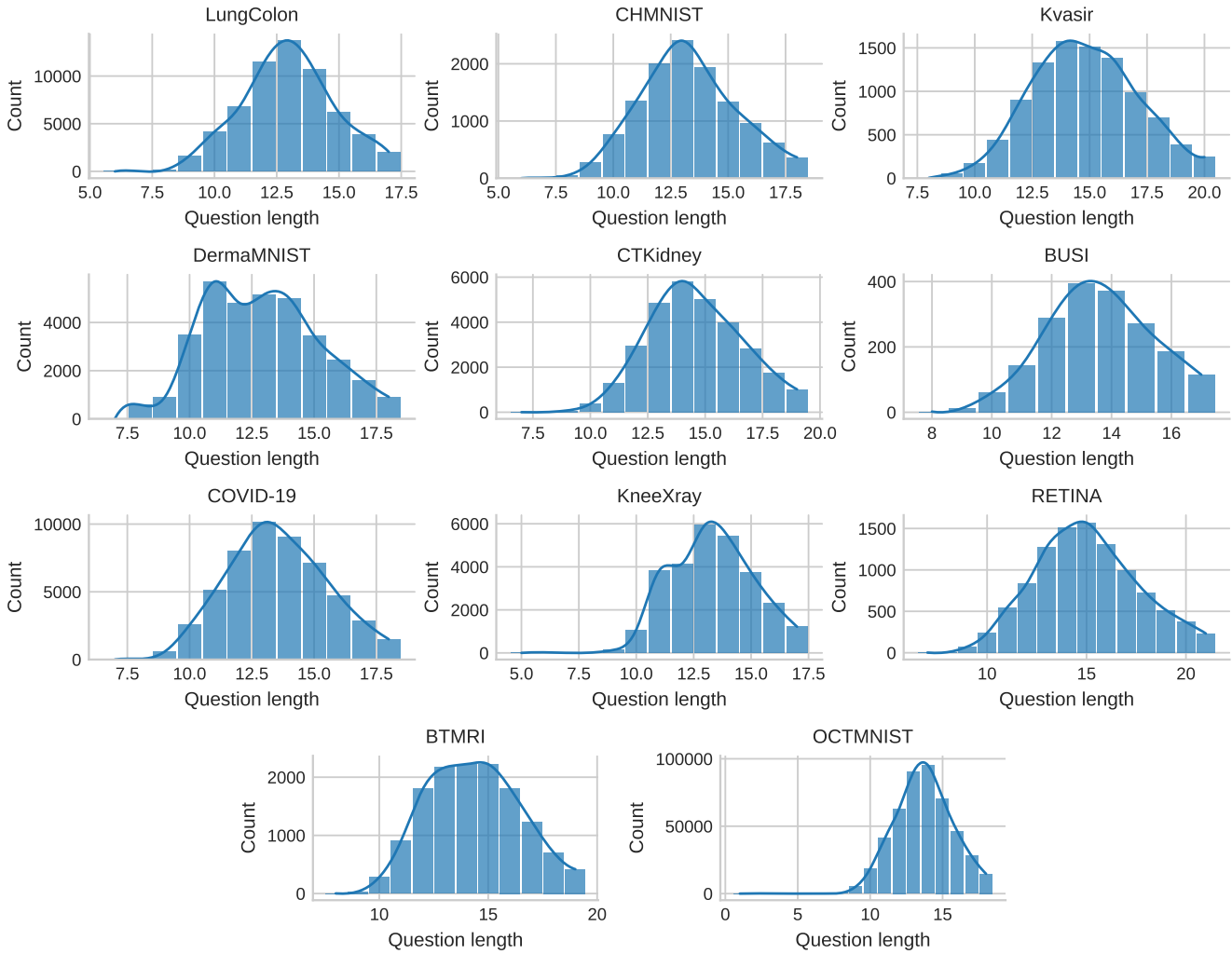


Figure 3. Detailed distribution of question lengths.

Table 2. Performance under different evaluation metrics.

Metric	A	B	C	D	N.A.	Avg.
P	0.5933	0.6702	0.6570	0.7042	0.8348	0.6919
R	0.7924	0.7313	0.7515	0.7113	0.3491	0.6671
F1	0.6786	0.6995	0.7011	0.7077	0.4923	0.6558

(0.692), Recall (0.667), and F1-score (0.656) further indicate balanced discrimination across answer options.

## 5. More Case Studies

Figures 6 and 7 present additional qualitative comparisons of model performance. Baseline models [4, 7, 8, 10, 11] tend to be verbose or speculative, often including unnecessary details or incorrect diagnoses. Conversely, our model delivers concise, clinically aligned answers that match the ground truth. This demonstrates better factual accuracy, fewer hallucinations, and a sharper focus on the diagnostic elements relevant to the query.

#### BTMRI Multiple-choice Questions Generation Prompt

**System:**

You are a senior medical imaging expert specializing in brain MRI interpretation and brain tumor diagnosis. You are tasked with creating professional multiple-choice Visual Question Answering (VQA) items based on provided brain MRI images and their tumor category. Follow strict medical common sense, use professional terminology, and ensure the questions are clinically meaningful.

**User:**

I will provide you with one brain MRI image and its tumor category.

Your task:

1. Generate FIVE different multiple-choice questions (VQA) based on the given image and tumor category.
2. Each question must focus on one of these aspects:
  - Image features (signal intensity, location, enhancement pattern, shape, etc.)
  - Anatomical structures visible in the image (sellar region, ventricles, white matter, etc.)
  - Pathological manifestations (mass effect, edema, invasion, compression of adjacent structures)
  - Clinical significance (possible symptoms, functional impact, treatment considerations)
3. The question stem must be clear, concise, and clinically meaningful.  
Do NOT directly mention the exact tumor category name. Instead, describe the imaging finding or relevant location.
4. Provide TEN answer options in an array of strings.
  - Each option should be a full statement (not starting with A/B/C...).
  - Exactly FIVE must be correct answers.
  - The incorrect options must be plausible but wrong.
5. The output must be valid JSON as an array of 5 objects:

```
{  
  "question": "...",  
  "options": ["...", "...", "...", "...", "...", "...", "...", "...", "...", "..."],  
  "answers": ["...", "...", "...", "...", "..."] # exactly 5 correct options  
}
```

Important:

- Do not include explanations or additional text outside the JSON.
- Do not mention dataset names or refer to the task itself.

Now, generate five questions for the following:

Tumor category: <CATEGORY\_NAME>

Figure 4. Example prompt for generating multiple-choice questions on the BTMRI dataset. Each question includes ten options: five correct answers and five distractors.

#### Q&A Generation Prompt

**System:**

You are a senior medical imaging expert with broad experience across multiple modalities including MRI, CT, X-ray, ultrasound, fundus photography, and histopathology.

You are tasked with creating professional question-answer pairs (VQA) based on provided medical images. Follow strict medical knowledge, use professional terminology, and ensure clinical relevance.

User:

I will provide one medical image and its category.

Your task:

1. Generate FIVE different question-answer pairs based on the given image and its category.
2. Each question should focus on one of these aspects:
  - Image Basics: modality, plane/section, anatomical region, and image quality.
  - Anatomical/Structural Features: key visible structures relevant to the image.
  - Pathological/Abnormal Features: lesions, signal/echogenicity changes, structural distortion, mass effect.
  - Possible Diagnosis: suggest likely diagnosis or lesion type and reasoning based on imaging findings.
  - Clinical Implications: functional impact, severity, treatment considerations, or recommended follow-up.
3. Ensure that at least one question addresses image basics and at least one question asks for possible diagnosis with reasoning.
4. Each question-answer pair must be a JSON object:

```
[  
  {  
    "question": "...",  
    "answer": "..."  
  },  
  ...  
]
```

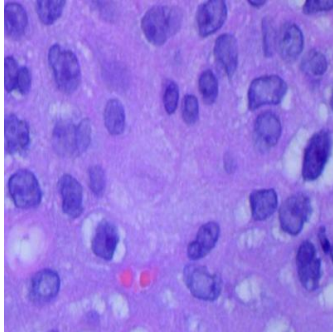
Important:

- Do NOT directly mention the category name in the question; describe imaging findings or anatomical context instead.
- Use professional, clinically relevant terminology.
- Output must be valid JSON with exactly five question-answer objects.

Now, generate five question-answer pairs for the following:

Image category: <CATEGORY\_NAME>

Figure 5. Prompt for generating five open-ended Q&A pairs per image, formatted in JSON.



What is the distinctive structural feature observed within the pulmonary lesion that suggests a malignant tumor?



### Ground Truth

The lesion shows irregular, spiculated margins with central necrosis and associated mass effect, which are characteristic features of squamous cell carcinoma of the lung.



### MiniCPM-V

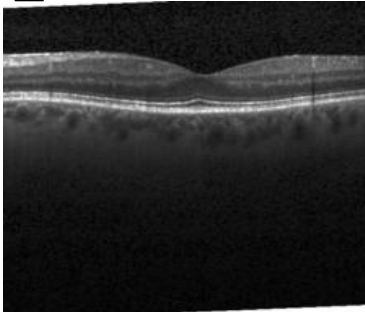
The distinctive structural feature observed within the pulmonary lesion that suggests a malignant tumor is the presence of pleomorphic nuclei. Pleomorphism refers to significant variations in size, shape, and staining characteristics among the nuclei of the cells. This irregularity is often associated with cancerous growths due to the uncontrolled division and differentiation of cells. The image shows multiple cells with varying nuclear morphology, indicating potential malignancy.



### Ours

A solitary, spiculated mass with irregular borders is visible within the lung tissue, a feature that is characteristic of malignant lesions.

(a)



What imaging modality and section of the eye are depicted in this scan?



### Ground Truth

This is an optical coherence tomography (OCT) scan of the posterior pole of the eye, displaying a cross-sectional view of the retina.



### InternVL 3.5

This image is from an OCT (Optical Coherence Tomography) scan, which is a cross-sectional imaging technique used to visualize the retina and detect abnormalities. The specific section shown is likely representing the retinal layers above the central fovea, which is critical for measuring parameters like the retinal thickness and assessing the structural integrity of the retina.



### Ours

This is an Optical Coherence Tomography (OCT) scan of the retina.

(b)



What are the potential clinical implications of these retinal findings for visual function?



### Ground Truth

These microvascular changes can lead to retinal ischemia, edema, and potentially cause visual acuity decline if they progress to more advanced stages or involve the macula significantly.



### Gemma3

Okay, let's analyze the retinal image and discuss the potential clinical implications.

#### Observations from the Image:

**Macular Occlusion (Dark Macula):** The most prominent feature is the large, dark area within the macula. This is a macular occlusion.....

Given these findings, here's a breakdown of the potential implications and possible diagnoses:

1. **Macular Hole:** .....

#### Recommendations:

**Thorough Ophthalmological Examination:** This patient needs an immediate and comprehensive eye exam by an ophthalmologist. This should include: .....



### Ours

These findings suggest potential visual field loss in the affected eye, and prompt management of diabetic retinopathy to prevent further visual deficits.

(c)

Figure 6. Qualitative comparison of model predictions on pulmonary histopathology, retinal OCT, and fundus photography.



Are there any signs of atypical features or malignant transformation visible in the lesion?



### Ground Truth

No, the lesion displays symmetrical shape, uniform color, and well-defined borders, with no irregular pigmentation, asymmetry, or surface ulceration, indicating a low likelihood of malignancy.



### Qwen3-VL

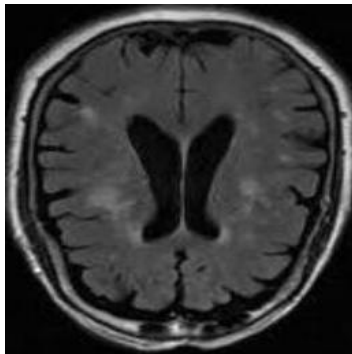
No, there are no signs of atypical features or malignant transformation visible in the lesion. The lesion appears as a well-defined, slightly elevated, pinkish-red area with a smooth surface and no visible ulceration, bleeding, or irregular borders. There are no signs of pigmentation changes, scaly patches, or asymmetric morphology that would suggest malignancy. The surrounding skin appears normal, and the lesion does not exhibit the classic features of malignant lesions such as invasion, irregular borders, or color variation. This appearance is consistent with a benign lesion, such as a seborrheic keratosis or a benign papule, and does not indicate malignant transformation. However, a definitive diagnosis should be made by a dermatologist, especially if there are any concerns or if the lesion changes over time.



### Ours

No, the lesion lacks asymmetry, irregular borders, color variegation, or atypical network, that would raise suspicion for malignancy.

(a)



Based on the imaging features, what is the most likely clinical significance of this scan?



### Ground Truth

This MRI represents a normal brain, indicating the absence of intracranial pathology, which suggests no obvious neurological deficits related to structural brain abnormalities.



### InternVL 3.5

The presence of multiple subcortical lesions with restricted diffusion in the right and left cerebral hemispheres, as observed in this MRI scan, is highly suggestive of a primary malignant brain tumor, such as glioblastoma or an undifferentiated glioma. This pattern of restricted diffusion on diffusion-weighted imaging (DWI) is a characteristic feature of malignant gliomas, which tend to have reduced water molecule diffusion due to a high nuclear-to-cytochrome c ratio. The location and hyperintense signal on DWI MRI indicate the mass effect of the lesion, as well as possible surrounding edema or hemorrhage. Further clinical correlation and additional imaging modalities.....



### Ours

Given the normal brain appearance, the most likely clinical significance is the absence of neurological signs of disease in this brain tissue.

(b)



What are the key considerations for post-procedural surveillance after treating this type of lesion?

- Option A: Long-term follow-up is unnecessary as recurrence risk is minimal.
- Option B: Repeat imaging with CT colonography is sufficient for future surveillance.
- Option C: Routine stool-based screenings are sufficient after initial removal.
- Option D: A tattoo mark should be placed near the site for easier follow-up visualization.

D  
Ground Truth

(c)

B



D  
Ours

Figure 7. Qualitative comparison of model performance across three benchmark datasets: DermaMNIST, BTMRI, and Kvasir.

## References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 1
- [3] Andrew A Borkowski, Marilyn M Bui, L Brannon Thomas, Catherine P Wilson, Lauren A DeLand, and Stephen M Mastorides. Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142*, 2019. 1
- [4] 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. *NeurIPS*, 36:28541–28564, 2023. 2
- [5] Bo Liu, Ke Zou, Li-Ming Zhan, Zexin Lu, Xiaoyu Dong, Yidi Chen, Chengqiang Xie, Jiannong Cao, Xiao-Ming Wu, and Huazhu Fu. Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis. In *ICCV*, pages 21310–21320, 2025. 1
- [6] Anas M Tahir, Muhammad EH Chowdhury, Amith Khadakdar, Tawsifur Rahman, Yazan Qiblawey, Uzair Khurshid, Serkan Kiranyaz, Nabil Ibtehaz, M Sohel Rahman, Somaya Al-Maadeed, et al. Covid-19 infection localization and severity grading from chest x-ray images. *Comput. Biol. Med.*, 139:105002, 2021. 1
- [7] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 2
- [8] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 2
- [9] Luis Wiedmann, Aritra Roy Gosthipaty, and Andrés Marafioti. nanovlm. <https://github.com/huggingface/nanoVLM>, 2025. 1
- [10] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 2
- [11] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 2