Sparse Spectral LoRA: Routed Experts for Medical VLMs

Supplementary Material

7. Pseudocode

The pseudocode is provided below.

Algorithm 1 MedQwen

Require: Input vector x, input dimension n, hyperparameters η, ρ, number of experts N

Ensure: Output $y = \tilde{W}^{(0)}x + \sum_{i=1}^{N} R(\mathbf{x})_i \, s \, B_i^{(0)} A_i^{(0)}$ x

procedure INITIALIZATION
  Scaling factor: $s \leftarrow \sqrt{3n\eta/r}$
  SVD decomposition: $W^{(0)} = U \, S \, V^\top$
  for $i = 1$ to N **do**
    $B_i^{(0)} \leftarrow \sqrt{1/(s\rho)}\, U' S'^{1/2}$
    $A_i^{(0)} \leftarrow \sqrt{1/(s\rho)}\, S'^{1/2} V'^\top$
  end for
  $W_{\text{res}}^{+} \leftarrow \frac{s}{N} \sum_{i=1}^{N} B_i^{(0)} A_i^{(0)}$
  $\tilde{W}^{(0)} \leftarrow W^{(0)} - W_{\text{res}}^{+}$
  return $\tilde{W}^{(0)}, \{B_i^{(0)}, A_i^{(0)}\}$
end procedure
procedure FORWARD(x)
  Compute gating weights $R(\mathbf{x})_i$
  return $\tilde{W}^{(0)}x + \sum_{i=1}^{N} R(\mathbf{x})_i \, s \, B_i^{(0)} A_i^{(0)}$ x
end procedure

8. Experiment Details

8.1. Dataset Information

Tables S1 and S2 summarize the datasets used in this study, covering a wide range of biomedical imaging modalities, such as MRI, CT, ultrasound, X-ray, and others. Each dataset is described in terms of its imaging modality, number of images, and question–answer text.

Visual Question Answering: VQA-RAD [37] contains 3,515 question–answer (QA) pairs and 315 radiology images, with questions spanning 11 categories and including both closed-ended and open-ended types. SLAKE [47] comprises 642 radiology images and over 7,000 QA pairs, along with segmentation masks and object detection bounding boxes. PathVQA [20] includes 4,998 pathology images with 32,799 QA pairs focusing on attributes such as location, shape, color, and appearance, categorized into open-ended and closed-ended types. OmniMedVQA [26] comprises 118,010 medical images and 127,995 QA pairs collected from 73 different medical datasets, encompassing 12 imaging modalities and covering more than 20 distinct anatomical regions. Importantly, all images in this benchmark originate from authentic clinical scenarios, ensuring alignment with real-world medical requirements.

Report Generation: MIMIC-CXR [32] includes 371,920 chest X-rays associated with 227,943 imaging studies from 65,079 patients. Following RadFM [83] and R2Gen [7], we use 337,292 cases for training. IU-Xray [11] consists of 7,470 chest X-ray images paired with corresponding diagnostic reports; following R2Gen [7], we use 4,730 cases from the training split.

Classification: We use the UniMed [34] dataset for pretraining and evaluate our model on standard zero-shot classification benchmarks commonly used for medical VLM evaluation, covering four imaging modalities: X-ray, MRI, CT, and ultrasound. For evaluation, we use the test sets from these widely recognized medical VQA datasets and additionally assess classification performance.

8.2. Implementation Details and Hyperparameters

Visual question answering and image classification experiments are conducted on a single NVIDIA A100 GPU with 40 GB of RAM. Additional ablation and report generation experiments are performed on four A100 GPUs, each with 40 GB of RAM. All models are trained and evaluated using bfloat16 precision.

We fine-tune our model on each task using carefully selected hyperparameters to ensure optimal performance. Detailed configurations, including learning rate, batch size, number of epochs, and other training settings, are provided to ensure reproducibility and consistency across experiments. These hyperparameters are summarized in Table S3 and Table S4. We set $\rho = 10$. The ratio between the full fine-tuning learning rate and the LoRA learning rate (η) is empirically set to 1 for ViT. In the Qwen experiments, when using a learning rate at the 1×10^{-4} level, we set $\eta = 0.1$; when using a learning rate at the 1×10^{-5} level, we set $\eta = 1$. This configuration follows common practice, where LoRA-based tuning typically employs a learning rate around 1×10^{-4}, while full fine-tuning methods operate at a lower rate near 1×10^{-5}. For LoRA-MoE experiments, we set the balance loss coefficient to 1×10^{-3}. We adopt a top-k routing strategy with $k = 2$, which outperforms other routing strategies as shown in Fig. 7. The same routing strategy is applied consistently across all LoRA-MoE baselines.

8.3. Evaluation Metrics

Following [46, 85], we use Accuracy, F1 Score and AUROC for evaluating the medical VQA task, and BLEU Score [61], ROUGE-L [43], and METEOR [1] for evaluating the report generation task. In alignment with existing hallucination benchmarks in both general and medical domains, accuracy (Acc) is employed as the primary metric for evaluating close-ended hallucination. Assessing open-ended hallucinations in generated reports, we follow CheXpert [28] and measure hallucination rates using CHAIR [41], which evaluates key symptom-centered visual findings. CHAIR is defined as: $\text{CHAIR} = \frac{|\mathcal{G} - \mathcal{S}|}{|\mathcal{G}|}$, where $\mathcal{G}$ represents the set of findings extracted from the generated report using CheXbert [65], and $\mathcal{S}$ represents the set of findings extracted from the real report using the same method. For a more comprehensive evalua-

Stage	Data Source	Sample Size
Stage 1	llava_med_alignment_500k.json	500K
Stage 2	instruct_60k_inline_mention	60K
Stage 3	RAD-VQA, SLAKE, Path-VQA, OmniMedVQA, IU-Xray, MIMIC-CXR, Harvard-FairVLMed, Quilt-1M, PMC-OA	2.4M

Table S1. Summary of Data Utilized Across Training Stages

Category	Name	Modality	# Image/QA text
Visual Question Answering	SLAKE [47]	CT, MRI, X-ray	642/14028
	VQA-RAD [37]	CT, MRI, X-ray	315/3515
	PathVQA [20]	Histopathology	4998/32799
	OmniMedVQA [26]	CT, MRI, X-ray, Ultrasound, Fundus, Histopathology, OCT, Dermoscopy, Colposcopy, Digital Photography, Infrared Reflectance Imaging, Endoscopy, Microscopy Images	118010/127995
	Quilt-1M [27]	Histopathology	1M/1M
	PMC-OA [45]	CT, MRI, X-ray, Ultrasound, Endoscopy, Microscopy Images	1.6M/1.6M
	Harvard-FairVLMed [55]	Fundus	10000/10000
Report Generation	IU-Xray [11]	X-ray	8121/3996
	MIMIC-CXR [32]	X-ray	227827/227835
Image Classification	UniMed [34]	CT, MRI, X-ray, Ultrasound, Histopathology	5.3M/5.3M

Table S2. An overview of the datasets used in this study.

Hyperparameter	Visual Question Answering
Batch Size	16
Rank	32
Alpha	64
Optimizer	AdamW
Warmup Steps	100
Dropout	0.05
Learning Rate	1e-4

Table S3. Hyperparameters of the VQA task for MedQwen.

Hyperparameter	Medical image classification
Batch Size	256
Rank	8
Alpha	16
Optimizer	AdamW
Warmup Steps	100
Dropout	0.05
Learning Rate	1e-4

Table S4. Hyperparameters of the image classification task.

tion, we additionally report key findings recall (Recall) and assess overall report quality using specialized metrics such as CheXbert [65], RadGraph [30], and RaTEScore [96], which have been specifically developed for medical report generation. These metrics align closely with radiologists' assessments, making them particularly suitable for evaluating the generation of open-ended medical reports, as demonstrated by RaTEScore data.

We evaluate the effectiveness of MedQwen in mitigating hallucinations in medical LVLMs across three medical benchmarks [4]:

- **Visual misinterpretation hallucination:** This category is evaluated using two datasets — Multi-Modality Visual Hallucination (MM-VisHal) and Chest X-ray Visual Hallucina-

tion (CXR-VisHal).

- **Knowledge deficiency hallucination:** This dataset is constructed from the MIMIC-CXR test set, where imaging reports are used as interpretations to prompt GPT-4 for generating diagnostic questions.
- **Context misalignment hallucination:** This benchmark links MIMIC-CXR data with the de-identified MIMIC-IV-EHR dataset [33] via subject IDs, providing comprehensive medical notes corresponding to each chest X-ray.

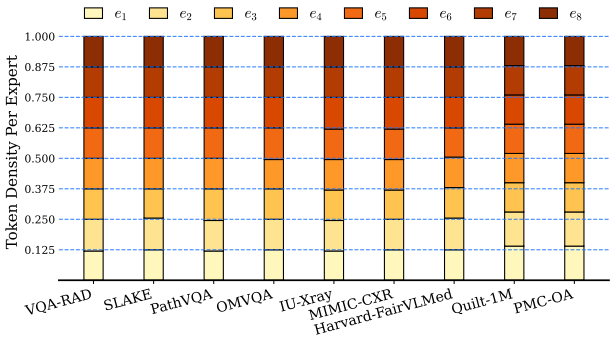


Figure S1. Expert load distribution across medical datasets.

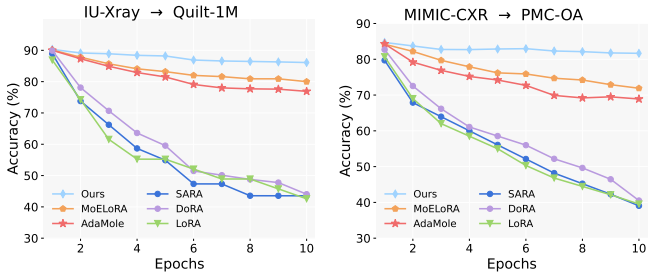


Figure S2. Catastrophic forgetting results.

9. Additional Details

9.1. Routing Analysis

Fig. S1 reports the expert load during training on nine datasets. With eight experts and two activated per token, the expected token density per expert is 0.125. The observed routing remains well balanced: no expert becomes inactive, and the load fluctuates around 0.125 within approximately $\pm$ 10%. Because each expert is initialized from a distinct SVD-derived subspace, this stable and non-degenerate routing suggests that the spectral priors remain functionally distinct and continue to guide specialization, rather than behaving like indistinguishable zero-initialized adapters, even after variance-reducing damping (via ρ and scaling). If the damping had erased the spectral structure, we would expect pronounced routing skew or expert collapse (i.e., a subset of experts dominating), a phenomenon commonly reported in prior MoE analyses.

9.2. Catastrophic Forgetting

To further evaluate the robustness of our method against forgetting, we extend the continual-learning setting to four datasets and multiple task sequences, as illustrated in Fig. S2. This expanded evaluation provides a broader view of retention performance across different sequential learning scenarios. We additionally include DoRA [51] and SARA [25] as representative SVD-based baselines for comparison. Across all dataset sequences, our MoE-SVD consistently maintains performance degradation below 5%, indicating stable knowledge retention even when the model is exposed to multiple domain shifts.

Table S5. Comparison of routing strategies on average performance.

Routing Strategy	Avg
Ours (top-$k = 2$)	**68.24**
Top-p ($p = 0.25$)	66.40
Top-k + Shared Expert	65.67

9.3. Routing Techniques

To further investigate the effect of routing strategies, we conduct additional experiments comparing several alternatives, including top-p routing and a top-k routing variant with shared experts (Table S5). Among the evaluated configurations, the top-k strategy with $k = 2$ consistently provides the strongest performance, outperforming the other routing schemes.

Table S6. MedQwen-e vs properly-scaled MedQwen.

Method	IU-Xray	MIMIC-CXR	Harvard-FVLM	Quilt-1M	PMC-OA
MedQwen	90.33	84.68	88.43	73.74	66.32
MedQwen-e	90.29	84.53	88.38	73.51	66.22

9.4. Proper Scaling

MedQwen assumes that the LoRA-MoE adapters are properly scaled at initialization. However, this assumption may not hold in practice. To address this issue, we extend the formulation to unscaled settings by aligning the scaling factors of the experts. In particular, we treat the first expert, which corresponds to the dominant low-rank spectral component and is analogous to the first expert in full MoE fine-tuning, as the reference with scaling factor s_1. To reduce the gap between our method and full MoE fine-tuning, we align the scaling factors s_i of the remaining experts so that their effective magnitudes match the reference expert. Formally, the scaling factors must satisfy: $s_1^2\sigma_0 = s_i^2\sigma_i$, which yields: $s_i = s_1\sqrt{\frac{\sigma_0}{\sigma_i}}$. Here, σ_i denotes the spectral mass of the corresponding segment, computed as the sum of singular values when the segment rank exceeds one. Importantly, only the scaling factors $s_{i>1}$ are adjusted, while the adapter initialization itself follows Eq. 22. We refer to this variant as **MedQwen-e**. Empirically, MedQwen-e achieves performance comparable to MedQwen across benchmarks (Table S6), indicating that the proposed scaling strategy stabilizes the method even in unscaled initialization settings.

LVLM	CheXbert ↑	RadGraph ↑	RaTEScore ↑	Recall ↑	CHAIR ↓
GPT-4o	21.71	10.28	45.39	33.73	11.99
LLaVA-NeXT 7B	16.31	4.41	39.93	10.88	16.08
LLaVA-NeXT 13B	14.76	5.34	38.59	6.38	14.82
MiniGPT-4	17.71	7.18	39.90	10.54	18.02
LLaVA-Med	19.72	7.31	39.86	25.17	20.85
LLaVA-Med-1.5	18.44	4.96	39.47	13.27	19.74
LLM-CXR	24.34	7.57	38.53	29.85	9.18
Med-Flamingo	17.50	5.83	35.87	17.52	23.96
RadFM	23.74	6.69	37.04	24.66	6.89
CheXagent	30.32	12.35	43.18	33.93	6.88
XrayGPT	25.63	12.88	44.45	30.87	12.84
MedQwen	**35.80**	**13.78**	**49.77**	**37.24**	**6.21**

Table S7. Open-ended evaluation on visual misinterpretation hallucination (underlined: second-best, **Bold**: best)

.

10. Hallucination Evaluation

10.1. Visual Misinterpretation Hallucination

A visual misinterpretation hallucination occurs when the model interprets fundamental visual components that are factually incorrect or unsupported by medical evidence.

Hallucination Evaluation on Closed-Ended Evaluations. The close-ended evaluation results on MM-VisHal and CXR-VisHal are detailed in Table S11. Our method, MedQwen, significantly outperforms existing state-of-the-art models across all metrics and benchmarks, indicating the effectiveness of MoE in reducing medical hallucinations in close-ended medical tasks. GPT-4o, owing to its large-scale training and strong cross-domain instruction following, generally exhibits superior resistance to hallucinations compared to other (Med)-LVLMs. In contrast, general-domain LVLMs demonstrate particular weaknesses with hallucination sub-types like Symptom and Measurement across all tested modalities. Within the Med-LVLM category, CheXagent shows better overall accuracy on the CXR-VisHal benchmark. However, despite specialized medical training, these models frequently display higher hallucination rates on the MM-VisHal benchmark, where their accuracy is markedly lower than on the single-modality CXR-VisHal dataset. Notably, models such as LLM-CXR and CheXagent achieve exceptional performance on datasets aligned with their primary training domains, such as chest X-rays.

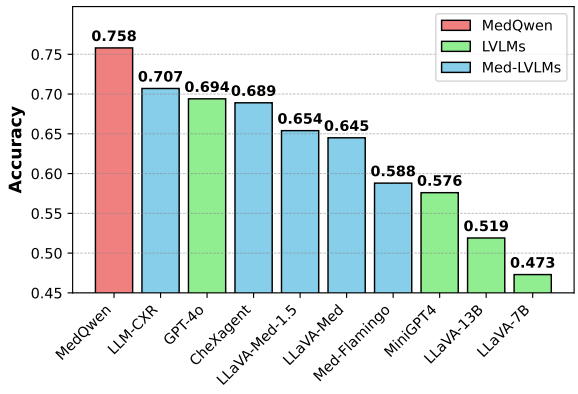


Figure S3. Close-ended evaluation of knowledge deficiency hallucination in (Med)-LVLMs and the effectiveness of hallucination mitigation methods.

Hallucination Evaluation on Open-Ended Tasks. Table S7 presents the hallucination rates (CHAIR) from our open-ended evaluation, indicating that the majority of Med-LVLMs, including CheXagent, struggle to resist hallucinations when generating critical medical findings. Notably, our proposed method, MedQwen, achieves a lower hallucination rate and higher recall, outperforming both Med-LVLMs and general-domain LVLMs. Furthermore, the LLaVA-Med series demonstrates suboptimal performance in open-ended evaluations, exhibiting higher hallucination rates despite achieving higher recall compared to general-domain LVLMs. Report-specific metrics consistently reflect the overall quality of generated outputs: models with lower CHAIR scores and higher recall, such as CheXagent, GPT-4o, and XrayGPT, generally achieve superior performance.

10.2. Knowledge Deficiency Hallucination

Hallucination can also occur when the model correctly interprets the image, such as recognizing key organs and visual features, but lacks the comprehensive medical knowledge required for accurate diagnosis or clinical decision-making.

Hallucination Evaluation on Closed-Ended Tasks. The results presented in Fig. S3 indicate that our proposed method, MedQwen, achieves a significant accuracy of 75.8%, representing a 5.1% improvement over LLM-CXR and a 6.4% improvement over GPT-4o. Med-LVLMs, benefiting from their specialized medical knowledge, typically exhibit superior accuracy compared to general-domain LVLMs, with certain Med-LVLMs achieving performance comparable to GPT-4o. Despite this, their overall performance against knowledge-based hallucinations remains inadequate. These observations highlight that, even with training on varied multimodal medical datasets, conventional Med-LVLMs are prone to generating hallucinations when responding to diagnostic inquiries requiring specific domain knowledge. Therefore, our findings suggest that our MoE approach offers a more robust solution for mitigating knowledge hallucinations than traditional medical tuning methods.

Hallucination Evaluation on Open-Ended Tasks. As evidenced in Table S8, our proposed method, MedQwen, achieves the highest performance among the twelve evaluated LVLMs, exhibiting a remarkably low hallucination score of $S_h = 0.63$. Most Med-LVLMs demonstrate an increased propensity for hallucinations when interpreting complex medical knowledge, which aligns with observations from generation-focused metrics. Conversely, the LLaVA-Med series exhibits greater robustness against knowledge-based hallucinations. Regarding generation quality, most Med-LVLMs show relatively lower word-level coverage of ground truth compared to general-domain LVLMs such as LLaVA-NeXT 7B and 13B, indicating suboptimal content consistency. While GPT-4o consistently ranks second across the majority of metrics and surpasses certain specialized Med-LVLMs, XrayGPT performs poorly across most evaluation metrics, frequently generating irrelevant text and extraneous details in a medical-report style, which reflects its confined training focus primarily on medical summary generation.

LVLM	Generation Metrics						Hallucination Score
	BertScore ↑	BLEU ↑	METEOR ↑	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	$\mathcal{S}_h$ ↓
GPT-4o	<u>91.71</u>	<u>14.60</u>	33.24	<u>47.77</u>	<u>28.76</u>	<u>40.76</u>	<u>0.77 ± 0.81</u>
LLaVA-NeXT 7B	90.16	13.81	38.34	44.75	22.04	33.83	2.02 ± 1.20
LLaVA-NeXT 13B	89.56	12.03	39.46	41.59	20.16	30.70	2.09 ± 1.09
MiniGPT-4	86.93	7.49	32.51	34.62	14.42	24.55	3.29 ± 1.25
LLaVA-Med	89.51	11.59	40.68	41.44	19.99	30.44	1.92 ± 1.02
LLaVA-Med-1.5	89.86	12.98	<u>41.30</u>	43.52	21.37	32.27	1.76 ± 0.96
LLM-CXR	87.98	4.15	16.55	28.71	13.05	23.28	2.52 ± 1.62
Med-Flamingo	84.52	5.33	22.74	26.24	10.17	21.00	3.69 ± 1.10
RadFM	79.66	7.99	25.51	32.63	14.22	24.14	2.30 ± 1.56
CheXagent	87.82	4.65	16.85	28.07	15.38	23.75	2.08 ± 1.50
XrayGPT	83.82	1.91	17.49	21.62	3.31	13.99	4.78 ± 0.63
MedQwen	**92.75**	**17.49**	**44.60**	**49.37**	**31.91**	**43.12**	**0.63 ± 0.21**

Table S8. Results on open-ended evaluation of knowledge deficiency hallucination. (<u>underlined</u>: second-best, **Bold**: best)

10.3. Context Misalignment Hallucination

In addition to the evaluations outlined in previous sections, clinical practice necessitates that medical image interpretation aligns with the patient's comprehensive medical history. This includes critical factors such as treatment plans, diagnostic records, family history, and other relevant clinical data. However, existing benchmarks for assessing hallucination in medical imaging predominantly focus on isolated image analysis, neglecting the broader clinical context integral to real-world practice. To address this gap and better align with the practical demands of the medical field, we evaluate the model's susceptibility to hallucinations by contextualizing medical images within the patient's holistic medical background.

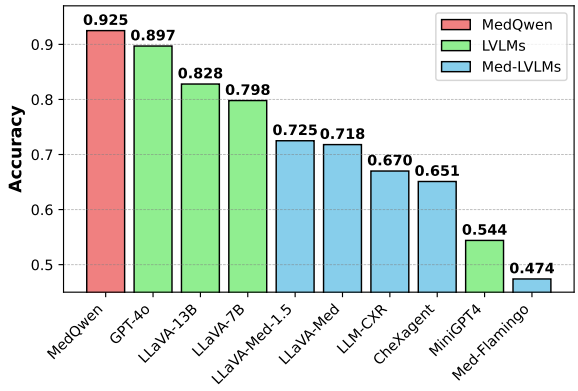


Figure S4. Close-ended evaluation of context misalignment hallucination in (Med)-LVLMs and the effectiveness of hallucination mitigation methods.

Mitigation Evaluation Results.
Despite the superior performance of our proposed method (MedQwen) illustrated in Fig. S4, general-domain LVLMs such as GPT-4o and LLaVA-NeXT 13B achieve higher accuracy than Med-LVLMs when responding to close-ended contextual questions. Notably, Med-Flamingo exhibits a below-average performance, suggesting that conventional fine-tuning on multimodal medical data might inadvertently impair the inherent reasoning capabilities of the foundational LVLMs. Consequently, Med-LVLMs become susceptible to hallucinations when confronted with intricate clinical contexts. Concurrently, these results underscore the efficacy of our proposed mixture-of-LoRA-experts approach in addressing the most challenging hallucination benchmarks.

11. Parameter and FLOPs Analysis

We present a comprehensive parameter analysis comparing the complexity of various baseline models and our proposed method across different backbone architectures. The notation used for the architectural parameters is defined as follows:

- **H**: Hidden dimension.
- **r**: Rank of the low-rank adaptation components.
- **e**: Number of experts in the MoE layer.
- **L**: Total number of layers in the model.
- **V**: Vocabulary size.
- **P**: Patch size in the Vision Transformer model.
- **C**: Number of input channels in the ViT model.

The detailed parameter budget analysis for the MedQwen and BiomedCLIP architectures is provided below:

MedQwen-7B: $H = 4096, r = 32, e = 2, L = 28, V = 151646$. The activation parameters are `q, k, v, up, down`.

1. **FFT**:
   - **Total Parameters**: $(10.25H^2 + 2H)L + H + 2HV$
     - Embedding layer and LM head: $2HV$
     - Attention mechanism: $2.25H^2$
     - MLP layer: $8H^2$
     - RMSNorm (2 layers): $2H$
     - Additional RMSNorm (last layer): H
     - Total per layer: $10.25H^2 + 2H$
2. **LoRA/MiLoRA/PiSSA/KASA**:
   - **Total Parameters**: $11.58HLr$
   - **Proportion**: 0.78%
3. **DoRA**:
   - **Total Parameters**: $(11.58Hr + 5)L$
   - **Proportion**: 0.78%
4. **HydraLoRA**:
   - **Total Parameters**: $(4.91Hr + 6.66Hr/e + 6.66He)L$
   - **Proportion**: 0.58%

5. **AdaMoLE**:
   - **Total Parameters**: $(11.58Hr + 6.66He + 6.66H)L$
   - **Proportion**: 0.82%
6. **MoELoRA/Ours**:
   - **Total Parameters**: $(11.58Hr + 6.66He)L$
     - Attention mechanism: $4.25Hr + 3He$
     - MLP layer: $7.33Hr + 3.66He$
     - Total per layer: $6.66He + 11.58Hr$
   - **Proportion**: 0.81%

BiomedCLIP: $H = 768, e = 2, r = 8, P = 32, L = 12, C = 3$. The activation parameters include `q, k, v, o, fc1, fc2`.

1. **FFT**:
   - **Total Parameters**: $(C+1)HP^2 + (12H^2 + 2H)L + H^2 + 3H + PH$
   - **Breakdown**:
     - Embedding layer: $PH + H + (C+1)P^2H$
     - encoder (L layers): $(12H^2 + 2H)L$
     - LayerNorm (1 layers): $2H$
     - Pooler: H^2
2. **Full FT MoE**:
   - **Total Parameters**: $(C+1)P^2H + (12eH^2 + 2H + 9He)L + 3H + PH + H^2$
   - **Proportion**: 760%
3. **LoRA/PiSSA/MiLoRA**:
   - **Total Parameters**: $18HLr$
   - **Proportion**: 1.49%
4. **LoRA (rank=16)**:
   - **Total Parameters**: $18HLr$
   - **Proportion**: 2.99%
5. **LoRA (rank=32)**:
   - **Total Parameters**: $18HLr$
   - **Proportion**: 5.98%
6. **HydraLoRA**:
   - **Total Parameters**: $(9Hr + 9He + 9Hr/e)L$
   - **Proportion**: 1.58%
7. **AdaMoLE**:
   - **Total Parameters**: $(18Hr + 9He + 9H)L$
   - **Proportion**: 2.33%
8. **MoLoRA/Ours**:
   - **Total Parameters**: $(18Hr + 9He)L$
   - **Breakdown**:
     - Attention mechanism: $8Hr + 4He$
     - MLP layer: $10Hr + 5He$
     - Total per layer: $18Hr + 9He$
   - **Proportion**: 2.24%

11.1. FLOPs Analysis

We conduct a detailed parameter analysis for each baseline and our proposed method, considering the underlying architectural backbones. The analysis utilizes the following set of variables: H, representing the model hidden dimension; e, indicating the number of experts; r, denoting the LoRA rank; L, representing the number of layers; P, denoting the patch size in ViT; V, signifying the vocabulary size; and C, representing the number of input channels in ViT. The subsequent sections provide the specific analysis for the MedQwen and BiomedCLIP architectures.

FLOPs for FT MoE

1. MoE linear for q and o: The FLOPs are calculated as
$$2 \cdot \big(2BsHe + k \cdot 2BsH^2\big).$$
2. MoE linear for k and v: Since MedQwen 7B's GQA reduces the number of heads for k and v to $1/8$ of q's heads, the FLOPs are
$$2 \cdot \big(2BsHe + k \cdot 2BsH^2/8\big).$$
3. FLOPs for $q \cdot k$ and score $\cdot\, v$: These remain independent of k, as we only upcycle the linear projection to e copies. The FLOPs are
$$2Bs^2H + 2Bs^2H.$$
4. MoE linear for down and gate: Since MedQwen 7B uses SwiGLU FFN, the FLOPs are
$$2 \cdot \big(2BsHe + k \cdot 2BsH \cdot \tfrac{8}{3}H\big).$$
5. MoE linear for up: The FLOPs are
$$2Bs \cdot \tfrac{8}{3}He + k \cdot 2Bs \cdot \tfrac{8}{3}H^2.$$

Across L layers, including the vocabulary embedding transformation, the total FLOPs of Full FT MoE are:
$$\text{FLOPs} = BL\Big(\tfrac{52}{3}esH + \tfrac{41}{2}ksH^2 + 4s^2H\Big) + 2BsHV.$$

FLOPs for Ours/MoLoRA/HydraLoRA

1. MoE linear for q and o: The FLOPs are calculated as
$$2B \cdot \big(2sH^2 + 2esH + 2k(sHd + sHd)\big).$$
2. MoE linear for k and v: Considering the effect of MedQwen 7B's GQA on k and v, the FLOPs are
$$2B \cdot \big(2sH^2/8 + 2esH + 2k(sHd + sHd/8)\big).$$
3. FLOPs for $q \cdot k$ and score $\cdot\, v$: The FLOPs for these operations are
$$2Bs^2H + 2Bs^2H.$$
4. MoE linear for down and gate: Since MedQwen 7B uses SwiGLU FFN, the FLOPs are
$$2B \cdot \big(2sH \cdot \tfrac{8}{3}H + 2esH + 2k(sHd + sd\tfrac{8}{3}H)\big).$$
5. MoE linear for up: The FLOPs are
$$2BsH \cdot \tfrac{8}{3}H + 2Bs\tfrac{8}{3}He + 2k\big(Bs\tfrac{8}{3}Hd + BsrH\big).$$

Across L layers, including the vocabulary embedding transformation, the total FLOPs of LoRA-MoE are:
$$\text{FLOPs} = BL\Big(\tfrac{52}{3}esH + \tfrac{41}{2}sH^2 + 4s^2H + \tfrac{69}{2}ksHd\Big) + 2BsHV$$

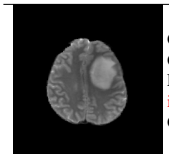
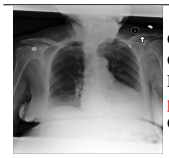
Case 1: Hallucination and mitigation in MM-VisHal	
	**Question:** "Does the brain appear abnormal in the image?" **Ground truth:** "Yes." **LLaVA-Med:** "No, there is no evidence of edema in the brain in the MRI." **Ours:** "Yes, there is evidence of edema in brain tissue."
Case 2: Hallucination and mitigation in CXR-VisHal	
	**Question:** "Does the image show the presence of a pneumothorax?" **Ground truth:** "No." **LLaVA-Med:** "Yes, this image shows signs consistent with a pneumothorax." **Ours:** "No, there is no clear evidence of a pneumothorax."

Table S9. Examples of hallucination and mitigation cases.

Case studies on hallucination and mitigation in report generation.
 **Ground truth report:** "The endotracheal tube is seen terminating 5.6 cm above the carina, indicating proper placement. The right internal jugular line ends at mid superior vena cava (SVC). A feeding tube is visible extending below the diaphragm into the stomach, although the distal end is not within the radiographic view. There are opacities in the left lower lung, suggesting a combination of atelectasis and mild pleural effusion, which is unchanged. The heart and mediastinal contours appear mildly enlarged but stable."
LLaVA-Med: "Chest radiograph shows the cardiomediastinal silhouette is broadened, suggesting possible cardiomegaly or volume overload. Hilar and mediastinal contours otherwise appear unremarkable. No pulmonary edema, free intraperitoneal air, acute rib or clavicular fracture is identified. There is no right internal jugular catheter visible, and no pneumothorax is observed."
GPT-4o: "The chest X-ray shows relevant measurements including the volume of the lungs, the size of the heart, and the position of the diaphragm. Additionally, the image may show any abnormalities or signs of symptoms related to the chest, such as enlarged cardiac silhouette, fluid accumulation, or superimposed pneumonia."
Ours: "The chest radiograph shows cardiomegaly, evidenced by an enlarged cardiac silhouette, along with radiographic signs of pulmonary edema. A central venous catheter is visible, its tip projecting into the mid-to-lower portion of the superior vena cava."

Table S10. Case studies highlighting hallucination and mitigation in medical report generation. The report generated by GPT-4o is generic and misses many details.

12. Ablation details

Here, we provide a detailed explanation of the construction of each initialization method. Suppose $h = min(m, n), t = \frac{h}{N}$

1. **Ours (O)**:

$$\mathcal{E}_r = \Big\{ \big(U_{[:,\, k:k+d]}, S_{[k:k+d,\, k:k+d]}, V^\top_{[k:k+d,\, :]}\big) \Big| \; k = (j-1)t, j = 1, \ldots, N \Big\}$$

2. **Principal (P)**:

$$\mathcal{E}_r = \Big\{ \big(U_{[:,\, k:k+d]}, S_{[k:k+d,\, k:k+d]}, V^\top_{[k:k+d,\, :]}\big) \Big| \; k = (j-1)d, j = 1, \ldots, N \Big\}$$

3. **Minor (M)**:

$$\mathcal{E}_r = \Big\{ \big(U_{[:,\, k:k+d]}, S_{[k:k+d,\, k:k+d]}, V^\top_{[k:k+d,\, :]}\big) \Big| \; k = h - jd, j = 1, \ldots, N \Big\}$$

4. **Random (R)**:

$$\mathcal{E}_r = \Big\{ \big(U_{[:,\, k:k+d]}, S_{[k:k+d,\, k:k+d]}, V^\top_{[k:k+d,\, :]}\big) \Big| \; k = tj, t = \text{random}\Big(0, \frac{h}{d} - 1\Big), j = 0, \ldots, N-1 \Big\}$$

12.1. Case Study for Hallucination and Mitigation

Table S9 presents case studies from close-ended datasets on visual misinterpretation hallucinations. We provide the produced replies from our technique and baseline approach in Cases 1 and 2 to visually show the efficacy of our strategy. our method which enhances visual grounding through attention modification, successfully corrects the hallucinated responses. While it does not work for LLaVA-Med. This study reinforces our findings that our mitigation method exhibit strength in hallucination mitigation, emphasizing the need for task-specific approach to improve Med-LVLM performance.

Also, Table S10 shows case studies of open-ended report generation on visual misinterpretation hallucinations. In this example, our method effectively mitigates the hallucination while even improving the recall of key findings. However, GPT-4o, while aiming to lower hallucination rates, significantly impact generation quality and recall, demonstrating the challenges of balancing hallucination mitigation and report completeness.

13. Load Balancing Loss

In standard MoE architectures [10, 14], a balance loss, $\mathcal{L}_b$, is commonly employed to prevent routing collapse, thereby ensuring a uniform distribution of tokens across the available experts. This loss is formally defined as the dot product between the fractional load and the average routing probability for each expert:

$$\mathcal{L}_b = \sum_{i=1}^{E} f_i P_i \tag{25}$$

$$f_i = \frac{E}{kT} \sum_{t=1}^{T} \mathbf{1}\{\text{token } x_t \text{ is assigned to expert } i\}, \tag{26}$$

$$P_i = \frac{1}{T} \sum_{t=1}^{T} \text{softmax}(z^i(x_t)) \tag{27}$$

Here, T denotes the total number of tokens, and $\mathbf{1}(\cdot)$ is the indicator function. The term f_i represents the normalized fraction of tokens explicitly routed to expert i, while P_i quantifies the average router probability for that expert. By minimizing $\mathcal{L}_b$, the training objective explicitly promotes an equitable utilization of all experts.

LVLM	MM-VisHal					CXR-VisHal				
	Acc-A ↑	Acc-M↑	Acc-S ↑	Acc-R ↑	Acc ↑	Acc-A ↑	Acc-M ↑	Acc-S ↑	Acc-R ↑	Acc ↑
GPT-4o	0.775	0.697	0.708	0.846	0.741	0.880	0.595	0.788	0.921	0.794
LLaVA-NeXT 7B	0.576	0.426	0.507	0.451	0.494	0.817	0.430	0.474	0.362	0.518
LLaVA-NeXT 13B	0.577	0.430	0.551	0.445	0.510	0.776	0.391	0.486	0.563	0.534
MiniGPT-4	0.483	0.537	0.553	0.430	0.512	0.341	0.301	0.573	0.354	0.483
LLaVA-Med	0.525	0.357	0.584	0.485	0.499	0.698	0.452	0.725	0.800	0.698
LLaVA-Med-1.5	0.619	0.397	0.499	0.483	0.499	0.840	0.494	0.651	0.845	0.684
LLM-CXR	0.486	0.460	0.513	0.314	0.461	0.681	0.504	0.743	0.403	0.675
Med-Flamingo	0.523	0.497	0.588	0.327	0.507	0.361	0.324	0.576	0.332	0.489
CheXagent	0.524	0.516	0.572	0.464	0.529	0.782	0.576	0.739	0.851	0.739
MedQwen	**0.883**	**0.798**	**0.896**	**0.847**	**0.854**	**0.893**	**0.702**	**0.839**	**0.916**	**0.825**

Table S11. Results on close-ended evaluation of visual misinterpretation hallucination. We report Accuracy for each sub-type: Anatomy (**Acc-A**), Measurement (**Acc-M**), Symptom (**Acc-S**), Radiology Knowledge (**Acc-R**). We also report the overall accuracy (**Acc**). Higher accuracy in these evaluations indicates a stronger ability to resist hallucination. (underlined: second-best, **Bold**: best)

14. Proof of Theoretical Results

14.1. Proof of Theorem 1

Theorem 1. *Let η_{FFT} and η_{LoRA} denote the learning rates for Full Fine-Tuning (FFT) and LoRA. LoRA and Full FT behave equivalently when their initial weights satisfy $\tilde{W}_0 \approx W_0$ and their scaled gradients match at every step, i.e., $\eta_{\text{LoRA}}\tilde{g}_t \approx \eta_{\text{FFT}} g_t$. See Eq. 4 for formal definitions.*

Proof. Define the LoRA effective weight as $\tilde{W}_t = W_{\text{init}} + sB_tA_t$ and its gradient as $\tilde{g}_t$. Using SGD, the updates are:

$$W_{t+1} = W_t - \eta_{\text{FFT}} g_t, \tag{28}$$
$$\tilde{W}_{t+1} = \tilde{W}_t - \eta_{\text{LoRA}} \tilde{g}_t. \tag{29}$$

Base Case. At $t = 0$, we have $\tilde{W}_0 = W_0$.
Inductive Step. Assume $\tilde{W}_t = W_t$ and the scaled gradients satisfy the alignment condition. Then:

$$\tilde{W}_{t+1} = \tilde{W}_t - \eta_{\text{LoRA}} \tilde{g}_t \tag{30}$$
$$= W_t - \eta_{\text{FFT}} g_t \tag{31}$$
$$= W_{t+1}. \tag{32}$$

Thus, the weights remain identical for all t, establishing the alignment property. □

14.2. Proof of Theorem 2

Theorem 2. *Let η_{FFT} and η_{LoRA} denote the learning rates employed in Full FT MoE and LoRA MoE training. For each expert $i \in \{1, \ldots, N\}$, the two training procedures remain aligned when their initial effective weights satisfy $\tilde{W}_i^{(0)} \approx W_i^{(0)}$ and their scaled gradients satisfy $\eta_{\text{LoRA}}\tilde{g}_i^t \approx \eta_{\text{FFT}} g_i^t$ at each optimization step.*

Proof. We demonstrate that these conditions ensure that LoRA MoE replicates the behavior of Full FT MoE, particularly with respect to the routing mechanism.

Base Case ($t = 0$). Because the Full FT MoE is constructed by upcycling, all expert weights satisfy $W_i^{(0)} = W^{(0)}$. Thus the initialization condition implies $\tilde{W}_i^{(0)} \approx W^{(0)}$. As both models use the same initialization seed, the router parameters at $t = 0$ are identical, and both architectures produce the same routing assignments.

Inductive Hypothesis. Assume that at iteration t the equality $\tilde{W}_i^t = W_i^t$ holds for all experts and that the routers coincide.

Inductive Step. Using the scaled gradient alignment condition, we obtain

$$\tilde{W}_i^{t+1} = \tilde{W}_i^t - \eta_{\text{LoRA}} \tilde{g}_i^t \tag{33}$$
$$\approx W_i^t - \eta_{\text{FFT}} g_i^t \tag{34}$$
$$= W_i^{t+1}. \tag{35}$$

Since the routers receive identical inputs and expert outputs, their updated parameters remain equal:

$$\text{MoE}(\mathbf{x}) = \sum_{i=1}^{N} R(\mathbf{x})_i W_i(\mathbf{x}) = \sum_{i=1}^{N} R(\mathbf{x})_i \tilde{W}_i(\mathbf{x}), \tag{36}$$

which matches the LoRA MoE output. Therefore, the routers remain aligned at step $t+1$. Induction confirms that this holds for all t, and hence the two MoE models exhibit equivalent behavior. □

14.3. Proof of Theorem 3

Theorem 3. *Consider the i.i.d. logits $z^i(\mathbf{x})$ and let $S_k(x)$ denote the indices of the k largest among them, where $k \leq N/2$. Define the MoE weights as*

$$R(\mathbf{x})_i = \begin{cases} \frac{\exp(z^i(\mathbf{x}))}{\sum_{j \in S_k(x)} \exp(z^j(\mathbf{x}))}, & \text{if } i \in S_k(x), \\ 0, & \text{if } i \notin S_k(x). \end{cases} \tag{37}$$

Then, for any pair $i \neq j$, we have:

$$\mathbb{E}[R(\mathbf{x})_i] = \frac{1}{N}, \tag{38}$$

$$\mathrm{Var}(R(\mathbf{x})_i) = \frac{N-k}{kN^2}. \tag{39}$$

Proof. Because the logits are identically distributed and independent, permutations of their indices do not alter the joint distribution. Since the top-k selection also respects such symmetry, the induced weights $R(\mathbf{x})_i$ are exchangeable, which implies

$$\mathbb{E}[R(\mathbf{x})_i] = \mathbb{E}[R(\mathbf{x})_j], \quad \forall i, j. \tag{40}$$

Using $\sum_{i=1}^N R(\mathbf{x})_i = 1$, we obtain:

$$\sum_{i=1}^N \mathbb{E}[R_i] = 1 \quad \Longrightarrow \quad \mathbb{E}[R_i] = \frac{1}{N}. \tag{41}$$

For the variance, observe:

$$\mathrm{Var}(R_i) = \mathbb{E}[R_i^2] - \frac{1}{N^2}. \tag{42}$$

Expanding the identity $\left(\sum_i R_i\right)^2 = 1$ yields

$$1 = N\,\mathbb{E}[R_i^2] + N(N-1)\,\mathbb{E}[R_i R_j]. \tag{43}$$

To compute $\mathbb{E}[R_i R_j]$, define

$$y_i = \begin{cases} \exp(z_i), & i \in S_k, \\ 0, & i \notin S_k, \end{cases} \quad \text{so that} \quad R_i = \frac{y_i}{\sum_{\ell \in S_k} y_\ell}. \tag{44}$$

Thus,

$$R_i R_j = \frac{y_i y_j}{\left(\sum_{\ell \in S_k} y_\ell\right)^2}. \tag{45}$$

Since the probability that both i and j appear in the top-k is $\binom{k}{2}/\binom{N}{2}$, and upon selection, each is normalized by a sum of k exponentials, symmetry gives

$$\mathbb{E}[R_i R_j] = \frac{k-1}{N(N-1)k}. \tag{46}$$

Substituting this into Eq (43),

$$N\,\mathbb{E}[R_i^2] = 1 - \frac{k-1}{k}, \tag{47}$$

hence

$$\mathbb{E}[R_i^2] = \frac{1}{Nk}. \tag{48}$$

Inserting into Eq (42) yields

$$\mathrm{Var}(R_i) = \frac{1}{Nk} - \frac{1}{N^2} = \frac{N-k}{kN^2}. \tag{49}$$

□

14.4. Proof of Theorem 4

Theorem 4. *The solution to the optimization problem for the residual weight W_{res}:*

$$W_{res}^+ = \arg\min_{W_{res}} \mathbb{E}_{\mathbf{x}} \left\| W_{res} - s \sum_{i=1}^N R(\mathbf{x})_i B_i^{(0)} A_i^{(0)} \right\|^2. \tag{50}$$

Its closed-form minimizer is given by:

$$W_{\text{res}}^+ = \frac{s}{N} \sum_{i=1}^N B_i^{(0)} A_i^{(0)}.$$

Proof. The symbol W_{res}^+ denotes the optimizer of the stated problem. By applying the linearity of expectation, the solution can be expressed as

$$W_{\text{res}}^+ = s\,\mathbb{E}_{\mathbf{x}} \left[\sum_{i=1}^N R(\mathbf{x})_i B_i^{(0)} A_i^{(0)} \right] \tag{51}$$

$$= s \sum_{i=1}^N \mathbb{E}_{\mathbf{x}}[R(\mathbf{x})_i]\, B_i^{(0)} A_i^{(0)} \tag{52}$$

$$= \frac{s}{N} \sum_{i=1}^N B_i^{(0)} A_i^{(0)}, \tag{53}$$

where Eq (51) follows from linearity and Eq (52) uses Theorem 3. □

14.5. Proof of Theorem 5

Theorem 5. *Given the zero-initialization condition $B_0 = 0$ and $A_0 \sim U\left(-\sqrt{\frac{6}{n}}, \sqrt{\frac{6}{n}}\right)$, and the effective LoRA gradient:*

$$\tilde{g}_t^i = s^2 \left(B_t^i {B_t^i}^\top g_t^i + g_t^i {A_t^i}^\top A_t^i \right),$$

the optimal scaling factor s that minimizes the gradient mismatch $\|\tilde{g}_t^i - \eta g_t^i\|$ is:

$$s = \sqrt{\frac{3n\eta}{r}},$$

where $\eta = \eta_{FFT}/\eta_{LoRA}$ is the learning rate ratio required for update alignment.

Proof. We seek the optimal s by solving the minimization problem

$$s^* = \arg\min_s \|\tilde{g}_t^i - \eta g_t^i\|. \tag{54}$$

We analyze the base case $t = 0$, where $B_0 = 0$. The objective simplifies to

$$\arg\min_s \left\| s^2 g_0^i A_0^\top A_0 - \eta g_0^i \right\|. \tag{55}$$

To obtain a closed-form solution, we invoke the Law of Large Numbers and replace the random matrix $A_0^\top A_0$ with its expected value. The initialization scheme (Leaky ReLU variance [86]) provides entries in A_0 with variance $\sigma_A^2 = 1/(3n)$. The expectation of the matrix product is

$$\mathbb{E}_{A_0}[A_0^\top A_0] = r\sigma_A^2 \mathbf{I}_{n\times n} = \frac{r}{3n}\mathbf{I}_{n\times n}. \tag{56}$$

Substituting this expected value into the minimization objective and assuming the optimal solution corresponds to setting the error term to zero, we require

$$s^2 g_0^i \left(\frac{r}{3n}\mathbf{I}\right) \approx \eta g_0^i. \tag{57}$$

For this approximate equality to hold, the scalar coefficients must match:

$$s^2 \frac{r}{3n} = \eta \quad \Longrightarrow \quad s = \sqrt{\frac{3n\eta}{r}}. \tag{58}$$

This result, derived from the first step and based on a strong expectation approximation, provides the theoretically optimal scaling factor for gradient alignment. Its applicability can be extended to subsequent steps due to the typically small magnitude of relative weight changes in PEFT [23]. □

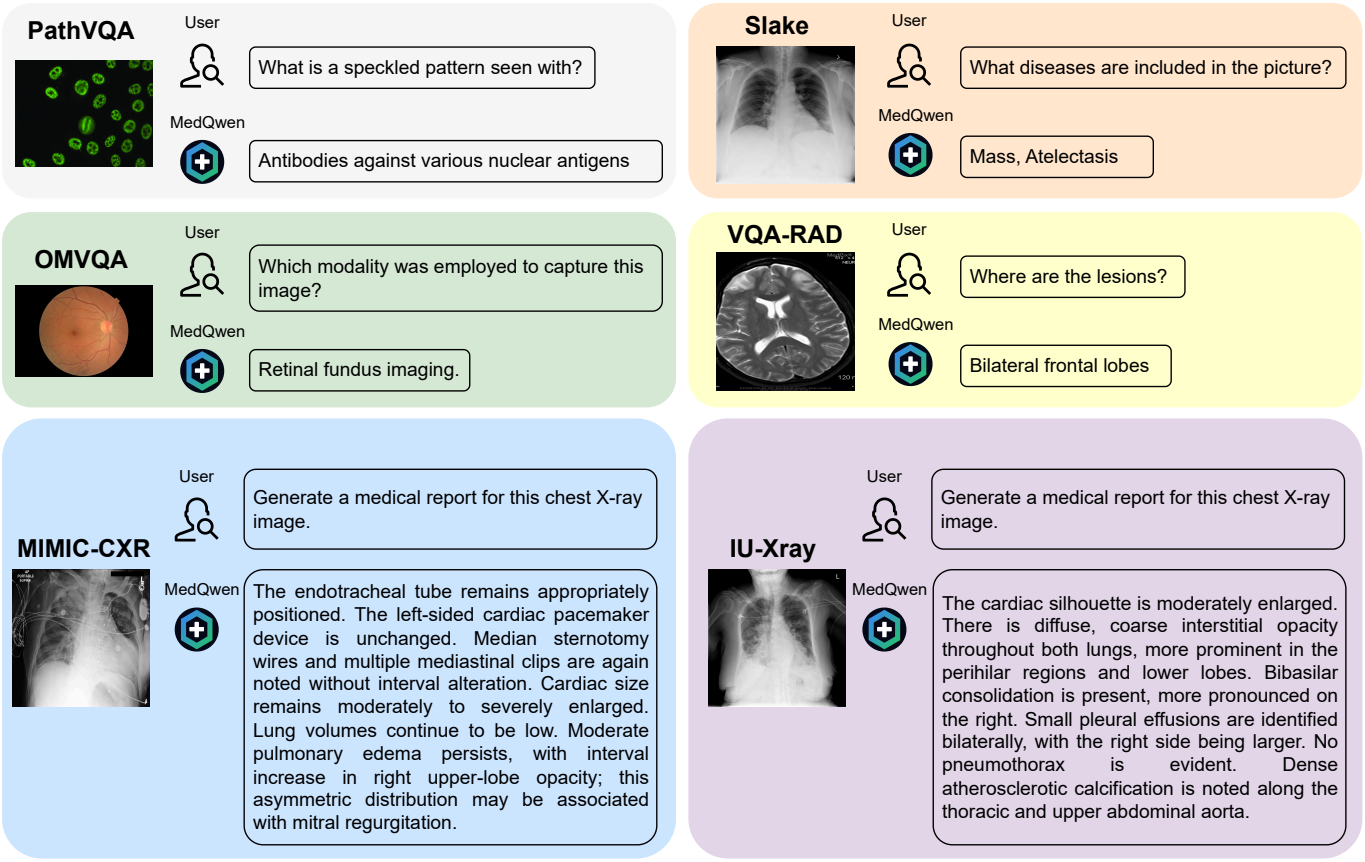


Figure S5. A visualization of MedQwen, demonstrating its ability to process multiple modalities. The top two rows show VQA results, and the bottom row shows report generation.