MEDVESS: FROM SEEING TO UNDERSTANDING MEDICAL IMAGES WITH TRAINING-FREE LATENT STEERING

Anonymous authors
Paper under double-blind review

ABSTRACT

General Vision-Language Models (VLMs) offer a promising foundation for medical image understanding without requiring task-specific model design. However, reliable medical image understanding remains challenging, with insufficient reliance on medical visual evidence and limited medical semantic interpretation of the visual content. To address these limitations, we propose **MedVESS**, a training-free latent steering framework that constructs two complementary, modality-specific steering directions. Visual Evidence Steering (VES) contrasts original medical images with fully masked counterparts to strengthen reliance on medical visual evidence, while Clinical Semantic Steering (CSS) contrasts complete medical reports with semantic-depleted counterparts under the same visual input to reinforce associations between medical visual patterns and corresponding clinical concepts. The two steering vectors are constructed offline and jointly injected into a selected intermediate layer at inference time without updating model parameters. Experiments on four medical VQA benchmarks spanning radiology, fundus photography, and pathology show consistent improvements across two general VLMs, achieving competitive performance against several specialized medical VLMs without medical-domain fine-tuning. These results highlight the potential of training-free latent steering as a lightweight approach to improving reliable medical image understanding in general VLMs.

1 INTRODUCTION

General Vision-Language Models (VLMs) (Liu et al., 2023; Bai et al., 2025b) have achieved substantial progress across a wide range of visual understanding tasks. Nevertheless, their reliability remains a concern in medical imaging, where erroneous predictions may affect clinical decision-making and patient safety (Xia et al., 2024a; Chang et al., 2025). In particular, prior studies suggest that VLMs may not sufficiently rely on the visual evidence provided during medical image reasoning (Zhu et al., 2025; Liang et al., 2026). Consistent with this observation, our preliminary analysis also reveals a similar pattern, as shown in the middle panel of Fig. 1: VLMs retain substantial VQA performance even when the medical image input is removed. By contrast, the same intervention causes substantially larger performance degradation on natural-image tasks, indicating weaker reliance on visual evidence in the medical domain. However, sufficient reliance on visual evidence represents only one aspect of reliable medical image understanding, which also requires accurate medical semantic interpretation of the visual content. Prior studies have shown that general VLMs without explicit medical-domain adaptation remain limited in recognizing medical findings and understanding clinically meaningful visual information (Hu et al., 2024; Yan et al., 2025). This raises a fundamental question for reliable medical image understanding: *How can we simultaneously strengthen VLMs' reliance on medical visual evidence and improve their medical semantic interpretation of the visual content?*

Existing approaches improve the reliability of VLMs in medical settings through domain-specific fine-tuning and alignment (Zhu et al., 2025; Jiang et al., 2025), retrieval- or tool-augmented generation (Xia et al., 2025; Zhang et al., 2026), and specialized inference strategies (Liang et al., 2026). Although effective, these methods often incur substantial overhead. Training-based approaches typically require large-scale, high-quality medical data and additional optimization of large VLMs,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

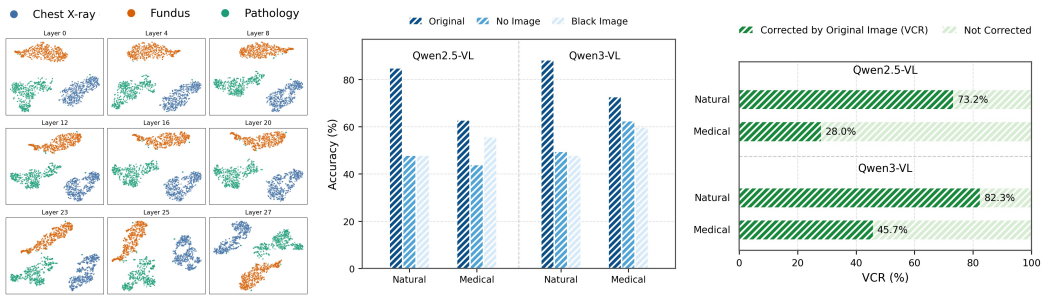

Figure 1: Preliminary analyses of medical visual representations and visual reliance in VLMs. (Left) Hidden visual representations of radiology, fundus, and pathology images across decoder layers. (Middle) VQA accuracy on natural and medical images under the original-image, no-image, and black-image conditions. (Right) Visual Correction Rate (VCR) on natural and medical images, measuring the proportion of black-image errors corrected after restoring the original image.

resulting in considerable data and computational costs (Zhu et al., 2025); retrieval, tool use, and contrastive or iterative decoding further increase system complexity and inference overhead (Xia et al., 2025; Zhang et al., 2026; Liang et al., 2026). These costs motivate a lightweight inference-time approach that keeps model parameters frozen, requires no additional training, and operates without external knowledge, retrieval, or repeated decoding.

Activation steering can be a lightweight alternative to fine-tuning by exploiting directions in the latent space that capture task-relevant representational differences (Liu et al., 2025; Li et al., 2025; Wu et al., 2026a; Yin et al., 2026). Typically, steering vectors are derived from contrasts between hidden representations under different input conditions and injected into selected hidden states at inference time, shifting internal representations toward a desired direction without updating model parameters. In VLMs, such directions are commonly constructed from visual contrasts, such as perturbed versus intact images (Liu et al., 2025), image-present versus image-absent contexts (Li et al., 2025), or image- and object-level differences (Chen et al., 2025). However, directly applying these constructions to medical imaging overlooks two key challenges. First, medical visual representations are highly modality-dependent. Radiographs, fundus photographs, and pathology images differ substantially in their visual distributions, anatomical structures, and diagnostic cues. Our representation analysis (Fig. 1, left panel) further reveals pronounced differences in hidden-state distributions across modalities, suggesting that a single shared visual vector may be insufficient to characterize heterogeneous medical images. Second, reliable medical image understanding depends on sufficient reliance on visual evidence and effective medical semantic interpretation of the visual content. However, generic visual steering primarily increases the influence of visual input, without explicitly reinforcing the associations between medical visual patterns and corresponding clinical concepts. These observations motivate two complementary, modality-specific steering directions that enhance reliance on *medical visual evidence* and facilitate the *medical semantic interpretation of the visual content*.

To this end, we propose **MedVESS** (**Med**ical **V**isual **E**vidence and **S**emantic **S**teering), a training-free latent steering framework that constructs two complementary, modality-specific steering vectors to enhance reliance on medical visual evidence and facilitate the medical semantic interpretation of the visual content. We first construct *Visual Evidence Steering* (VES) by contrasting the original medical image with a fully masked counterpart, yielding a visual direction that encourages greater reliance on the current image. We further construct *Clinical Semantic Steering* (CSS) by contrasting a complete medical report with a semantic-depleted counterpart for the same image, yielding a semantic direction that reinforces associations between medical visual patterns and corresponding clinical concepts. Both vectors are constructed offline and jointly injected into a selected intermediate layer at inference time, without model retraining. During inference, MedVESS operates without external knowledge, retrieval, or additional decoding procedures.

We evaluate MedVESS on four medical VQA benchmarks spanning radiology (Johnson et al., 2019; Demner-Fushman et al., 2016), fundus photography (Luo et al., 2024), and pathology (Lin et al., 2023). Across Qwen2.5-VL-7B (Bai et al., 2025b) and Qwen3-VL-8B (Bai et al., 2025a), Med-

VESS consistently improves medical VQA performance. Notably, without any medical-domain fine-tuning, the steered VLMs achieve performance comparable to several medically adapted models, highlighting the effectiveness of our method as a lightweight adaptation strategy for medical vision-language tasks. Our contributions can be summarized as follows:

1. We reveal substantially weaker visual reliance in medical-image tasks than in natural-image tasks, together with pronounced modality-dependent differences in medical visual representations.
2. We propose **MedVESS**, a training-free latent steering framework that combines modality-specific *Visual Evidence Steering* (VES) and *Clinical Semantic Steering* (CSS) to jointly strengthen visual reliance and medical semantic interpretation of the visual content.
3. MedVESS consistently improves two general VLMs across four medical VQA benchmarks without medical-domain fine-tuning, achieving competitive performance against several medically adapted models.

2 Related Work

2.1 Medical Adaptation of Vision-Language Models

Recent work has increasingly adapted general VLMs to the medical domain to improve medical image understanding. Training-based approaches leverage medical multimodal data for domain-specific optimization. LLaVA-Med (Li et al., 2023a), HuatuoGPT-Vision (Chen et al., 2024), and Lingshu (Xu et al., 2025) adapt general VLMs through biomedical visual instruction tuning and multi-stage medical-domain training. Preference-based methods such as MMedPO (Zhu et al., 2025) and HSCR (Jiang et al., 2025) further introduce clinically informed preference alignment to better align model predictions with medical visual evidence. Another line of work augments generation with retrieved medical knowledge. RULE (Xia et al., 2024b) and MMed-RAG (Xia et al., 2025) develop retrieval-augmented frameworks that incorporate relevant external medical evidence during generation. More recently, ARCD (Liang et al., 2026) performs anatomy-guided contrastive decoding using diagnostically relevant regions to reduce visually unsupported predictions. These studies motivate lightweight inference-time adaptation that better leverages medical visual evidence and clinical semantics, without additional training or external retrieval.

2.2 Activation Steering in Vision-Language Models

Activation steering provides a training-free means of controlling VLMs by directly modifying their internal representations at inference time. Early methods mainly construct directions that restore visual information: VTI (Liu et al., 2025) derives intervention signals from representations stabilized under visual perturbations, while VISTA (Li et al., 2025) contrasts image-present and image-absent contexts to compensate for visual information loss during generation. ICT (Chen et al., 2025) further extends visual intervention from global image-level information to finer object-level details. More recent work incorporates hallucination-specific or semantic signals into steering construction. For example, SHARP (Wu et al., 2025) targets distinct representations associated with textual priors and visual conflicts, while AFTER (Wang et al., 2026a) introduces factual semantic guidance into activation steering. Recent studies further refine the intervention itself through purified directions, sparse-layer editing, or adaptive steering (Wu et al., 2026a;b; Yin et al., 2026; Lin et al., 2026; Liu et al., 2026). In the medical domain, MEDA (Wang et al., 2026b) introduces medical-oriented activation editing by incorporating query-decisive manifestations and diagnostic principles into activation construction. Building on these advances, our work extends activation steering to medical imaging by accounting for modality-dependent visual representations and jointly enhancing visual reliance and medical semantic interpretation of the visual content.

3 Preliminary Analysis

We conduct two empirical analyses to examine modality-dependent visual representations and VLMs' reliance on medical visual evidence.

Modality-Dependent Visual Representations. We first examine the hidden representations of medical images from different modalities, including radiology (Johnson et al., 2019), fundus photography (Luo et al., 2024), and pathology (Lin et al., 2023). Specifically, we extract visual hidden states across decoder layers and visualize their distributions with t-SNE. As shown in the left panel of Fig. 1, representations from different modalities are clearly separated across multiple layers, revealing pronounced modality-dependent structures in the latent space. This suggests that a single shared visual steering direction may be insufficient to characterize heterogeneous medical modalities.

Insufficient Visual Reliance in Medical Imaging. We further examine VLMs' reliance on medical visual evidence by comparing natural images from POPE (Li et al., 2023b) with medical images from MIMIC-CXR (Johnson et al., 2019) under three conditions: the original image, no image, and a black image. As shown in the middle panel of Fig. 1, removing informative visual input causes substantially less performance degradation on medical images than on natural images. These results suggest that, when reasoning over medical images, VLMs under-utilize visual evidence, while their predictions can remain supported by non-visual signals, such as textual cues and pretrained linguistic priors.

To further probe whether visual evidence can overturn predictions formed without informative images, we define the *Visual Correction Rate* (VCR) as the proportion of samples that are corrected by restoring the original image among those answered incorrectly under the black-image condition:

$$\mathrm{VCR} = P(\text{Original Correct} \mid \text{Black-Image Wrong}). \tag{1}$$

As shown in the right panel of Fig. 1, restoring the original image corrects a substantially smaller fraction of black-image errors for medical images than for natural images. This limited correction suggests that medical visual evidence has weaker corrective influence on predictions that can already be supported by non-visual information, further indicating insufficient visual reliance.

4 METHOD

In this section, we present **MedVESS**, a training-free latent steering framework that improves two complementary aspects of medical image understanding: visual reliance and medical semantic interpretation of visual content. For each medical imaging modality, we construct two complementary, modality-specific steering vectors: *Visual Evidence Steering* (VES), which strengthens the model's reliance on medical visual evidence, and *Clinical Semantic Steering* (CSS), which reinforces associations between medical visual patterns and corresponding clinical concepts. We then select an intervention layer offline and jointly inject the two vectors at that layer during inference.

4.1 PRELIMINARIES: LAYER-WISE ACTIVATION STEERING

During offline steering-vector construction, given an image I_i and a question Q_i, we append the corresponding report R_i to the textual input and feed the resulting image–text sequence into a vision-language model $\mathcal{M}$. For each decoder layer ℓ, we use the hidden state of the last valid token as the representation of sample i, denoted by $h_{i,\ell}$.

Activation steering constructs steering vectors from paired inputs that differ in a target property. Given a positive–negative pair (x_i^+, x_i^-), we extract their layer-wise representations $h_{i,\ell}^+$ and $h_{i,\ell}^-$ and compute the activation difference as follows:

$$\Delta h_{i,\ell} = h_{i,\ell}^+ - h_{i,\ell}^-. \tag{2}$$

For each layer, we collect the activation differences from all construction samples and perform principal component analysis (PCA). Following prior latent steering methods (Liu et al., 2025; Wu et al., 2026a), we use the first principal component to obtain the layer-wise steering vector v_ℓ, capturing the dominant variation shared across paired samples while reducing sample-specific noise.

During inference, steering is performed by adding the corresponding vector to the hidden representation at a selected decoder layer:

$$\widetilde{h}_\ell = h_\ell + \alpha v_\ell, \tag{3}$$

where α controls the steering strength. In the following sections, we describe how the positive and negative pairs are constructed to obtain the visual and semantic steering vectors.

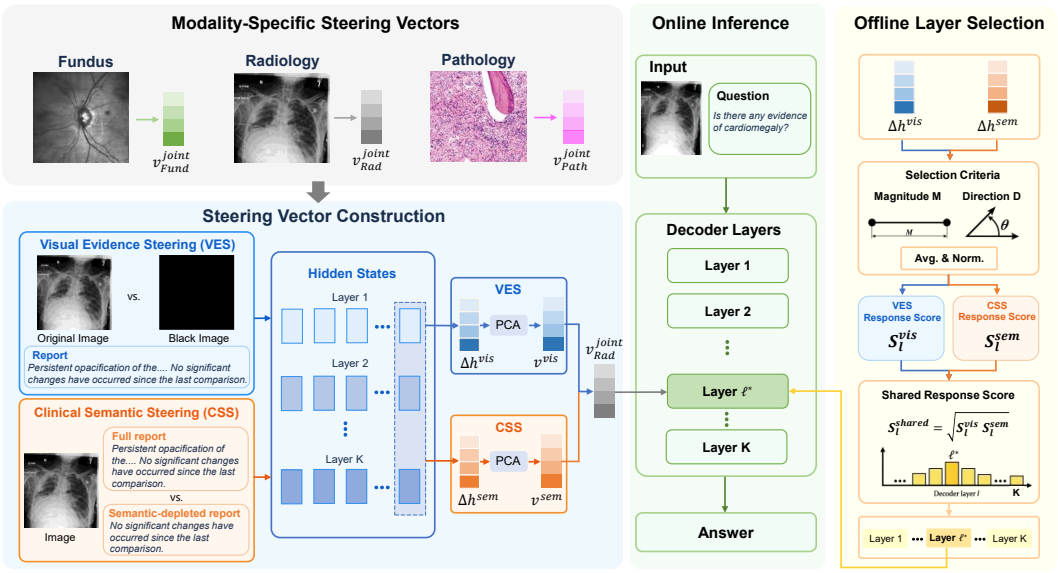

Figure 2: Overview of MedVESS, a training-free latent steering framework for medical image understanding. It constructs modality-specific steering vectors for different medical imaging modalities, with radiology used as an example in the subsequent pipeline. VES and CSS derive complementary visual and clinical-semantic steering vectors, while an offline criterion selects the intervention layer based on layer-wise responses. The selected vectors are then jointly injected during inference without parameter updates.

4.2 Modality-Specific Steering Construction

Different medical modalities differ substantially in visual structures and diagnostic patterns, while also exhibiting clearly distinct internal representation distributions, as observed in our preliminary analysis. We therefore construct steering vectors independently for each medical modality m. Specifically, the modality-specific construction set is defined as follows:

$$\mathcal{S}_m = \{(I_i, Q_i, R_i)\}_{i=1}^{N}, \tag{4}$$

where I_i, Q_i, and R_i denote the medical image, the corresponding question, and the paired medical report, respectively, and N denotes the number of construction samples used for each modality. Based on $\mathcal{S}_m$, we construct modality-specific visual and semantic steering vectors $v_{m,\ell}^{\text{vis}}$ and $v_{m,\ell}^{\text{sem}}$, respectively. For notational clarity, we omit the modality index m from sample-level variables and retain it for modality-specific quantities defined over $\mathcal{S}_m$.

4.3 Visual Evidence Steering

Given the weaker reliance of VLMs on visual evidence in medical imaging, we construct *Visual Evidence Steering* (VES) to strengthen the influence of the current image on the model's internal representations. For each construction sample (I_i, Q_i, R_i), we fully mask the original medical image I_i to obtain a black image I_i^{black} with no informative visual content, which serves as the visual negative sample while the question and report remain unchanged. We therefore construct the VES positive–negative pair as follows:

$$x_{i,\text{vis}}^{+} = (I_i, Q_i, R_i), \qquad x_{i,\text{vis}}^{-} = (I_i^{\text{black}}, Q_i, R_i). \tag{5}$$

This contrast attributes the resulting activation change to the presence of informative medical visual content. Following the activation steering procedure described above, we compute the layer-wise activation differences and obtain the modality-specific VES vector as follows:

$$\Delta h_{i,\ell}^{\text{vis}} = h_{i,\ell}^{+,\text{vis}} - h_{i,\ell}^{-,\text{vis}}, \qquad v_{m,\ell}^{\text{vis}} = \text{PC}_1\left(\{\Delta h_{i,\ell}^{\text{vis}}\}_{i=1}^{N}\right). \tag{6}$$

The resulting VES vector encodes the shared representation shift associated with informative visual input within each medical modality. During inference, it steers the hidden representations toward states that are more strongly conditioned on the current medical image.

4.4 Clinical Semantic Steering

VES strengthens the model's reliance on medical visual evidence, while reliable medical image understanding additionally requires accurate medical semantic interpretation of the visual content. A key aspect of such interpretation is associating relevant visual evidence with corresponding medical findings and clinical concepts. We therefore introduce *Clinical Semantic Steering* (CSS) to reinforce these associations.

For each full report R_i, we first identify the target clinical semantics and remove the corresponding content while preserving the remaining non-target textual context. The residual content is then rewritten into a coherent form to obtain a semantic-negative report R_i^-. We construct the CSS positive–negative pair as follows:

$$x_{i,\mathrm{sem}}^{+} = (I_i, Q_i, R_i), \qquad x_{i,\mathrm{sem}}^{-} = (I_i, Q_i, R_i^-). \tag{7}$$

This pair contrasts the presence and removal of target clinical semantics while keeping the visual input fixed. Following the activation steering procedure described above, we compute the modality-specific CSS vector as follows:

$$\Delta h_{i,\ell}^{\mathrm{sem}} = h_{i,\ell}^{+,\mathrm{sem}} - h_{i,\ell}^{-,\mathrm{sem}}, \qquad v_{m,\ell}^{\mathrm{sem}} = \mathrm{PC}_1\left(\left\{\Delta h_{i,\ell}^{\mathrm{sem}}\right\}_{i=1}^{N}\right). \tag{8}$$

The resulting CSS vector captures the dominant representational difference associated with the presence versus removal of target clinical semantics under the same medical image. During inference, CSS complements VES by strengthening the model's medical semantic interpretation of the visual content.

4.5 Offline Layer Selection

Different decoder layers exhibit distinct multimodal representation characteristics, making the effectiveness of steering sensitive to the intervention layer. We therefore select the intervention layer offline based on the layer-wise hidden-state changes of the construction samples. When the target information changes between the paired conditions, a stronger response in a layer's hidden representation indicates that the layer is more closely related to the corresponding information and is thus a more suitable intervention position. We quantify this layer-wise response from two complementary perspectives: change magnitude and representation direction. For steering type $c \in \{\mathrm{vis}, \mathrm{sem}\}$, we define the following two measures:

$$M_{i,\ell}^{c} = \frac{\left\| h_{i,\ell}^{+,c} - h_{i,\ell}^{-,c} \right\|_2}{\frac{1}{2}\left(\left\| h_{i,\ell}^{+,c} \right\|_2 + \left\| h_{i,\ell}^{-,c} \right\|_2 \right)}, \qquad D_{i,\ell}^{c} = 1 - \cos\left(h_{i,\ell}^{+,c}, h_{i,\ell}^{-,c} \right), \tag{9}$$

where $M_{i,\ell}^{c}$ measures the relative hidden-state change magnitude and $D_{i,\ell}^{c}$ measures the change in representation direction. We average both measures over the construction samples and normalize them independently across decoder layers, yielding the layer response score:

$$\bar{M}_{m,\ell}^{c} = \frac{1}{N}\sum_{i=1}^{N} M_{i,\ell}^{c}, \qquad \bar{D}_{m,\ell}^{c} = \frac{1}{N}\sum_{i=1}^{N} D_{i,\ell}^{c}, \qquad S_{m,\ell}^{c} = \frac{1}{2}\left(\frac{\bar{M}_{m,\ell}^{c}}{\max_{\ell'} \bar{M}_{m,\ell'}^{c}} + \frac{\bar{D}_{m,\ell}^{c}}{\max_{\ell'} \bar{D}_{m,\ell'}^{c}} \right). \tag{10}$$

Since VES and CSS are jointly applied at the same decoder layer, we combine their layer response scores using the geometric mean and select the layer that responds strongly to both visual and semantic changes:

$$S_{m,\ell}^{\mathrm{shared}} = \sqrt{S_{m,\ell}^{\mathrm{vis}} S_{m,\ell}^{\mathrm{sem}}}, \qquad \ell_m^{*} = \arg\max_{\ell} S_{m,\ell}^{\mathrm{shared}}. \tag{11}$$

After determining ℓ_m^{*}, we directly combine the corresponding VES and CSS steering vectors and inject the joint vector into the hidden representation during inference:

$$v_{m,\ell_m^{*}}^{\mathrm{joint}} = v_{m,\ell_m^{*}}^{\mathrm{vis}} + v_{m,\ell_m^{*}}^{\mathrm{sem}}, \qquad \widetilde{h}_{\ell_m^{*}} = h_{\ell_m^{*}} + \alpha v_{m,\ell_m^{*}}^{\mathrm{joint}}. \tag{12}$$

Table 1: Comparison of MedVESS with state-of-the-art medical VLMs and representative medical adaptation and retrieval methods on MIMIC-CXR, IU-Xray, PMC-OA, and Harvard-FairVLMed (abbreviated as FairVLMed). The best results are shown in **bold**, and the second-best results are underlined.

Methods	MIMIC-CXR		IU-Xray		PMC-OA		FairVLMed	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Medical Adaptation and Retrieval Methods								
RULE	62.04	38.21	58.07	38.33	–	–	86.50	65.84
MMed-RAG	63.82	39.41	70.91	47.42	–	–	66.22	64.66
MMedPO	60.94	37.76	68.39	45.07	–	–	–	–
ExGra-Med	45.55	30.34	49.30	34.24	51.10	47.36	76.84	59.20
Specialized Medical VLMs								
LLaVA-Med	56.69	35.15	57.07	46.46	59.68	45.47	**88.40**	62.15
HuatuoGPT-Vision-7B	70.89	47.59	**92.98**	**81.20**	67.85	67.37	63.55	54.25
Lingshu-7B	76.02	**49.44**	91.32	74.61	**68.51**	**67.90**	74.33	**66.19**
Training-Free Latent Steering								
Qwen2.5-VL-7B	62.55	42.99	81.44	65.86	56.42	53.25	18.82	18.26
+ Ours	67.57	44.27	81.90	68.18	67.40	66.51	37.95	36.23
Δ	+5.02	+1.28	+0.46	+2.32	+10.98	+13.26	+19.13	+17.97
Qwen3-VL-8B	72.46	46.99	91.47	76.53	61.84	61.34	32.79	32.50
+ Ours	**79.00**	49.01	92.18	77.60	68.05	67.66	67.53	53.78
Δ	+6.54	+2.02	+0.71	+1.07	+6.21	+6.32	+34.74	+21.28

5 Experiment

5.1 Experimental Setup

Benchmarks and Metrics. We evaluate MedVESS on four medical VQA benchmarks spanning three imaging modalities: radiology, fundus photography, and pathology. For radiology, we use the MIMIC-CXR (Johnson et al., 2019) and IU-Xray (Demner-Fushman et al., 2016) question sets provided by MedHEval (Chang et al., 2025). For fundus and pathology, we follow the evaluation setting of MMed-RAG (Xia et al., 2025) and adopt Harvard-FairVLMed (Luo et al., 2024) and the pathology subset of PMC-OA (Lin et al., 2023), respectively. We report Accuracy and F1-score as the evaluation metrics. For Harvard-FairVLMed and PMC-OA, which only contain binary Yes/No questions, we report Macro-F1 over the two labels. For MIMIC-CXR and IU-Xray, which contain both binary and multiple-choice questions, we compute Macro-F1 separately for the two question types and report their equal-weight average. Additional details on dataset construction and evaluation splits are provided in Appendix A.

Baselines. We apply MedVESS to two general VLM backbones, Qwen2.5-VL-7B (Bai et al., 2025b) and Qwen3-VL-8B (Bai et al., 2025a). For comparison, we include three specialized medical VLMs: LLaVA-Med (Li et al., 2023a), HuatuoGPT-Vision-7B (Chen et al., 2024), and Lingshu-7B (Xu et al., 2025). We further compare against representative medical VLMs spanning retrieval-augmented, preference-based, and training-based paradigms, including RULE (Xia et al., 2024b), MMed-RAG (Xia et al., 2025), MMedPO (Zhu et al., 2025), and ExGra-Med (Nguyen et al., 2025).

Implementation Details. For each medical imaging modality, we construct the steering vectors using 500 construction samples. For radiology, we use construction samples from the MIMIC-CXR *Visual Misinterpretation* subset of MedHEval (Chang et al., 2025). For fundus photography and pathology, the construction samples are drawn from the training splits of Harvard-FairVLMed (Luo et al., 2024) and PMC-OA (Lin et al., 2023), respectively, following the data organization adopted by MMed-RAG (Xia et al., 2025), with no overlap with the evaluation sets. Further details on semantic-negative report generation and inference settings are provided in the Appendices C, E, and D.

5.2 Main Results

As shown in Table 1, MedVESS consistently improves both Qwen2.5-VL-7B (Bai et al., 2025b) and Qwen3-VL-8B (Bai et al., 2025a) across all four medical VQA benchmarks, with particularly notable gains on MIMIC-CXR, PMC-OA, and Harvard-FairVLMed. On IU-Xray (Demner-Fushman et al., 2016), where the vanilla backbones already achieve strong performance, MedVESS provides smaller but consistent improvements. Harvard-FairVLMed exhibits the largest absolute gain, with Qwen3-VL-8B improving from 32.79% to 67.53% in accuracy and from 32.50% to 53.78% in Macro-F1. Given the highly imbalanced binary label distribution, we interpret this result jointly with class-wise performance rather than raw accuracy alone. Notably, the resulting Macro-F1 exceeds that of an always-Yes baseline (53.78% vs. 46.36%), showing that the improvement cannot be explained by trivial majority-class prediction alone. Detailed class-wise analysis is provided in Appendix F.

Compared with specialized medical VLMs (Li et al., 2023a; Chen et al., 2024; Xu et al., 2025) and existing medical adaptation and retrieval methods (Xia et al., 2024b; 2025; Zhu et al., 2025; Nguyen et al., 2025), MedVESS remains highly competitive. Qwen3-VL-8B with MedVESS achieves the best accuracy on MIMIC-CXR and closely matches the strongest medical-specialized model on PMC-OA, while maintaining competitive performance on IU-Xray. Notably, these comparisons are made against models that have undergone dedicated medical-domain training. For example, Lingshu (Xu et al., 2025) explicitly incorporates MIMIC-CXR, IU-Xray, PMC-OA, and Harvard-FairVLMed into its training corpus, while LLaVA-Med (Li et al., 2023a) and HuatuoGPT-Vision (Chen et al., 2024) are also trained on large-scale biomedical multimodal data. In contrast, MedVESS requires no medical-domain fine-tuning or parameter updates, relying only on lightweight offline steering-vector construction and inference-time intervention while achieving competitive performance across multiple medical benchmarks.

5.3 Ablation Studies

Effect of MedVESS Components. We first evaluate the individual contributions of VES and CSS. Applying either component alone consistently improves Qwen3-VL-8B (Bai et al., 2025a) across all four benchmarks, demonstrating the effectiveness of both steering signals. Combining VES and CSS yields further improvements on every dataset and achieves the best overall performance, indicating that the two components play complementary roles. CSS yields larger gains than VES, particularly on PMC-OA (Lin et al., 2023) and Harvard-FairVLMed (Luo et al., 2024).

Table 2: Component ablation on Qwen3-VL-8B. MIMIC, IU, PMC, and FairVL denote MIMIC-CXR, IU-Xray, PMC-OA, and Harvard-FairVLMed, respectively.

Components		Performance			
VES	**CSS**	**MIMIC**	**IU**	**PMC**	**FairVL**
–	–	72.46	91.47	61.84	32.79
✓	–	74.57	91.57	62.89	33.87
–	✓	78.46	91.83	67.70	65.54
✓	✓	**79.00**	**92.18**	**68.05**	**67.53**

Effect of Modality-Specific Construction. We further examine whether the modality differences observed in our representation analysis affect steering-vector construction. For a controlled comparison, we construct shared steering vectors from balanced radiology, fundus, and pathology samples while keeping the total number of construction samples identical to the modality-specific setting. As shown in Table 3, modality-specific construction consistently outperforms shared construction across all four benchmarks, with the largest improvement on Harvard-FairVLMed (Luo et al., 2024). This result is consistent with the modality-dependent representation differences identified in our preliminary analysis, indicating that representation characteristics differ across medical imaging modalities and supporting modality-specific steering construction.

Table 3: Ablation of modality-specific steering construction on Qwen3-VL-8B. Dataset abbreviations follow Table 2.

Construction	**MIMIC**	**IU**	**PMC**	**FairVL**
Vanilla	72.46	91.47	61.84	32.79
Shared	77.64	91.78	66.30	54.36
Per-Modality	**79.00**	**92.18**	**68.05**	**67.53**

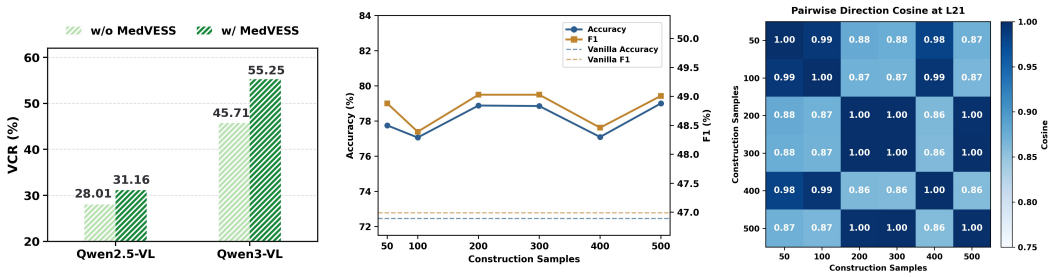

Figure 3: Further analysis of MedVESS. (Left) VCR with and without MedVESS. (Middle) Accuracy and F1 across construction-set sizes. (Right) Pairwise cosine similarity between steering vectors across construction-set sizes.

Effectiveness of Offline Layer Selection. To assess the effectiveness of our offline layer-selection strategy, we determine the intervention layer offline from the construction set and then compare the selected layer with other decoder layers during evaluation. As shown in Table 4, performance varies substantially across intervention layers, while the layer selected by our offline criterion, $L21$, achieves the best performance among all evaluated candidates. We additionally test an all-layer setting, with steering vectors injected into every decoder layer following a commonly used intervention strategy (Liu et al., 2025; Li et al., 2025). Under this setting, both metrics collapse, while the generated responses degenerate into repetitive and semantically meaningless outputs. These results validate the effectiveness of offline layer selection in identifying suitable intervention layers.

Table 4: Layer-selection analysis on Qwen3-VL-8B over MIMIC-CXR.

Intervention	Acc	F1
Vanilla	72.46	46.99
L13	72.91	46.95
L17	72.82	46.81
L21 (Selected)	**79.00**	**49.01**
L25	74.16	48.09
L29	71.16	46.38
All Layers	0.05	0.17

5.4 Further Analysis

Utilization of Visual Evidence. Our preliminary analysis reveals that restoring the original medical image corrects only a limited fraction of errors made under the black-image condition. We therefore examine whether MedVESS enables more effective use of the restored visual evidence. As shown in the left panel of Fig. 3, VCR increases from 28.01% to 31.16% on Qwen2.5-VL and from 45.71% to 55.25% on Qwen3-VL. This increase indicates that MedVESS enables the restored image to correct a larger fraction of predictions made without informative visual input, suggesting improved visual reliance in medical-image reasoning.

Data Efficiency of Steering Construction. We analyze the data requirement for steering-vector construction using Qwen3-VL-8B (Bai et al., 2025a) on MIMIC-CXR (Johnson et al., 2019). As shown in Fig. 3, accuracy and F1 remain stable as the construction set varies from 50 to 500 samples and consistently outperform the vanilla baseline, indicating that effective steering can be achieved with a relatively small construction set. Steering vectors constructed with different sample sizes also exhibit high pairwise cosine similarity, showing consistent directions across construction-set sizes.

6 Conclusion

In this work, we identify two major limitations of general VLMs in medical image understanding: insufficient reliance on medical visual evidence and limited medical semantic interpretation of the visual content. Motivated by these limitations, we introduce MedVESS, a training-free latent steering framework that combines modality-specific Visual Evidence Steering and Clinical Semantic Steering with offline layer selection. Experiments across three modalities demonstrate consistent improvements on two general VLMs. As a lightweight and effective approach to reliable medical image understanding, MedVESS achieves competitive performance against several specialized medical VLMs while requiring no medical-domain fine-tuning or parameter updates.

Reproducibility Statement

We have taken steps to facilitate the reproducibility of our work. The construction of MedVESS, including modality-specific steering vectors, Visual Evidence Steering (VES), Clinical Semantic Steering (CSS), and the offline layer-selection criterion, is described in Sections 4.2–4.5. The experimental benchmarks, model backbones, evaluation metrics, and construction-set setup are provided in Section 5.1. Appendix A provides detailed information on the evaluation and construction data, while Appendix B specifies the PCA-based steering-vector construction and inference-time intervention procedure. The construction of semantic-negative reports, including the complete prompts used for semantic extraction and rewriting, is provided in Appendix C. Additional analyses of intervention-layer selection and steering strength, together with the configurations used in the main experiments, are reported in Appendices D and E. We will publicly release the source code and configuration files upon publication to facilitate reproduction and further research.

AI Use Statement

We used generative AI tools for assistance with language editing, literature retrieval, and parts of code generation and debugging. Gemini-3.1-Pro was additionally used to construct semantic-negative reports for Harvard-FairVLMed and PMC-OA, as described in Appendix C. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final content of this work.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo 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. *arXiv preprint arXiv:2502.13923*, 2025b.

Aofei Chang, Le Huang, Parminder Bhatia, Taha Kass-Hout, Fenglong Ma, and Cao Xiao. MedHEval: Benchmarking hallucinations and mitigation strategies in medical large vision-language models. *arXiv preprint arXiv:2503.02157*, 2025.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and Benyou Wang. Towards injecting medical visual knowledge into multimodal LLMs at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7346–7370. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.418.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. ICT: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4209–4221, 2025. doi: 10.1109/CVPR52734.2025.00398.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. doi: 10.1093/jamia/ocv080.

Google DeepMind. Gemini 3.1 pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, February 2026.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22170–22183, 2024.

Songtao Jiang, Yan Zhang, Yeying Jin, Zhihang Tang, Yangyang Wu, Yang Feng, Jian Wu, and Zuozhu Liu. HSCR: Hierarchical self-contrastive rewarding for aligning medical vision language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13853–13868. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.679.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6:317, 2019. doi: 10.1038/s41597-019-0322-0.

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. In *Advances in Neural Information Processing Systems*, volume 36, pp. 28541–28564, 2023a. doi: 10.52202/075280-1240.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.emnlp-main.20.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35799–35819. PMLR, 2025.

Xiao Liang, Chenxi Liu, Zhi Ma, Di Wang, Bin Jing, Quan Wang, and Yuanyuan Shi. Anatomical region-guided contrastive decoding: A plug-and-play strategy for mitigating hallucinations in medical VLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6871–6879, 2026. doi: 10.1609/aaai.v40i9.37620.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-CLIP: Contrastive language-image pre-training using biomedical documents. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pp. 525–536. Springer Nature Switzerland, 2023. doi: 10.1007/978-3-031-43993-3_51.

Yangguang Lin, Quan Fang, Yufei Li, Jiachen Sun, Junyu Gao, and Jitao Sang. HulluEdit: Single-pass evidence-consistent subspace editing for mitigating hallucinations in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11086–11095, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1516.

Sheng Liu, Haotian Ye, and James Y. Zou. Reducing hallucinations in large vision-language models via latent space steering. In *The Thirteenth International Conference on Learning Representations*, 2025.

Shuliang Liu, Songbo Yang, Dong Fang, Sihang Jia, Yuqi Tang, Lingfeng Su, Ruoshui Peng, Yibo Yan, Xin Zou, and Xuming Hu. Vision-language introspection: Mitigating overconfident hallucinations in MLLMs via interpretable bi-causal steering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38518–38543. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1784.

Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, Yi Fang, and Mengyu Wang. FairCLIP: Harnessing fairness in vision-language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12289–12301, 2024. doi: 10.1109/CVPR52733.2024.01168.

Duy M. H. Nguyen, Nghiem Diep, Trung Nguyen, Hoang-Bao Le, Tai D. Nguyen, Anh-Tien Nguyen, TrungTin Nguyen, Nhat Ho, Pengtao Xie, Roger Wattenhofer, Daniel Sonntag, James Y. Zou, and Mathias Niepert. ExGra-Med: Extended context graph alignment for medical vision-language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3657.

Tianbo Wang, Yuqing Ma, Kewei Liao, Zhange Zhang, Simin Li, Jinyang Guo, and Xianglong Liu. AFTER: Mitigating the object hallucination of LVLM via adaptive factual-guided activation editing. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Tianbo Wang, Yuqing Ma, Lingyan Meng, Zhange Zhang, Kewei Liao, Jian Yang, Simin Li, Jinyang Guo, and Xianglong Liu. MEDA: Medical-oriented activation editing for hallucination mitigation in medical large vision-language model. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=VzZxoRiU3G.

Jialin Wu, Wei Shi, Han Shen, Peigui Qi, Kunsheng Tang, Zhicong Huang, Binghao Wang, and Zhou Yang. REVIS: Sparse latent steering to mitigate object hallucination in large vision-language models. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=92tssj0P1S.

Jingwen Wu, Xijun Zhang, and Ge Song. Rethinking visual neglect: Steering via context-preference for MLLM hallucination mitigation, 2026b.

Junfei Wu, Yue Ding, Guofan Liu, Tianze Xia, Ziyue Huang, Dianbo Sui, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. SHARP: Steering hallucination in LVLMs via representation engineering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14346–14361. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.725.

Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, Wenhao Zheng, Zhaoyang Wang, Xiao Wang, Xuchao Zhang, Chetan Bansal, Marc Niethammer, Junzhou Huang, Hongtu Zhu, Yun Li, Jimeng Sun, Zongyuan Ge, Gang Li, James Zou, and Huaxiu Yao. CARES: A comprehensive benchmark of trustworthiness in medical vision language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. doi: 10.52202/079017-4455.

Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. RULE: Reliable multimodal RAG for factuality in medical vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1081–1093. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.emnlp-main.62.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Y. Zou, and Huaxiu Yao. MMed-RAG: Versatile multimodal RAG system for medical vision language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

Qianqi Yan, Xuehai He, Xiang Yue, and Xin Eric Wang. Worse than random? an embarrassingly simple probing evaluation of large multimodal models in medical vqa. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19188–19205, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.981.

Jianghao Yin, Qin Chen, Kedi Chen, Jie Zhou, Xingjiao Wu, and Liang He. Dynamic multimodal activation steering for hallucination mitigation in large vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Kai Zhang, Corey D. Barrett, Jangwon Kim, Lichao Sun, Tara Taghavi, and Krishnaram Kenthapadi. RadAgents: Multimodal agentic reasoning for chest x-ray interpretation with radiologist-like workflows. In *Proceedings of the 9th International Conference on Medical Imaging with Deep Learning*, volume 315 of *Proceedings of Machine Learning Research*, pp. 3496–3519. PMLR, 2026.

Kangyu Zhu, Peng Xia, Yun Li, Hongtu Zhu, Sheng Wang, and Huaxiu Yao. MMedPO: Aligning medical vision-language models with clinical-aware multimodal preference optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80207–80222. PMLR, 2025.

A Dataset and Construction Data Details

We evaluate MedVESS on four medical VQA benchmarks spanning radiology, fundus photography, and pathology. For radiology, we adopt the closed-ended VQA data from MedHEval (Chang et al., 2025). The MIMIC-CXR (Johnson et al., 2019) evaluation set consists of the context misalignment, knowledge deficiency, and visual misinterpretation closed-ended subsets, resulting in 7,410 evaluation samples. For IU-Xray (Demner-Fushman et al., 2016), we use the closed-ended visual misinterpretation subset, which contains 1,994 samples. MIMIC-CXR and IU-Xray share the same radiology steering vectors. The construction set is drawn from the independent open-ended subset of Visual Misinterpretation Hallucination in MedHEval. Each sample contains a chest X-ray image, a corresponding question, and a medical report. We use 500 samples for steering-vector construction.

For fundus photography and pathology, we follow the data organization of MMed-RAG (Xia et al., 2025) and use the VQA test splits of Harvard-FairVLMed (Luo et al., 2024) and PMC-OA (Lin et al., 2023) for evaluation, containing 4,266 and 1,994 samples, respectively. For PMC-OA, we use the pathology subset. The corresponding steering vectors are constructed from the fundus photography and pathology training data, respectively, using 500 samples with paired medical images and reports for each modality.

B Additional Method Details

We provide additional implementation details for clarity and reproducibility, including the PCA-based construction of steering vectors and the token-level intervention strategy used during inference.

PCA-Based Steering Vector Construction. For each decoder layer ℓ, we first compute the mean activation difference

$$\bar{\Delta h}_\ell = \frac{1}{N}\sum_{i=1}^{N} \Delta h_{i,\ell},$$

and perform PCA on the centered differences $\Delta h_{i,\ell} - \bar{\Delta h}_\ell$. Let u_ℓ denote the unit-norm first principal direction. Since the sign of a principal component is arbitrary, we align it with the mean activation difference:

$$u_\ell^\top \bar{\Delta h}_\ell < 0 \quad \Rightarrow \quad u_\ell \leftarrow -u_\ell.$$

We then restore the typical magnitude of the activation differences using their average ℓ_2 norm,

$$s_\ell = \frac{1}{N}\sum_{i=1}^{N} \|\Delta h_{i,\ell}\|_2, \qquad v_\ell = s_\ell u_\ell.$$

Accordingly, the notation $\mathrm{PC}_1(\{\Delta h_{i,\ell}\}_{i=1}^N)$ used in the main text denotes this sign-aligned and magnitude-restored first-principal-component steering vector v_ℓ, rather than the unit PCA direction u_ℓ alone. This construction preserves both the dominant shared direction captured by PCA and the typical scale of the activation shifts.

Inference-Time Token Intervention. At inference time, MedVESS applies steering only to the last-token hidden state at the selected decoder layer ℓ^*:

$$\tilde{h}_{\ell^*,-1} = h_{\ell^*,-1} + \alpha v_{\ell^*}^{\text{joint}},$$

where α controls the steering strength. The intervention is applied during both prefill and autoregressive decoding. During prefill, only the last token of the input sequence is modified; during decoding, the same intervention is applied to the current token at every generation step.

C Semantic-Negative Report Construction

We provide additional details on the construction of semantic-negative reports used for CSS. Since the available textual annotations differ across medical modalities, we adopt dataset-specific construction procedures. Representative examples of the original text, removed semantic targets, and resulting semantic-negative text are provided below.

Radiology. For MIMIC-CXR (Johnson et al., 2019), we use the open-ended subset of Visual Misinterpretation Hallucination from MedHEval (Chang et al., 2025). The radiology reports in this subset have already been preprocessed by MedHEval to remove noisy content, and each report is accompanied by structured symptom/finding annotations. We therefore directly use these structured findings as semantic targets and remove the corresponding sentences from the original report. Sentence matching considers both textual containment and token overlap to account for different surface forms. Sentences that do not match any semantic target are retained. If removing the target findings leaves no meaningful residual content, we use the neutral placeholder "No specific visual finding is provided." as the semantic-negative report. MIMIC-CXR and IU-Xray (Demner-Fushman et al., 2016) share the radiology CSS vector constructed using this procedure.

Fundus Photography and Pathology. Harvard-FairVLMed (Luo et al., 2024) and PMC-OA (Lin et al., 2023) do not provide structured finding annotations in the same form as MedHEval. We therefore adopt a two-stage semantic extraction and rewriting procedure using Gemini-3.1-Pro (Google DeepMind, 2026). First, modality-specific semantic targets are extracted from the original report or caption. For Harvard-FairVLMed, the extracted targets mainly include diagnoses, ophthalmic abnormalities, anatomical or structural findings, measurements, disease status, and clinically meaningful negative findings. For PMC-OA, the targets mainly include tissue or cell identity, morphology, biological structures, expression or staining patterns, spatial distributions, disease-related states, and quantitative biological observations. The extraction prompts require the model to use only information explicitly stated in the source text and to preserve attributes such as negation, uncertainty, laterality, and temporal status. A shared rewrite prompt is then used to remove all extracted semantic targets, including their paraphrases, measurements, and implied expressions, while preserving unrelated background or experimental context whenever possible. If no meaningful non-target content remains, a neutral placeholder is used as the semantic-negative text. The complete prompts are provided below.

Representative Examples. The following examples illustrate the semantic-negative construction across the three modalities.

Radiology (MIMIC-CXR).

> **Original report:** The image reveals persistent opacification of the right lower lung field, likely due to pleural effusion and atelectasis. There is also a small left pleural effusion. No significant changes have occurred since the last comparison. There is no evidence of pneumothorax. **Removed semantic targets:** persistent opacification; pleural effusion; atelectasis; small left pleural effusion; no pneumothorax **Semantic-negative report:** No significant changes have occurred since the last comparison.

Fundus Photography (Harvard-FairVLMed).

> **Original report:** Patient was referred for narrow angle evaluation, has CCT 646/645 with a Tmax of 17/16. Patient has potential glaucoma, but no family history. Possible optic nerve defects observed. No symptoms of angle closure. Gonio shows narrow angles.

Removed semantic targets: CCT 646/645; Tmax of 17/16; potential glaucoma; possible optic nerve defects; no symptoms of angle closure; narrow angles.

Semantic-negative report: Patient was referred for evaluation. Patient has no family history.

Pathology (PMC-OA).

Original caption: Microscopic observation of the gastric fundic mucosa layer from a rabbit inoculated with C. guttulatus at 200× magnification.

Removed semantic targets: gastric fundic mucosa layer; microscopic observation.

Semantic-negative caption: From a rabbit inoculated with C. guttulatus.

Prompt Design and Validation. For Harvard-FairVLMed and PMC-OA, we use modality-specific prompts for semantic-target extraction and a shared prompt for semantic-negative rewriting. The extraction prompts constrain Gemini-3.1-Pro to identify only explicitly stated medical or biomedical information without introducing additional inference. The rewriting prompt then removes all extracted targets while preserving only non-target content that does not reveal the same medical meaning.

FUNDUS PHOTOGRAPHY SEMANTIC EXTRACTION PROMPT

```
You are extracting medically meaningful information from an ophthalmology
    clinical report.

Your task is to separate the report into two types of medical information
    .

### 1. core_medical_semantics

Extract information that directly describes the patient's ophthalmic
    clinical state, including:

- diagnosis or suspected diagnosis
- symptoms
- structural or anatomical findings
- ophthalmic abnormalities
- OCT / RNFL / GCL findings
- visual field / HVF findings
- intraocular pressure findings
- measurements
- disease severity
- progression, improvement, worsening, or stability
- meaningful negative ophthalmic findings that explicitly state a normal
    or absent ophthalmic condition

### 2. auxiliary_clinical_context

Extract medically relevant contextual information that does not directly
    describe the core ophthalmic state, including:

- medication or treatment
- surgery or procedure
- ocular history
- other medical history
- family history
- follow-up or management plan
- demographic information
- technical or examination context

### Rules

- Only extract information explicitly stated in the report.
- Do not infer or add new medical information.
```

```
- Do not change the original clinical meaning.
- Preserve negation, uncertainty, severity, laterality, and temporal
    status.
- Keep each extracted item concise.
- For each item, provide the shortest exact text span from the original
    report that supports it.
- If no meaningful core ophthalmic information is present, return an
    empty core_medical_semantics list.
- If the report says only that a disease is not mentioned, not documented
    , or not indicated, do not treat that as a negative finding.
- Return valid JSON only.

### Input

REPORT:

{{REPORT}}
```

PATHOLOGY SEMANTIC EXTRACTION PROMPT

```
You are extracting biomedical information from a scientific figure
    caption.

Your task is to separate the caption into three types of information.

### 1. core_biomedical_semantics

Extract substantive biological or medical information explicitly stated
    in the caption, including:

- tissue or cell identity
- morphology
- biological structures
- staining or expression patterns
- molecular or cellular localization
- spatial distribution
- biological state
- disease-related state
- abnormalities
- normal biological state
- comparisons between biological groups
- quantitative biological observations

### 2. experimental_context

Extract information describing the experimental or imaging setup,
    including:

- staining method
- microscopy or imaging method
- magnification
- scale bar
- treatment or intervention
- dose or concentration
- time point
- experimental group
- sample size
- specimen or preparation information
- figure labels or annotations

### 3. non_biomedical_content

Use this field for content that is not meaningfully biomedical.
```

```
### Rules

- Only extract information explicitly stated in the caption.
- Do not infer or add biological conclusions.
- Do not infer disease or pathology unless explicitly stated.
- Do not infer morphology from the imaging method alone.
- Do not treat microscopy type, staining method, magnification, scale bar
    , treatment, or dose alone as core biomedical semantics.
- Preserve comparisons, uncertainty, and quantitative values.
- Keep each item concise.
- For each item, provide the shortest exact text span from the original
    caption that supports it.
- If no substantive biomedical semantics are present, return an empty
    core_biomedical_semantics list.
- Return valid JSON only.

### Input

CAPTION:

{{REPORT}}
```

SEMANTIC-NEGATIVE REWRITE PROMPT

```
You are constructing negative text pairs for medical vision-language
    activation steering.

You are given:
1. An original medical report or biomedical figure caption.
2. A list of semantic_targets extracted from the original text.

Your task is to rewrite the original text after removing ALL information
    expressed by the provided semantic_targets.

The negative text should preserve only non-target, non-finding, non-
    diagnostic, non-biomedical, or purely connective/background wording
    when such wording can remain grammatical and meaningful.

If removing all semantic_targets leaves no meaningful non-target content,
     set:
negative_report = "No specific finding is provided."

### Removal Rules

- Remove every semantic target in semantic_targets.
- Remove exact mentions, paraphrases, abbreviations, negations,
    measurements, laterality, uncertainty, and sentence-level
    implications of each target.
- Do not keep a synonym or restatement of any removed target.
- If a sentence mostly exists to express one or more semantic_targets,
    remove the whole sentence or rewrite it so that the target
    information is gone.
- If deleting spans would make a sentence ungrammatical, rewrite the
    remaining non-target text into fluent English.
- Preserve only content that is not in semantic_targets and does not
    reveal the same medical or biomedical meaning.

### Preservation Rules

- Do not invent new information.
- Do not add new normal findings, negative findings, diagnoses,
    procedures, measurements, tests, imaging methods, staining methods,
    or biological entities.
```

```
- Preserve generic non-semantic text only if it does not reveal any
    removed target.
- Preserve report style when possible, but prioritize removing target
    semantics over preserving style.
- The negative_report should be concise.
- If the result is empty, vague, or only broken fragments, use exactly:
No specific finding is provided.

Return valid JSON only.

### Input

sample_id:
{{SAMPLE_ID}}

original_report:
{{ORIGINAL_REPORT}}

semantic_targets:
{{TARGETS_JSON}}
```

D Additional Layer-Selection Analysis

To further evaluate the offline layer-selection strategy, we analyze Qwen3-VL-8B (Bai et al., 2025a) under MedVESS while varying only the intervention layer. All other settings, including the construction vectors, decoding procedure, and steering strength, are kept fixed. We apply the offline selection criterion separately for each dataset–model configuration, with the selected intervention layers summarized in Table 6.

Complete Layer Sweep on MIMIC-CXR. We perform an exhaustive sweep over all 36 decoder layers on MIMIC-CXR (Johnson et al., 2019). As shown in Fig. 4, performance remains relatively stable in early layers and rises sharply around $L20$–$L23$. The offline-selected layer $L21$ lies immediately adjacent to the empirical optimum at $L22$ and achieves nearly identical downstream performance, ranking second overall in the exhaustive layer sweep. This indicates that the proposed criterion effectively localizes the narrow high-performing intervention region and selects a near-optimal layer without requiring downstream layer-wise search.

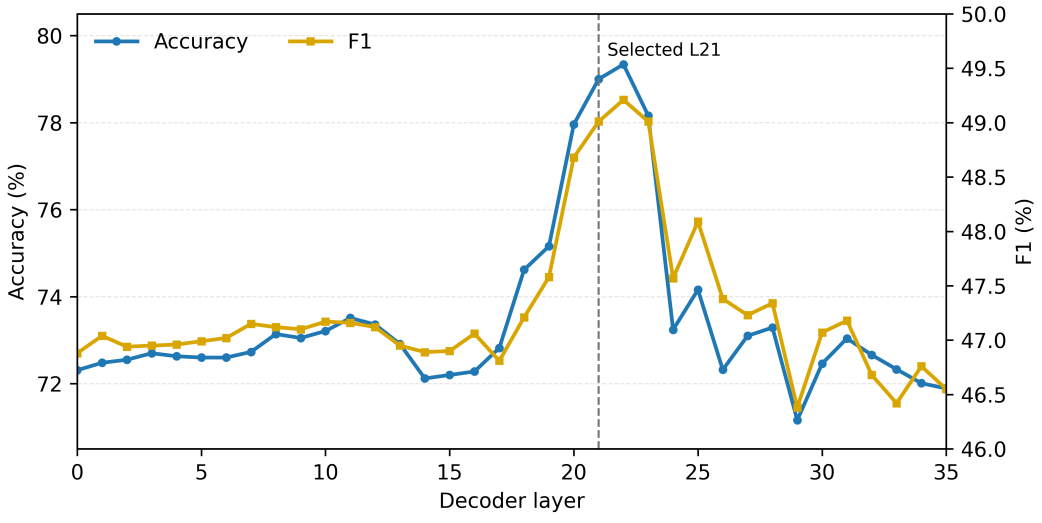

Figure 4: Complete layer-wise performance of Qwen3-VL-8B with MedVESS on MIMIC-CXR using $\alpha = 1.0$. Accuracy and F1 are shown across all decoder layers, with the offline-selected layer $L21$ marked by the dashed line.

Table 5: Candidate-layer performance of MedVESS on Harvard-FairVLMed and PMC-OA.

	Harvard-FairVLMed		PMC-OA	
Intervention Layer	Acc	F1	Acc	F1
L13	37.41	36.59	62.84	62.53
L17	20.68	20.37	61.94	60.94
L21 (Selected)	**67.53**	**53.78**	**68.05**	**67.66**
L25	47.28	43.82	65.15	65.12
L29	23.65	23.58	63.39	63.26
All Layers	0.00	0.00	0.00	0.00

Table 6: Intervention layers used in the main experiments.

Dataset	Qwen2.5-VL-7B	Qwen3-VL-8B
MIMIC-CXR	L18	L21
IU-Xray	L18	L21
Harvard-FairVLMed	L18	L21
PMC-OA	L27	L21

Candidate-Layer Analysis on Harvard-FairVLMed and PMC-OA. We further evaluate the offline layer-selection strategy on Harvard-FairVLMed (Luo et al., 2024) and PMC-OA (Lin et al., 2023) using representative candidate layers. As shown in Table 5, the offline-selected $L21$ achieves the best performance among the evaluated candidates on both datasets, further demonstrating the effectiveness of our selection criterion across different medical imaging datasets. In contrast, intervening at all decoder layers leads to a complete performance collapse on both datasets, consistent with the failure observed on MIMIC-CXR, where all-layer steering severely degrades both metrics and produces degenerate outputs. These results further support the importance of selecting an appropriate intermediate intervention layer.

E Steering-Strength Sensitivity

We further analyze the sensitivity of MedVESS to the steering strength α across different medical modalities and model backbones, while keeping the intervention layer and all other settings fixed. Specifically, we perform a sweep over $\alpha \in \{0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0, 2.5\}$ for Qwen2.5-VL-7B (Bai et al., 2025b) and Qwen3-VL-8B (Bai et al., 2025a) on all four benchmarks.

As shown in Fig. 5, MedVESS exhibits different sensitivity patterns across datasets and model backbones. MIMIC-CXR (Johnson et al., 2019) remains relatively stable over a moderate range of α, whereas IU-Xray (Demner-Fushman et al., 2016) is more sensitive to stronger intervention. Harvard-FairVLMed (Luo et al., 2024) generally benefits from larger steering strengths, while PMC-OA (Lin et al., 2023) degrades under excessive intervention, particularly for Qwen3-VL-8B. These observations indicate that the effect of steering magnitude is configuration-dependent rather than uniform across datasets and model backbones. For reference, the steering strengths used in the main experiments are marked in the corresponding panels of Fig. 5.

F Further Analysis on Harvard-FairVLMed

Harvard-FairVLMed (Luo et al., 2024) contains 3,687 Yes cases (86.43%) and 579 No cases (13.57%), resulting in a highly imbalanced binary label distribution. As shown in Table 7, the vanilla Qwen models also exhibit highly skewed prediction distributions, despite Yes being the dominant ground-truth label. We therefore complement raw accuracy with Macro-F1 and class-wise F1 to provide a more balanced view of the performance changes.

For Qwen3-VL-8B, MedVESS improves Macro-F1 from 32.50% to 53.78%, alongside the increase in accuracy from 32.79% to 67.53%. The class-wise results show a substantial increase in $F1_{\text{Yes}}$,

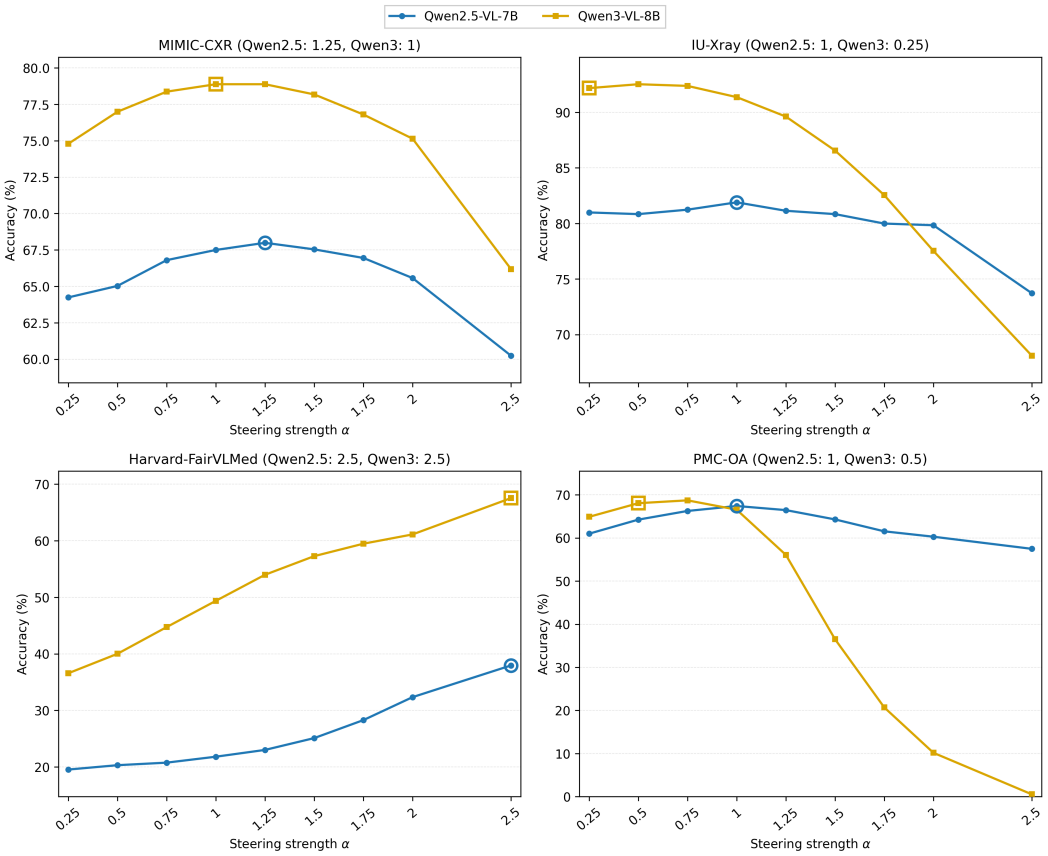

Figure 5: Sensitivity of MedVESS to the steering strength α across datasets and model backbones. The intervention layer and all other settings are kept fixed. Markers denote the steering strengths used in the main experiments.

Table 7: Class-wise analysis on Harvard-FairVLMed. The dataset contains 86.43% Yes and 13.57% No labels. Pred. Yes denotes the proportion of Yes predictions.

Setting	Acc.	Macro-F1	$F1_{\text{Yes}}$	$F1_{\text{No}}$	Pred. Yes (%)
Always Yes	86.43	46.36	92.72	0.00	100.00
Qwen2.5-VL-7B	18.82	18.26	11.50	25.03	5.30
+ MedVESS	37.95	36.23	46.71	25.75	30.00
Qwen3-VL-8B	32.79	32.50	36.95	28.06	20.16
+ MedVESS	67.53	53.78	78.99	28.57	68.12

while $F1_{\text{No}}$ remains similar. Importantly, the resulting Macro-F1 exceeds the 46.36% obtained by an always-Yes baseline. Thus, although the skewed label distribution contributes to the large change in raw accuracy, the improvement cannot be explained by trivial majority-class prediction alone.

G Qualitative Examples across Modalities

To complement the quantitative results, we provide representative qualitative examples from MIMIC-CXR (Johnson et al., 2019), Harvard-FairVLMed (Luo et al., 2024), and PMC-OA (Lin et al., 2023), spanning radiology, fundus photography, and pathology, respectively, in Fig. 6. For each case, we show the medical image, question and answer options, predictions from the comparison models, the MedVESS prediction, and the ground-truth answer, with incorrect predictions highlighted in red and correct predictions highlighted in green. In these selected cases, the com-

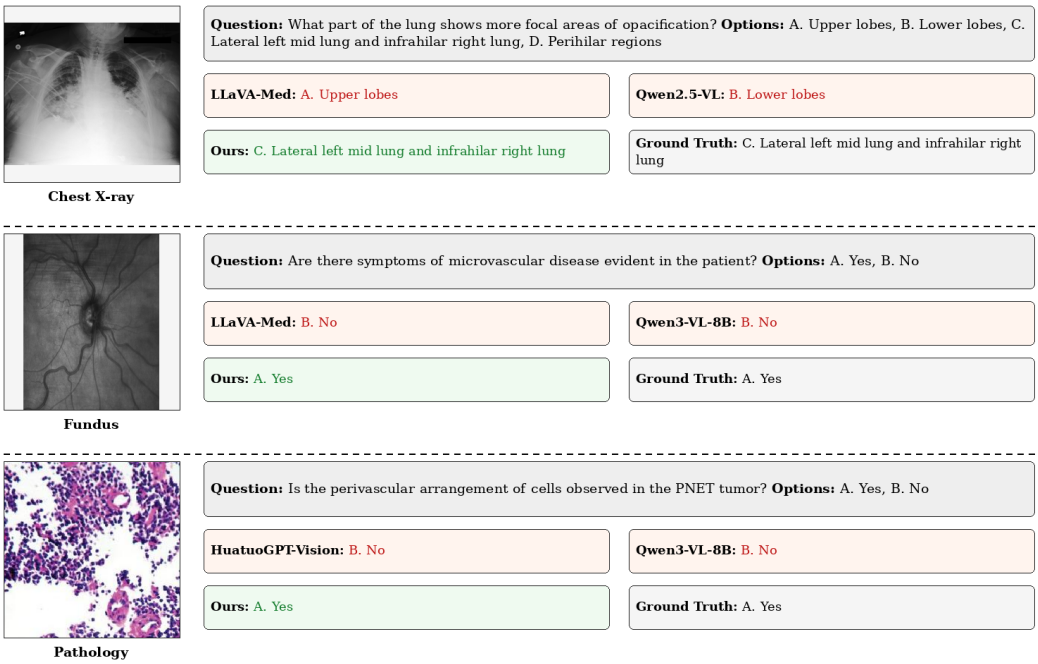

Figure 6: Representative qualitative examples of MedVESS across radiology, fundus photography, and pathology.

parison models produce incorrect predictions, whereas MedVESS recovers the ground-truth answer. The examples involve different forms of modality-specific visual evidence, including the spatial localization of focal opacification in chest X-rays, microvascular abnormalities in fundus imaging, and the perivascular arrangement of tumor cells in pathology images. These cases qualitatively illustrate that MedVESS improves the model's use of clinically relevant visual evidence across heterogeneous medical modalities.