Analyzing and Mitigating Cross-Lingual Degradation in Multilingual Medical VQA

Jingbo Wang, Sendong Zhao*, Haochun Wang, Bing Qin, Ting Liu
Research Center for Social Computing and Interactive Robotics
Harbin Institute of Technology, China
{jingbowang,sdzhao}@ir.hit.edu.cn

Abstract

Medical visual question answering (VQA) is a crucial task in clinical AI, yet its evaluation has so far centered almost exclusively on English, limiting its relevance to linguistically diverse patients and clinicians. Recent multilingual medical VQA benchmarks show that large vision-language models (LVLMs) degrade in non-English languages, but lack a fine-grained analysis of how cross-lingual variation affects the distinct capabilities that medical VQA requires. To this end, we construct a multilingual medical VQA benchmark over eight languages, organized into four representative scenarios that isolate the core capabilities medical VQA requires. Evaluating five open- and closed-source LVLMs, we find that cross-lingual degradation is not uniform but highly scenario-dependent. We therefore propose MedVL-XLRepE, a training-free scenario-aware representation engineering method, leveraging LVLMs' superior English medical VQA capability to steer non-English representations toward their English counterparts at inference time. Across three LVLMs and eight languages, MedVL-XLRepE consistently mitigates cross-lingual degradation, with gains of up to 6.33%.

1 Introduction

Medical visual question answering (VQA), which aims to interpret medical images and answer clinical queries, is a crucial research area in clinical AI (Lau et al., 2018; He et al., 2020; Li et al., 2023a). Driven by the rapid progress of large vision-language models (LVLMs) (Chen et al., 2024; DeepSeek-AI, 2026), recent systems have achieved substantial progress on medical VQA (Hu et al., 2024; Dong et al., 2025). Despite these remarkable advances, the medical evaluation of LVLMs has so far centered almost exclusively on English, which limits their applicability to a linguistically diverse population of patients and clinicians.

To evaluate whether LVLMs can serve multilingual populations, recent studies have constructed multilingual medical VQA benchmarks. These benchmarks draw medical images and questions from the national medical examinations and real-world clinical consultations of several countries, and evaluate the cross-lingual performance of LVLMs (Matos et al., 2024; Riccio et al., 2025; Yim et al., 2024b). Although these benchmarks show that the medical VQA performance of LVLMs degrades in non-English languages, their analysis of this degradation has two key limitations. First, they measure only the *overall success rate* on medical VQA, lacking fine-grained analysis of how cross-lingual variation affects the distinct capabilities that medical VQA requires. Second, their *language coverage is limited*, lacking analysis of how this degradation generalizes across a broader range of languages.

To address these limitations, we construct a multilingual medical VQA benchmark covering eight languages. Guided by the core capabilities that medical VQA requires (Lin et al., 2023; Pahud de Mortanges et al., 2024; Acosta et al., 2022), we organize the benchmark into four representative scenarios: Perceptual Recognition, Attribute-Aware Recognition, Sequential Images Understanding, and Vision-Text Integrated Reasoning, each isolating one such capability. Evaluations across five open- and closed-source LVLMs reveal that cross-lingual degradation is not uniform but highly scenario-dependent, indicating that language affects different medical reasoning capabilities unevenly. Such scenario-dependent cross-lingual degradation cannot be addressed by a single uniform correction, but requires a scenario-specific method instead.

To leverage the superior medical VQA capabilities of LVLMs in English for improving

*Corresponding author.

non-English performance, we propose MedVL-XLRepE, a training-free, scenario-aware cross-lingual representation engineering method. For each scenario, MedVL-XLRepE leverages a language vector and scenario-specific medical vectors from the representation gap between parallel English and target-language inputs, steering non-English representations toward their English counterparts at inference time. Experiments across three LVLMs and eight languages show that MedVL-XLRepE achieves consistent improvements, with gains of up to 6.33%, demonstrating its effectiveness in mitigating cross-lingual degradation in medical VQA.

Our contributions are summarized as follows:

- We construct a multilingual medical VQA benchmark covering eight languages and organized into four scenarios that isolate the core capabilities medical VQA requires.
- We present a comprehensive analysis of multilingual medical VQA across open- and closed-source LVLMs, revealing that cross-lingual degradation is strongly scenario-dependent.
- We propose MedVL-XLRepE, a scenario-aware representation engineering method that effectively mitigates cross-lingual degradation in medical VQA.

2 Related Work

2.1 Multilingual Medical VQA Benchmarks

Medical VQA has become a core task for evaluating clinical AI (Lau et al., 2018; He et al., 2020; Li et al., 2023a; Hu et al., 2024). Existing benchmarks, however, remain predominantly English-centric. Recent studies show that models excelling at English degrade substantially in other languages (Zhang et al., 2023; Schmidt et al., 2025), underscoring the need to evaluate medical VQA beyond English. To this end, a few multilingual medical VQA benchmarks have recently emerged. WorldMedQA-V (Matos et al., 2024) and MMMED (Riccio et al., 2025) draw image-based questions from the national medical examinations of several countries, whereas DermaVQA (Yim et al., 2024b) and MEDIQA-M3G (Yim et al., 2024a) target real-world dermatology consultations. Nevertheless, these benchmarks remain too coarse to reveal how cross-lingual variation affects the clinical capabilities that medical VQA requires.

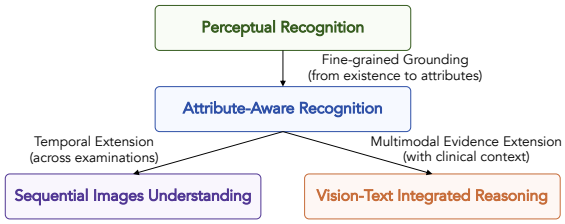


Figure 1: Medical VQA scenarios studied in this work.

2.2 Cross-lingual Representation Engineering

Prior works have demonstrated that semantically aligned inputs across languages occupy divergent regions in the latent space, a divergence associated with cross-lingual performance gaps (Chang et al., 2022; Zhao et al., 2024; Peng et al., 2025). Complementing this, studies in representation engineering have established that hidden-state interventions offer an effective means of steering model outputs at inference time (Zou et al., 2023; Turner et al., 2023; Li et al., 2023b). Motivated by these observations, recent work derives a steering vector that pulls non-English representations toward their English counterparts at inference, thereby narrowing the cross-lingual gap (Li et al., 2025; Chen et al., 2025). However, these methods rely on a single type of steering vector applied uniformly across all inputs, and thus cannot match cross-lingual degradation that differs across clinical capabilities.

3 Multilingual Medical VQA Benchmark

In this section, we study cross-lingual medical VQA in a controlled setting where visual evidence and clinical intent are comparable across languages. We first organize the task into representative clinical scenarios and construct a semantically aligned multilingual benchmark, then analyze performance variation across models, languages, and scenarios. The goal is to identify which forms of medical multimodal reasoning are most sensitive to language change, and to use these findings to motivate the method in the next section.

3.1 Medical VQA Problem Definition

Medical VQA requires the model to answer clinically meaningful questions from medical images, often by identifying relevant findings and their attributes (Lin et al., 2023), comparing current examinations with prior studies (Pahud de Mortanges et al., 2024), and interpreting imaging evidence together with accompanying clinical text (Acosta et al., 2022). In practice, these capability demands

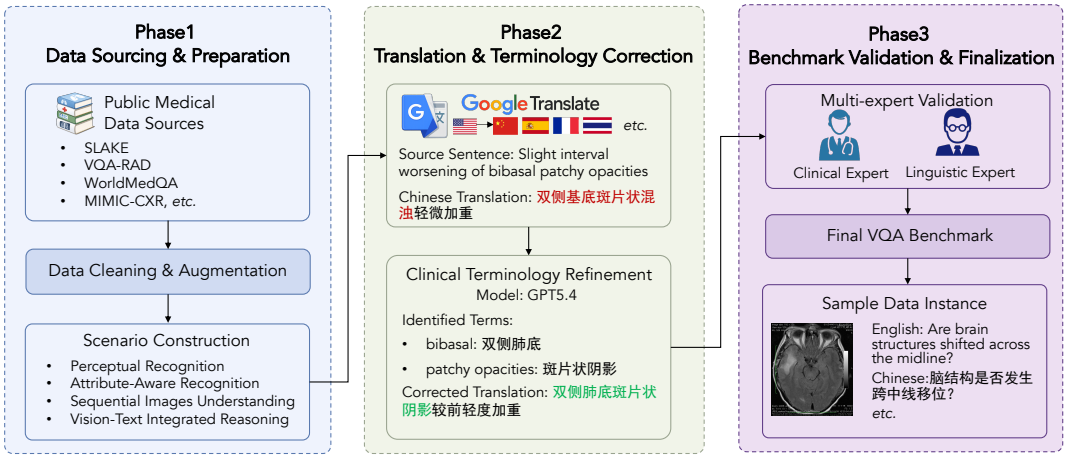


Figure 2: Construction pipeline of the multilingual medical VQA benchmark: data sourcing and scenario-wise reorganization, multilingual translation and terminology alignment, and benchmark validation and finalization.

recur in four clinical scenarios centered on finding recognition, attribute characterization, temporal comparison, and image-text interpretation. As illustrated in Figure 1, we therefore define four scenarios for medical VQA, referred to as Perceptual Recognition, Attribute-Aware Recognition, Sequential Images Understanding, and Vision-Text Integrated Reasoning.

Perceptual Recognition (PR) PR refers to single-image questions that ask whether a visible finding, structure, or abnormal pattern is present. It reflects the most basic perceptual requirement in medical VQA, since clinically meaningful answering often begins with identifying the relevant visual finding.

Attribute-Aware Recognition (AAR) AAR extends basic finding recognition by requiring the model to determine clinically relevant attributes of a finding. Such attributes include laterality, anatomical location, size, extent, severity, morphology, and spatial relation, all of which are essential to medically precise description and interpretation. The challenge is not only to identify the finding itself but also to relate the queried attribute to the visual finding.

Sequential Images Understanding (SIU) SIU concerns questions that require the model to compare temporally ordered studies and determine how a finding evolves over time. This scenario is needed when clinical interpretation depends not only on the current image, but also on change relative to prior examinations, where progression, improvement, or stability can be diagnostically decisive. The central challenge in SIU is to align corresponding medical content across time points and determine the relevant change between studies, rather than to interpret each image in isolation.

Vision-Text Integrated Reasoning (VTI) VTI extends medical VQA beyond image-only questions to cases where the answer depends on both the image and complementary clinical text, such as patient history or medical records. It reflects clinical questions for which visual content alone is insufficient and textual medical context is necessary for interpretation. This scenario places a greater demand on accurate understanding of medical text in addition to the image.

3.2 Benchmark Construction

To analyze cross-lingual effects, we construct a multilingual medical VQA benchmark in which visual evidence and clinical intent are controlled across languages, making language the primary variable under comparison. The construction goal is not merely to aggregate existing medical datasets, but to reorganize them into scenario-aligned probes that localize where language variation perturbs multimodal medical reasoning. As illustrated in Figure 2, we organize benchmark construction into three stages to preserve cross-lingual comparability while maintaining medical fidelity.

Step 1: Data Sourcing and Scenario-wise Reorganization We begin with public medical multimodal resources, including VQA-RAD (Lau et al., 2018), WorldMedQA-V (Matos et al., 2024), MMXU (Mu et al., 2025), and MIMIC-CXR (John-

Model	Scenario	EN	ZH	ES	FR	JA	TH	AR	BG
Open-source LVLMs									
Gemma3-12B-IT	PR	66.67	**67.49**	64.21	64.48	62.57	64.21	65.85	59.84
	AAR	**52.54**	50.85	44.07	39.83	46.61	47.46	46.61	38.98
	SIU	**43.66**	41.79	40.30	42.54	40.30	40.67	39.55	40.30
	VTI	**54.23**	46.48	48.24	47.18	44.01	45.42	40.14	46.83
Qwen3.5-9B	PR	**75.96**	72.95	**75.96**	74.86	71.31	75.41	73.50	72.68
	AAR	**73.73**	66.10	69.49	70.34	69.49	65.25	67.80	68.64
	SIU	**66.42**	61.19	61.57	57.09	61.19	61.57	58.21	60.45
	VTI	**64.79**	61.27	62.68	63.03	56.34	57.39	58.45	62.68
InternVL3.5-14B-Instruct	PR	**71.31**	65.57	66.67	65.85	58.74	62.57	56.01	66.39
	AAR	**66.10**	59.32	61.02	52.54	61.86	53.39	56.78	61.86
	SIU	**54.85**	49.63	53.36	53.73	47.01	51.49	45.90	50.37
	VTI	56.69	**57.75**	53.52	50.35	50.00	42.96	40.14	47.54
Closed-source LVLMs									
GPT-5.4-mini	PR	**78.69**	74.86	76.23	75.14	72.68	77.32	73.50	75.96
	AAR	75.42	74.58	72.03	69.49	74.58	72.03	**77.12**	73.73
	SIU	56.72	52.24	**57.46**	52.61	53.73	55.60	55.60	50.75
	VTI	**75.70**	75.35	74.65	73.59	72.89	73.24	74.65	72.89
Gemini-3-Flash-Preview	PR	**79.51**	77.05	78.69	77.32	78.42	**79.51**	77.32	77.60
	AAR	77.97	76.27	**81.36**	72.88	77.12	79.66	79.66	78.81
	SIU	**74.63**	70.15	73.88	72.39	69.40	70.90	71.27	71.64
	VTI	**89.08**	86.27	88.38	86.27	85.92	86.62	86.62	84.15

Table 1: Performance (%) of open- and closed-source LVLMs across four medical VQA scenarios and eight languages in our multilingual VQA benchmark. **Bold** and underlined values indicate the best and worst performance.

son et al., 2019), and retain only those samples for which the image evidence, clinical intent, and answer supervision remain comparable after multilingual translation. The retained samples are then reorganized by task mechanism rather than by original dataset identity, and the final benchmark is formulated using closed-form questions only, including binary yes/no and multiple-choice formats, to reduce answer-form variability across languages.

Step 2: Multilingual Translation and Terminology Alignment For the resulting English-source samples, we construct parallel multilingual questions spanning both mid-to-high-resource languages (Chinese, Spanish, French, and Japanese) and comparatively lower-resource languages (Thai, Arabic, and Bulgarian), while keeping the same underlying visual evidence and answer supervision across languages. We first use Google Translate to obtain draft translations at scale. However, as illustrated in Figure 2, these drafts do not always preserve medical terminology with sufficient precision. GPT-5.4 (OpenAI, 2026) is then used for terminology-focused refinement, with the goal of improving medical terminology accuracy while preserving the clinical intent of the original question.

Step 3: Benchmark Validation and Finalization After translation and terminology refinement, the multilingual items are reviewed by medical and linguistic experts before inclusion in the final benchmark. The review verifies that each item preserves the intended medical meaning across languages, remains grounded in the image evidence required for the question, and maintains a stable answer mapping after multilingual construction. The detailed scenario-wise data statistics are provided in the appendix A.1.

3.3 Experiment Setup

Baseline We select baselines from two model types: (1) **Open-source LVLMs**: Gemma3-12B-IT (Gemma Team, 2025), Qwen3.5-9B (Qwen Team, 2026), and InternVL3.5-14B-Instruct (Wang et al., 2025); (2) **Closed-source LVLMs**: GPT-5.4-mini (OpenAI, 2026) and Gemini-3-Flash-Preview (Google DeepMind, 2025).

Implementation Details For open-source models, evaluation is conducted on a server with 8 × NVIDIA H20 GPUs. For stable reproduction, we set the decoding temperature to 0 in all experiments.

3.4 Cross-lingual Evaluation Analysis

Overall Multilingual Gaps Are Clear, and Closed-Source Models Are More Stable As shown in Table 1, multilingual performance gaps

are clearly observable in the overall results. Across languages, the overall ranking is EN, ES, ZH, TH, BG, FR, JA, and AR. In particular, English attains the highest mean accuracy at 67.73%, whereas Arabic yields the lowest at 62.23%, corresponding to a gap of 5.50%. Thus, even when visual evidence and question intent are controlled across languages, language changes still lead to measurable performance differences in medical VQA. As shown in Table 5, closed-source models consistently outperform open-source models, achieving an average of 74.03% compared with 57.22% for open-source models. The same table further shows that closed-source models exhibit more stable cross-lingual performance: relative to their English average of 75.97%, the largest gap to a non-English language is 3.51%, whereas the corresponding English average for open-source models is 62.25%, with the largest gap reaching 8.17%.

Cross-Lingual Degradation Is Not Uniform Across Scenarios As shown in Table 6, cross-lingual degradation varies substantially across scenarios rather than appearing as a uniform reduction throughout the benchmark. PR shows a relatively small cross-lingual span of 5.68%. When the task shifts from recognizing whether a finding is present to determining medically relevant attributes of that finding, the span increases to 8.14% in AAR, indicating that language variation more strongly affects attribute grounding than basic finding recognition. When the task further requires integrating visual evidence with complementary clinical text, VTI still exhibits a large span of 8.10%, suggesting that cross-lingual differences remain pronounced in image-text evidence alignment. SIU, in contrast, yields a smaller span of 5.15%, indicating that introducing temporal comparison across multiple images does not by itself lead to the largest cross-lingual gap.

The Hardest Scenario Is Not the Most Language-Sensitive As shown in Table 6, multilingual fragility cannot be reduced to intrinsic task difficulty. SIU is the hardest scenario in the benchmark, with the lowest average accuracy at 55.80%, yet it is also the most stable across languages, with a span of only 5.15%. AAR and VTI show a different profile: both are easier than SIU in overall accuracy, at 65.08% and 63.61%, but they exhibit substantially larger cross-lingual spans of 8.14% and 8.10%, respectively. This direct contrast shows that the strongest multilingual instability does not arise in the intrinsically hardest scenario, but in scenarios whose required medical reasoning is more vulnerable to language change.

Language Resource Level Does Not Determine Scenario-Level Performance Language resource level does not map directly onto multilingual medical VQA performance. Although English is the strongest language overall at 67.73%, the distinction between mid-to-high-resource languages (Chinese, Spanish, French, and Japanese) and comparatively lower-resource languages (Thai, Arabic, and Bulgarian) is much less pronounced, with only a 1.01% difference in aggregate (63.83% vs. 62.82%). Moreover, relative performance can reverse across resource groups once scenario is taken into account: Japanese is stronger than Thai on AAR (65.93% vs. 63.56%), whereas Thai outperforms Japanese on PR (71.80% vs. 68.74%). The full scenario-wise language rankings are provided in Table 7.

4 MedVL-XLRepE

Section 3.4 shows that cross-lingual performance gaps in medical VQA are scenario-dependent rather than uniform. This suggests that language variation should be handled with selective alignment rather than a single global adjustment. Recent work on representation engineering has shown that lightweight inference-time interventions on internal activations can steer model behavior without updating model parameters (Turner et al., 2023; Zou et al., 2023); more recent multilingual studies further suggest that similar interventions can help narrow cross-lingual gaps in perception and reasoning (Li et al., 2025; Chen et al., 2025). Motivated by this, we propose MedVL-XLRepE, a scenario-aware Medical Vision-Language Cross-lingual Representation Engineering method for multilingual medical VQA.

4.1 Preliminaries

For each sample x, let I denote the medical image, and let T_{en} and T_{tgt} denote the English and target-language textual inputs. The corresponding multimodal inputs are defined as

$$X_{en} = (I, T_{en}), \qquad X_{tgt} = (I, T_{tgt}) \quad (1)$$

Due to the autoregressive nature of LVLMs, the final input token's hidden state provides a compact summary of the encoded multimodal context. We

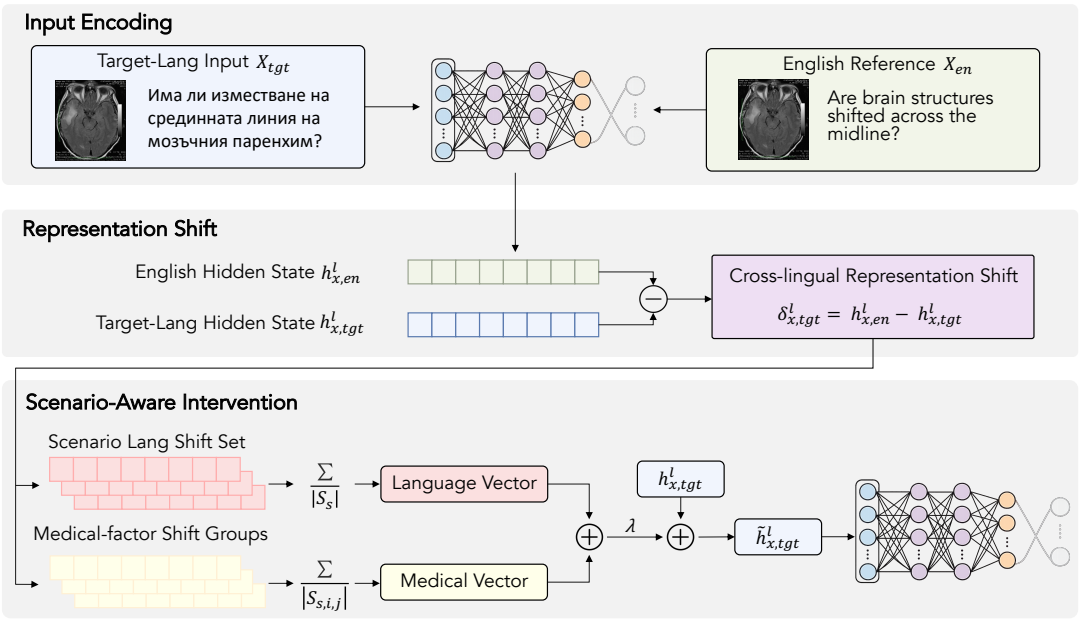


Figure 3: Overview of MedVL-XLRepE. The hidden-state difference between paired English and target-language inputs defines a cross-lingual shift $\delta^{(l)}_{x,tgt}$. Aggregating $\delta^{(l)}_{x,tgt}$ over the scenario set yields a language vector and a medical vector, whose sum is added to $h^{(l)}_{x,tgt}$ to obtain the intervened representation $\tilde{h}^{(l)}_{x,tgt}$.

therefore use this token as the anchor representation for cross-lingual comparison. Specifically, let $h^{(l)}_{x,en}$ and $h^{(l)}_{x,tgt}$ denote the hidden states of the final input token for sample x at layer l under X_{en} and X_{tgt}, respectively. We then define the cross-lingual representation shift as

$$\delta^{(l)}_{x,tgt} = h^{(l)}_{x,en} - h^{(l)}_{x,tgt} \tag{2}$$

This difference captures the sample-level English-target representation shift at layer l and serves as the basic quantity for the vector construction below.

4.2 Scenario-Aware Intervention Vector

The cross-lingual shift within each scenario stems from two distinct sources: a general language-level shift between English and the target language, and finer shifts shaped by the scenario's clinical factors.

Language-Level Vector For each target language and scenario s, we average the sample-level representation differences over all samples in that scenario:

$$v^{(l)}_{lang,s} = \frac{1}{|S_s|} \sum_{x \in S_s} \delta^{(l)}_{x,tgt} \tag{3}$$

where S_s denotes the set of samples belonging to scenario s. This vector captures the systematic English-target shift common to all samples in scenario s at layer l.

Medical Vectors To capture finer cross-lingual variation tied to medically meaningful structure, we partition the samples of each scenario along 2 medical factors that reflect the primary clinical axes along which the scenario's visual evidence and question are jointly organized. The first factor is *anatomy*, a shared factor across all scenarios that defines the spatial context of the visual evidence. The second factor is scenario-specific: *pathology* for PR and SIU, where the evidence is organized around a queried clinical finding (single-image recognition for PR, temporal comparison of findings for SIU); *radiological attribute* for AAR, where the question binds a finding to a queried attribute such as laterality, spatial extent, or severity; and *clinical context type* for VTI, where auxiliary clinical text provides a diagnostic prior over the same visual evidence.

For each scenario s and medical factor i, we group S_s according to factor i into disjoint subsets $\{S_{s,i,j}\}_j$. For each group we compute the group-mean cross-lingual shift

$$\bar{\delta}^{(l)}_{s,i,j} = \frac{1}{|S_{s,i,j}|} \sum_{x \in S_{s,i,j}} \delta^{(l)}_{x,tgt} \tag{4}$$

and group centroid of the target-language hidden states

$$\mu^{(l)}_{s,i,j} = \frac{1}{|S_{s,i,j}|} \sum_{x \in S_{s,i,j}} h^{(l)}_{x,tgt} \tag{5}$$

Model	Scenario	XLRepE	EN	ZH	ES	FR	JA	TH	AR	BG
Gemma3-12B-IT	PR	×	66.67	67.49	64.21	64.48	62.57	64.21	65.85	59.84
		✓	—	68.58 ↑1.09	65.03 ↑0.82	64.48 ↑0.00	63.11 ↑0.54	66.94 ↑2.73	68.03 ↑2.18	63.93 ↑4.09
	AAR	×	52.54	50.85	44.07	39.83	46.61	47.46	46.61	38.98
		✓	—	52.54 ↑1.69	44.92 ↑0.85	44.92 ↑5.09	46.61 ↑0.00	50.00 ↑2.54	47.46 ↑0.85	44.07 ↑5.09
	SIU	×	43.66	41.79	40.30	42.54	40.30	40.67	39.55	40.30
		✓	—	43.66 ↑1.87	43.28 ↑2.98	44.03 ↑1.49	42.54 ↑2.24	44.03 ↑3.36	42.54 ↑2.99	43.28 ↑2.98
	VTI	×	54.23	46.48	48.24	47.18	44.01	45.42	40.14	46.83
		✓	—	47.54 ↑1.06	51.76 ↑3.52	49.30 ↑2.12	47.54 ↑3.53	46.13 ↑0.71	41.55 ↑1.41	47.18 ↑0.35
Qwen3.5-9B	PR	×	75.96	72.95	75.96	74.86	71.31	75.41	73.50	72.68
		✓	—	74.32 ↑1.37	77.60 ↑1.64	76.23 ↑1.37	75.41 ↑4.10	77.32 ↑1.91	74.59 ↑1.09	77.60 ↑4.92
	AAR	×	73.73	66.10	69.49	70.34	69.49	65.25	67.80	68.64
		✓	—	69.49 ↑3.39	71.19 ↑1.70	76.27 ↑5.93	75.42 ↑5.93	71.19 ↑5.94	69.49 ↑1.69	72.03 ↑3.39
	SIU	×	66.42	61.19	61.57	57.09	61.19	61.57	58.21	60.45
		✓	—	63.43 ↑2.24	66.42 ↑4.85	57.09 ↑0.00	65.67 ↑4.48	66.79 ↑5.22	61.57 ↑3.36	64.55 ↑4.10
	VTI	×	64.79	61.27	62.68	63.03	56.34	57.39	58.45	62.68
		✓	—	63.38 ↑2.11	66.20 ↑3.52	65.49 ↑2.46	59.15 ↑2.81	58.80 ↑1.41	62.32 ↑3.87	69.01 ↑6.33
InternVL3.5-14B-Instruct	PR	×	71.31	65.57	66.67	65.85	58.74	62.57	56.01	66.39
		✓	—	66.39 ↑0.82	68.03 ↑1.36	66.94 ↑1.09	61.20 ↑2.46	63.93 ↑1.36	58.20 ↑2.19	67.49 ↑1.10
	AAR	×	66.10	59.32	61.02	52.54	61.86	53.39	56.78	61.86
		✓	—	61.86 ↑2.54	62.71 ↑1.69	55.93 ↑3.39	63.56 ↑1.70	56.78 ↑3.39	60.17 ↑3.39	62.71 ↑0.85
	SIU	×	54.85	49.63	53.36	53.73	47.01	51.49	45.90	50.37
		✓	—	50.75 ↑1.12	54.85 ↑1.49	54.85 ↑1.12	48.51 ↑1.50	52.61 ↑1.12	47.76 ↑1.86	52.61 ↑2.24
	VTI	×	56.69	57.75	53.52	50.35	50.00	42.96	40.14	47.54
		✓	—	60.21 ↑2.46	54.58 ↑1.06	52.11 ↑1.76	51.41 ↑1.41	42.96 ↑0.00	42.96 ↑2.82	51.06 ↑3.52

Table 2: Effectiveness of MedVL-XLRepE across models, scenarios, and target languages. Each cell reports accuracy (%) for the baseline (×) and MedVL-XLRepE (✓).

At inference, for a target-language input X_{tgt} with hidden state $h^{(l)}_{x,tgt}$, we select for each factor i the group whose centroid is most similar to $h^{(l)}_{x,tgt}$ by maximizing cosine similarity:

$$j_{sel} = \arg\max_j \frac{(h^{(l)}_{x,tgt})^\top \mu^{(l)}_{s,i,j}}{\|h^{(l)}_{x,tgt}\|_2 \, \|\mu^{(l)}_{s,i,j}\|_2} \tag{6}$$

The medical vector for factor i is then the group-mean shift of the selected group:

$$v^{(l)}_{med,s,i} = \bar{\delta}^{(l)}_{s,i,j_{sel}} \tag{7}$$

4.3 Cross-lingual Representation Engineering

For X_{tgt} in scenario s, the intervention vector $v^{(l)}_s$ jointly captures the general language-level shift and the scenario-specific medical correction, averaging the latter over the two factors to balance their contributions:

$$v^{(l)}_s = v^{(l)}_{lang,s} + \frac{1}{2}\sum_{i=1}^{2} v^{(l)}_{med,s,i} \tag{8}$$

Prior work has shown that activation-level vector additions along specific directions can effectively steer model behavior (Turner et al., 2023). To apply this directional correction without distorting the activation magnitude, the intervention adds $v^{(l)}_s$ with strength λ to $h^{(l)}_{x,tgt}$ and renormalizes to the original ℓ_2 norm:

$$\tilde{h}^{(l)}_{x,tgt} = \left\|h^{(l)}_{x,tgt}\right\|_2 \cdot \frac{h^{(l)}_{x,tgt} + \lambda\, v^{(l)}_s}{\left\|h^{(l)}_{x,tgt} + \lambda\, v^{(l)}_s\right\|_2} \tag{9}$$

The corrected hidden state $\tilde{h}^{(l)}_{x,tgt}$ replaces $h^{(l)}_{x,tgt}$ in the forward pass, closing the cross-lingual gap along the scenario-conditioned medical axes while leaving all other activations unchanged.

5 Experiments

5.1 Experimental Setup

We evaluate MedVL-XLRepE on the three open-source LVLMs used in Section 3.4: Gemma3-12B-IT, Qwen3.5-9B, and InternVL3.5-14B-Instruct. Within each scenario, the benchmark is split into disjoint halves for calibration and evaluation, with detailed statistics provided in Appendix A.1. We use intervention strength $\lambda = 0.1$, and adopt the two scenario-specific medical factors described in Section 4.2.

5.2 Main Results

All Clinical Scenarios Benefit from MedVL-XLRepE Table 2 shows that MedVL-XLRepE yields a positive mean gain in all four clinical scenarios. AAR, which exhibits the largest baseline cross-lingual span, receives the largest mean gain

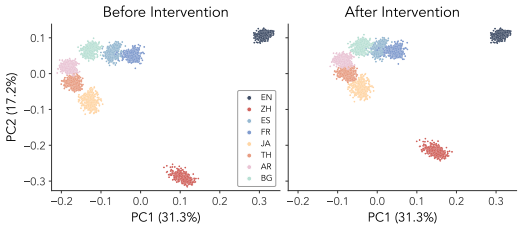


Figure 4: PCA of multilingual hidden states before and after applying MedVL-XLRepE.

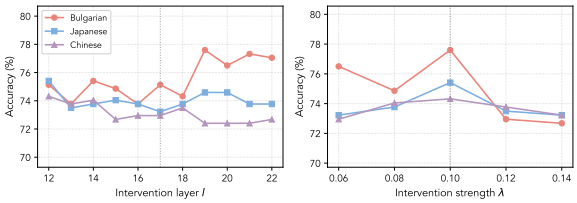


Figure 5: Performance under varying intervention layer l (left) and strength λ (right) on Qwen3.5-9B/PR.

(2.91%). Even SIU, the most language-stable scenario, still attains a 2.51% improvement.

Consistent Gains across All Three Backbones Averaged across all scenarios and target languages, MedVL-XLRepE raises cross-lingual accuracy by 2.38%. The improvement holds across all three backbones, with per-model mean gains of 2.08% on Gemma3-12B-IT, 3.25% on Qwen3.5-9B, and 1.81% on InternVL3.5-14B-Instruct.

Lower-Resource Target Languages Benefit Most The lower-resource group (Bulgarian, Thai, Arabic) obtains a mean gain of 2.68%, exceeding the 2.16% of the mid-to-high-resource group (Chinese, Spanish, French, Japanese). Bulgarian on Qwen3.5 (VTI) records the largest improvement of 6.33%, and the resulting accuracy of 69.01% even surpasses the corresponding English baseline of 64.79%.

5.3 Cross-lingual Alignment Visualization

To visualize how MedVL-XLRepE reshapes the model's internal representations, we apply PCA to the final input token's hidden states at the intervention layer of Qwen3.5-9B in the PR scenario. As shown in Figure 4, before applying MedVL-XLRepE the eight languages form clearly separated clusters, with English isolated on one side and the seven target-language clusters lying on the opposite side. After applying MedVL-XLRepE, every non-English cluster shifts toward the English anchor along PC1, while the English cluster itself remains unchanged by design. Averaged over the seven target languages, the centroid distance to English in PC space drops by 19.3%.

5.4 Hyperparameter Sensitivity Analysis

We analyse the sensitivity of MedVL-XLRepE to its hyper-parameters—the intervention layer l and the strength λ—on Qwen3.5-9B in the PR scenario, with Bulgarian, Japanese, and Chinese as relatively low-, mid-, and high-resource target languages.

Intervention Layer Analysis Multilingual transformers concentrate language-sensitive structure and English-pivoted semantics in the middle layers (Chang et al., 2022; Wendler et al., 2024). We therefore restrict the intervention layer l to this middle-layer range. For Qwen3.5-9B, the range corresponds to layers 12–22, and Figure 5 (left) reports the sensitivity of accuracy as l is varied over these candidate layers.

Intervention Strength Analysis Activation-shift methods exhibit an inverted-U over the strength axis: too small a value fails to displace the target component, while too large a value drives the hidden state off-manifold (Li et al., 2023b). We therefore evaluate λ from 0.06 to 0.14 in steps of 0.02. Figure 5 (right) shows that accuracy peaks at λ=0.10 simultaneously for all three target languages, so we adopt λ=0.10 as the default value of this hyper-parameter.

6 Conclusion

In this work, we study cross-lingual degradation in multilingual medical VQA. To enable a fine-grained analysis, we construct a benchmark covering eight languages and organized into four representative scenarios. Evaluating five open- and closed-source LVLMs on this benchmark, we find that cross-lingual degradation is not uniform but highly scenario-dependent, and therefore cannot be addressed by a single global correction. Motivated by this finding, we propose MedVL-XLRepE, a training-free, scenario-aware cross-lingual representation engineering method that steers non-English representations toward their English counterparts at inference time. Across three LVLMs and eight languages, MedVL-XLRepE consistently mitigates cross-lingual degradation, with gains of up to 6.33%. We hope our benchmark and method will support future research toward equitable, multilingually reliable clinical LVLMs.

Limitations

While MedVL-XLRepE consistently mitigates cross-lingual degradation, our study has several limitations. Although our benchmark spans eight languages and four representative clinical scenarios, it still covers only a small portion of the world's languages and does not capture the full diversity of clinical tasks. In addition, MedVL-XLRepE requires white-box access to the model's internal activations, so it applies only to open-source LVLMs and cannot be used to improve closed-source LVLMs. Finally, the method depends on paired English and target-language inputs and a held-out calibration split to construct the intervention vectors, and its effectiveness when only the target language is available, or under a severe scarcity of calibration samples, is not yet characterized.

Ethical Considerations

Our benchmark and MedVL-XLRepE are research artifacts for studying and mitigating cross-lingual degradation of LVLMs on medical VQA, and any patient-facing use would require additional clinical validation and human oversight. The benchmark is reorganized from publicly released medical VQA resources under their original licenses and access terms. Multilingual items are produced by machine translation followed by review from human experts to preserve clinical intent across languages. Equitable performance across languages is itself an ethical concern for clinical AI, and our benchmark and method are aimed at narrowing this gap so that medical LVLMs can serve patients and clinicians beyond English-speaking populations.

References

Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. 2022. Multimodal biomedical ai. *Nature medicine*, 28(9):1773–1784.

Tyler Chang, Zhuowen Tu, and Benjamin Bergen. 2022. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and 1 others. 2024. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 7346–7370.

Ruihan Chen, Qiming Li, Xiaocheng Feng, Xiaoliang Yang, Weihong Zhong, Yuxuan Gu, Zekun Zhou, and Bing Qin. 2025. Mpr-gui: Benchmarking and enhancing multilingual perception and reasoning in gui agents. *arXiv preprint arXiv:2512.00756*.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

Wenjie Dong, Shuhao Shen, Yuqiang Han, Tao Tan, Jian Wu, and Hongxia Xu. 2025. Generative models in medical visual question answering: A survey. *Applied Sciences*, 15(6):2983.

Google DeepMind Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Google DeepMind. 2025. Gemini 3 flash: Frontier intelligence built for speed.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

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

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

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

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Qiming Li, Xiaocheng Feng, Yixuan Ma, Zekai Ye, Ruihan Chen, Xiachong Feng, and Bing Qin. 2025. Unlocking multilingual reasoning capability of llms and lvlms through representation engineering. *arXiv preprint arXiv:2511.23231*.

Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. 2023. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143:102611.

João Matos, Shan Chen, Siena Placino, Yingya Li, Juan Carlos Climent Pardo, Daphna Idan, Takeshi Tohyama, David Restrepo, Luis F. Nakayama, Jose M. M. Pascual-Leone, Guergana Savova, Hugo Aerts, Leo A. Celi, A. Ian Wong, Danielle S. Bitterman, and Jack Gallifant. 2024. Worldmedqa-v: a multilingual, multimodal medical examination dataset for multimodal language models evaluation. *Preprint*, arXiv:2410.12722.

Linjie Mu, Zhongzhen Huang, Shengqian Qin, Yakun Zhu, Shaoting Zhang, and Xiaofan Zhang. 2025. Mmxu: A multi-modal and multi-x-ray understanding dataset for disease progression. *arXiv preprint arXiv:2502.11651*.

OpenAI. 2026. Introducing GPT-5.4.

Aurélie Pahud de Mortanges, Haozhe Luo, Shelley Zixin Shu, Amith Kamath, Yannick Suter, Mohamed Shelan, Alexander Pöllinger, and Mauricio Reyes. 2024. Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging. *NPJ digital medicine*, 7(1):195.

Qiwei Peng, Guimin Hu, Yekun Chai, and Anders Søgaard. 2025. Debiasing multilingual llms in cross-lingual latent space. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22593–22604.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Giuseppe Riccio, Antonio Romano, Mariano Barone, Gian Marco Orlando, Diego Russo, Marco Postiglione, Valerio La Gatta, and Vincenzo Moscato. 2025. A multilingual multimodal medical examination dataset for visual question answering in healthcare. In *2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 435–440. IEEE.

Fabian David Schmidt, Florian Schneider, Chris Biemann, and Goran Glavaš. 2025. Mvl-sib: a massively multilingual vision-language benchmark for cross-modal topical matching. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16285–16312.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394.

Wen-wai Yim, Asma Ben Abacha, Yujuan Fu, Zhaoyi Sun, Fei Xia, Meliha Yetisgen-Yildiz, and Martin Krallinger. 2024a. Overview of the mediqa-m3g 2024 shared task on multilingual multimodal medical answer generation. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 581–589.

Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Asma Ben Abacha, Meliha Yetisgen, and Fei Xia. 2024b. Dermavqa: A multilingual visual question answering dataset for dermatology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 209–219. Springer.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36:5484–5505.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *Advances in Neural Information Processing Systems*, 37:15296–15319.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Benchmark Construction Details

A.1 Scenario-wise Data Statistics

Table 3 reports the benchmark scale used in our main experiments. After the data sourcing, scenario-wise reorganization, multilingual translation, terminology refinement, and expert validation steps described in Section 3.2, the final benchmark contains 2,071 original English-source items and 16,568 multilingual item-language pairs under the 8-language setting. At the scenario level, PR contains 731 original items (5,848 multilingual instances), AAR contains 236 (1,888 multilingual instances), SIU contains 536 (4,288 multilingual instances), and VTI contains 568 (4,544 multilingual instances).

Within each scenario, we split the items evenly into two disjoint halves: one half is reserved as the calibration set used by MedVL-XLRepE to estimate the language-level and medical vectors in Section 4.2, and the other half is used as the test set on which all reported results are obtained. The split is stratified by language and by medical factor so that both halves preserve the original distribution, and no item from the calibration half ever appears in evaluation.

Scenario	Original	Multilingual
Perceptual Recognition	731	5,848
Attribute-Aware Recognition	236	1,888
Sequential Images Understanding	536	4,288
Vision-Text Integrated Reasoning	568	4,544
Total	2,071	16,568

Table 3: Scenario-wise benchmark statistics under the 8-language setting.

A.2 Terminology Refinement Prompt

To refine the initial Google Translate outputs described in Step 2 of Section 3.2, we use GPT-5.4 with a terminology-focused prompt. The prompt is designed to preserve the original medical intent while correcting non-standard or inaccurate medical wording in the draft translation. The full prompt template is shown in *Terminology Refinement Prompt*.

A.3 Empirical Motivation for Terminology Refinement

To support the terminology-refinement step in Section 3.2, we quantify how often GPT-5.4 revises the initial Google Translate output in the PR and AAR scenarios. Table 4 reports these revision rates. The results show substantial modification rates across all seven target languages used in our main experiments, indicating that raw machine translation frequently leaves terminology that is further corrected during refinement. This pattern provides empirical support for including a dedicated terminology-alignment stage in the benchmark construction pipeline.

A.4 Expert Validation Protocol

The expert review described in Step 3 of Section 3.2 is conducted by researchers from our research group with medical and linguistics backgrounds, participating as part of the internal research collaboration. Each multilingual item is independently checked under medical and linguistic criteria, and items that fail either review are either corrected with a minimal edit or discarded from the final benchmark. The brief instructions are shown in *Medical Reviewer Instruction* and *Linguistic Reviewer Instruction*.

B Additional Analysis Results

B.1 Aggregated Performance Statistics

To support the analysis of cross-lingual variation at a level more interpretable than the raw model-by-language results, we further aggregate performance by source group and by scenario. Table 5 shows that closed-source models are not only stronger overall than open-source models, but also more stable across languages. Table 6 shows that the cross-lingual gap is not uniform across scenarios: AAR and VTI display the largest spans, whereas PR and SIU are more stable. These aggregated results provide the direct numerical basis for the main-text comparisons between source-group robustness and scenario sensitivity.

B.2 Language Rankings by Scenario

To examine whether language resource level is sufficient to predict multilingual medical VQA performance, we also report explicit language rankings for the overall benchmark and for each scenario separately. Table 7 shows that, although English remains strongest overall, the relative ordering among the remaining languages is not fixed once the scenario is specified. This table therefore complements the aggregate averages by making visible where scenario-specific rankings no longer align with a simple resource-based expectation.

C MedVL-XLRepE Implementation Details

C.1 Medical Factor Extraction

MedVL-XLRepE partitions calibration samples by scenario-specific medical factors (Section 4.2). Factor labels are obtained by combining benchmark metadata fields, rule-based extraction from question text, and LLM-based classification where metadata is absent. These labels are required only on the calibration split to form the groups $\{S_{s,i,j}\}_j$; at test time the group is selected purely by cosine similarity between the input hidden state and the precomputed group centroids, so no ground-truth medical labels are needed on the evaluation samples.

Scenario	Chinese	Spanish	French	Japanese	Thai	Arabic	Bulgarian
PR	53.76	35.84	55.54	72.64	81.40	69.49	50.48
AAR	65.25	37.29	57.63	83.90	82.20	64.41	58.47

Table 4: Percentage (%) of items whose Google Translate draft was modified by GPT-5.4 in the VQA-RAD PR and AAR subsets.

Group	EN	ZH	ES	FR	JA	TH	AR	BG	Avg	Max–Min
Open-source	62.25	58.37	58.42	56.82	55.79	55.65	54.08	56.38	57.22	8.17
Closed-source	75.97	73.35	75.33	72.46	73.09	74.36	74.47	73.19	74.03	3.51

Table 5: Aggregated language performance (%) by source group.

Anatomy is shared across all scenarios. Labels are derived from benchmark metadata where available (e.g., *chest*, *head*, *abdomen*) and supplemented by rule-based extraction from question text for scenarios where explicit organ metadata is absent.

Pathology (PR, SIU) groups samples by the specific clinical finding being queried. For PR, samples are partitioned by the finding type targeted in the question — e.g., pneumothorax, cardiomegaly, or atelectasis — since different findings induce different visual targets and may be expressed with language-specific clinical terminology. For SIU, the same factor is applied within a temporal comparison context, where the finding type reflects what is being tracked across serial studies.

Radiological attribute (AAR) groups samples by the type of visual property the question asks the model to assess. Representative sub-types include laterality (e.g., left vs. right positioning), spatial extent (e.g., mass size or inspiratory effort), morphological attributes (e.g., contour symmetry), and density or signal intensity (e.g., hyperattenuation).

Clinical context type (VTI) groups samples by the type of auxiliary clinical text accompanying the image. Representative text types include structured patient case descriptions with clinical history, ECG interpretation contexts, imaging-based diagnostic scenarios, and procedural case descriptions, each providing a different form of diagnostic prior over the visual evidence.

C.2 Component Ablation: Language Vector vs. Medical Vector

MedVL-XLRepE applies two corrections: a language-level vector that steers target-language representations toward English, and a medical vector that re-centres them within the scenario-specific factor subspace. Table 8 isolates each component on Qwen3.5-9B in the PR and AAR scenarios. Applied alone, each component yields limited average gains and occasionally hurts individual languages. For instance, the medical vector alone decreases Japanese by 1.69% in AAR, and the language vector alone decreases Arabic by 0.28% in PR. MedVL-XLRepE eliminates these isolated regressions and achieves +2.34% on PR and +3.99% on AAR. The AAR gain notably exceeds the sum of the two isolated contributions (+1.09%), indicating that the two corrections reinforce each other when applied jointly.

D Statistical Significance of MedVL-XLRepE Gains

This appendix examines whether the cross-lingual improvements reported in Table 2 are directionally consistent across the evaluated model–scenario–language configurations.

D.1 Test Protocol

All evaluations use greedy decoding (temperature 0), and MedVL-XLRepE is a fixed inference-time intervention with no stochastic component; each (model, scenario, target-language) configuration therefore yields a single deterministic accuracy that repeated runs reproduce exactly. The relevant question is consequently not run-to-run variance, but whether the improvement is *systematic* across the population of evaluated configurations.

We treat each of the 84 configurations in Table 2 (3 backbones $\times$ 4 scenarios $\times$ 7 target languages; English is the anchor and is not intervened) as a paired observation, pairing the baseline accuracy with the MedVL-XLRepE accuracy, and define the per-configuration gain as their difference. On these 84 paired gains we apply three complementary tests: a paired t-test (parametric), a Wilcoxon signed-rank test (non-parametric, no normality assumption), and a sign test (distribution-free). To

Setting	EN	ZH	ES	FR	JA	TH	AR	BG	Avg	Max–Min
Overall	67.73	64.36	65.19	63.08	62.71	63.13	62.23	63.10	63.94	5.50
PR	74.43	71.58	72.35	71.53	68.74	71.80	69.24	70.49	71.27	5.68
AAR	69.15	65.42	65.59	61.02	65.93	63.56	65.59	64.40	65.08	8.14
SIU	59.26	55.00	57.31	55.67	54.33	56.05	54.11	54.70	55.80	5.15
VTI	68.10	65.42	65.49	64.08	61.83	61.13	60.00	62.82	63.61	8.10

Table 6: Scenario-wise aggregated language performance (%).

Scenario	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8
Overall	EN (67.73)	ES (65.19)	ZH (64.36)	TH (63.13)	BG (63.10)	FR (63.08)	JA (62.71)	AR (62.23)
PR	EN (74.43)	ES (72.35)	TH (71.80)	ZH (71.58)	FR (71.53)	BG (70.49)	AR (69.24)	JA (68.74)
AAR	EN (69.15)	JA (65.93)	ES (65.59)	AR (65.59)	ZH (65.42)	BG (64.40)	TH (63.56)	FR (61.02)
SIU	EN (59.26)	ES (57.31)	TH (56.05)	FR (55.67)	ZH (55.00)	BG (54.70)	JA (54.33)	AR (54.11)
VTI	EN (68.10)	ES (65.49)	ZH (65.42)	FR (64.08)	BG (62.82)	JA (61.83)	TH (61.13)	AR (60.00)

Table 7: Scenario-wise language rankings averaged over all evaluated models.

ensure that significance is not an artefact of pooling across heterogeneous backbones, we additionally report each test separately for every model (28 configurations each).

D.2 Results

Table 9 summarizes the analysis. Across all 84 configurations the mean gain is $+2.38\%$ with a 95% confidence interval of $[+2.05, +2.71]$ that lies well above zero, and the paired t-test strongly rejects the no-effect null ($t(83) = 14.37$, $p < 10^{-15}$). The non-parametric tests agree: the Wilcoxon signed-rank test gives $z = 7.77$ ($p = 7.8 \times 10^{-15}$), and of the 84 configurations 80 show a positive gain, 4 show exactly no change, and none shows a decrease, so the sign test likewise rejects at $p < 10^{-15}$.

The effect holds within every backbone. The per-model mean gains are $+2.08\%$, $+3.25\%$, and $+1.81\%$ for Gemma3-12B-IT, Qwen3.5-9B, and InternVL3.5-14B-Instruct respectively, each with a 95% confidence interval strictly above zero and each individually significant under all three tests. Significance is therefore not an artefact of pooling across models. Moreover, since no configuration in the entire evaluation grid exhibits a negative gain, the improvement delivered by MedVL-XLRepE is directionally consistent across all models, scenarios, and target languages.

Scenario	REPE	ZH	ES	FR	JA	TH	AR	BG
PR	Baseline	72.95	75.96	74.86	71.31	75.41	73.50	72.68
	Language	73.77 ↑0.82	76.23 ↑0.27	77.05 ↑2.19	73.22 ↑1.91	75.41 ↑0.00	73.22 ↓0.28	73.77 ↑1.09
	Medical	73.50 ↑0.55	75.41 ↓0.55	78.69 ↑3.83	73.22 ↑1.91	74.59 ↓0.82	74.59 ↑1.09	75.41 ↑2.73
	MedVL-XLRepE	74.32 ↑1.37	77.60 ↑1.64	76.23 ↑1.37	75.41 ↑4.10	77.32 ↑1.91	74.59 ↑1.09	77.60 ↑4.92
AAR	Baseline	66.10	69.49	70.34	69.49	65.25	67.80	68.64
	Language	67.80 ↑1.70	69.49 ↑0.00	72.03 ↑1.69	70.34 ↑0.85	66.10 ↑0.85	66.95 ↓0.85	68.64 ↑0.00
	Medical	66.95 ↑0.85	68.64 ↓0.85	72.88 ↑2.54	67.80 ↓1.69	66.95 ↑1.70	68.64 ↑0.84	68.64 ↑0.00
	MedVL-XLRepE	69.49 ↑3.39	71.19 ↑1.70	76.27 ↑5.93	75.42 ↑5.93	71.19 ↑5.94	69.49 ↑1.69	72.03 ↑3.39

Table 8: Component ablation of MedVL-XLRepE on Qwen3.5-9B. Each cell lists, from top to bottom: Baseline / Language only / Medical only / MedVL-XLRepE. Colored markers show change relative to Baseline.

Terminology Refinement Prompt

You are a medical translation reviewer. You are given an English medical source text and a draft translation produced by Google Translate. Your task is to produce the final translation in the target language.

Requirements:
1) Preserve the exact medical meaning, clinical intent, and level of specificity of the English source. Do not add, omit, weaken, or reinterpret information.
2) Correct medical terminology using standard, clinically appropriate wording in the target language.
3) Keep the original discourse form. If the source is a question, the output must remain a question. If the source is an answer option or fragment, preserve the same sentence style and granularity.
4) Revise the Google Translate draft conservatively: keep wording that is already correct and natural, and change only expressions that are medically inaccurate, linguistically awkward, or non-standard in clinical usage.
5) Pay particular attention to terminology that can affect downstream medical reasoning, including disease names, anatomical structures, laterality, severity descriptors, procedures, and pathology-related expressions.
6) Return only the final corrected translation. Do not output explanations, notes, JSON, or markdown.

English source text: {source_text}
Draft target-language translation: {google_translation}

Group	n	Mean (95% CI)	Paired t	Wilcoxon	Sign
Overall	84	+2.38 [2.05, 2.71]	$14.37^{\dagger}$	7.8×10^{-15}	80 / 0
Gemma3-12B-IT	28	+2.08 [1.53, 2.63]	$7.73^{\dagger}$	8.3×10^{-6}	26 / 0
Qwen3.5-9B	28	+3.25 [2.58, 3.93]	$9.89^{\dagger}$	5.6×10^{-6}	27 / 0
InternVL3.5-14B	28	+1.81 [1.46, 2.17]	$10.61^{\dagger}$	5.6×10^{-6}	27 / 0

Table 9: Statistical significance of MedVL-XLRepE cross-lingual gains, computed over the paired per-configuration gains in Table 2. Mean gain is in accuracy points (%) with a 95% confidence interval. Wilcoxon reports the two-sided signed-rank p-value; Sign reports the number of configurations with a positive / negative gain. † denotes $p < 10^{-5}$ for the paired t-test ($p < 10^{-15}$ for the Overall row).

Medical Reviewer Instruction

You are given an English source item from a medical VQA benchmark and its target-language version produced by machine translation and terminology refinement. Your task is to decide whether the target-language version is medically faithful to the English source.

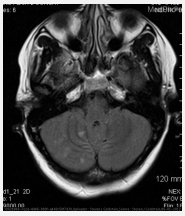
For each item, verify the following:
1) **Medical meaning.** The target-language question and answer options preserve the exact medical meaning, clinical intent, and level of specificity of the English source. No clinical information is added, omitted, weakened, or reinterpreted.
2) **Terminology.** Medical terms (disease names, anatomical structures, laterality, severity descriptors, procedures, and pathology-related expressions) are standard and clinically appropriate in the target language.
3) **Image grounding.** The question remains answerable from the same visual evidence as the English source. The translation does not introduce visual cues, anatomical references, or attributes that are not present in the image.
4) **Answer correctness.** The identity of the correct answer remains the same as in the English source.

Linguistic Reviewer Instruction

You are given an English source item from a medical VQA benchmark and its target-language version produced by machine translation and terminology refinement. Your task is to decide whether the target-language version is linguistically natural and structurally consistent with the source.

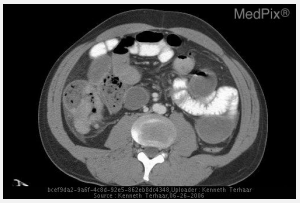
For each item, verify the following:
1) **Grammatical correctness.** The target-language question and answer options are grammatically well-formed in the target language.
2) **Naturalness.** The phrasing reads naturally to a native speaker. Awkward or literally translated constructions are revised into natural target-language expressions, without changing the meaning.
3) **Discourse form.** If the English source is a question, the target-language version remains a question; answer options remain options of the same sentence style and granularity.
4) **Answer-mapping stability.** Option order, label letters, and the structural mapping between the question and its options are preserved exactly as in the English source.

Perceptual Recognition (PR)



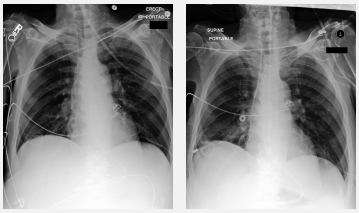
Question: Is the Right vertebral artery normal?
Answer: No

Attribute-Aware Recognition (AAR)



Question: Is the mass surrounding the aorta?
Answer: No

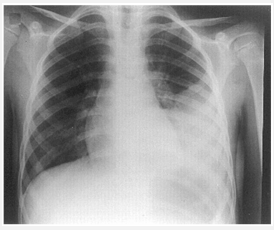
Sequential Images Understanding (SIU)



Prior Current

Question: How has the aeration changed in the base of the right lung when comparing both CXR images?
Options:
A. There is worsened aeration in the right lung base.
B. There is no change in aeration in the right lung base.
C. There is improved aeration in the right lung base.
D. Aeration has completely improved in both lung bases.
Answer: C. There is improved aeration in the right lung base.

Vision-Text Integrated Reasoning (VTI)



Question: A three-year-old child, malnourished, who had been hospitalized ten days ago, is taken to the Medical Emergency Department. The child has had an unchecked fever, cough and difficulty breathing for two days. The mother reports that the patient is unable to drink liquids and has vomited several times in the last 24 hours. Upon physical examination, the doctor observed that the child had a regular general condition, fever of 38.5°C, mild dehydration, tachydyspnea, with intercostal insufficiency, presence of crackling rales and decreased breath sounds in the left hemithorax; heart rate = 130 bpm, respiratory rate = 64 bpm and oxygen saturation = 91%. The chest x-ray is shown below. The etiological agent and treatment of pneumonia presented by the child are:
Options:
A. *Haemophilus influenzae*; crystalline penicillin.
B. *Streptococcus pneumoniae*; procaine penicillin.
C. *Staphylococcus aureus*; ceftriaxone associated with oxacillin.
D. *Mycoplasma pneumoniae*; antibiotic therapy with macrolides.
Answer: C. *Staphylococcus aureus*; ceftriaxone associated with oxacillin.

Figure 6: Examples of the four clinical scenarios.