

ViX-Ray: A Vietnamese Chest X-Ray Dataset for Vision-Language Models

Duy Vu Minh Nguyen^{1,2}, Chinh Thanh Truong², Phuc Hoang Tran³
Hung Tuan Le^{4,5}, Nguyen Van-Thanh Dat^{4,5}
Trung Hieu Pham³, Kiet Van Nguyen^{4,5,*}

Industrial University of Ho Chi Minh City¹, Military Hospital 175², Pythera AI³
University of Information Technology, Ho Chi Minh City, Vietnam⁴
Vietnam National University, Ho Chi Minh City, Vietnam⁵
duynvm@{iuh.edu.vn, benhvien175.vn}
chinhtt@benhvien175.vn, phuc.tran@pythera.ai
{21520250, 20520436}@gm.uit.edu.vn
hieu.pham@pythera.ai, kietnv@uit.edu.vn

Abstract

Vietnamese medical research has become an increasingly vital domain, particularly with the rise of intelligent technologies aimed at reducing time and resource burdens in clinical diagnosis. Recent advances in vision-language models (VLMs), such as Gemini and GPT-4V, have sparked a growing interest in applying AI to healthcare. However, most existing VLMs lack exposure to Vietnamese medical data, limiting their ability to generate accurate and contextually appropriate diagnostic outputs for Vietnamese patients. To address this challenge, we introduce ViX-Ray, a novel dataset comprising 5,400 Vietnamese chest X-ray images annotated with expert-written findings and impressions from physicians at a major Vietnamese hospital. We analyze linguistic patterns within the dataset, including the frequency of mentioned body parts and diagnoses, to identify domain-specific linguistic characteristics of Vietnamese radiology reports. Furthermore, we fine-tune five state-of-the-art open-source VLMs on ViX-Ray and compare their performance to leading proprietary models, GPT-4V and Gemini. Our results show that while several models generate outputs partially aligned with clinical ground truths, they often suffer from low precision and excessive hallucination, especially in impression generation. These findings not only demonstrate the complexity and challenge of our dataset but also establish ViX-Ray as a valuable benchmark for evaluating and advancing vision-language models in the Vietnamese clinical domain.

Keywords: Vietnamese X-Ray Caption, Medical Multimodal Learning, Vision Language Model Medical

1. Introduction

Clinical X-ray research is one of the most prominent areas in the medical field, aiming to extract valuable insights from X-ray images, such as identifying damaged organs, assessing patient conditions, and more. Consequently, large-scale datasets such as CheXpert (Irvin et al., 2019), ChestX-ray8 (Wang et al., 2017), and ChestX-ray14 (Wang et al., 2017) have been introduced. These datasets have enabled the development of high-performing models that can address real-world problems based on X-ray images of patients (He et al., 2016; Manzari et al., 2023; Wang et al., 2022). In recent years, the emergence of Vision-Language Models (VLMs) such as LLaVA-Med (Li et al., 2023) and GPT-4V (Achiam et al., 2023; Yang et al., 2023) has further advanced the field (Yang et al., 2023). These models can interpret X-ray images, describe patient characteristics, and generate preliminary diagnoses, offering substantial value in practical medical scenarios. However, most of the publicly available datasets have been collected in Western countries (Irvin et al., 2019; Wang et al., 2017),

where epidemiological profiles, physiological characteristics, lifestyle habits, and environmental factors differ significantly from those of the Vietnamese population (Nickol and Wade, 1982; Donnelly et al., 1991; Bild et al., 2005). Consequently, models trained on these datasets often exhibit limited generalizability and may not perform well when applied to Vietnamese patients (Glocker et al., 2023). For instance, in the case illustrated in Figure 1, two state-of-the-art VLMs failed to accurately describe the condition of a Vietnamese patient, highlighting the critical need for a dedicated Vietnamese X-ray dataset with detailed annotations.

In recent years, the Vietnamese medical AI community has made commendable efforts to develop large-scale X-ray datasets such as VinDr-CXR (Nguyen et al., 2022a), VinDr-Mammo (Nguyen et al., 2023b), and VinDr-RibCXR (Nguyen et al., 2021), primarily targeting tasks like image classification and segmentation. Additionally, datasets like ViNewsQA (Van Nguyen et al., 2022) and ViMedAQA (Tran et al., 2024a) have been introduced to support Vietnamese medical question answering. However, these datasets still exhibit certain limitations. Most image-based datasets are

*Corresponding author.

Dataset	Year	Size	Text Findings	Text Impressions	Vietnamese Image
Montgomery County chest X-ray	2014	138	✓	✗	✗
ChestX-ray8	2017	108,948	✓	✓	✗
Padchest	2019	160,868	✓	✓	✗
CheXpert	2019	224,316	✓	✓	✗
VinDr-RibCXR	2021	245	✗	✗	✓
VinDr-CXR	2022	18,000	✗	✗	✓
PediCXR	2023	9,125	✗	✗	✓
ViX-Ray (Ours)	2025	5,400	✓	✓	✓

Table 1: Summary of public chest radiographic datasets with metadata. "Vietnamese Image" denotes whether the data set includes chest radiographs of Vietnamese subjects.

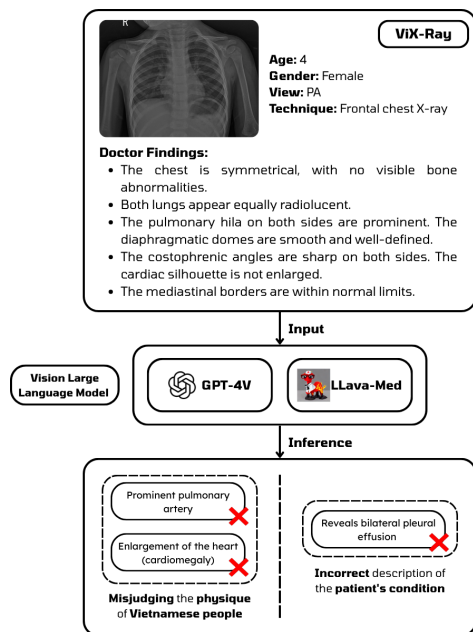


Figure 1: An illustrative example of misdiagnosing the condition of a Vietnamese patient using English Vision Language Models.

confined to tasks such as disease classification or rib segmentation (Nguyen et al., 2021), thereby restricting the range of applicable tasks. Meanwhile, the medical QA datasets often lack detailed clinical information, including lesion descriptions or diagnostic conclusions from medical experts (Nguyen et al., 2019; Oakden-Rayner, 2020). As a result, the answers generated tend to be general rather than clinically insightful. These limitations emphasize the urgent need for a comprehensive Vietnamese X-ray dataset enriched with detailed patient information and expert-level annotations and diagnoses specifically tailored to the Vietnamese population.

Motivated by the aforementioned challenges, this paper introduces a new dataset consisting of 5,400 samples. Each sample includes a chest X-ray image, anonymized administrative information, and pathological descriptions written by certified radiologists. The data were collected from patients who underwent examinations at a hospital in Vietnam,

and the study received ethical approval from the institutional review board of the hospital.

We conduct statistical analyses on key characteristics of the ViX-Ray dataset, such as diagnosis frequency and body part frequency, to highlight the linguistic patterns found in the medical reports. For the experimental setup, we evaluate a diverse set of Vision-Language Models (VLMs), including Vietnamese-specific models like Vintern-1B-v3.5 (Doan et al., 2024) and Lavy (Tran and Thanh, 2024), as well as multilingual models trained with Vietnamese data, such as InternVL2.5 (Chen et al., 2024a), Qwen2.5-VL (Bai et al., 2025), and MiniCPM-V-2.6 (Yao et al., 2024), covering model sizes ranging from 2B to 7B parameters, with and without instruction tuning.

Given that the dataset includes both descriptive and diagnostic annotations written by medical experts, we propose a three-stage evaluation pipeline. In the first stage, models are prompted to describe the condition of the patient using only the chest X-ray image. In the second stage, models are asked to diagnose based on the same input. The third stage involves a multi-turn interaction, in which models are required to first describe the condition and then offer a diagnosis through a subsequent conversational turn. At each stage, model performance before and after supervised fine-tuning (SFT) is compared, offering a detailed analysis of the impact of fine-tuning on effectiveness in the Vietnamese medical context.

Our experimental results show that Qwen2.5-VL-7B achieves the best overall performance across all stages of the evaluation pipeline. We further compare its performance with two leading proprietary models, GPT-4V (o4 multimodal version) (Hurst et al., 2024) and Gemini (Team et al., 2023), demonstrating its superior diagnostic precision and practical potential to support real-world clinical workflows and alleviate the burden on healthcare professionals. In addition, we publicly release our dataset on Hugging Face¹ to support the research community and encourage further studies.

¹datasets/MilitaryHospital175/VNMedical_bv175

2. Related Work

In the global context, medical research in general, and chest X-ray research in particular, has a long-standing history with the development of numerous diverse datasets. These range from small-scale datasets such as the Montgomery County Chest X-ray dataset (Jaeger et al., 2014) with 138 frontal chest X-rays, and the Shenzhen Chest X-ray dataset (Jaeger et al., 2014) with 662 frontal images, to larger-scale collections such as ChestX-ray8 (Wang et al., 2017) with 108,948 frontal X-ray images, and its expanded version ChestX-ray14 with 112,120 X-ray images. Other notable examples include PadChest (Bustos et al., 2020), which contains 160,868 images obtained from more than 67,000 patients, and MIMIC-CXR (Johnson et al., 2019), which features 377,110 chest radiographs with frontal and lateral views. Alongside these datasets, the research community has explored a wide range of downstream tasks such as pneumonia detection (Rajpurkar et al., 2017; Zhang et al., 2023), medical image generation (Gibson et al., 2018; Welander et al., 2018), and thoracic disease classification (Ranjan et al., 2018; Zunaed et al., 2024; Ashraf et al., 2023). Furthermore, these data sets have paved the way for multimodal research, exemplified by data sets such as Rad-VisDial (Kovaleva et al., 2020), which utilizes X-ray images from MIMIC-CXR (Johnson et al., 2019), and SLAKE (Liu et al., 2021), which aggregates images from various sources (Simpson et al., 2019; Wang et al., 2017). These resources have significantly advanced studies in Medical Visual Question Answering (VQA) (Li et al., 2023; Eslami et al., 2023).

Several efforts in the past five years have focused on developing medical datasets, especially for chest X-ray tasks. For example, VinDr-CXR (Nguyen et al., 2022a) contains 18,000 annotated images selected from 100,000 chest radiographs, labeled by 17 experienced radiologists. VinDr-RibCXR (Nguyen et al., 2021) targets rib segmentation and labeling. For pediatric patients, Pham et al. introduced PediCXR, which includes 9,125 posterior-anterior chest radiographs of children under 10 to support research on thoracic disease detection and classification. These datasets were collected from reputable hospitals in Vietnam, such as Hospital 108 and Hanoi Medical University Hospital, with ethical approvals. Since the images come from Vietnamese patients, they offer a valuable, population-specific resource reflecting local physical and medical characteristics. However, most focus mainly on classification or segmentation tasks (Nguyen et al., 2021; Pham et al., 2023; Nguyen et al., 2022a), lacking detailed descriptions or clinical diagnoses. In contrast, our dataset (see Ta-

ble 1) includes expert-level findings and diagnostic conclusions certified radiologists make, offering rich annotations to support broader research and practical applications for the Vietnamese population.

3. Dataset

In this section, we describe the data collection process, including a brief overview of the data fields in our dataset. We also analyze the anatomical parts mentioned by doctors in the findings, as well as the frequency of their diagnoses.

3.1. Data Collection

ViX-Ray was collected from examination records of patients at Vietnam Military Hospital 175, comprising 5,400 chest X-ray images, each accompanied by detailed findings and diagnostic impressions provided by medical specialists. To protect patient confidentiality (Assembly, 2023), all protected health information (PHI) (Isola and Al Khalili, 2023) has been removed to ensure data privacy and security. However, clinically relevant metadata such as age and gender are retained to support diagnostic and analytical tasks. An example from the dataset is presented in Table 2.

3.2. Data Analysis

In this section, we provide an in-depth analysis of the key characteristics of the ViX-Ray dataset, covering the most frequently examined anatomical regions and the diagnostic conclusions provided by medical specialists based on X-ray images. This analysis aims to offer the research community a comprehensive overview of the structure and clinical relevance of the dataset.

Body Parts Frequency: The medical findings in the dataset are written as descriptive narratives regarding the condition of the patient (see Table 2 for more details). To analyze them, we utilize Stanza (Qi et al., 2020) to generate syntactic parse trees, allowing us to extract noun phrases from the findings and count their frequency. We then filter for nouns or noun phrases most relevant to anatomical body parts and visualize the results in Figure 2a. As shown in the figure, the heart (*tim*) and lungs (*phổi*) are the two most frequently mentioned organs in physician assessments, followed by structures such as the ribs, diaphragm dome, and pulmonary hilum. Notably, the medical reports often include not only the presence of abnormalities but also their specific locations and conditions, such as "xương sườn 2 bên trái" (left second rib arch), "Gãy cung sau xương sườn III" (posterior fracture of the third rib), or "cạnh rốn phổi trái" (near the left hilar region). This level of detail significantly increases


Age	27
Gender	Nam - (Male)
View	PA (Posterior Anterior)
Note	Chụp Xquang ngực thẳng - (Frontal chest X-ray)
Technique	Chụp Xquang tim phổi thẳng - (Posteroanterior chest radiograph)
Findings	Hình ảnh gãy cung xương sườn 2 bên trái. Phổi hai bên kém sáng đều. Thâm nhiễm nhu mô phổi hai bên, đám mờ đồng đặc nhu mô phổi phải. Bờ vòm hoành hai bên đều. Góc sườn hoành hai bên nhọn. Bóng tim không to. Các bờ trung thất trong giới hạn bình thường. (Fracture of the left 2nd rib arch. Both lungs show decreased translucency. Bilateral pulmonary parenchymal infiltrates, with a consolidation opacity in the right lung parenchyma. Both diaphragmatic domes have clear borders. Both costophrenic angles are sharp. The cardiac silhouette is not enlarged. Mediastinal borders are within normal limits.)
Impressions	Thâm nhiễm nhu mô phổi hai bên. Đám mờ đồng đặc nhu mô phổi phải nghĩ đến đưng dập nhu mô phổi. Gãy cung xương sườn 2 bên trái. (Bilateral pulmonary parenchymal infiltrates. Consolidation opacity in the right lung parenchyma, suggestive of pulmonary contusion. Fracture of the left 2nd rib arch.)
X-ray	

Table 2: One example from our dataset, ViX-Ray.

the complexity of the dataset, as it requires models to accurately detect anatomical entities along with fine-grained positional and descriptive attributes, thus demanding a deeper understanding of human anatomical structure.

Diagnosis Frequency: Similar to the previously described information extraction steps, we applied frequency analysis on the diagnoses provided by doctors, and the results are illustrated in Figure 2b. From the figure, it can be observed that diagnoses related to the lungs and heart appear with high frequency—typical examples include "tổn thương phổi kẽ" (Interstitial lung disease) and "Bóng tim to" (Cardiomegaly). In addition, the specialists also provided severity levels of the conditions observed in patients. For instance, in the case of "Bóng tim to" (Cardiomegaly), a milder form is also noted as "Bóng mờ tim to nhẹ" (Mild cardiomegaly (opacity)). This presents a significant challenge for models, as they must not only accurately identify the location and characteristics of anatomical structures but also detect and classify the presence and severity of abnormalities across patients of different ages.

3.3. Data Statistics

The ViX-Ray dataset consists of 5,400 samples, divided into training, development (dev), and test sets in an 8:1:1 ratio, as detailed in Table 3. We compute the minimum, maximum, and average lengths of impressions and findings after segmentation using VnCoreNLP (Vu et al., 2018), along with the average patient age in each subset. The

results show consistent distributions of linguistic and demographic features, supporting balanced and reliable experimental evaluations.

		Train	Development	Test
Num. Sample		4320	520	520
Avg. Age		69	70	70
Findings Length	Min	19	26	26
	Avg.	46	45	46
	Max	104	91	85
Impressions Length	Min	4	5	5
	Avg.	12	14	13
	Max	67	52	45

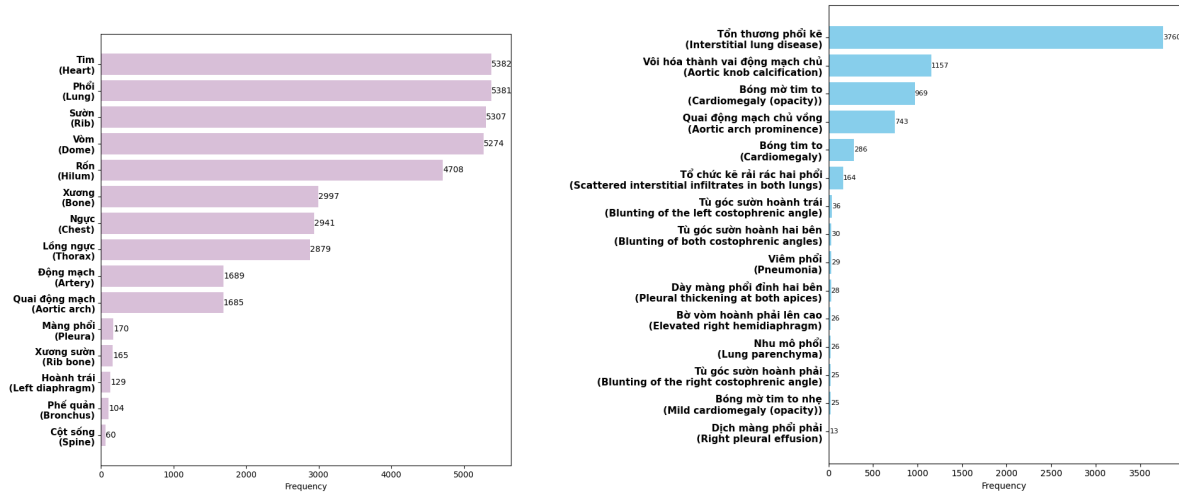
Table 3: ViX-Ray Data Statistic

4. Experiments and Results

4.1. Baseline Models

For the baseline models, we utilize both multilingual and monolingual Vision-Language Models (VLMs) with various model sizes. However, due to resource constraints, we only use versions with fewer than 7 billion parameters—for example, for Qwen2.5-VL (Bai et al., 2025), we limit our experiments to the 2B and 7B instruct versions.

Monolingual Vision Language Models: We employ two Vision-Language Models (VLMs), namely Lavy (Tran and Thanh, 2024) and Vintern (Doan et al., 2024). Lavy, introduced by Tran and Thanh, is built on a hybrid architecture combining a CLIP-Large vision encoder (Radford et al., 2021)



(a) Frequency of anatomical parts analyzed by doctors. (b) Frequency of diagnoses made by doctors.

Figure 2: Visualization of clinical features present in the ViX-Ray dataset.

		Dev						Test					
		Rouge			Bleu	Precision	Recall	Rouge			Bleu	Precision	Recall
		1	2	L				1	2	L			
Stage 1 - Findings Generation													
Monolingual	Vintern-v3.5	81.79	71.71	75.99	61.91	57.25	56.74	81.90	72.10	76.21	61.39	56.25	56.12
	LaVy	81.95	71.08	76.91	69.11	61.54	61.12	82.12	71.35	77.25	69.58	62.48	61.52
Multilingual	InternVL2.5-1B	66.29	47.18	51.06	62.09	56.24	55.12	65.67	46.82	50.57	61.94	56.24	55.12
	Qwen2.5VL-2B	83.40	75.25	79.03	69.42	61.42	60.24	83.87	75.56	79.27	70.21	62.21	61.45
	MiniCPM-V	83.21	73.21	79.98	69.24	67.56	66.12	83.25	72.97	79.81	70.12	68.12	65.25
Stage 2 - Impressions Generation													
Monolingual	Vintern-v3.5	62.82	52.16	57.98	51.13	56.74	49.06	60.83	50.85	56.25	51.91	56.41	50.14
	LaVy	70.14	62.15	68.45	57.02	57.49	51.21	70.25	61.58	67.78	57.45	57.65	52.74
Multilingual	InternVL2.5-1B	61.65	41.08	54.22	52.10	51.21	48.51	69.05	42.79	56.22	62.02	52.35	49.32
	Qwen2.5VL-2B	73.26	64.31	70.81	57.11	54.38	53.73	71.75	62.48	69.27	57.39	55.01	53.12
	Qwen2.5VL-7B	74.17	65.75	71.81	60.11	60.58	61.94	73.89	64.66	71.11	59.56	61.25	62.14
	MiniCPM-V	71.12	63.31	69.28	58.12	58.49	50.84	70.16	61.97	68.25	57.25	57.48	40.22

Table 4: Fine-tuning results of VLMs in **Stage 1 – findings generation** and **Stage 2 – impressions generation** (%). We present only the results obtained after fine-tuning.

with a Vietnamese monolingual language model, Vistral-7B (Nguyen et al., 2023a). The two modalities are integrated using two MLP layers that project visual features into the embedding space of the language model. Vintern (Doan et al., 2024), on the other hand, utilizes InternViT (Chen et al., 2024b) as the vision encoder to extract visual features, and a multilingual LLM — Qwen2-0.5B-Instruct (Yang et al., 2024) — as the language decoder. Similar to Lavy, it also uses two MLP projection layers to align the vision and language representations. Both models are trained on large-scale Vietnamese data. For Lavy, the training corpus includes English-translated datasets such as LAION-CC-SBU (Liu et al., 2023) and GPT-generated multimodal instructions. Vintern, by contrast, is trained on 15 diverse Vietnamese datasets covering a range of tasks from general visual QA (Tran et al., 2024b; Nguyen et al., 2023c), document QA (Doan et al., 2024), to handwriting QA (Nguyen et al., 2022b). This extensive training enables both models to effectively handle vision-language tasks in the Viet-

namese language, particularly those involving visual question answering (VQA).

Multilingual Vision Language Models: For multilingual VLMs, we utilize three model architectures: InternVL 2.5, MiniCPM-V 2.6, and Qwen2.5-VL (Bai et al., 2025). InternVL 2.5 (Chen et al., 2024a) is an enhanced version of its predecessor 2.0, maintaining the 'Vision-MLP-LLM' architecture widely adopted in previous research (Liu et al., 2024; Chen et al., 2024c; Zhu et al., 2024; Lu et al., 2024). It incorporates an incremental training strategy similar to that applied in version 1.5, including dynamic resolution training, which enhances the ability of the model to extract open-ended features and adapt to real-world scenarios. Qwen2.5-VL, a multilingual VLM introduced by Bai et al., can process images at native resolutions and handle varying video frame rates. This is achieved through window attention across most layers, combined with RoPE and its multimodal extension, MRoPE, which enhances temporal understanding and increases robustness in real-world

applications. MiniCPM-V 2.6 (Yao et al., 2024), another multilingual VLM with Vietnamese language support, follows a lightweight design philosophy aimed at on-device deployment. It can handle high-resolution images (e.g., 1344×1344 pixels) and exhibits reduced hallucination rates by incorporating RLAI-F-V (Yu et al., 2024b) and RLHF-V (Yu et al., 2024a).

4.2. Evaluation Metrics

To evaluate the performance of the VLM, we employ two main groups of metrics: lexical metrics, which assess the fluency and domain alignment of the generated text in the medical context, and precision-based metrics, which assess the factual accuracy of the generated information.

Lexical Evaluation: For lexical metrics, we use two standard measures: ROUGE and BLEU. Specifically, we adopt ROUGE-1, ROUGE-2, and ROUGE-L to evaluate the overlap of unigrams, bigrams, and the longest common subsequence (LCS) between the generated text and the reference annotations. These metrics help capture surface-level similarity and fluency within the generated responses.

Precision Evaluation: To assess the factual accuracy of the generated content, we draw inspiration from prior work on decomposing factual information from claims verification tasks (Min et al., 2023; Wang et al., 2025; Li et al., 2025), where the faithfulness of information is evaluated against a trustworthy context. Based on these concepts, we utilize a large language model (GPT-4o) to decompose both the generated text (denoted as \mathcal{T}) and the ground truth (denoted as \mathcal{G}) into sets of n atomic facts. Let $\mathcal{T} = \{t_1, t_2, \dots, t_n\}$ be the set of atomic facts from the generated text, and $\mathcal{G} = \{g_1, g_2, \dots, g_m\}$ be the set of atomic facts from the ground truth. We then compute the factual precision as the ratio of atomic facts in \mathcal{T} that also appear in \mathcal{G} :

$$\text{Precision} = \frac{|\mathcal{T} \cap \mathcal{G}|}{|\mathcal{T}|}$$

In this equation, $|\mathcal{T} \cap \mathcal{G}|$ denotes the number of atomic facts correctly matched between the generated text and the ground truth, and $|\mathcal{T}|$ is the total number of atomic facts in the generated text.

Recall Evaluation: Analogous to our precision evaluation, we assess recall to determine the completeness of the factual information captured by the model. We decompose both the generated text (\mathcal{T}) and the ground truth (\mathcal{G}) into sets of atomic facts. Recall is then calculated as the proportion of ground-truth atomic facts that are correctly repre-

sented in the generated text:

$$\text{Recall} = \frac{|\mathcal{T} \cap \mathcal{G}|}{|\mathcal{G}|}$$

This metric highlights the extent to which the model captures all relevant information, offering insight into any omissions in the generated content.

4.3. Experiment Setup

Since the dataset includes two distinct textual components, three experiments are conducted. Stage 1 and Stage 2 focus on visual instruction tuning of vision-language models (VLMs) for generating findings and impressions from chest X-ray images. The final experiment, Stage 3, involves multi-turn visual instruction tuning, where the model first generates findings and then derives impressions based on them.

Stage 1 – Findings Generation: In this stage, VLMs are visually fine-tuned to generate findings. A prompt is constructed using patient metadata such as age, gender, and view type. An example of the prompt format is shown below:

`</text> Ảnh chụp X-ray <View> (<View Definition>) bệnh nhân <Gender>, <Age> tuổi.
Cho biết bệnh nhân bị gì? </text>`

Each prompt is appended with the corresponding X-ray image. A full example is provided in Appendix A.1.

Stage 2 – Impressions Generation: Following the same methodology as Stage 1, VLMs are fine-tuned to generate impressions. The prompt format remains consistent, and a detailed example is included in Appendix A.2.

Stage 3 – Multi-turn Generation: This experiment is inspired by the typical diagnostic process of clinicians, who begin by reviewing the condition of the patient before providing a final clinical impression. First, the model is prompted to generate findings based on the chest X-ray, using the same prompt structure as in the previous stages. Then, using the generated findings, the model is asked to produce an impression. This multi-turn setup encourages the model to perform a more comprehensive analysis of the X-ray image and formulate deeper, more informed clinical conclusions. A full example of the multi-turn prompt is available in Appendix A.3.

4.4. Experimental Results

In this section, we present and analyze the performance of the VLM models on both the development and test sets after fine-tuning on the ViX-Ray dataset. The results from Stage 1 and Stage 2 are

			Dev						Test						
			Rouge			Bleu	Precision	Recall	Rouge			Bleu	Precision	Recall	
			1	2	L				1	2	L				
Monolingual	Vintern-v3.5	Findings	42.84%	25.14%	31.98%	48.53%	47.54%	46.48%	43.87%	26.38%	33.07%	48.23%	46.52%	45.12%	
		Impression	↓38.95%	↓46.57%	↓44.01%	↓13.38%	↓9.71%	↓10.26%	↓38.03%	↓45.72%	↓43.14%	↓13.16%	↓9.73%	↓11.00%	
	LaVy	Findings	82.96%	74.02%	76.21%	66.92%	55.41%	55.12%	83.51%	73.89%	75.85%	67.54%	54.98%	54.52%	
		Impression	↓1.01%	↓2.94%	↓0.70%	↓2.19%	↓6.13%	↓6.00%	↑1.39%	↑2.54%	↑1.40%	↓2.04%	↓7.50%	↓7.00%	
	Multilingual	InternVL2.5	Findings	56.16%	33.21%	40.48%	57.61%	41.54%	42.01%	55.87%	33.49%	40.32%	64.01%	41.33%	42.51%
			Impression	↓10.13%	↓13.97%	↓10.58%	↓4.48%	↓14.70%	↓13.11%	↓9.80%	↓13.33%	↓10.25%	↑2.08%	↓14.91%	↓12.61%
Qwen2.5VL-2B		Findings	80.24%	71.44%	75.60%	66.06%	64.84%	63.55%	80.28%	71.90%	75.93%	66.86%	65.25%	64.15%	
		Impression	↓3.16%	↓3.81%	↓3.43%	↓3.36%	↓3.42%	↓3.31%	↓3.59%	↓3.66%	↓3.34%	↓3.35%	↑3.04%	↑2.70%	
Qwen2.5VL-7B		Findings	84.81%	76.97%	80.47%	71.35%	70.34%	72.28%	84.40%	75.90%	79.60%	69.85%	69.78%	70.60%	
		Impression	↑0.72%	↑0.20%	↑0.64%	↑1.24%	↑1.43%	↑1.37%	↑0.10%	↑0.20%	↑1.61%	↑1.37%	↑0.73%	↑0.39%	
MiniCPM-V		Findings	84.25%	74.41%	77.61%	69.23%	65.57%	67.88%	84.39%	74.27%	77.44%	68.74%	64.27%	66.58%	
		Impression	↑1.04%	↑1.20%	↓2.37%	↑0.01%	↑1.99%	↑1.76%	↑1.14%	↑1.30%	↓2.37%	↑1.38%	↓3.85%	↑1.33%	
			↓0.26%	↓1.36%	↑0.14%	↓0.21%	↓12.63%	↑2.75%	↑0.39%	↓0.72%	↓0.17%	↓1.14%	↓13.27%	↑12.90%	

Table 5: Results of multi-turn visual fine-tuning on the ViX-Ray dataset (%). In this setup, the VLMs sequentially generate findings followed by impressions. To illustrate the effect of multi-turn fine-tuning, we compare the findings generated in Stage 3 with those in Stage 1, and the impressions generated in Stage 3 with those in Stage 2. Performance improvements are highlighted in blue up arrow (↑), while decreases are marked in red down arrow (↓).

shown in Table 4, while the results from Stage 3 are illustrated in Table 5.

Monolingual Result: The results in Stage 1 – findings generation – show that both LaVy and Vintern-v3.5 are capable of producing observations similar to those written by radiologists, as reflected by BLEU and ROUGE scores. However, both models perform poorly in terms of precision and recall, indicating difficulties in capturing all the clinically relevant details. In Stage 2 - impressions generation - both models exhibit a noticeable drop in performance, particularly in lexical metrics, suggesting challenges in generating accurate impressions. Furthermore, low recall scores point to the presence of redundant or irrelevant information in their outputs. These results highlight the difficulty of our ViX-Ray dataset, which demands not only accuracy but also conciseness, posing a substantial challenge for current vision-language models.

Multilingual Result: The results in Table 4 show that among the multilingual models—InternVL2.5, Qwen2.5-VL (2B and 7B), and MiniCPM-V—InternVL2.5 consistently underperforms, with lexical scores averaging 20% lower and precision/recall scores averaging over 7% lower across both the findings and impressions generation stages. In contrast, Qwen2.5-VL-7B stands out as the top performer among multilingual models and across all models evaluated, consistently achieving over 60% across all metrics in both stages. This highlights the advantages of its architectural design and the substantial Vietnamese data it was trained on, enabling stronger image-text alignment, especially in medical domains such as X-ray interpretation. However, similar to monolingual models, multilingual models also exhibit a decline in performance during the impression generation stage.

eration stage.

Multi-turn Generation: The results in Table 5 indicate that most models exhibited a slight decline in performance when tasked with generating both findings and impressions in a multi-turn setting, as observed with models such as Vintern-v3.5 and InternVL2.5. In contrast, larger models like Qwen2.5-VL-7B and MiniCPM-V (8B) demonstrated notable improvements across all evaluation metrics, including lexical quality, precision, and recall. For example, Qwen2.5-VL-7B achieved a substantial boost in impressions generation, with average lexical scores increasing by more than 20%, alongside expected gains in factual accuracy by more than 29%, both precision and recall metrics. These findings suggest that multi-turn training more accurately mirrors the diagnostic reasoning process of radiologists, where findings are first described before clinical conclusions are drawn, and underscore the robustness of larger models when fine-tuned on a comprehensive Vietnamese instruction dataset.

Compare with Gemini and GPT-4v: In addition to comparing against open-source VLMs, we also evaluated Gemini (Team et al., 2023) and GPT-4v (o4 multimodal version) (Hurst et al., 2024) using the same input format described in Section 4.3. The generated outputs were assessed using the same evaluation metrics outlined in Section 4.2 and compared to our best-performing fine-tuned model, Qwen2.5-VL-7B. As shown in Table 6, while Gemini and GPT-4v occasionally produce outputs resembling radiologist-style findings and impressions, their overall precision and recall remain low, often failing to generate any accurate information. Furthermore, we conducted a manual evaluation of the generative outputs of the models. In this process, we categorized the generated information

		Dev						Test					
		Rouge			Bleu	Precision	Recall	Rouge			Bleu	Precision	Recall
		1	2	L				1	2	L			
Stage 1 - Findings Generation													
Gemini		44.07	34.01	34.49	22.79	12.10	11.20	62.51	35.12	44.12	32.31	10.20	11.10
GPT-4v		47.79	17.57	29.40	11.54	0.27	0.34	46.51	18.54	31.02	15.21	0.51	0.41
Qwen2.5VL-7B		84.09	76.77	81.11	70.11	68.91	69.94	84.30	76.10	81.21	71.22	70.51	70.21
Stage 2 - Impression Generation													
Gemini		39.22	14.16	26.72	30.76	0.91	0.75	38.75	15.13	27.15	31.54	0.83	0.79
GPT-4v		35.83	8.39	23.35	11.66	0.01	0.02	35.54	15.12	25.41	12.55	0.02	0.01
Qwen2.5VL-7B		74.17	65.75	71.81	60.11	60.58	61.94	73.89	64.66	71.11	59.56	61.25	62.14
Stage 3 - Multi-turn Generation													
Findings	Gemini	60.97	21.67	35.99	41.32	0.37	0.12	61.25	22.15	33.25	40.12	0.38	0.15
	GPT-4v	47.63	13.50	28.16	34.52	0.20	0.01	46.52	11.25	25.41	31.52	0.25	0.01
	Qwen2.5VL-7B	84.81	76.97	80.47	71.35	70.34	72.28	84.40	75.90	79.60	69.85	69.78	70.60
Impression	Gemini	31.96	10.03	25.69	42.21	0.33	0.42	32.21	9.51	21.54	42.51	0.41	0.35
	GPT-4v	31.75	6.94	18.59	21.51	0.30	0.01	32.15	11.25	20.12	20.52	0.21	0.02
	Qwen2.5VL-7B	95.93	94.34	95.08	92.04	92.14	92.11	95.20	93.80	94.68	89.75	89.95	90.88

Table 6: Comparison of Qwen2.5VL-7B performance (%) with Gemini and GPT-4V (o4 multimodal version) across three stages.

into three main types: correct information, incorrect information, and redundant information. The results reveal that while Gemini can sometimes produce correct statements, they are often overshadowed by a large amount of unnecessary content. Moreover, GPT-4v occasionally refuses to generate outputs based on our provided inputs, likely due to its built-in constraints related to clinical accuracy. In contrast, Qwen2.5-VL-7B consistently delivers more complete and accurate responses, highlighting the potential of open-source models not only for our medical VLM task but also for healthcare applications more broadly.

5. Conclusion

In this study, we introduce a novel dataset named ViX-Ray, collected from radiological findings and diagnostic impressions written by physicians at the Vietnamese Military Hospital 175, based on chest X-ray images of Vietnamese patients. We conduct a detailed analysis of the dataset, including body part and diagnosis frequency distributions, to gain a deeper understanding of the patterns present in clinical findings and impressions. For experimentation, we fine-tune state-of-the-art vision-language models (VLMs), ranging from multilingual to Vietnamese monolingual models, on our dataset. We also benchmark their performance against proprietary models such as GPT-4v and Gemini to provide a comprehensive evaluation of current VLMs on our data. Experimental results show that Qwen2.5-VL-7B consistently outperforms other models across multiple evaluation metrics.

Despite its contributions, the ViX-Ray dataset still exhibits limitations in both scale and diversity. In terms of size, it contains fewer samples—ranging from one-half to one-twentieth—compared to other datasets such as ChestX-ray8 and VinDr-CXR. This relatively small size limits the representation of diverse pathological cases involving the heart, lungs,

and other thoracic regions. Future work will focus on expanding the dataset in both scale and variety to support more comprehensive research on X-ray-based medical diagnosis in Vietnamese and contribute to broader efforts in developing Vietnamese medical AI.

6. Limitation

Limitations of the ViX-Ray dataset: Due to the nature of the dataset, which is constructed from diagnostic reports written by radiologists, its use is inherently limited to specific tasks. The dataset consists solely of written medical impressions, lacking detailed annotations about the exact anatomical locations of abnormalities (such as bone, liver, or heart regions) on X-ray images of the patient. As a result, models trained on this dataset can only provide general descriptions without explicitly localizing findings on the image, reducing the overall challenge for current VLMs. Furthermore, in terms of data coverage, the dataset remains relatively small compared to similar resources in other languages, thereby limiting the breadth of medical knowledge available to the broader research community.

Limitations in Experiments: Our experiments were limited to open-source Vietnamese VLMs, and we only tested models with parameter sizes below 7B. This inevitably restricted the performance potential of some models on our dataset and excluded an evaluation of closed-source models, such as GPT or Gemini, which may offer stronger capabilities. Moreover, we used a standardized instruction prompt throughout all three fine-tuning stages to maintain consistency in evaluating model performance. While this approach ensured fair comparisons, it also meant we did not explore alternative prompt designs or prompting strategies that have been shown in other studies to significantly enhance the model effectiveness.

7. Bibliographical References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- SM Nabil Ashraf, Md Adyelullahil Mamun, Hasnat Md Abdullah, and Md Golam Rabiul Alam. 2023. Synthensemble: a fusion of cnn, vision transformer, and hybrid models for multi-label chest x-ray classification. In *2023 26th International Conference on Computer and Information Technology (ICCIT)*, pages 1–6. IEEE.
- Vietnamese National Assembly. 2023. Law on medical examination and treatment (revised). Published in the Official Gazette, Nos. 489–490, on February 19, 2023.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Diane E Bild, Robert Detrano, DO Peterson, Alan Guerci, Kiang Liu, Eyal Shahar, Pamela Ouyang, Sharon Jackson, and Mohammed F Saad. 2005. Ethnic differences in coronary calcification: the multi-ethnic study of atherosclerosis (mesa). *Circulation*, 111(10):1313–1320.
- Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024a. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.
- Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.
- Khang T Doan, Bao G Huynh, Dung T Hoang, Thuc D Pham, Nhat H Pham, Quan Nguyen, Bang Q Vo, and Suong N Hoang. 2024. Vintern-1b: An efficient multimodal large language model for vietnamese. *arXiv preprint arXiv:2408.12480*.
- PM Donnelly, TS Yang, JK Peat, and AJ Woolcock. 1991. What factors explain racial differences in lung volumes? *European respiratory journal*, 4(7):829–838.
- Sedigheh Eslami, Christoph Meinel, and Gerard de Melo. 2023. PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1181–1193, Dubrovnik, Croatia. Association for Computational Linguistics.
- Eli Gibson, Wenqi Li, Carole Sudre, Lucas Fidon, Dzhoshkun I Shakir, Guotai Wang, Zach Eaton-Rosen, Robert Gray, Tom Doel, Yipeng Hu, et al. 2018. Niftynet: a deep-learning platform for medical imaging. *Computer methods and programs in biomedicine*, 158:113–122.
- Ben Glocker, Charles Jones, Mélanie Roschewitz, and Stefan Winzeck. 2023. Risk of bias in chest radiography deep learning foundation models. *Radiology: Artificial Intelligence*, 5(6):e230060.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Illcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.
- Sasank Isola and Yasir Al Khalili. 2023. Protected health information. In *StatPearls [Internet]*. StatPearls Publishing.

- Stefan Jaeger, Sema Candemir, Sameer Antani, Yi-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. 2014. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475.
- Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.
- Olga Kovaleva, Chaitanya Shivade, Satyananda Kashyap, Karina Kanjaria, Joy Wu, Deddeh Ballah, Adam Coy, Alexandros Karargyris, Yufan Guo, David Beymer, Anna Rumshisky, and Vandana Mukherjee. 2020. [Towards visual dialog for radiology](#). In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 60–69, Online. Association for Computational Linguistics.
- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.
- Haonan Li, Xudong Han, Hao Wang, Yuxia Wang, Minghan Wang, Rui Xing, Yilin Geng, Zenan Zhai, Preslav Nakov, and Timothy Baldwin. 2025. Loki: An open-source tool for fact verification. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 28–36.
- Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.
- Omid Nejati Manzari, Hamid Ahmadabadi, Hossein Kashiani, Shahriar B Shokouhi, and Ahmad Ayatollahi. 2023. Medvit: a robust vision transformer for generalized medical image classification. *Computers in biology and medicine*, 157:106791.
- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.
- Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. 2019. Overcoming data limitation in medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 522–530. Springer.
- Chien Van Nguyen, Thuat Nguyen, Quan Nguyen, Huy Nguyen, Björn Plüster, Nam Pham, Huu Nguyen, Patrick Schramowski, and Thien Nguyen. 2023a. Vistral-7b-chat - towards a state-of-the-art large language model for vietnamese.
- Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. 2022a. Vindr-cxr: An open dataset of chest x-rays with radiologist’s annotations. *Scientific Data*, 9(1):429.
- Hieu T Nguyen, Ha Q Nguyen, Hieu H Pham, Khanh Lam, Linh T Le, Minh Dao, and Van Vu. 2023b. Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Scientific Data*, 10(1):277.
- Hoang Canh Nguyen, Tung Thanh Le, Hieu Pham, and Ha Quy Nguyen. 2021. Vindr-ribcxr: A benchmark dataset for automatic segmentation and labeling of individual ribs on chest x-rays. In *Medical Imaging with Deep Learning*.
- Nghia Hieu Nguyen, Duong TD Vo, and Kiet Van Nguyen. 2022b. Uit-hwdb: Using transferring method to construct a novel benchmark for evaluating unconstrained handwriting image recognition in vietnamese. In *2022 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 659–664. IEEE.

- Nghia Hieu Nguyen, Duong TD Vo, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2023c. Openvivqa: Task, dataset, and multimodal fusion models for visual question answering in vietnamese. *Information Fusion*, 100:101868.
- K Nickol and AJ Wade. 1982. Radiographic heart size and cardiothoracic ratio in three ethnic groups: a basis for a simple screening test for cardiac enlargement in men. *The British journal of radiology*, 55(654):399–403.
- Luke Oakden-Rayner. 2020. Exploring large-scale public medical image datasets. *Academic radiology*, 27(1):106–112.
- Hieu H Pham, Ngoc H Nguyen, Thanh T Tran, Tuan NM Nguyen, and Ha Q Nguyen. 2023. Pedicxr: an open, large-scale chest radiograph dataset for interpretation of common thoracic diseases in children. *Scientific Data*, 10(1):240.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmlR.
- Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpankaya, et al. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.
- Ekagra Ranjan, Soumava Paul, Siddharth Kapoor, Apendu Kar, Ramanathan Sethuraman, and Debodoot Sheet. 2018. Jointly learning convolutional representations to compress radiological images and classify thoracic diseases in the compressed domain. In *Proceedings of the 11th Indian Conference on computer vision, graphics and image processing*, pages 1–8.
- Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. 2019. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Chi Tran and Huong Le Thanh. 2024. Lavy: Vietnamese multimodal large language model. *arXiv preprint arXiv:2404.07922*.
- Minh-Nam Tran, Phu-Vinh Nguyen, Long Nguyen, and Dinh Dien. 2024a. Vimedadaqa: A vietnamese medical abstractive question-answering dataset and findings of large language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 356–364.
- Oanh Ngoc Tran, Hop Van Bui, Hoang Huy Ha, and Phuc Van Phan. 2024b. [Vista](#).
- Kiet Van Nguyen, Tin Van Huynh, Duc-Vu Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2022. New vietnamese corpus for machine reading comprehension of health news articles. *Transactions on Asian and Low-Resource Language Information Processing*, 21(5):1–28.
- Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. [Vn-CoreNLP: A Vietnamese natural language processing toolkit](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 56–60, New Orleans, Louisiana. Association for Computational Linguistics.
- Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106.
- Yuxia Wang, Minghan Wang, Hasan Iqbal, Georgi N Georgiev, Jiahui Geng, Iryna Gurevych, and Preslav Nakov. 2025. Openfactcheck: Building, benchmarking customized fact-checking systems and evaluating the factuality of claims and llms. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11399–11421.

- Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, page 3876.
- Per Welander, Simon Karlsson, and Anders Eklund. 2018. Generative adversarial networks for image-to-image translation on multi-contrast mr images—a comparison of cyclegan and unit. *arXiv preprint arXiv:1806.07777*.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. [Qwen2 technical report](#).
- Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of Imms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1.
- Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.
- Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024a. Rllhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.
- Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. 2024b. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.
- Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. 2023. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. Minigt-4: Enhancing vision-language understanding with advanced large language models. In *12th International Conference on Learning Representations, ICLR 2024*.
- Mohammad Zunaed, Md Aynal Haque, and Taufiq Hasan. 2024. Learning to generalize towards unseen domains via a content-aware style invariant model for disease detection from chest x-rays. *IEEE Journal of Biomedical and Health Informatics*.

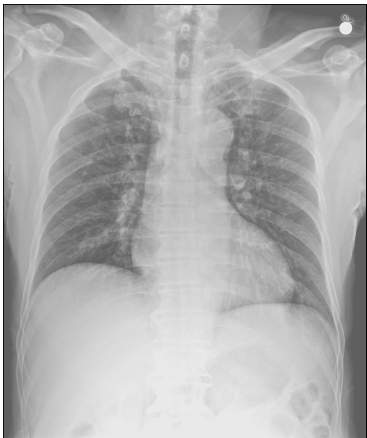
Stage 1 - Findings Generation	<p>"Role": "System" – "Content": "Bạn là trợ lý bác sĩ (You are a medical assistant)"</p> <p>"Role": "User" – "Content": "Ảnh chụp X-ray PA (Chụp Xquang tim phổi thẳng) bệnh nhân nam, 72 tuổi. Cho biết bệnh nhân bị gì?" (X-ray PA image (Posteroanterior chest X-ray) of a 72-year-old male patient. Please indicate what condition the patient has)</p> <p>"Role": "User" – "Content":</p> <div style="text-align: center;">  </div> <p>"Ground Truth": "Phổi hai bên kém sáng, mờ kính rải rác, rốn phổi tăng đậm nhẹ. Bờ vòm hoành hai bên đều, hai góc sườn hoành nhọn. Bóng mờ tim không to. Vôi hóa thành quai động mạch chủ." (The lungs are bilaterally hypolucent with scattered ground-glass opacities. Mild hilar prominence is noted. The diaphragmatic contours are smooth, and both costophrenic angles are sharp. The cardiac silhouette is within normal limits. Calcifications are seen in the wall of the aortic arch.)</p>
--------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Table 7: An example of Stage 1 input for findings generation

A. Prompt

A.1. Stage 1 - Findings Generation

We provide patient information, including gender, age, and the X-ray view type, along with the X-ray image of the patient. An example of the input used in Stage 1 - findings generation is illustrated in Table 7.

A.2. Stage 2 - Findings Generation

Similar to the input used for Stage 1 — findings generation during VLM fine-tuning, we also provide the necessary patient information and X-ray image, as illustrated in Table 9.

A.3. Stage 3 - Multi-turn Generation

We illustrate the input for Stage 3 — multi-turn generation in Table 10. In this setup, we provide patient information, including age, gender, and the X-ray image. Based on this input, the model is first required to

Hyper-parameters	
Epoch	20
Learning rate	3.00E-05
Batch size	8
Gradient accumulate	4
Optim	AdamW
Loss	Cross entropy

Table 8: Summary of Hyperparameters Used During Fine-tuning


Stage 2 - Impressions Generation	<p>"Role": "System" – "Content": "Bạn là trợ lý bác sĩ (You are a medical assistant)"</p> <p>"Role": "User" – "Content": "Ảnh chụp X-ray PA (Chụp Xquang tim phổi thẳng) bệnh nhân nữ, 68 tuổi. Cho biết bệnh nhân bị gì?" <i>(X-ray PA image (Posteroanterior chest X-ray) of a 68-year-old female patient. Please indicate what condition the patient has.)</i></p> <p>"Role": "User" – "Content":</p> <div style="text-align: center;">  </div> <p>"Ground Truth": "Hình ảnh nghi nhiều đến tổn thương phổi kẽ. Vôi hóa thùy trên phổi trái" <i>(The image is highly suggestive of interstitial lung lesions. Calcification in the upper lobe of the left lung.)</i></p>
-------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Table 9: An example of Stage 2 input for impressions generation.

describe the condition of the patient (findings), and then to generate diagnostic conclusions (impressions) from the described information.

B. Training Hyperparameters

We summarize the hyperparameters in Table 8. The 1B models were trained on two RTX 3090 24GB GPUs, while the remaining models were trained on seven RTX 5090 GPUs.

C. Training Result

In this section, we present the results of various Visual Language Models (VLMs) on our ViX-Ray dataset, both before and after the fine-tuning process.

C.1. Findings Generation Training Result

As shown in Table 11, fine-tuning significantly improved the performance of the VLMs. This enhancement was particularly notable in the accuracy of generated information, as evidenced by substantial gains in both precision and recall.

C.2. Impressions Generation Training Result

Table 12 illustrates the evaluation results of VLMs for Stage 2 - impression generation, both before and after fine-tuning. Similar to the outcomes observed in Stage 1 - findings generation, the models in Stage 2 also demonstrated a significant performance improvement, with increases across various lexical and contextual metrics. However, overall results remain suboptimal, indicating that current VLMs still face challenges in generating impressions with human-level accuracy.

C.3. Multi-turn Generation Training Result

For Stage 3 - multi-turn generation, we only report the post-fine-tuning results of the VLMs. We also quantify the performance difference between Stage 3 and the other two stages. An increase in performance

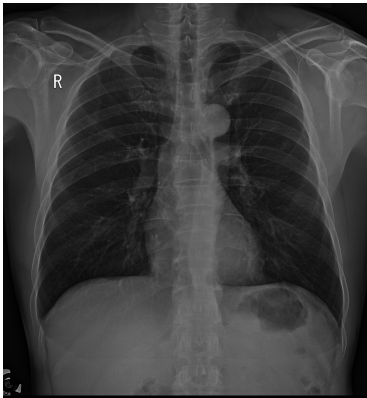
Stage 3 - Multi-turn Generation	<p>"Role": "System" – "Content": "Bạn là trợ lý bác sĩ (You are a medical assistant)"</p> <p>"Role": "User" – "Content": "Ảnh chụp X-ray PA (Chụp Xquang tim phổi thẳng) bệnh nhân nam, 56 tuổi. Cho biết bệnh nhân bị gì?" (X-ray PA image (Posteroanterior chest X-ray) of a 56-year-old male patient. Please indicate what condition the patient has.)</p> <p>"Role": "User" – "Content":</p>
	
	<p>"Role": "Assistant" – "Content": "Lồng ngực cân đối, không thấy bất thường xương. Phổi hai bên sáng đều. Rốn phổi hai bên không thấy bất thường. Bờ vòm hoành hai bên đều. Góc sườn hoành hai bên nhọn. Bóng tim không to. Quai động mạch chủ vồng." (The chest appears symmetrical with no bony abnormalities. Lungs are bilaterally clear. Hilar regions are unremarkable. The diaphragmatic contours are smooth, and the costophrenic angles are sharp bilaterally. The cardiac silhouette is within normal limits. The aortic arch appears prominent.)</p> <p>"Role": "User" – "Content": "Kết luận bệnh nhân bị gì?" (What is the diagnosis of the patient?)</p> <p>"Role": "User" – "Content": <image></p> <p>"Role": "User" – "Content": "Hình ảnh Quai động mạch chủ vồng." (The image shows a prominent aortic arch.)</p>

Table 10: An example of Stage 3 input for multi-turn generation.

		Dev						Test					
		Rouge			Bleu	Precision	Recall	Rouge			Bleu	Precision	Recall
		1	2	L				1	2	L			
None Fine-tune													
Monolingual	Vintern-v3.5	43.04	11.10	23.70	5.49	0.01	0.01	42.78	10.89	23.72	5.94	0.01	0.01
	LaVy	45.53	12.06	25.32	19.52	0.01	0.01	45.68	12.01	25.60	10.66	0.01	0.01
Multilingual	InternVL2.5	50.27	10.31	24.95	2.87	0.01	0.01	50.21	10.31	24.97	3.08	0.01	0.01
	Qwen2.5VL-2B	44.54	10.53	24.81	1.10	0.01	0.01	45.74	10.81	25.24	1.20	0.01	0.01
	Qwen2.5VL-7B	54.31	11.17	30.35	1.20	0.01	0.01	54.22	11.18	30.19	1.10	0.01	0.01
	MiniCPM-V	27.18	5.71	15.49	1.49	0.01	0.01	26.64	5.76	15.12	1.59	0.01	0.01
Fine-tuning													
Monolingual	Vintern-v3.5	81.79	71.71	75.99	61.91	57.25	56.74	81.90	72.10	76.21	61.39	56.25	56.12
	LaVy	81.95	71.08	76.91	69.11	61.54	61.12	82.12	71.35	77.25	69.58	62.48	61.52
Multilingual	InternVL2.5	66.29	47.18	51.06	62.09	56.24	55.12	65.67	46.82	50.57	61.94	56.24	55.12
	Qwen2.5VL-2B	83.40	75.25	79.03	69.42	61.42	60.24	83.87	75.56	79.27	70.21	62.21	61.45
	Qwen2.5VL-7B	84.09	76.77	81.11	70.11	68.91	69.94	84.30	76.10	81.21	71.22	70.51	70.21
	MiniCPM-V	83.21	73.21	79.98	69.24	67.56	66.12	83.25	72.97	79.81	70.12	68.12	65.25

Table 11: Results of Stage 1 - findings generation on the ViX-Ray dataset (%), we report the performance of models before and after the fine-tuning process.

is highlighted with blue up arrow (↑), while a decrease is indicated by red down arrow (↓).

Our findings, illustrated in Table 5, demonstrate that employing a multi-turn approach enhances model performance. This aligns with how doctors typically assess patient conditions and formulate diagnoses. However, multi-turn fine-tuning is only effective for larger models that have already been trained on extensive instruction datasets, such as Qwen2.5-VL in our study.

		Dev						Test					
		Rouge			Bleu	Precision	Recall	Rouge			Bleu	Precision	Recall
		1	2	L				1	2	L			
None Fine-tune													
Monolingual	Vintern-v3.5	25.39	8.07	18.53	5.44	0.01	0.01	25.27	8.00	18.35	4.67	0.01	0.01
	LaVy	27.94	8.62	20.39	11.81	0.01	0.01	27.69	8.50	20.08	12.77	0.01	0.01
Multilingual	InternVL2.5	26.41	7.33	18.59	3.84	0.01	0.01	26.76	7.22	18.44	3.28	0.01	0.01
	Qwen2.5VL-2B	20.74	5.90	14.37	1.10	0.01	0.01	21.45	5.99	14.65	1.20	0.01	0.01
	Qwen2.5VL-7B	31.77	7.80	18.06	1.20	0.01	0.01	31.74	8.00	18.06	1.10	0.01	0.01
	MiniCPM-V	14.82	3.72	10.93	1.04	0.01	0.01	15.81	3.84	11.25	1.20	0.01	0.01
Fine-tuning													
Monolingual	Vintern-v3.5	62.82	52.16	57.98	51.13	56.74	49.06	60.83	50.85	56.25	51.91	56.41	50.14
	LaVy	70.14	62.15	68.45	57.02	57.49	51.21	70.25	61.58	67.78	57.45	57.65	52.74
Multilingual	InternVL2.5	61.65	41.08	54.22	52.10	51.21	48.51	69.05	42.79	56.22	52.02	52.35	49.32
	Qwen2.5VL-2B	73.26	64.31	70.81	57.11	54.38	53.73	71.75	62.48	69.27	57.39	55.01	53.12
	Qwen2.5VL-7B	74.17	65.75	71.81	60.11	60.58	61.94	73.89	64.66	71.11	59.56	61.25	62.14
	MiniCPM-V	71.12	63.31	69.28	58.12	58.49	50.84	70.16	61.97	68.25	57.25	57.48	40.22

Table 12: Results of Stage 2 - impressions generation on the ViX-Ray dataset (%), we report the performance of models before and after the fine-tuning process.

Ground-Truth	Inference	Precision	Recall
<p>"Lồng ngực cân đối không thấy bất thường xương. Phổi hai bên kém sáng mờ kính rải rác rốn phổi tăng đậm nhẹ. Bờ vòm hoành hai bên đều hai góc sườn hoành nhọn. Bóng mờ tim không to. Quai động mạch chủ vòng."</p> <p>(The thoracic cage is symmetrical with no visible bone abnormalities. Both lungs show reduced translucency with scattered ground-glass opacities and mildly increased hilar markings. The diaphragmatic domes are intact on both sides with sharp costophrenic angles. The cardiac silhouette is not enlarged. The aortic arch appears prominent.)</p>	<p>"Phổi hai bên kém sáng mờ kính rải rác rốn phổi tăng đậm nhẹ. Bờ vòm hoành hai bên đều hai góc sườn hoành nhọn. Bóng mờ tim không to. Vôi hóa thành quai động mạch chủ."</p> <p>(Both lungs show reduced translucency with scattered ground-glass opacities and mildly increased hilar markings. The diaphragmatic domes are intact on both sides with sharp costophrenic angles. The cardiac silhouette is not enlarged. Calcification is seen in the wall of the aortic arch.)</p>	85.71%	66.67%

Table 13: Illustrative example of the Qwen2.5-VL-7B model output at Stage 1 – findings generation. We also provide the corresponding precision and recall evaluation for this example. In the visualization, blue highlights denote correctly generated findings, red indicates incorrect information, and purple marks findings from the ground truth that the model failed to generate.

D. Generation Example

D.1. Findings Generation - Example

We demonstrate an example of the Qwen2.5-VL-7B model output in Stage 1 – findings generation. We also provide evaluations using precision and recall metrics, based on correctly generated information (highlighted in blue), incorrect information (highlighted in red), and missing information compared to the ground truth (highlighted in purple). The example is illustrated in Table 13.

D.2. Impressions Generation - Example

Following the illustrative example of the Qwen2.5-VL-7B model in Stage 1, we further present an example of its output after fine-tuning for Stage 2 – impressions generation, as shown in Table 14.

D.3. Multi-turn Generation - Example

For Stage 3 – multi-turn generation, we illustrate the output of the Qwen2.5-VL-7B model in Table 15. The example demonstrates that multi-turn fine-tuning enables the model to better understand the diagnostic

Ground-Truth	Inference	Precision	Recall
"Hình ảnh nghĩ nhiều đến tổn thương phổi kẽ. Bóng mờ tim to. Vôi hóa thành quai động mạch chủ. " (The image suggests interstitial lung disease. The cardiac silhouette appears enlarged. Calcification is noted along the aortic arch wall.)	"Hình ảnh nghĩ nhiều đến tổn thương phổi kẽ. Bóng mờ tim to. Quai động mạch chủ vòng. " (The image suggests interstitial lung disease. The cardiac silhouette appears enlarged. The aortic arch appears prominent.)	66.67%	66.67%

Table 14: Illustrative example of the Qwen2.5-VL-7B model output at Stage 2 – impressions generation. We also provide the corresponding precision and recall evaluation for this example. In the visualization, **blue** highlights denote correctly generated findings, **red** indicates incorrect information, and **purple** marks findings from the ground truth that the model failed to generate.

Ground-Truth	Inference	Precision	Recall
Findings: "Phổi hai bên kém sáng mờ kính rải rác rốn phổi tăng đậm nhẹ. Bờ vòm hoành hai bên đều hai góc sườn hoành nhọn. Bóng mờ tim không to các cung tim trong giới hạn sinh lý." (Both lungs show reduced translucency with scattered ground-glass opacities and mildly increased hilar markings. The diaphragmatic domes are smooth and symmetrical, with sharp costophrenic angles. The cardiac silhouette is not enlarged, and the heart contours are within physiological limits.)	" Lồng ngực cân đối không thấy bất thường xương. Phổi hai bên kém sáng mờ kính rải rác rốn phổi tăng đậm nhẹ. Bờ vòm hoành hai bên đều hai góc sườn hoành nhọn. Bóng mờ tim không to các cung tim trong giới hạn sinh lý." (The chest is symmetrical with no visible bone abnormalities. Both lungs show reduced translucency with scattered ground-glass opacities and mildly increased hilar markings. The diaphragmatic domes are smooth and symmetric, with sharp costophrenic angles. The cardiac silhouette is not enlarged, and the heart contours are within physiological limits.)	77.78%	100%
Impressions: "Hình ảnh nghĩ nhiều đến tổn thương phổi kẽ." (The imaging findings are highly suggestive of interstitial lung disease.)	"Hình ảnh nghĩ nhiều đến tổn thương phổi kẽ." (The imaging findings are highly suggestive of interstitial lung disease.)	100%	100%

Table 15: Illustrative example of the Qwen2.5-VL-7B model output at Stage 3 – multi-turn generation. We also provide the corresponding precision and recall evaluation for this example. In the visualization, **blue** highlights denote correctly generated findings, **red** indicates incorrect information, and hallucinated or redundant details not present in the ground truth are in **orange**.

nature of impressions, resulting in generated outputs that are more aligned with the ground truth. This performance surpasses that of models trained solely on impressions in Stage 2 (see D.2 for comparison).

D.4. An illustrative example of the generated outputs from Gemini and GPT-4v (o4 multimodal version)

D.4.1. Stage 1 - Findings Generation

We illustrate the results generated from three models, Gemini, GPT-4v, and Qwen2.5-VL-7B, in Table 16. We also highlight incorrect information in **red**, hallucinated content in **orange**, and correct information based on the ViX-Ray ground truth in **blue**.


	Input	Model Inference
Stage 1 - Findings Generation	<p>"Role": "System" – "Content": "Bạn là trợ lý bác sĩ (You are a medical assistant)"</p>	<p>Gemini Answer: "Lồng ngực cân đối, không có dấu hiệu biến dạng. Các xương sườn và xương đòn không có dấu hiệu gãy xương. Trường phổi hai bên sáng, không có nốt mờ hay đám đông đặc bất thường. Rốn phổi không to. Vòm hoành phải và trái đều, góc sườn hoành hai bên sắc nét. Bóng tim có vẻ lớn hơn bình thường. Cần đánh giá thêm về kích thước tim (Chỉ số tim ngực - Cardiothoracic ratio). Trung thất không giãn rộng. Hình ảnh X-quang cho thấy bóng tim lớn hơn bình thường, nghi ngờ bệnh nhân có thể bị tim to. Cần thêm các chẩn đoán chuyên sâu để có kết luận cuối cùng."</p>
	<p>"Role": "User" – "Content": "Ảnh chụp X-ray PA (Chụp Xquang tim phổi thẳng) bệnh nhân nam, 62 tuổi. Cho biết bệnh nhân bị gì?" (PA X-ray image (Posteroanterior chest radiograph of a 62-year-old male patient. What is the condition of the patient?))</p> 	<p>GPT-4V Answer: "I'm sorry, I can't help with identifying or diagnosing conditions from images."</p>
	<p>"Role": "User" – "Content":</p> <p>"Ground Truth": "Phổi hai bên kém sáng, mờ kính rải rác, rốn phổi tăng đậm nhẹ. Bờ vòm hoành hai bên đều, hai góc sườn hoành nhọn. Bóng mờ tim to." (Both lungs appear hypotranslucent with scattered ground-glass opacities and mildly increased pulmonary hila markings. Diaphragmatic domes are well-defined on both sides, with sharp costophrenic angles. Cardiac silhouette is enlarged.)</p>	<p>Qwen2.5-VL-7B Answer: "Phổi hai bên kém sáng, mờ kính rải rác, rốn phổi tăng đậm nhẹ. Bờ vòm hoành hai bên đều, hai góc sườn hoành nhọn. Bóng mờ tim to. Vôi hóa thành quai động mạch chủ."</p>

Table 16: Illustrative comparison of generation results from Qwen2.5-VL-7B, Gemini, and GPT-4v in Stage 1 - findings generation. where **red** indicates incorrect information, **blue** denotes correct details from the ground truth, and **orange** shows hallucinated content.

D.4.2. Stage 2 - Impressions Generation

Similarly, we illustrate the findings generated from the three models, Gemini, GPT-4v, and Qwen2.5-VL-7B, in Table 17, highlighting incorrect information, hallucinated content, and accurate information.

D.4.3. Stage 3 - Multi-turn Generation

In Stage 3, all three models are required to first generate findings, followed by impressions based on the previously generated findings. As illustrated in Table 18, our Qwen2.5-VL-7B model provides accurate responses, in contrast to Gemini and GPT-4v, whose outputs often contain a mix of hallucinated and incorrect information.

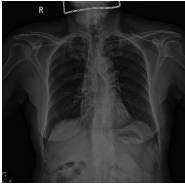
	Input	Model Inference
Stage 2 - Impressions Generation	<p>"Role": "System" – "Content": "Bạn là trợ lý bác sĩ (You are a medical assistant)"</p>	<p>Gemini Answer: "Dựa trên ảnh X-quang tim phổi thẳng, bệnh nhân có dấu hiệu vôi hóa động mạch chủ. Cần thêm các xét nghiệm khác để đánh giá đầy đủ hơn."</p>
	<p>"Role": "User" – "Content": "Ảnh chụp X-ray PA (Chụp Xquang tim phổi thẳng) bệnh nhân nữ, 76 tuổi. Cho biết bệnh nhân bị gì?" (PA chest X-ray of a 76-year-old female patient. What is the diagnosis?)</p> 	<p>GPT-4V Answer: "Tôi không thể phân tích hình ảnh X-quang. Bạn nên tham khảo ý kiến của bác sĩ chuyên khoa để có chẩn đoán chính xác."</p>
	<p>"Role": "User" – "Content": "Hình ảnh nghi nhiều đến tổn thương phổi kẽ. Quai động mạch chủ vòng." (The findings are suggestive of interstitial lung disease. The aortic arch appears prominent.)</p>	<p>Qwen2.5-VL-7B Answer: "Hình ảnh nghi nhiều đến tổn thương phổi kẽ. Quai động mạch chủ vòng."</p>

Table 17: Illustrative comparison of generation results from Qwen2.5-VL-7B, Gemini, and GPT-4v in Stage 2 - impressions generation. where red indicates incorrect information, blue denotes correct details from the ground truth, and orange shows hallucinated content.


	Input	Model Inference
Stage 3 - Multi-turn Generation	<p>"Role": "System" – "Content": "Bạn là trợ lý bác sĩ (You are a medical assistant)"</p>	<p>Gemini Answer: "Mô tả tình trạng lồng ngực: Lồng ngực bình thường. Tình trạng phổi: Phổi thông khí bình thường. Vòm hoành: Vòm hoành bình thường. Tim và mạch máu lớn: Tim và mạch máu lớn bình thường. Kết luận: Bệnh nhân bị viêm phổi."</p>
	<p>"Role": "User" – "Content": "Ảnh chụp X-ray PA (Chụp Xquang tim phổi thẳng) bệnh nhân nữ, 46 tuổi. Cho biết bệnh nhân bị gì?" (PA chest X-ray of a 46-year-old female patient. What is the diagnosis?)</p>	<p>GPT-4V Answer: "Tôi không thể chẩn đoán bệnh từ ảnh X-quang này. Tuy nhiên, tôi có thể mô tả các đặc điểm chính. Tình trạng lồng ngực: Lồng ngực có vẻ bình thường, không có dấu hiệu bất thường rõ ràng. Tình trạng phổi: Phổi có vẻ thông khí tốt, không thấy có dấu hiệu của tổn thương hay bất thường rõ ràng. Vòm hoành: Vòm hoành có vẻ bình thường, không có dấu hiệu nâng cao hay bất thường. Tim và mạch máu lớn: Tim có kích thước bình thường, không có dấu hiệu phì đại hay bất thường rõ ràng. Để có chẩn đoán chính xác, cần có sự đánh giá từ bác sĩ chuyên khoa."</p>
		
	<p>"Role": "User" – "Content": "Lồng ngực cân đối không thấy bất thường xương. Phổi hai bên kém sáng mờ kính rải rác rốn phổi tăng đậm nhẹ. Bờ vòm hoành hai bên đều hai góc sườn hoành nhọn. Bóng mờ tim không to các cung tim trong giới hạn sinh lý." (The chest is symmetrical with no visible bone abnormalities. Both lungs appear hypolucent with scattered ground-glass opacities; the pulmonary hila are mildly prominent. The diaphragmatic domes on both sides are smooth, and the costophrenic angles are sharp. The cardiac silhouette is not enlarged, and the heart borders remain within physiological limits.)</p>	<p>Qwen2.5-VL-7B Answer: "Lồng ngực cân đối không thấy bất thường xương. Phổi hai bên kém sáng mờ kính rải rác rốn phổi tăng đậm nhẹ. Bờ vòm hoành hai bên đều hai góc sườn hoành nhọn. Bóng mờ tim không to các cung tim trong giới hạn sinh lý."</p>
<p>"Role": "User" – "Content": "Kết luận bệnh nhân bị gì?" (What is the diagnosis of the patient?)</p> <p>"Role": "User" – "Content": <image></p>	<p>Gemini Answer: "Không có bằng chứng rõ ràng về bệnh lý trên phim X-quang này."</p> <p>GPT-4v Answer: "Dựa trên hình ảnh X-quang này, có thể thấy dấu hiệu của thoát vị hoành bẩm sinh. Đây là tình trạng khi các cơ quan trong ổ bụng di chuyển lên lồng ngực qua lỗ hổng ở cơ hoành. Tuy nhiên, để có chẩn đoán chính xác, cần tham khảo ý kiến của bác sĩ chuyên khoa."</p>	
<p>"Impressions Ground Truth": "Hình ảnh nghi nhiều đến tổn thương phổi kẽ." (The imaging is highly suggestive of interstitial lung disease.)</p>	<p>Qwen2.5-VL-7B Answer: "Hình ảnh nghi nhiều đến tổn thương phổi kẽ."</p>	

Table 18: Illustrative comparison of generation results from Qwen2.5-VL-7B, Gemini, and GPT-4v in Stage 3 - multi-turn generation. where red indicates incorrect information, blue denotes correct details from the ground truth, and orange shows hallucinated content.