Toward a Vision-Language Foundation Model for Medical Data: Multimodal Dataset and Benchmarks for Vietnamese PET/CT Report Generation

Huu Tien Nguyen[1*] **Dac Thai Nguyen**[1*] **The Minh Duc Nguyen**[1]
Trung Thanh Nguyen[2] **Thao Nguyen Truong**[3] **Huy Hieu Pham**[4] **Johan Barthelemy**[5]
Minh Quan Tran[5] **Thanh Tam Nguyen**[6] **Quoc Viet Hung Nguyen**[6]
Quynh Anh Chau[7] **Hong Son Mai**[8] **Thanh Trung Nguyen**[8] **Phi Le Nguyen**[1†]
[1]AI4LIFE, Hanoi University of Science and Technology, Vietnam
[2]Nagoya University, Japan [3]AIST, Japan [4]VinUniversity, Vietnam
[5]NVIDIA, United States [6]Griffith University, Australia
[7]Hanoi Medical University, Vietnam [8]108 Military Central Hospital, Vietnam

Abstract

Vision-Language Foundation Models (VLMs), trained on large-scale multimodal datasets, have driven significant advances in Artificial Intelligence (AI) by enabling rich cross-modal reasoning. Despite their success in general domains, applying these models to medical imaging remains challenging due to the limited availability of diverse imaging modalities and multilingual clinical data. Most existing medical VLMs are trained on a subset of imaging modalities and focus primarily on high-resource languages, thus limiting their generalizability and clinical utility. To address these limitations, we introduce a novel Vietnamese-language multimodal medical dataset consisting of 2,757 whole-body PET/CT volumes from independent patients and their corresponding full-length clinical reports. This dataset is designed to fill two pressing gaps in medical AI development: (1) the lack of PET/CT imaging data in existing VLMs training corpora, which hinders the development of models capable of handling functional imaging tasks; and (2) the underrepresentation of low-resource languages, particularly the Vietnamese language, in medical vision-language research. To the best of our knowledge, this is the first dataset to provide comprehensive PET/CT-report pairs in Vietnamese. We further introduce a training framework to enhance VLMs' learning, including data augmentation and expert-validated test sets. We conduct comprehensive experiments benchmarking state-of-the-art VLMs on downstream tasks, including medical report generation and visual question answering. The experimental results show that incorporating our dataset significantly improves the performance of existing VLMs. However, despite these advancements, the models still underperform on clinically critical criteria, particularly the diagnosis of lung cancer, indicating substantial room for future improvement. We believe this dataset and benchmark will serve as a pivotal step in advancing the development of more robust VLMs for medical imaging, particularly in low-resource languages, and improving their clinical relevance in Vietnamese healthcare.

[*]Equal contribution.
[†]Corresponding author: lenp@soict.hust.edu.vn.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1 Introduction

Vision-Language Foundation Models (VLMs) have rapidly evolved as a cornerstone of modern Artificial Intelligence (AI), capable of jointly modeling information across visual and textual modalities. These models are typically pre-trained on diverse datasets encompassing billions of image-text pairs, enabling them to acquire generalized and transferable representations [1–5]. This cross-modal alignment allows VLMs to bridge the semantic gap between images and language, facilitating downstream tasks such as image captioning [6, 7], visual question answering [4–7], report generation [8, 9], and even zero-shot image classification [1, 10], with minimal task-specific supervision.

While VLMs such as CLIP [1], Flamingo [11], and GPT-4o [12] have demonstrated exceptional performance on natural image benchmarks, transferring this success to the medical domain remains an ongoing challenge. The primary barrier lies in the domain shift: medical images fundamentally differ from natural images in terms of texture, structure, semantics, and purpose [13, 14]. Furthermore, the textual annotations associated with medical images are typically richer, more technical, and context-sensitive, often demanding expert-level knowledge for accurate interpretation [15–17].

In response to these limitations, recent efforts have focused on developing medical-specific VLMs, including MedCLIP [18] and MedFlamingo [19]. These models aim to adapt general-purpose VLMs' architecture to the medical domain by retraining or fine-tuning on domain-specific datasets. However, existing models are still constrained in several critical aspects. *First,* the visual modality coverage in current medical VLMs is narrow. Most existing works focus on well-established imaging types such as chest X-rays [20], Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) scans [21], and histopathology slides [22]. In contrast, functional imaging modalities such as Positron Emission Tomography (PET), which are essential in oncology, cardiology, and neurology, remain significantly underrepresented in current datasets and VLMs. *Second,* the linguistic side of existing VLMs, and vision-language datasets, is largely monolingual, overwhelmingly dominated by English [23]. Very few resources consider linguistic inclusivity, resulting in low-resource languages such as Vietnamese being severely underrepresented in developing and evaluating VLMs. In addition, most existing datasets include only brief image captions [24, 25] or limited diagnostic annotations [26], which are insufficient to fully capture the complex and nuanced information embedded in medical images.

Our preliminary experiments highlight significant limitations of current medical VLMs when applied to PET/CT imaging data. Table 1 presents the results of a Vietnamese clinical report generation task using PET/CT images from a dataset we curated, comprising 1,725 clinical case studies. As shown, models such as LLaVA-Med [27], M3D [28], and RadFM [29] yield near-zero BLEU-4 scores and low ROUGE and BERT score metrics, reflecting their poor capacity to generate coherent and clinically relevant reports. While GPT-4o [12] demonstrates relatively better performance, its BLEU-4 score remains at a modest 31%, indicating inadequate generation quality. These findings underscore the pressing need for more diverse training data regarding imaging modalities and linguistic representation to enhance the robustness and generalizability of medical VLMs.

Table 1: Limitations of current VLMs in Vietnamese reports generating from PET images. R-1 and R-L denote ROUGE-1 and ROUGE-L scores. **GPT-4o is evaluated under few-shot prompting.*

Model	**BLEU-4** ↑	**R-1** ↑	**R-L** ↑	**BERT** ↑
LLaVA-Med [27]	0.01	50.08	27.89	64.63
M3D [28]	0.04	41.01	23.53	67.21
RadFM [29]	0.06	54.23	28.33	69.49
GPT-4o* [12]	31.12	67.96	52.76	81.09

To address this challenge, we introduce the first large-scale paired dataset of PET/CT images and corresponding clinical reports in Vietnamese, a language with limited medical AI resources. Our focus is motivated by these factors:

- **Clinical significance:** PET/CT scan is indispensable in modern diagnostic workflows, especially in oncology, where it enables non-invasive assessments of tumor metabolism and spread [30–33]. Its importance in early diagnosis, staging, and treatment monitoring is unparalleled, yet it remains underutilized in AI due to data scarcity.

- **Data scarcity and accessibility:** Public PET/CT datasets are rare. To our knowledge, no existing dataset offers paired PET/CT images with detailed clinical reports. PET/CT scans are also among the most expensive imaging procedures [34, 35], further limiting data availability and open access.

Table 2: Comparison of our ViMed-PET dataset with existing medical vision-language datasets. Our dataset is the first large-scale PET/CT dataset with clinical reports in Vietnamese, stored in standard medical DICOM format. *"Multiple"* indicates more than one modality (e.g., CT, MRI, X-ray). *"K"* is thousand. *(*)* Values show the number of 2D slices/images extracted from 3D volumes.

Dataset Name	**Text**		**Image**			**Modality Size**	
	PET-related	Type	3D Volume	PET/CT	Others	PET/CT	Others
MIMIC-CXR [26]	✗	Report	✗	✗	X-ray	–	227K
PMC-OA [24]	✗	Caption	✗	✓	Multiple	600K	1,646K
ROCOv2 [25]	✗	Caption	✗	✓	Multiple	432	79K
CT-RATE [37]	✗	Report	✓	✗	CT	–	50K
M3D-Data [28]	✗	Report	✓	✗	CT	–	120K
MedMD-3D [29]	✗	Caption/Report	✓	✗	Multiple	–	500K
RIDER Lung PET-CT [38]	✗	–	✓	✓	–	266K (*)	–
Head-Neck PET-CT [39]	✗	–	✓	✓	–	123K (*)	–
Lung-PET-CT-Dx [40]	✗	–	✓	✓	–	251K (*)	–
FDG-PET-CT-Lesions [41]	✗	–	✓	✓	–	917K (*)	–
Our ViMed-PET Dataset	✓	Report	✓	✓	–	1,567K (*)	–

For our ViMed-PET dataset, $1,567K$ *paired slices correspond to 2,757 paired whole-body PET/CT volumes.*

- **Language equity:** The lack of medical image-report datasets in Vietnamese exacerbates health data inequity. With over 100 million native speakers [36], Vietnamese represents a substantial user base that remains excluded from AI-enabled healthcare technologies.

The main contributions of this study are as follows:

1. We introduce a comprehensive dataset comprising 2,757 whole-body PET/CT volumes from independent patients along with full-length Vietnamese clinical reports. The dataset spans a demographically and pathologically diverse patient population, reflecting real-world clinical variability. It provides a valuable resource for advancing the training of medical VLMs, with the potential to support a broader range of modalities and enable multilingual development, especially for low-resource languages.
2. We develop a data augmentation framework that enriches the visual and textual components of the dataset, improving its effectiveness for model training and generalization.
3. We leverage our newly curated PET/CT image-report dataset, named ViMed-PET, to fine-tune state-of-the-art VLMs and evaluate their performance on tasks such as medical report generation and visual question answering. Experimental results show notable gains, enhancing the capabilities of pre-trained medical VLMs.
4. We collaborate with medical domain experts to develop a clinically validated test set specifically tailored for lung cancer diagnosis. This test set incorporates structured, clinically relevant evaluation metrics that assess model performance in real-world diagnostic scenarios. Rather than relying solely on conventional Natural Language Processing metrics that emphasize lexical matching, our benchmark provides a comprehensive evaluation of the ability of a model to address the nuanced and complex demands of clinical lung cancer diagnosis. This offers a more holistic and meaningful assessment of medical VLMs' effectiveness.

2 ViMed-PET: The proposed Vietnamese Vision-Language Medical Dataset of PET/CT Images and Clinical Reports

2.1 Existing Medical Multimodal Datasets

Recent advances in medical VLMs have been driven by datasets that align medical images with associated textual annotations. Table 2 summarizes existing datasets, which primarily cover CT, MRI, and X-ray modalities and provide either image captions or diagnostic reports. For 2D imaging, representative examples include MIMIC-CXR [26], PMC-OA [24], and ROCOv2 [25]. Several datasets have introduced volumetric data to support 3D understanding, including CT-RATE [37], M3D-Data [28], and MedMD [29]. However, functional imaging modalities such as PET/CT remain largely absent from current benchmarks. Although datasets like RIDER Lung PET-CT [38], Lung-

Table 3: The proposed ViMed-PET dataset. (a) Statistics of the original data. (b) Augmented datasets for training and evaluation various downstream tasks.

(a) Our ViMed-PET dataset (M: Male, F: Female).

Year	Studies (M, F)	Age (years)	Height (cm)	Weight (kg)	# Slices
2017	215 (137, 78)	53.55 ± 15.25	160.63 ± 11.81	55.81 ± 12.02	126,766
2018	462 (308, 154)	56.77 ± 13.56	161.55 ± 8.27	56.72 ± 9.97	270,668
2019	339 (227, 112)	57.35 ± 12.94	161.89 ± 7.89	58.18 ± 9.79	200,660
2023	1741 (1144, 597)	58.69 ± 13.61	161.25 ± 8.68	57.53 ± 10.10	968,968
Total	2757 (1816, 941)	57.81 ± 13.73	161.33 ± 8.81	57.34 ± 10.22	1,567,062

(b) Task-specific augmented dataset.

Subset	Size
VQA	8,271 conversations
Report Generation	5,571 reports
Study Comparison	10,000 pairs
Medical Test Set	398 lesions

PET-CT-Dx [40], Head-Neck-PET-CT [39], and FDG-PET-CT-Lesions [41] provide 3D PET/CT scans, they do not include aligned clinical reports, which limits their use for generative modeling and multimodal reasoning. This gap motivates the need for PET/CT imaging and structured clinical language datasets to enable training and evaluation of VLMs in functional imaging contexts.

To overcome current limitations, we present ViMed-PET, a comprehensive PET/CT image-report dataset composed of two main parts, as summarized in Table 3. The first part consists of the original dataset, including paired PET/CT images and their corresponding clinical reports, collected directly from a hospital (Table 3a). The second part comprises a series of augmented datasets derived from the original data. These augmentations aim to increase the diversity and richness of the dataset while enabling more fine-grained alignment between the visual and textual modalities (Table 3b). The following sections detail the structure of ViMed-PET and the methodology used to construct it.

2.2 Data Description

The proposed ViMed-PET dataset is collected exclusively from a national tertiary general hospital in Vietnam, one of the country's largest medical centers. As a high-volume referral institution receiving patients from across all regions, its data reflect broad clinical diversity and ensure high representativeness and reliability. It consists of 2,757 paired CT-PET volumes (equivalent to 1,567,062 paired CT-PET slices) collected over four years, each accompanied by a corresponding full-length clinical report. Note that the dataset does not contain complete data for all 12 months of each year, as detailed in the statistics shown in Table 3. Each study includes approximately 250–500 paired CT and PET slices, covering the area from the head to the upper thighs (just above the knees). The dataset encompasses various disease cases such as lung cancer, thyroid cancer, and other conditions, representing a broad range of clinical scenarios. The images are stored in the Digital Imaging and Communications in Medicine (DICOM) format, including pixel data and relevant metadata such as patient age, sex, body weight, radiotracer activity, and other acquisition parameters. Acquired using GE Discovery 710 PET/CT and GE Discovery STE PET/CT systems, the images provide high-quality data for analysis. Furthermore, the PET images have undergone attenuation correction using the corresponding CT data to ensure accurate representation. Each medical report, stored in DOCX format, corresponds to a single PET/CT study and contains detailed patient information, clinical status, medical history, scanning methods, and physician observations, making the dataset rich in both imaging and clinical information. All data are obtained under the oversight of an Institutional Ethics Committee (Ethics Approval No. 6184/CN-HDDDBV). Informed consent for anonymized research use is obtained by the hospital, in accordance with institutional policy and national regulations.

2.3 Data Pre-processing

The pre-processing pipeline consists of several key steps to ensure consistency, accuracy, and usability for model training and evaluation. Figure 1 illustrates an example from the ViMed-PET dataset, showing the workflow from raw input to the PET/CT image-report pair.

De-identification. In accordance with privacy regulations, we remove all patient-identifiable information, such as patient name and patient ID, from both the PET/CT images and the associated reports. Additionally, to protect confidentiality, we remove details related to doctors and hospitals, including the institution name, the referring physician's name, the names of the physicians reading the study, and the operator's name.

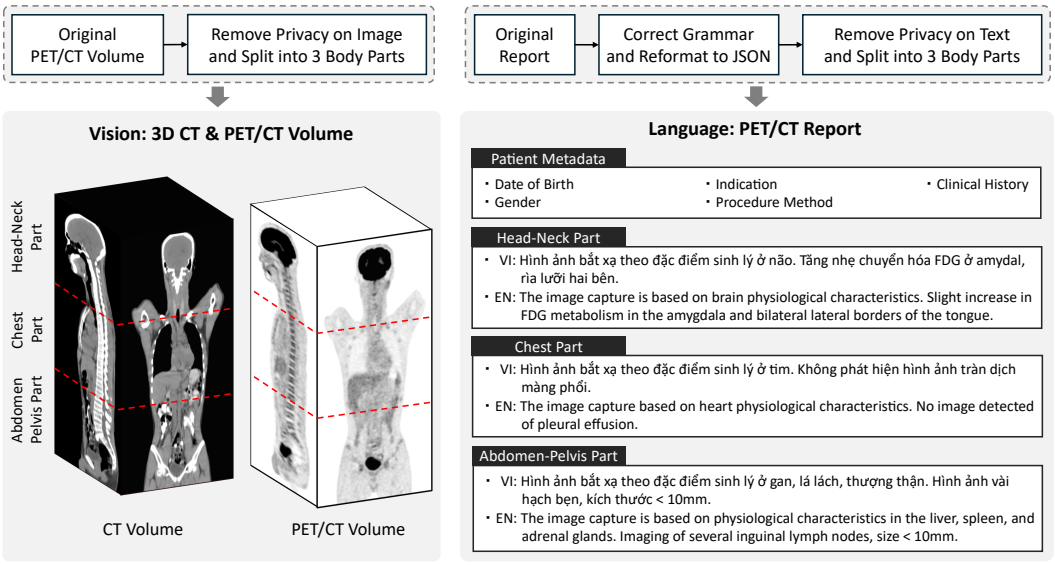


Figure 1: An example from our ViMed-PET dataset. The visual input consists of aligned 3D CT and PET/CT volumes, segmented by anatomical regions: head–neck, chest, and abdomen–pelvis. The corresponding report includes patient metadata and structured descriptions for each region in Vietnamese (VI), with English (EN) as the translation.

Report parsing and standardization. We structure the reports using a predefined template from the hospital, ensuring consistent formatting. A keyword-based search algorithm extracts key information, after which the original DOCX format is converted into a standardized JSON format for easy integration with the corresponding images. After automated extraction, we conduct manual checks to correct grammatical or spelling errors, ensuring the integrity of the text data.

Body part-based data partition. To expand the dataset and improve the data quality used for training and evaluating the model, we divide each study, comprising a PET/CT volume and its corresponding report, into three anatomically distinct regions: head-neck, chest, and abdomen-pelvis. This approach results in a total of 8,271 paired image-report samples. To preserve continuity across adjacent regions and avoid the loss of critical contextual information at segment boundaries, we introduce a 20-slice overlap between neighboring segments. The proportional boundaries of each region are adaptively determined based on the patient's body height, ensuring flexibility across subjects of different statures. The specific ratios for region division follow expert guidelines, with the head–neck typically covering approximately the first 20% of the body length, the chest starting around 15% below the last slice of the head–neck region, and the abdomen–pelvis including the remainder from the chest down to the pelvis. This segmentation strategy increases the number of training samples and significantly improves the alignment between the visual and textual modalities. By localizing image content and clinical descriptions to specific anatomical regions, we enable more precise fine-tuning of models, ultimately enhancing their ability to learn region-specific patterns and improving overall performance.

2.4 Dataset Construction

We construct four specialized subsets to facilitate various stages of model development and evaluation: (1) Visual Question Answering (VQA), (2) Report generation, (3) Study comparison dataset, and (4) Medical test set. The first three subsets are used to fine-tune the 3D vision encoder and the large language model. The final subset is specifically curated to assess the clinical efficacy of the proposed framework in the real-world medical context of a specific task: lung cancer diagnosis.

Visual question answering dataset. The VQA dataset aims to fine-tune VLMs by enabling context-aware, multi-turn dialogue about biomedical images for tasks like diagnostic reasoning and clinical decision-making. The dataset is composed of two parts: single-turn and multi-turn conversations. First, we created a set of 27,855 image-related questions that prompt descriptive answers. These questions were then randomly sampled and paired with each PET/CT image-report pair from the

original dataset to generate single-turn VQA samples. Next, we employ the few-shot prompting strategy to guide the GPT-4o [12] in generating multi-turn conversations based on the clinical reports from the original dataset (detailed in the Appendix B.1). In total, we construct 8,271 multi-turn conversations as part of this dataset.

Report generation dataset. To increase the diversity of the dataset, we augment the textual modality. Specifically, we use GPT-4o [12] to paraphrase the original clinical reports in the initial dataset. For each original report, we generate one corresponding paraphrased version. To ensure the clinical accuracy of the paraphrased content, a subset of the generated reports is randomly reviewed by medical experts. As a result, starting from the original set of 5,571 paired PET/CT image-report samples, we create an augmented dataset of 5,571 additional PET/CT image-report pairs, which we subsequently use for fine-tuning the model.

Study comparison dataset. To further expand the dataset, we introduce a novel augmentation method. In this approach, instead of aligning a single PET/CT image with its corresponding report, each data instance aligns both the similarity and difference between two PET/CT images with the similarity and difference between their corresponding reports. Specifically, we construct a study comparison dataset consisting of tuples in the form: $(X_i^u \| X_i^v), \text{Comp}(X_r^u, X_r^v)$, where $(X_i^u \| X_i^v)$ is the concatenation of two PET images from studies u and v, and $\text{Comp}(X_r^u, X_r^v)$ is a text description that highlights the comparison between their corresponding reports, including similarities and differences. The comparison descriptions $\text{Comp}(X_r^u, X_r^v)$ are generated using GPT-4o [12] (detailed in the Appendix B.2). This dataset contains a total of 10,000 samples.

Medical test set. One of the most challenging aspects of evaluating medical VLMs is quantifying their clinical accuracy. In a generated clinical report, the medical importance and semantic weight of each word vary significantly. For instance, key elements such as lesion type and lesion location carry far more clinical relevance than other details. Therefore, standard NLP evaluation metrics that compare model output with the original report often fail to reflect the true clinical quality of the model's output. To address this challenge, we create a specialized ground-truth dataset that extracts the most clinically significant information from the original reports (refer as *medical-important information*). Such information is then represented in a structured JSON format includes details such as lesion type, lesion location, and key PET/CT metabolism parameters, including SUVmax, FDG metabolism, and the invasiveness of the lesion. The dataset construction follows a two-step process: First, medical experts manually curate a small set of medical-important information from the original reports. Next, we use this curated dataset along with few-shot prompting techniques to guide GPT-4o [12] in automatically extracting medical-important information from the full set of clinical reports. To ensure the reliability of the extracted data, all outputs generated by the model are independently verified by two experienced physicians. As this study is constrained by available resources, this dataset only focuses on lung cancer cases. In total, we construct a dataset of 80 instances, corresponding to 80 lung cancer patients, covering 398 individual lesions.

3 Model Selection and Fine-tuning Flow

In the following, we describe the models used for benchmarking in Section 3.1, and the details of our process to fine-tune these models using ViMed-PET dataset in Section 3.2.

3.1 Model Selection

A typical VLM consists of two main components: a vision encoder and a text encoder. In our framework, we adopt two 3D vision encoders: CT-ViT [42] and Cosmos Tokenizer [43], each selected for their complementary strengths in medical image modeling. CT-ViT is chosen as it is the only publicly available vision transformer pretrained on 3D medical imaging data, specifically 3D CT scans. In contrast, Cosmos Tokenizer is originally pretrained on general-purpose tasks. Although not specifically trained on medical images, its architectural design, optimized for handling sequential inputs, makes it naturally compatible with 3D medical imaging, where volumetric scans can be viewed as ordered slices. We adapt the Cosmos Tokenizer for 3D PET/CT data by removing the causality-based attention mechanisms that are essential for video modeling but irrelevant for spatially coherent medical volumes. For the language component, we utilize Mistral-7B [44] and LLaMA-2-7B [45], language models recommended by the state-of-the-art VLMs as LLaVA-Med [27] and M3D [28], respectively. These models are fine-tuned on biomedical instruction-following datasets,

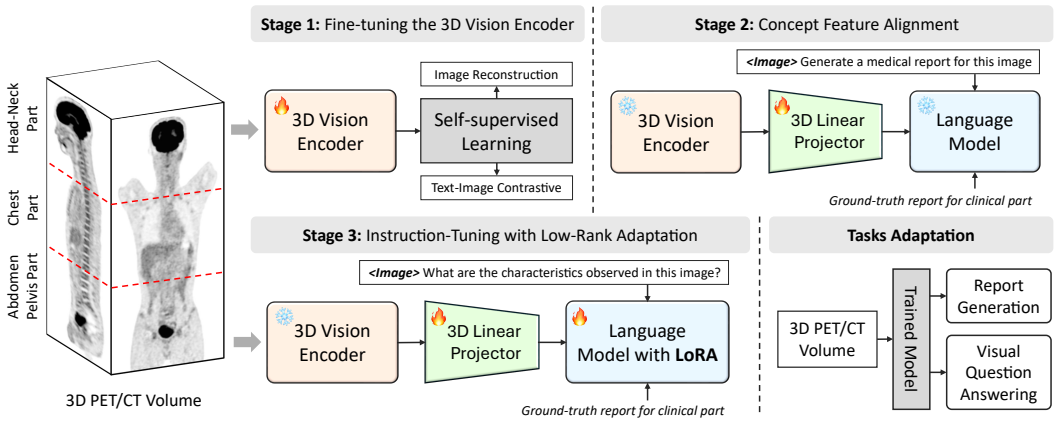


Figure 2: Overview of the fine-tuning pipeline. Stage 1: Fine-tuning the 3D Vision Encoder with PET/CT volumes and associated reports; Stage 2: Aligning concept features between the 3D image and textual embeddings; Stage 3: Instruction tuning of the complete architecture using Low-Rank Adaptation. Finally, the trained model is used for Report Generation and VQA tasks.

enhancing their ability to interpret complex clinical narratives and engage in nuanced, medically relevant dialogue.

3.2 Model Fine-tuning Flow

We employ our curated dataset to fine-tune the baseline models through a structured approach. As shown in Figure 2, our process consists of three stages: (1) Adaptation of the vision encoder to the PET/CT imaging modality, (2) Alignment of visual and textual embedding spaces, and (3) Instruction tuning to optimize the model for downstream multimodal clinical tasks.

Stage 1: 3D vision encoder fine-tuning. For the encoder based on CT-ViT [42], we use a self-supervised learning approach based on text–image contrastive learning, inspired by CLIP [1]. Here, the model is trained to align visual representations of PET images with textual embeddings derived from their corresponding clinical reports. This cross-modal supervision enables the encoder to learn semantically meaningful and clinically relevant visual features. In contrast, for the encoder based on Cosmos Tokenizer [43], we employ an image reconstruction objective. Specifically, a decoder within the Cosmos Tokenizer architecture is used to reconstruct the original PET/CT volume from the latent representations produced by the encoder. The model is trained with a reconstruction loss, which encourages it to capture the underlying structural and semantic features of the 3D medical images.

Stage 2: Concept feature alignment. In this stage, we fine-tune a linear projection layer that maps visual feature representations into the embedding space of the text encoder. We use single-turn image-text pairs from our VQA dataset, where each sample includes a PET/CT image and an instruction-based question. Examples include prompts such as "<image> What are the main findings in this medical image?" or "<image> Please write a detailed medical report for this image.". The model is trained to generate the original textual response based on the given image and prompt. During this process, we freeze the weights of the visual encoder and the language model, allowing updates only to the linear projector. This design ensures efficient and stable alignment between visual and textual embeddings while minimizing overfitting and preserving pre-trained representations.

Stage 3: Instruction-tuning with Low-Rank Adaptation. In the final stage, we fine-tune the VLM using our VQA dataset, which includes single-turn question–answer pairs and multi-turn conversational interactions. During this phase, the parameters of the image encoder are kept frozen to preserve previously learned visual representations. We update only the parameters of the visual–text alignment projector and the language model, applying Low-Rank Adaptation (LoRA) [46] for efficient fine-tuning. This stage strengthens the capacity of the model to interpret and respond to a broad spectrum of medical queries by integrating visual and textual modalities. The inclusion of simple and complex dialogue formats further improves its robustness in biomedical VQA tasks.

3.3 Training and Evaluation

We integrate two vision encoders (CT-ViT and a customized Cosmos Tokenizer) with two text encoders (Mistral-7B and LLaMA-2-7B) to construct four VLMs. We evaluate the performance of these models on two tasks: medical report generation and VQA. The dataset is divided into three subsets: a training set with 5,571 image-report pairs, a validation set with 975 pairs, and a test set with 1,725 pairs. Each subset contains samples from all four years of our collection to reduce temporal bias. To assess the impact of data augmentation, we fine-tune our framework under three configurations: (1) O: using only the Original dataset, (2) O-G: using the Original dataset and the report Generation dataset, (3) O-G-C: using the Original dataset, the report Generation dataset, and the study Comparison dataset. In addition, we benchmark our models against baselines, including LLaVA-Med [27], M3D [28], RadFM [29], and GPT-4o [12]. Further details regarding model architectures, training configurations, and optimization settings are provided in Appendix B.3.

3.4 Evaluation Metrics

Common natural language processing metrics. We evaluate model performance using standard Natural Language Processing (NLP) metrics, including BLEU-4 [47], ROUGE [48], and BERTScore [49]. BLEU-4 measures 4-gram precision in generated text, ROUGE-1 captures unigram recall for summarization tasks, and ROUGE-L evaluates fluency based on the longest common subsequence. BERTScore assesses semantic similarity between generated and reference texts by leveraging contextual embeddings from a pre-trained BERT model.

Proposed clinical metrics. Conventional NLP metrics are insufficient for evaluating the clinical accuracy of generated text, particularly in terms of medical relevance. For example, in auto-generated clinical reports, details such as tumor location or FDG uptake carry significantly more diagnostic weight than general descriptions. To address this limitation, we introduce targeted evaluation metrics that focus specifically on clinically meaningful content. We use our expert-curated Medical Test Set (see Section 2.4) to evaluate model performance on the report generation task, emphasizing the extraction of key attributes including lesion Type, lesion Position, and FDG uptake. Specifically, for each generated report, we use GPT-4o [12] to extract structured content aligned with these three clinical attributes. The extracted information is converted into categorical variables (detailed in Appendix B.4) and verified by medical professionals. We then compute F1-scores by comparing these model-generated outputs with the ground truth labels in our test set. We report four clinical evaluation metrics: F1-T, based on lesion Type; F1-TP, based on Type and Position; F1-TF, incorporating Type and FDG uptake; and F1-TPF, which evaluates all three aspects of Type, Position, and FDG uptake.

4 Evaluation Results

This section presents the experimental results assessing the performance of the VLMs fine-tuned on our ViMed-PET dataset, with respect to two key tasks: clinical report generation (Table 4) and medical VQA (Table 5). Additional benchmarks and evaluation results are provided in Appendix C.

4.1 PET/CT Report Generation Task

Comparison with existing baselines. Our results first demonstrate that fine-tuning VLMs on our proposed ViMed-PET dataset leads to substantial performance improvements across both standard NLP metrics and clinically specific evaluation metrics. For instance, when LLaMA-2-7B is paired with either CT-ViT or our customized Cosmos Tokenizer and fine-tuned on ViMed-PET, it significantly outperforms the pretrained LLaMA-2-7B model used in M3D across all key metrics (i.e., BLEU-4, ROUGE-1, ROUGE-L, and BERT score). Notably, the BLEU-4 score improves significantly over baseline medical LLMs following fine-tuning, reflecting a dramatic enhancement in generation quality. Compared to GPT-4o, which is evaluated under few-shot prompting, models fine-tuned with ViMed-PET also yield substantial performance gains. Specifically, BLEU-4, ROUGE-1, ROUGE-L, and BERT score increase by up to 89.17%, 17.88%, 40.09%, and 11.67%, respectively. Furthermore, clinical metrics, including F1-T, F1-TP, F1-TF and F1-TPF, also exhibit notable improvements, increasing by more than 1.8 times, underscoring the clinical relevance and robustness of the generated reports. These findings highlight the effectiveness of fine-tuning with ViMed-PET in enhancing the performance of VLMs for clinical report generation.

Table 4: Performance on report generation task. We define training configurations as: **O**-**O**riginal dataset, **G**-Report **G**enerate dataset, **C**-Study **C**omparison dataset. R-1 and R-L denote ROUGE-1 and ROUGE-L scores. ↑ means higher values are better. The best and second-best results are emphasized using **bold** and underline, respectively. **GPT-4o is evaluated under few-shot prompting.*

	Model		Settings			NLP Metrics ↑				Clinical F1-Score (%) ↑			
	Vision	Language	O	G	C	BLEU-4	R-1	R-L	BERT	F1-T	F1-TP	F1-TF	F1-TPF
Baseline	LLaVA-Med [27]			–		0.01	50.08	27.89	64.63	–	–	–	–
	M3D [28]			–		0.04	41.01	23.53	67.21	–	–	–	–
	RadFM [29]			–		0.06	54.23	28.33	69.49	–	–	–	–
	GPT-4o* [12]			–		31.12	67.96	52.76	81.09	24.21	13.62	20.57	7.87
Fine-tuned	CT-ViT	Mistral-7B	✓			53.30	77.79	68.60	88.35	43.49	24.97	29.80	18.26
			✓	✓		58.07	**80.11**	72.74	89.98	51.11	30.66	**37.02**	**22.65**
			✓	✓	✓	58.05	80.08	72.70	89.92	**51.96**	30.17	35.47	21.23
		LLaMA-2-7B	✓			57.91	79.87	72.89	89.88	48.28	29.44	33.42	20.42
			✓	✓		56.15	79.09	71.39	89.20	46.24	24.15	32.40	17.38
			✓	✓	✓	56.05	78.95	71.43	89.12	47.97	29.17	33.66	20.76
	Cosmos Tokenizer	Mistral-7B	✓			53.66	78.04	69.93	88.69	46.29	30.00	31.14	19.71
			✓	✓		55.80	78.93	71.68	89.43	46.51	25.68	31.10	17.40
			✓	✓	✓	**58.87**	80.10	**73.91**	**90.55**	47.93	22.78	32.84	15.68
		LLaMA-2-7B	✓			57.59	79.26	**73.91**	90.35	45.05	18.21	32.27	14.06
			✓	✓		57.98	79.83	73.49	90.04	45.59	27.13	30.21	17.62
			✓	✓	✓	57.14	79.37	73.13	90.01	49.73	**31.42**	33.61	22.13

Comparison between LLMs. Comparing the performance of the two LLMs using common NLP metrics, we observe that when fine-tuning is performed solely on the original dataset (setting O), LLaMA2-7B outperforms Mistral-7B. This can be attributed to the fact that the M3D backbone used in LLaMA2-7B has been pretrained on 3D medical imaging data, allowing it to better capture spatial features and align 3D PET/CT representations with textual descriptions. In contrast, Mistral-7B is pretrained on 2D image–text data, which limits its capacity to model 3D spatial context in low-data settings. However, when the training dataset is expanded to include both the original and augmented data (settings O-G and O-G-C), Mistral-7B demonstrates superior performance, likely due to its more efficient architecture compared to LLaMA2-13B. This finding is consistent with results reported in the Mistral-7B paper [44], which highlights the robustness of the model and scalability in large-scale learning settings. For clinical evaluation metrics, Mistral-7B outperforms LLaMA2-7B when paired with the CT-ViT encoder. However, when combined with the Cosmos Tokenizer, LLaMA2-7B shows a slight advantage over Mistral-7B.

Comparison between vision encoders. The results show that CT-ViT and Cosmos Tokenizer achieve comparable performance on standard NLP metrics across various settings, including integration with different LLMs and the use of augmented data. However, CT-ViT consistently outperforms Cosmos Tokenizer on clinical metrics across all four evaluation criteria and training configurations. This indicates that CT-ViT, which is specifically designed and pretrained on 3D medical imaging data, is more effective in improving the clinical accuracy of VLMs than Cosmos Tokenizer, which is pretrained on general-purpose tasks.

4.2 PET/CT VQA Task

Table 5 shows the results on the VQA task under the O-G-C training setting. Fine-tuning with our ViMed-PET dataset significantly improves performance across all evaluation metrics compared to the baseline GPT-4o model. Specifically, the best-performing fine-tuned model, CT-ViT paired with Mistral-7B, achieves substantial gains, outperforming GPT-4o by factors of $10.3\times$, $1.3\times$, $1.7\times$, and $1.1\times$ on BLEU-4, ROUGE-1, ROUGE-L, and BERT score, respectively. When comparing different combinations of vision and language encoders, the results on the VQA task follow a pattern similar to those in the report generation task. The combination of CT-ViT and Mistral-7B consistently delivers the highest performance, followed by Cosmos Tokenizer paired with LLaMA-2-7B. A comprehensive analysis of the VQA task is provided in the Appendix C.

Table 5: Performance on the VQA task under the O-G-C training setting. R-1 and R-L denote ROUGE-1 and ROUGE-L scores. ↑ means higher values are better. The best and second-best results are emphasized using **bold** and underline, respectively. **GPT-4o is evaluated under few-shot prompting.*

Model		NLP Metrics ↑			
Vision	Language	BLEU-4	R-1	R-L	BERT
GPT-4o* [12]		3.01	49.35	30.09	71.92
CT-ViT	Mistral-7B	<u>31.14</u>	**65.61**	**51.22**	**82.50**
	LLaMA-2-7B	**31.36**	59.14	48.00	76.72
Cosmos Tokenizer	Mistral-7B	28.09	62.92	48.37	79.25
	LLaMA-2-7B	28.40	<u>63.29</u>	<u>48.76</u>	<u>79.35</u>

5 Conclusion

In this study, we introduced ViMed-PET, a high-quality dataset comprising 2,757 paired whole-body PET/CT volumes and 2,757 Vietnamese clinical reports, covering a wide range of patient cases. We also developed a clinically validated lung cancer test set to support meaningful evaluation beyond conventional NLP metrics. Additionally, we proposed a data augmentation strategy that enhances both visual and textual inputs to improve the fine-tuning of medical vision-language models. Models fine-tuned on ViMed-PET demonstrated substantial gains, achieving improvements in both standard NLP metrics and clinical evaluation scores compared to pretrained baselines.

Limitations and Societal Impacts. We acknowledge that clinical reports often follow standardized formats, which can limit output diversity. Clinical results with F1 scores around 50% highlight the challenges in modeling PET/CT data and performing complex medical reasoning. Although the proposed ViMed-PET dataset contains PET and CT volumes, our benchmark focuses exclusively on PET/CT imaging. This decision is based on the observation that report content is primarily driven by PET information, with minimal reference to CT anatomical details. In future work, we plan to extend our approach to better incorporate CT data and further improve the accuracy of VLMs. This study provides a foundation for such enhancements. We believe ViMed-PET serves as a valuable resource for advancing medical vision-language modeling in the low-resource Vietnamese language and the underexplored PET/CT domain, supporting more equitable AI development in healthcare.

Acknowledgments and Disclosure of Funding

This research is funded by Hanoi University of Science and Technology (HUST) under grant number T2024-TĐ-002. We would like to thank the NVIDIA Academic Grant Program 2025 for providing the computing resources used in this research.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

[2] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of LMMs: Preliminary explorations with GPT-4V(ision). *arXiv preprint arXiv:2309.17421*, pages 1–166, 2023.

[3] Anthropic. Claude: An AI Assistant by Anthropic. `https://www.anthropic.com/index/claude`, 2024. Accessed: 2025-05-11.

[4] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[6] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[7] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023.

[8] Sixing Yan, William K Cheung, Ivor W Tsang, Keith Chiu, Terence M Tong, Ka Chun Cheung, and Simon See. Ahive: Anatomy-aware hierarchical vision encoding for interactive radiology report retrieval. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14324–14333, 2024.

[9] Trong Thang Pham, Ngoc-Vuong Ho, Nhat-Tan Bui, Thinh Phan, Patel Brijesh, Donald Adjeroh, Gianfranco Doretto, Anh Nguyen, Carol C Wu, Hien Nguyen, and Le Ngang. Fg-cxr: A radiologist-aligned gaze dataset for enhancing interpretability in chest x-ray report generation. In *Proceedings of the 2024 Asian Conference on Computer Vision*, pages 941–958, 2024.

[10] Sajid Javed, Arif Mahmood, Iyyakutti Iyappan Ganapathi, Fayaz Ali Dharejo, Naoufel Werghi, and Mohammed Bennamoun. Cplip: zero-shot learning for histopathology with comprehensive vision-language alignment. In *Proceedings of the IEEE/CVF 2024 Conference on Computer Vision and Pattern Recognition*, pages 11450–11459, 2024.

[11] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Arthur Mensch, Katie Milln, Matthew Reynolds, Rebecca Ring, Matthew Tancik, Xiuye Zhai, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

[12] OpenAI. Gpt-4o: Openai's multimodal model with vision, audio, and text capabilities. `https://openai.com/index/gpt-4o`, 2024. Accessed: 2025-04-30.

[13] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

[14] Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics*, 26(12):6070–6080, 2022.

[15] Manar Aljabri, Manal AlAmir, Manal AlGhamdi, Mohamed Abdel-Mottaleb, and Fernando Collado-Mesa. Towards a better understanding of annotation tools for medical imaging: a survey. *Multimedia Tools and Applications*, 81(18):25877–25911, 2022.

[16] Alexander Davis, Rafael Souza, and Jia-Hao Lim. Knowledge-augmented language models interpreting structured chest x-ray findings. *arXiv preprint arXiv:2505.01711*, 2025.

[17] Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical Image Analysis*, 80: 102510, 2022.

[18] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing.*, volume 2022, page 3876, 2022.

[19] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health*, pages 353–367. PMLR, 2023.

[20] Pierre Chambon, Christian Bluethgen, Jean-Benoit Delbrouck, Rogier Van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P Langlotz, and Akshay Chaudhari. Roentgen: vision-language foundation model for chest x-ray generation. *arXiv preprint arXiv:2211.12737*, 2022.

[21] Lijian Xu, Hao Sun, Ziyu Ni, Hongsheng Li, and Shaoting Zhang. Medvilam: A multi-modal large language model with advanced generalizability and explainability for medical data understanding and generation. *arXiv preprint arXiv:2409.19684*, 2024.

[22] Yu Xin, Gorkem Can Ates, Kuang Gong, and Wei Shao. Med3dvlm: An efficient vision-language model for 3d medical image analysis. *arXiv preprint arXiv:2503.20047*, 2025.

[23] Thao Nguyen, Matthew Wallingford, Sebastin Santy, Wei-Chiu Ma, Sewoong Oh, Ludwig Schmidt, Pang Wei W Koh, and Ranjay Krishna. Multilingual diversity improves vision-language representations. *Advances in Neural Information Processing Systems*, 37:91430–91459, 2024.

[24] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer, 2023.

[25] Johannes Rückert, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia S Schmidt, Sven Koitka, Obioma Pelka, Asma Ben Abacha, Alba G. Seco de Herrera, et al. Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. *Scientific Data*, 11(1):688, 2024.

[26] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

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

[28] Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

[29] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data, 2023.

[30] Michelle M Kim, Abhijit Parolia, Mark P Dunphy, and Sriram Venneti. Non-invasive metabolic imaging of brain tumours in the era of precision medicine. *Nature Reviews Clinical Oncology*, 13(12):725–739, 2016.

[31] David Y Lewis, Dmitry Soloviev, and Kevin M Brindle. Imaging tumor metabolism using positron emission tomography. *The Cancer Journal*, 21(2):129–136, 2015.

[32] Sanjiv Sam Gambhir. Molecular imaging of cancer with positron emission tomography. *Nature Reviews Cancer*, 2(9):683–693, 2002.

[33] Johannes Schwenck, Dominik Sonanini, Jonathan M Cotton, Hans-Georg Rammensee, Christian la Fougère, Lars Zender, and Bernd J Pichler. Advances in pet imaging of cancer. *Nature Reviews Cancer*, 23(7):474–490, 2023.

[34] Ian Alberts, Stuart More, Karen Knapp, Riccardo Mei, Stefano Fanti, Clemens Mingels, Lorenzo Nardo, Nii Boye Hammond, Harish Nagaraj, Axel Rominger, et al. Is long–axial-field-of-view pet/ct cost-effective? an international health–economic analysis. *Journal of Nuclear Medicine*, 2025.

[35] Mohammad Naghavi-Behzad, Oke Gerke, Annette Raskov Kodahl, Marianne Vogsen, Jon Thor Asmussen, Wolfgang Weber, Malene Grubbe Hildebrandt, and Kristian Kidholm. Cost-effectiveness of 2-[18f] fdg-pet/ct versus ce-ct for response monitoring in patients with metastatic breast cancer: a register-based comparative study. *Scientific Reports*, 13(1):16315, 2023.

[36] World Population Review. Explore the world population through data (2025). https://worldpopulationreview.com/, 2025. Accessed: 2025-05-12.

[37] Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. CT2REP: Automated radiology report generation for 3d medical imaging. In *Proceedings of the 2024 International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 476–486. Springer, 2024.

[38] Peter Muzi, Matthew Wanner, and Paul Kinahan. Rider lung pet-ct: Data for quantitative imaging biomarker evaluation. *The Cancer Imaging Archive*, 2015.

[39] Martin Vallières, Emily Kay-Rivest, Léo Jean Perrin, Xavier Liem, Christophe Furstoss, Nader Khaouam, Phuc Félix Nguyen-Tan, Chang-Shu Wang, and Khalil Sultanem. Data from head-neck-pet-ct. *The Cancer Imaging Archive*, 2017.

[40] P. Li, S. Wang, T. Li, J. Lu, Y. HuangFu, and D. Wang. A large-scale ct and pet/ct dataset for lung cancer diagnosis (lung-pet-ct-dx). *The Cancer Imaging Archive*, 2020.

[41] S. Gatidis and T. Kuestner. A whole-body FDG-PET/CT dataset with manually annotated tumor lesions (FDG-PET-CT-Lesions). *The Cancer Imaging Archive*, 2022.

[42] Ibrahim Ethem Hamamci, Sezgin Er, Anjany Sekuboyina, Enis Simsar, Alperen Tezcan, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Furkan Almas, Irem Doğan, Muhammed Furkan Dasdelen, et al. Generatect: Text-conditional generation of 3d chest ct volumes. In *European Conference on Computer Vision*, pages 126–143. Springer, 2024.

[43] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[44] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[45] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[46] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 1(2):3, 2022.

[47] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[48] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, 2004.

[49] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[50] Jialong Zuo, Jiahao Hong, Feng Zhang, Changqian Yu, Hanyu Zhou, Changxin Gao, Nong Sang, and Jingdong Wang. Plip: Language-image pre-training for person representation learning. *Advances in Neural Information Processing Systems*, 37:45666–45702, 2024.

[51] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

[52] Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Cesar Truyts, et al. Merlin: A vision language foundation model for 3d computed tomography. *Research Square*, pages rs–3, 2024.

[53] Omkar Chakradhar Thawakar, Abdelrahman M Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Khan. Xraygpt: Chest radiographs summarization using large medical vision-language models. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 440–448, 2024.

[54] Shawn Xu, Lin Yang, Christopher Kelly, Marcin Sieniek, Timo Kohlberger, Martin Ma, Wei-Hung Weng, Atilla Kiraly, Sahar Kazemzadeh, Zakkai Melamed, et al. Elixr: Towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders. *arXiv preprint arXiv:2308.01317*, 2023.

[55] Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024.

[56] Junling Liu, Ziming Wang, Qichen Ye, Dading Chong, Peilin Zhou, and Yining Hua. Qilin-med-vl: Towards chinese large vision-language model for general healthcare. *arXiv preprint arXiv:2310.17956*, 2023.

[57] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*, 2023.

[58] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[59] Dat Quoc Nguyen and Anh Tuan Nguyen. Phobert: Pre-trained language models for vietnamese. *arXiv preprint arXiv:2003.00744*, 2020.

[60] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[61] Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. Open-vocabulary action localization with iterative visual prompting. *IEEE Access*, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have already outlined this in the Abstract and Introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of this study are presented in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provided detailed descriptions of the training data, network architecture, and experimental details to ensure the reproducibility of the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the public URL to our dataset, and the source code is submitted along with this paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We describe the data splitting in the main text, while the hyperparameters used for the experiments are detailed in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report the error bar due to the high cost for training LLMs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Each experiment was conducted on a machine with four A100 GPUs, with training and evaluation completed within three days (see Appendix for details).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This study complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: This study aims to benchmark existing LLMs on a medical dataset and does not present any negative societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We indicate licenses of public datasets in Appendix.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide the documentation along with the code.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: We remove the private information from the dataset.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Our paper uses GPT-4o to generate augmented datasets based on a few-shot examples created by domain experts. We also leverage GPT-4o to post-process model outputs for evaluating clinical metrics. In addition, we utilize pretrained Mistral and LLaMA-2 models, combined with vision encoders, to conduct experiments on our dataset.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Medical Vision-Language Models

Recent advances in Vision-Language Models (VLMs) have opened new possibilities for medical image analysis by enabling multimodal reasoning across visual and textual inputs. In the medical domain, VLMs are broadly categorized into two types: CLIP-based and Large-Language-Model-based (LLM-based) models. CLIP-based models, such as MedCLIP [18], PLIP [50], and BiomedCLIP [51], leverage contrastive learning to align images with textual descriptions, performing well on classification and retrieval tasks. However, their lack of generative capability limits their use in applications such as report generation. In contrast, LLM-based models, including M3D-LaMed [28], CT2Rep [37], Merlin [52], and RadFM [29], combine image encoders with language models to support complex reasoning and text generation.

Despite recent advances, most existing VLMs are trained primarily on 2D medical images (e.g., X-rays, Dermatology, Pathology), with the models such as XrayGPT [53], ELIXR [54], and CheX-agent [55]. This focus limits their capacity to process 3D imaging modalities like PET/CT, which require spatial and intensity-aware reasoning across volumetric data. In addition, most VLMs are developed for English, with limited support for other languages due to a lack of multilingual annotated datasets. Recent models like M3D-LaMed [28] and RadFM [29] introduce architectures capable of handling 3D inputs, improving performance across imaging modalities. For multilingual contexts, Qilin-Med-VL [56] and HuatuoGPT-Vision [57] show potential in Chinese and bilingual applications. However, these VLMs perform poorly on PET/CT imaging, often confusing it with MRI or SPECT and failing to produce accurate, medically grounded outputs. However, these efforts have yet to address the needs of low-resource languages such as Vietnamese, where both medical imaging and language data remain scarce.

B Technical Appendices

B.1 Visual Question Answering Dataset

```
messages = [{"role": "system", "content": "You are a medical assistant and are being provided with information related to a medical image. This information comes in the form of a short clinical report, which includes the location of the image and some preliminary diagnostic findings. Based on this, you are expected to answer the given questions as if you are directly viewing the image. You should generate a dialogue between yourself, acting as a medical assistant, and a patient, focusing on the content of the image. Both the questions and answers in the dialog must reflect the assumption that you are visually inspecting the image. The questions must be diverse, and your answers must be based solely on the available information. Also the questions should cover various aspects of the image content, including the anatomical location where the image was taken, possible diagnoses in image, size or characteristics of any lesions and other observable clinical features. Only ask questions that can be answered with certainty, based on either: the visual information directly present in the image or clearly inferable information that obviously evident, even if not explicitly visible in the image. Do not include questions that can not be answered with certainty. The dialogue may include complex questions but them must be grounded in clearly evident and justifiable information.That is, the complexity of the question is acceptable only if the answer can still be reasoned with confidence based on what is explicitly or obviously present. When answering complex questions, provide detailed, well-reasoned responses. The answers should refer credible clinical sources of appropriate, clearly explain your logical reasoning. Be especially careful to avoid asking or answering anything based on ambiguous, assumed or unverifiable details. Below is an example for you to follow."}]

for sample in fewshot_samples:
    messages.append({"role": "user", "content": sample['context']})
    messages.append({"role": "assistant", "content": sample['response']})
messages.append({"role": "user", "content": query})
```

Figure 3: Message used to prompt GPT-4o to generate our medical VQA conversations. Manually curated few-shot examples are included in the prompt, where each example consists of an input `sample['context']` and an output `sample['response']`. See Figure 4 for a sample few-shot example.

To construct the Visual Question Answering (VQA) dataset, we follow the methodology introduced by LLaVA [4], adopting two types of response formats: **detailed description** and **conversation**. For the detailed description format, we directly use the original clinical report as the textual response corresponding to the input image. For the conversational format, we employ instruction-based prompting in conjunction with few-shot prompting. Specifically, GPT-4o [12] is guided to generate coherent question–answer pairs by providing clear instructions that constrain the generated content to align with the given report. Additionally, we include several few-shot examples to illustrate the

Example Input:
This image captures the patient's thoracic (chest) region. Physiological FDG uptake is observed in the heart, consistent with normal metabolic activity. There is no evidence of pleural effusion on either side and no pericardial effusion detected. A few small mediastinal lymph nodes measuring approximately 10 mm are noted in the pretracheal, subaortic, and subcarinal regions. These nodes show no increased FDG uptake. There is a linear opacity in the right middle lobe and a ground-glass opacity in the right lower lobe, both without increased FDG uptake, suggesting a benign nature. A few axillary lymph nodes measuring 10 mm are observed bilaterally, also without increased FDG uptake.

Example Output:

Question	Which part of the body does this image capture?
Answer	This image captures the chest (thoracic) region of the patient.
Question	Are there any abnormal lymph nodes detected in the chest?
Answer	Yes, there are a few small mediastinal lymph nodes measuring approximately 10 mm, located in the pretracheal, subaortic, and subcarinal regions. However, these nodes show no increased FDG uptake, suggesting they are likely benign.
Question	Are there any abnormalities in the pleura or pericardium?
Answer	No, the image does not show any pleural effusion on either side, nor is there any pericardial effusion detected.
Question	Are there any parenchymal lung lesions?
Answer	Yes, there is a linear opacity in the right middle lobe and a ground-glass opacity in the right lower lobe. However, both findings show no increased FDG uptake, suggesting these are likely benign lesions.
Question	Are there any abnormal lymph nodes detected in the axillary region?
Answer	Yes, a few axillary lymph nodes approximately 10 mm in size are seen in both axillae, but they do not exhibit increased FDG uptake, indicating no signs of malignancy.

Figure 4: Few-shot examples included in our prompt to construct the VQA conversation dataset.

desired structure and tone, enhancing the ability of the model to produce contextually appropriate outputs. The full prompting details are shown in Figures 3 and 4. Note that while the figures are translated into English for illustration purposes, the actual prompts and data are in Vietnamese.

B.2 Study Comparison Dataset

```
messages = [{"role": "system", "content": "You are a medical expert, and your current task is to compare two CT scan images of a patient. Please indicate the similarities and differences between the two images. Make sure not to add or omit any information from the two images. I will provide you with two reports, please treat them as if you are looking directly at the two images. Also, please do not make any inference to get the conclusion about the status of patient. Below are example comparisons for you to follow: "}]

for sample in fewshot_samples:
        messages.append({"role": "user", "content": sample['context']})
        messages.append({"role": "assistant", "content": sample['response']})
messages.append({"role": "user", "content": query})
```

Figure 5: System message used to prompt GPT-4o for generating the study comparison dataset. The prompt includes few-shot examples, where each example consists of an input `sample['content']` (a pair of reports to compare) and an output `sample['response']` that describes the similarities and differences between the two reports. See Figure 6 for a sample few-shot example.

To construct the study comparison dataset, we adopt a few-shot prompting approach using GPT-4o [12]. Few-shot examples are created by randomly sampling three pairs of medical reports, each annotated by domain experts to highlight similarities and differences between the reports. To ensure meaningful comparisons, all report pairs are selected from the same anatomical region. The full prompting setup is illustrated in Figures 5 and 6. Note that while the figures are translated into English for illustration purposes, the actual prompts and data are in Vietnamese.

Example Input:

Report 1	FDG uptake observed according to physiological characteristics in the brain. Mild FDG uptake at the amygdala and both lateral edges of the tongue, consistent with physiological uptake. A few lymph nodes at the mandibular angle and medial to the sternocleidomastoid muscle (group IIB) on both sides, small in size (10 mm), no increased FDG uptake, likely due to inflammation or benign hyperplasia.
Report 2	FDG uptake observed according to physiological characteristics in the brain. Mild FDG uptake at the amygdala and both lateral edges of the tongue, consistent with physiological uptake. Mucosal thickening in the right maxillary sinus, no increased FDG uptake, likely due to inflammation. No abnormal focal FDG uptake in the thyroid gland. A few lymph nodes at the mandibular angle and medial to the sternocleidomastoid muscle (group IIB) on both sides, small in size (10 mm), no increased FDG uptake, likely due to inflammation or benign hyperplasia.

Example Output:

The similarities between the two images are that both show a few small cervical lymph nodes on both sides, specifically at the mandibular angle and medial to the sternocleidomastoid muscle (group IIB). These nodes are around 10 mm in size and do not exhibit increased FDG uptake, suggesting benign reactive changes or inflammation. Both images also show mild FDG uptake at the amygdala and the lateral edges of the tongue, as well as physiological brain metabolism, with no indication of malignancy. The difference lies in the fact that the second image additionally reveals mucosal thickening in the right maxillary sinus without increased FDG uptake, suggesting non-active mucosal inflammation. In contrast, the first image does not report any sinus abnormalities, indicating no signs of sinusitis at the time of scanning.

Figure 6: Few-shot example used in our prompt for generating the study comparison dataset. The yellow highlights indicate the differences between the two reports.

B.3 Training and Model Configurations

B.3.1 Fine-tuning Vision Encoders

We select CT-ViT [42] and Cosmos Tokenizer [43] as the vision encoders for our VLMs, as they are well-suited for processing 3D volumetric inputs with depths of up to 200 slices and have been pre-trained on large-scale datasets in prior work. Details on model selection are discussed in Section 3.1.

CT-ViT. We employ a specialized ViT model, CT-ViT [42], as the vision encoder in our VLMs. CT-ViT is designed to effectively process 3D chest CT volumes and is pre-trained on a large-scale medical dataset comprising 25,701 non-contrast 3D chest CT volumes from 21,314 unique patients. These volumes vary in resolution and contain between 100 and 600 axial slices. To align visual and textual modalities, we adopt a CLIP-based [1] training approach. The model is fine-tuned for up to 30 epochs using the AdamW optimizer [58], with a learning rate of 1.25×10^{-6} and a batch size of 8 per GPU across four NVIDIA A100 GPUs (80 GB each). Early stopping is applied based on the convergence of training loss, ensuring efficient optimization. For the text modality, we integrate PhoBERT [59], a state-of-the-art Vietnamese language model pre-trained on a large-scale Vietnamese corpus. PhoBERT has demonstrated superior performance over multilingual models such as XLM-R across several Vietnamese natural language processing (NLP) tasks, including part-of-speech tagging, dependency parsing, named entity recognition, and natural language inference. Its linguistic compatibility with clinical texts in our dataset enables effective semantic representation and understanding.

Cosmos Tokenizer. We leverage the architecture of the Cosmos Tokenizer [43], originally designed for image and video reconstruction tasks. To adapt it for 3D PET/CT imaging, we remove causality-based attention mechanisms, which are essential for modeling temporal dependencies in video but unnecessary for spatially coherent volumetric medical scans. This modification allows us to retain the benefits of pre-trained weights while enabling effective processing of 3D medical data. We fine-tune the customized Cosmos Tokenizer using a single-phase reconstruction approach. The total loss function $\mathcal{L}_{\text{total}}$ combines two terms: an L_1 reconstruction loss $\mathcal{L}_1$ and an inverted Structural Similarity Index Measure (SSIM) loss $\mathcal{L}_{\text{iSSIM}}$, defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_1 + \lambda \mathcal{L}_{\text{iSSIM}} = \|\hat{x}_{0:T} - x_{0:T}\|_1 + \lambda\big(1 - \text{SSIM}(\hat{x}_{0:T}, x_{0:T})\big) \tag{1}$$

where $\hat{x}_{0:T}$ is the reconstructed volume, $x_{0:T}$ is the ground-truth volume, and λ is the trade-off coefficient (set to 1×10^{-2} across all experiments). Training is performed for up to 20 epochs using a Cosine Annealing Scheduler [60], with an initial learning rate of 1×10^{-4} and a batch size of 8 across four NVIDIA A100 GPUs (80 GB each). Since the Cosmos Tokenizer requires a fixed number of input frames, we standardize all PET/CT volumes to 120 slices. This value is chosen based on the distribution in our dataset, and we apply zero-padding or linear interpolation to achieve this fixed size.

Table 6: Performance comparison of causal attention settings on ViMed-PET across two tasks: Report Generation and Visual Question Answering (VQA). R-1 and R-L denote ROUGE-1 and ROUGE-L scores. ↑ indicates higher is better.

Model		Causal Attention	Report Generation				VQA			
Vision	Language		BLEU-4↑	R-1↑	R-L↑	BERT↑	BLEU-4↑	R-1↑	R-L↑	BERT↑
Cosmos Tokenizer	LLaMA-2-7B	✓	54.99	77.82	68.88	87.86	19.87	55.60	41.10	76.49
		×	57.59	79.26	73.91	90.35	28.40	63.29	48.76	79.35

The original Cosmos Tokenizer uses causal attention for video tasks, modeling forward-only temporal relationships. However, 3D PET volumes have bidirectional spatial relationships, making causal attention less appropriate. To assess this, we compared both settings (with vs. without causal masking) using `LLaMA-2-7B` on the ViMed-PET dataset across two tasks: Report Generation and Visual Question Answering (VQA). As shown in Table 6, removing causal attention yields significantly better performance across both tasks.

B.3.2 Fine-tuning VLMs

After fine-tuning the vision encoders, we integrate each with two language models derived from state-of-the-art medical multimodal foundation models: LLaMA-2-7B from M3D [28] and Mistral-7B from LLaVA-Med [27]. The integration is facilitated by a linear projection layer that aligns the visual and textual embedding spaces.

Conceptual Alignment. We use single-turn data composed of prompts such as "<image> What are the main findings in this medical image?" and "<image> Please write a detailed medical report for this image.", paired with the corresponding medical report as the target output. During training, the weights of both the LLM and the vision encoder are frozen, allowing updates only to the linear projection layer. Training is conducted using a batch size of 16 per GPU across 4 A100 GPUs (80 GB each), with gradient accumulation over 4 steps. We employ the AdamW optimizer [58] with a warmup ratio of 0.03 and an initial learning rate of 2×10^{-3}, followed by a Cosine Annealing Scheduler [60]. Training runs for up to 20 epochs, and the checkpoint with the lowest validation loss is selected for evaluation.

LoRA Fine-tuning. We employ both single-turn and multi-turn conversational data to continue fine-tuning the linear projector and to update the LLM using the Low-Rank Adaptation (LoRA) [46] method. This method efficiently adapts the pretrained LLM by injecting trainable low-rank matrices into selected linear layers, substantially reducing the number of trainable parameters and computational overhead. The LoRA configuration is set as follows: rank (r) = 64, scaling factor (α) = 16, and dropout rate = 0.05. The task type is defined as `CAUSAL_LM`, aligning with the LLM's causal language modeling objective. Training is conducted with a batch size of 8 per GPU across 4 NVIDIA A100 GPUs (80 GB each), using gradient accumulation over 4 steps. We use the AdamW optimizer [58] with a warmup ratio of 0.03 and an initial learning rate of 2×10^{-5}, followed by a Cosine Annealing Scheduler [60]. Training is performed for 20 epochs, and the checkpoint with the lowest validation loss is selected for evaluation.

B.3.3 Fine-tuning Resources

We report the training time and GPU memory consumption for fine-tuning VLMs across different stages, using a setup of four NVIDIA A100 GPUs with 80 GB memory each, as summarized in Table 7. The GPU memory values in the table reflect the peak consumption observed across all four GPUs. All measurements were recorded under a consistent software environment: Python 3.8.20, CUDA nvcc 12.8.61, Accelerate 1.0.1, DeepSpeed 0.16.2, PyTorch 2.1.0, Transformers 4.46.3, and PEFT 0.4.0. Our results show that VLMs utilizing the Cosmos Tokenizer as the vision encoder are more efficient in both training time and memory usage compared to those based on the CT-ViT architecture. This suggests that the architectural design of the Cosmos Tokenizer offers a more resource-efficient training process, which is particularly advantageous in large-scale or resource-constrained settings.

Table 7: Training resource consumption of VLMs on the Original dataset. Memory (Mem) values indicate the peak GPU memory usage (in GB) across four A100 GPUs.

	Model		Computational Resources			
			Concept Alignment		LoRA Fine-tuning	
	Vision	Language	Time (Hours)	Mem (GB)	Time (Hours)	Mem (GB)
Fine-tuned	CT-ViT	Mistral-7B	2.00	61.0	12.00	76.0
		LLaMA-2-7B	2.00	62.0	12.00	73.0
	Cosmos Tokenizer	Mistral-7B	1.83	47.5	11.00	70.0
		LLaMA-2-7B	1.75	46.5	11.00	71.6

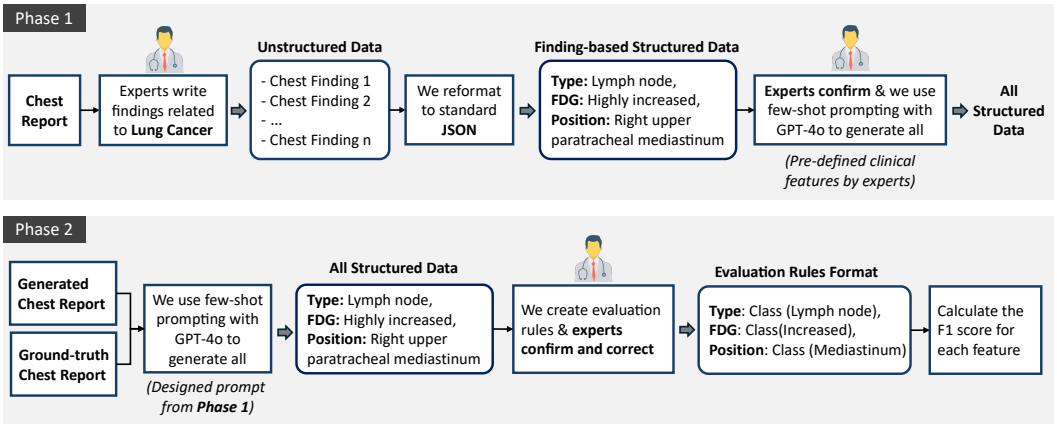


Figure 7: Clinical evaluation pipeline. Phase 1: Experts define structured clinical attributes from reports, which are validated and used to construct prompts for GPT-4o. Phase 2: GPT-4o extracts structured outputs from generated and ground-truth reports, which are mapped to clinical classes for F1-score evaluation.

B.4 Clinical Evaluation Metrics

To clinically evaluate the performance of reports generated by VLMs, we propose a metric computation process developed in collaboration with medical experts. The overall workflow is illustrated in Figure 7, focusing on the extraction of key clinical attributes: lesion type, lesion location, and FDG uptake values.

From the reports generated by the VLMs, we apply a few-shot prompting strategy with GPT-4o [12] to structure the outputs, enabling systematic evaluation of each model's performance. The prompt used for this task is illustrated in Figures 8 and 9. The extracted information is subsequently mapped into categorical variables, which are validated by medical experts. The categories are defined as follows:

- **Type**: {lymph node, pulmonary nodule, ground-glass opacity, pulmonary mass, pleural thickening, interstitial thickening, consolidation, effusion, soft tissue nodule, wall thickening, calcified nodule, hypermetabolic lesion}
- **FDG**: {increase, not increase}
- **Position**: {mediastinum, lung, abdomen, axilla, cervical region}

Subsequently, based on rules manually constructed in collaboration with domain experts, extracted values are grouped into semantically equivalent categories. If two values belong to the same group, they are considered equivalent for evaluation purposes. To preserve evaluation integrity, any extracted value that cannot be confidently assigned to a predefined category is labeled as other and excluded from positive prediction counts. We compute F1-scores by comparing model-generated attributes against ground truth annotations in our medical test set. The evaluation comprises four metrics: F1-T, which measures the F1 score based solely on lesion **T**ypes; F1-TP, which considers both lesion **T**ypes

```
messages = [ {"role": "system", "content": """Assume you are an AI specialized in extracting information from medical reports. Please provide accurate, complete, and detailed information directly from the report, without fabricating any answers. Return only a list containing JSON objects as requested - no unrelated characters are allowed. Assume you are an AI specialized in extracting information from medical reports. Follow these steps:

1. Identify and read the relevant sections of the report to answer the following questions:
- Are there any tumors, lymph nodes, lesions, or abnormalities in the lungs or metastasized to the lungs?
- What is the size of the lesion, tumor, lymph node, or abnormality?
- What is the shape of the lesion, tumor, lymph node, or abnormality?
- What is the FDG uptake level?

2. From the identified segments, extract the important information. For any information that is not available, record it as 'Not available'. Return a list where each item is one JSON object in the following format:
[ {    "Size of tumor/lesion/abnormality": ...,
       "Shape of tumor/lesion/abnormality": ...,
       " Position of tumor/lesion/abnormality": ...,
       "FDG uptake": {"SUVmax": ..., "FDG metabolism": ...},
        "Invasion": ...,
        "Metabolic stage": ...    }]

I will provide some examples for you. """}]

for sample in fewshot_samples:
        messages.append({"role": "user", "content": sample['context']})
        messages.append({"role": "assistant", "content": sample['response']})
messages.append({"role": "user", "content": query})
```

Figure 8: Message used to prompt GPT-4o for structuring VLM-generated reports into JSON format. Manually curated few-shot examples are included in the prompt, where each example consists of an input `sample['context']` and an output `sample['response']`. See Figure 9 for an example.

Example Input:
A spiculated mass opacity in subsegment I of the right upper lung lobe, measuring 74 x 56 mm, with increased FDG uptake (SUVmax: 14.9).

Example Output:

```
[{ "Size of tumor/lesion/abnormality": "74 x 56 mm",
   "Shape of tumor/lesion/abnormality": "Spiculated mass opacity",
   "Position of tumor/lesion/abnormality": "Subsegment I of the right upper lung lobe",
   "FDG uptake": {    "SUVmax": "14.9",    "FDG metabolism": "Increased"    },
   "Invasion": "Not available",
   "Metabolic stage": "Not available"}]
```

Figure 9: Example of a few-shot prompt used to guide GPT-4o in extracting structured JSON data from VLM-generated reports.

and **P**osition; F1-TF, which evaluates lesion **T**ypes together with **F**DG uptake; and F1-TPF, which assesses all three attributes: **T**ype, **P**osition, and **F**DG uptake.

B.5 Task Evaluation by GPT-4o

Due to the inability of GPT-4o to directly analyze 3D inputs, inspired by [61], we adopt a flattening strategy that converts all slices of a 3D volume into multiple 2D slice images, each labeled with a numerical order in the top-right corner. These slices are then arranged into a 5 by 5 (25 slices) 2D grid image, as illustrated in Figure 10. Each 3D volume is thus represented by approximately 5 to 8 such grid images, which are subsequently input into GPT-4o with a prompt shown in Figure 11. From these inputs, GPT-4o generates corresponding reports, which we then evaluate using NLP-based metrics and clinical F1 scores against the ground truth reports.

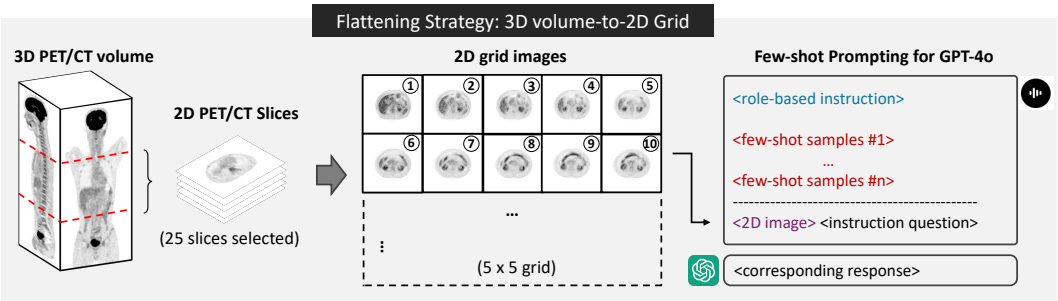


Figure 10: Visualization of the flattening strategy. Consecutive 2D slices from a 3D medical volume are arranged in numerical order (top-right corner) and concatenated into a 5×5 grid image to enable input into GPT-4o.

Assume you are an AI specialized in analyzing 3D medical images, including the regions: head-neck, chest, abdomen-pelvis. You are provided with 3D images as multiple consecutive 2D slices, numbered in order to form a complete 3D volume. Please analyze the 3D images by each region, detect physiological features and abnormal lesions. Provide the most detailed and accurate medical diagnosis possible.

Requirements:
- Identify the anatomical region (head and neck, chest, abdomen and pelvis).
- Give a brief diagnosis based on physiological features and abnormal findings.
- Correlate with relevant clinical examination methods (if needed).

The response format I need is as follows:
- This is an image of the ... region.
- Medical Diagnostic Report: ...

Below are example formats of the analysis I need:
- Example #1 for 3D head-neck image: <head-neck report examples>
- Example #2 for 3D chest image: <chest report examples>
- Example #3 for 3D abdomen-pelvis image: <abdomen-pelvis report examples>

Please provide medical diagnoses based on the images I provide.

<2D image> You are provided with a 3D image input (in the format divided into multiple 2D images, each 2D image is numbered to indicate its order within the 3D image). The provided 3D image belongs to one of three regions: abdomen and pelvis, chest, or head and neck. Please provide the medical diagnosis for this 3D image.

Figure 11: Prompt template used with GPT-4o to analyze concatenated 2D grid images and generate structured medical report outputs. Manually curated few-shot examples are included to guide the model.

C Additional Results

C.1 PET/CT VQA Task

We provide additional qualitative results comparing the baseline and various fine-tuning strategies on the VQA task in Table 8. These findings reveal several key insights that are consistent with trends observed in the report generation task.

Comparison with existing baselines: Fine-tuning VLMs on our proposed ViMed-PET dataset leads to significant improvements across all NLP evaluation metrics in the VQA task.

Comparison between LLMs: The relative performance of LLMs mirrors observations from the report generation task. When fine-tuning is limited to the original dataset (setting O), LLaMA2-7B outperforms Mistral-7B. However, with large-scale training on augmented data (settings O-G and O-G-C), Mistral-7B demonstrates superior performance, likely due to its more efficient architecture compared to LLaMA2-13B. Notably, when integrated with the Cosmos Tokenizer, LLaMA2-7B shows a modest performance advantage over Mistral-7B.

Comparison between vision encoders: Across all training settings, CT-ViT consistently outperforms the Cosmos Tokenizer. This indicates that CT-ViT, which is specifically designed and pretrained on

Table 8: Performance on VQA task. We define training configurations as: **O**-**O**riginal dataset, **G**-Report **G**enerate dataset, **C**-Study **C**omparison dataset. R-1 and R-L denote ROUGE-1 and ROUGE-L scores. ↑ means higher values are better. The best and second-best results are emphasized using **bold** and underline, respectively. **GPT-4o is evaluated under few-shot prompting.*

	Model		Settings			NLP Metrics ↑			
	Vision	Language	O	G	C	BLEU-4	R-1	R-L	BERT
Baseline	LLaVA-Med [27]			–		3.39	47.83	33.82	75.86
	M3D [28]			–		0.03	11.80	9.66	59.87
	RadFM [29]			–		0.04	11.71	12.24	61.93
	GPT-4o* [12]			–		3.01	49.35	30.09	71.92
Fine-tuned	CT-ViT	Mistral-7B	✓			23.22	57.61	43.80	77.06
			✓	✓		<u>31.33</u>	**65.61**	**51.22**	**82.05**
			✓	✓	✓	31.14	<u>65.10</u>	<u>50.33</u>	<u>81.80</u>
		LLaMA-2-7B	✓			26.93	56.01	42.28	75.31
			✓	✓		26.36	56.48	42.79	77.73
			✓	✓	✓	**31.36**	59.14	48.00	76.72
	Cosmos Tokenizer	Mistral-7B	✓			20.01	58.17	42.54	76.47
			✓	✓		25.71	61.05	46.59	78.49
			✓	✓	✓	28.09	62.92	48.37	79.25
		LLaMA-2-7B	✓			25.83	61.80	46.58	78.87
			✓	✓		26.11	62.26	47.05	79.39
			✓	✓	✓	28.40	63.29	48.76	79.35

3D medical imaging data, provides greater clinical relevance and effectiveness in improving VLM performance compared to the Cosmos Tokenizer, which was pretrained on general-purpose datasets.

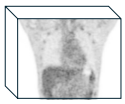
C.2 Report Generation and VQA Samples

We present examples of generated PET/CT reports and VQA interactions using the CT-ViT + Mistral-7B combination. Figure 12 illustrates a sample from the report generation task, highlighting both exact matches and discrepancies between predicted and ground truth reports for the chest and abdomen–pelvis regions. Additionally, Figures 13 and 14 show examples of VQA conversations.

Report Generation Task

<Question>: <image> Please write a detailed medical report for this image.

<PET/CT volume>



Chest volume

<Ground Truth>

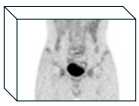
- VI: Hình ảnh bắt xạ theo đặc điểm sinh lý ở tim. Không phát hiện hình ảnh tràn dịch màng phổi hai bên và tràn dịch màng tim. Hình ảnh nốt mờ bán đặc có kích thước 9 x 10 mm, tăng chuyển hóa FDG (SUVmax: 3,5) ở ngoại vi thùy trên phổi trái. Hình ảnh vài hạch ở dưới quai động mạch chủ và bên phải khí quản đoạn thấp, kích thước 10 mm, không tăng chuyển hóa FDG.
- EN: **Imaging shows uptake according to physiological characteristics in the heart**. **No evidence of bilateral pleural effusion or pericardial effusion**. **A semi-solid pulmonary nodule** seen in the periphery of the left upper lobe, measuring **9 x 10 mm** with **increased FDG uptake** (**SUVmax: 3.5**). **Several lymph nodes** are present below the aortic arch and on the right side of the lower trachea, **measuring 10 mm**, **without increased FDG uptake**.

<Generated Report>

- VI: Hình ảnh bắt xạ theo đặc điểm sinh lý ở tim. Không phát hiện hình ảnh tràn dịch màng phổi hai bên và tràn dịch màng tim. Hình ảnh vài hạch trung thất ở cạnh khí quản đoạn thấp, trước carina, cùng với hạch rốn phổi hai bên (hạch lớn nhất có kích thước 11 x 9 mm), tăng chuyển hóa FDG (SUVmax: 5,3). Chưa phát hiện hình ảnh tăng chuyển hóa FDG khu trú, bất thường tại nhu mô hai phổi. Hình ảnh vài hạch nách, kích thước 10 mm, không tăng chuyển hóa FDG ở hố nách hai bên.
- EN: **Imaging shows uptake according to physiological characteristics in the heart**. **No evidence of bilateral pleural effusion or pericardial effusion**. **Several mediastinal lymph nodes** adjacent to the lower trachea, anterior to the carina, along with bilateral hilar lymph nodes (the largest measuring (**11 x 9 mm**), **show increased FDG uptake** (**SUVmax: 5.3**). No focal abnormal FDG uptake is detected in the pulmonary parenchyma of both lungs. **Several** axillary **lymph nodes measuring 10 mm** are seen bilaterally **without increased FDG uptake**.

<Question>: <image> What are the main findings in this medical image?

<PET/CT volume>



Abdomen-Pelvis volume

<Ground Truth>

- VI: Hình ảnh bắt xạ theo đặc điểm sinh lý ở gan, lách, thượng thận và đài bể thận 2 bên. Hình ảnh sỏi thận trái. Hình ảnh tăng chuyển hóa FDG ở khung đại tràng theo đặc điểm sinh lý. Hình ảnh vài hạch bẹn, kích thước 10 mm, không tăng chuyển hóa FDG ở hố bẹn hai bên.
- EN: **Imaging shows uptake according to physiological characteristics in the liver, spleen, adrenal glands, and bilateral renal collecting systems**. **A left renal calculus** is identified. **FDG uptake in the colon is consistent with physiological pattern**. **Several inguinal lymph nodes measuring 10 mm are seen bilaterally without increased FDG uptake.**

<Generated Report>

- VI: Hình ảnh bắt xạ theo đặc điểm sinh lý ở gan, lách, thượng thận và đài bể thận 2 bên. Hình ảnh tăng chuyển hóa FDG lan tỏa ở niêm mạc dạ dày khả năng do viêm. Hình ảnh tăng chuyển hóa FDG ở khung đại tràng theo đặc điểm sinh lý. Hình ảnh vài hạch bẹn, kích thước 10 mm, không tăng chuyển hóa FDG ở hố bẹn hai bên.
- EN: **Imaging shows uptake according to physiological characteristics in the liver, spleen, adrenal glands, and bilateral renal collecting systems**. Diffuse increased **FDG uptake is observed in the gastric mucosa**, likely due to inflammation. **FDG uptake in the colon is consistent with physiological pattern**. **Several inguinal lymph nodes measuring 10 mm are seen bilaterally without increased FDG uptake.**

Figure 12: Ground truth and generated PET/CT reports for the chest and abdomen-pelvis regions using the CT-ViT + Mistral-7B model. Reports are originally written in Vietnamese (VI), with English (EN) translations shown for illustration. Exact matches between the generated and reference text are shown in **bold**, while key details that are missing or incorrectly added in the generated reports are marked in **red**.

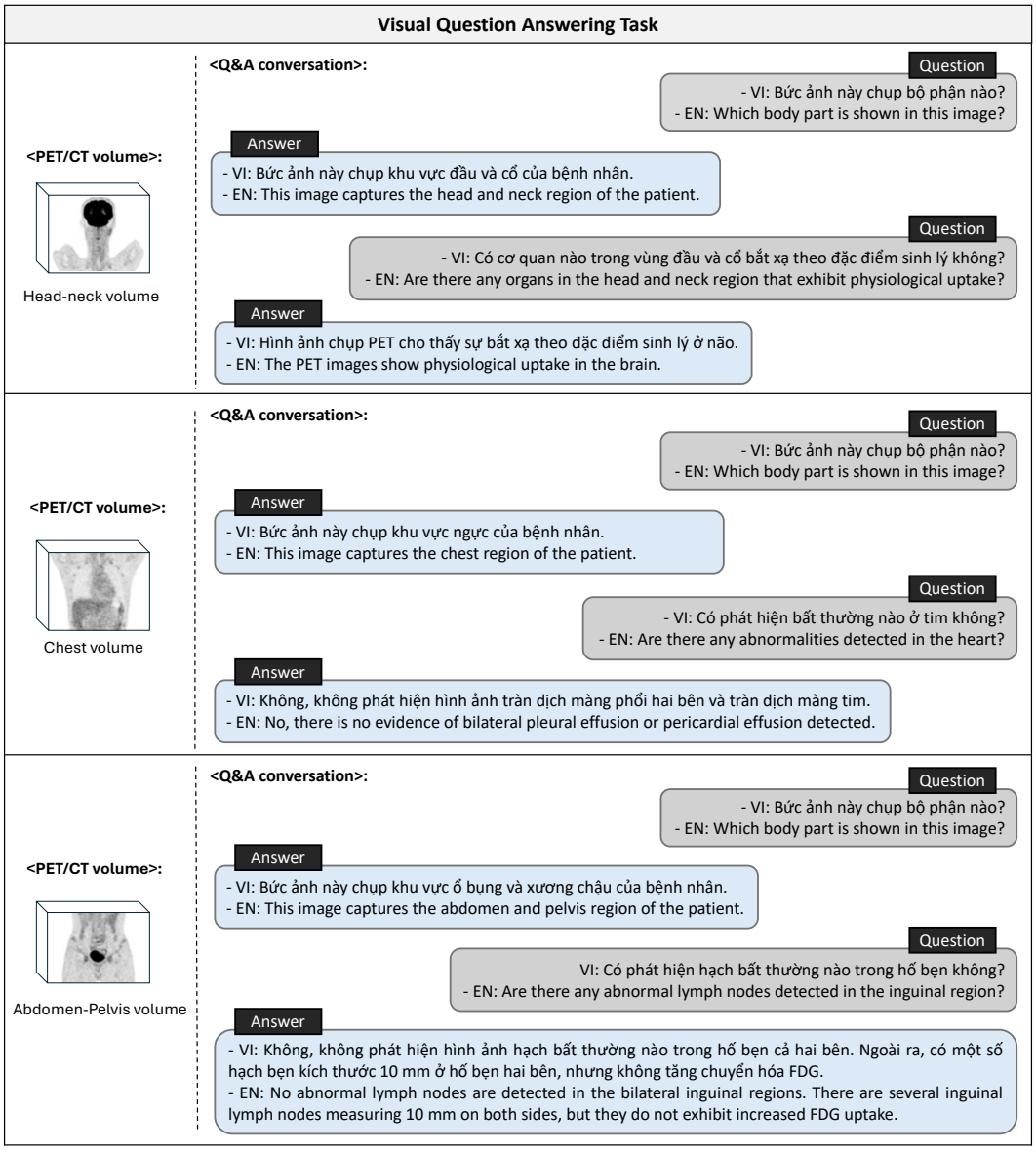


Figure 13: Short-form VQA interaction in Vietnamese (EN: translated) between a user and the CT-ViT + Mistral-7B model. The example illustrates concise factual queries and direct responses.

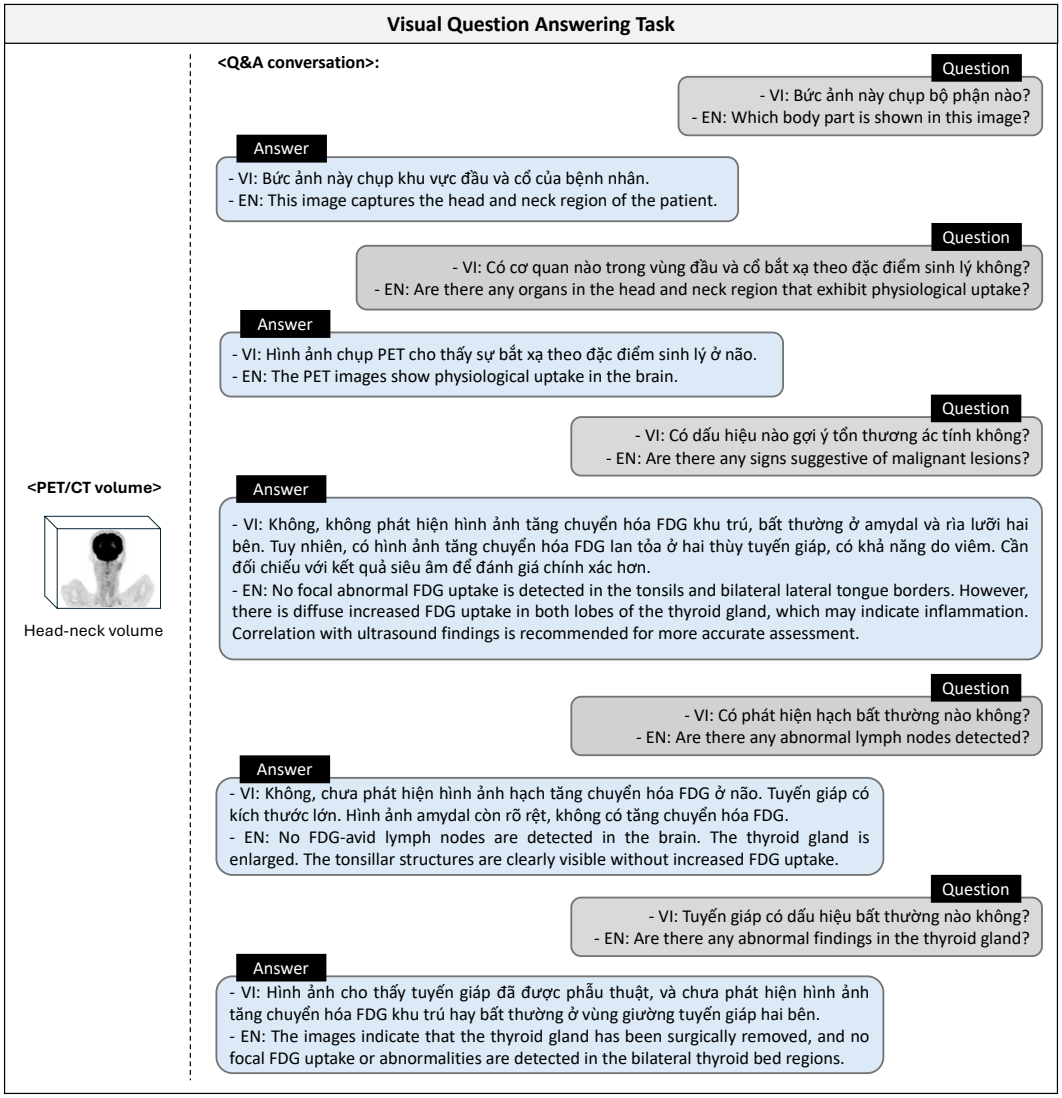


Figure 14: Long-form VQA interaction in Vietnamese (EN: translated) using the CT-ViT + Mistral-7B model. The conversation includes complex multi-sentence reasoning and detailed medical explanation.