

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 [1], PLIP [2], and BiomedCLIP [3], 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 [4], CT2Rep [5], Merlin [6], and RadFM [7], 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 [8], ELIXR [9], and CheXagent [10]. 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 [4] and RadFM [7] introduce architectures capable of handling 3D inputs, improving performance across imaging modalities. For multilingual contexts, Qilin-Med-VL [11] and HuatuoGPT-Vision [12] 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 1: 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 2 for a sample few-shot example.

To construct the Visual Question Answering (VQA) dataset, we follow the methodology introduced by LLaVA [13], 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 [14] 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 2: 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 1 and 2. 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 3: 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['context']` (a pair of reports to compare) and an output `sample['response']` that describes the similarities and differences between the two reports. See Figure 4 for a sample few-shot example.

To construct the study comparison dataset, we adopt a few-shot prompting approach using GPT-4o [14]. 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 3 and 4. 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 4: Few-shot example used in our prompt for generating the study comparison dataset. The yellow highlights indicate the differences between the two reports.

42 B.3 Training and Model Configurations

43 B.3.1 Fine-tuning Vision Encoders

44 We select CT-ViT [15] and Cosmos Tokenizer [16] as the vision encoders for our VLMs, as they are
 45 well-suited for processing 3D volumetric inputs with depths of up to 200 slices and have been pre-
 46 trained on large-scale datasets in prior work. Details on model selection are discussed in Section 3.1.

47 **CT-ViT.** We employ a specialized ViT model, CT-ViT [15], as the vision encoder in our VLMs.
 48 CT-ViT is designed to effectively process 3D chest CT volumes and is pre-trained on a large-scale
 49 medical dataset comprising 25,701 non-contrast 3D chest CT volumes from 21,314 unique patients.
 50 These volumes vary in resolution and contain between 100 and 600 axial slices. To align visual
 51 and textual modalities, we adopt a CLIP-based [17] training approach. The model is fine-tuned
 52 for up to 30 epochs using the AdamW optimizer [18], with a learning rate of 1.25×10^{-6} and a
 53 batch size of 8 per GPU across four NVIDIA A100 GPUs (80 GB each). Early stopping is applied
 54 based on the convergence of training loss, ensuring efficient optimization. For the text modality, we
 55 integrate PhoBERT [19], a state-of-the-art Vietnamese language model pre-trained on a large-scale
 56 Vietnamese corpus. PhoBERT has demonstrated superior performance over multilingual models such
 57 as XLM-R across several Vietnamese natural language processing (NLP) tasks, including part-of-
 58 speech tagging, dependency parsing, named entity recognition, and natural language inference. Its
 59 linguistic compatibility with clinical texts in our dataset enables effective semantic representation
 60 and understanding.

61 **Cosmos Tokenizer.** We leverage the architecture of the Cosmos Tokenizer [16], originally designed
 62 for image and video reconstruction tasks. To adapt it for 3D PET/CT imaging, we remove causality-
 63 based attention mechanisms, which are essential for modeling temporal dependencies in video but
 64 unnecessary for spatially coherent volumetric medical scans. This modification allows us to retain
 65 the benefits of pre-trained weights while enabling effective processing of 3D medical data. We
 66 fine-tune the customized Cosmos Tokenizer using a single-phase reconstruction approach. The total
 67 loss function $\mathcal{L}_{\text{total}}$ combines two terms: an L_1 reconstruction loss \mathcal{L}_1 and an inverted Structural
 68 Similarity Index Measure (SSIM) loss $\mathcal{L}_{\text{SSIM}}$, defined as:

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

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

76 B.3.2 Fine-tuning VLMs

77 After fine-tuning the vision encoders, we integrate each with two language models derived from
 78 state-of-the-art medical multimodal foundation models: LLaMA-2-7B from M3D [4] and Mistral-7B
 79 from LLaVA-Med [21]. The integration is facilitated by a linear projection layer that aligns the visual
 80 and textual embedding spaces.

81 **Conceptual Alignment.** We use single-turn data composed of prompts such as “<image> What
 82 are the main findings in this medical image?” and “<image> Please write a detailed medical report
 83 for this image.”, paired with the corresponding medical report as the target output. During training,
 84 the weights of both the LLM and the vision encoder are frozen, allowing updates only to the linear
 85 projection layer. Training is conducted using a batch size of 16 per GPU across 4 A100 GPUs (80
 86 GB each), with gradient accumulation over 4 steps. We employ the AdamW optimizer [18] with
 87 a warmup ratio of 0.03 and an initial learning rate of 2×10^{-3} , followed by a Cosine Annealing
 88 Scheduler [20]. Training runs for up to 20 epochs, and the checkpoint with the lowest validation loss
 89 is selected for evaluation.

90 **LoRA Fine-tuning.** We employ both single-turn and multi-turn conversational data to continue
 91 fine-tuning the linear projector and to update the LLM using the Low-Rank Adaptation (LoRA) [22]
 92 method. This method efficiently adapts the pretrained LLM by injecting trainable low-rank matrices
 93 into selected linear layers, substantially reducing the number of trainable parameters and computa-
 94 tional overhead. The LoRA configuration is set as follows: rank (r) = 64, scaling factor (α) = 16, and
 95 dropout rate = 0.05. The task type is defined as CAUSAL_LM, aligning with the LLM’s causal language
 96 modeling objective. Training is conducted with a batch size of 8 per GPU across 4 NVIDIA A100
 97 GPUs (80 GB each), using gradient accumulation over 4 steps. We use the AdamW optimizer [18]
 98 with a warmup ratio of 0.03 and an initial learning rate of 2×10^{-5} , followed by a Cosine Annealing
 99 Scheduler [20]. Training is performed for 20 epochs, and the checkpoint with the lowest validation
 100 loss is selected for evaluation.

101 B.3.3 Fine-tuning Resources

102 We report the training time and GPU memory consumption for fine-tuning VLMs across different
 103 stages, using a setup of four NVIDIA A100 GPUs with 80 GB memory each, as summarized in Table 1.
 104 The GPU memory values in the table reflect the peak consumption observed across all four GPUs.
 105 All measurements were recorded under a consistent software environment: Python 3.8.20, CUDA
 106 nvcc 12.8.61, Accelerate 1.0.1, DeepSpeed 0.16.2, PyTorch 2.1.0, Transformers 4.46.3, and PEFT
 107 0.4.0. Our results show that VLMs utilizing the Cosmos Tokenizer as the vision encoder are more
 108 efficient in both training time and memory usage compared to those based on the CT-ViT architecture.
 109 This suggests that the architectural design of the Cosmos Tokenizer offers a more resource-efficient
 110 training process, which is particularly advantageous in large-scale or resource-constrained settings.

Table 1: 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 |

111 B.4 Clinical Evaluation Metrics

112 To clinically evaluate the performance of reports generated by VLMs, we propose a metric computa-
 113 tion process developed in collaboration with medical experts. The overall workflow is illustrated in
 114 Figure 5, focusing on the extraction of key clinical attributes: lesion type, lesion location, and FDG
 115 uptake values.

116 From the reports generated by the VLMs, we apply a few-shot prompting strategy with GPT-4o [14]
 117 to structure the outputs, enabling systematic evaluation of each model’s performance. The prompt

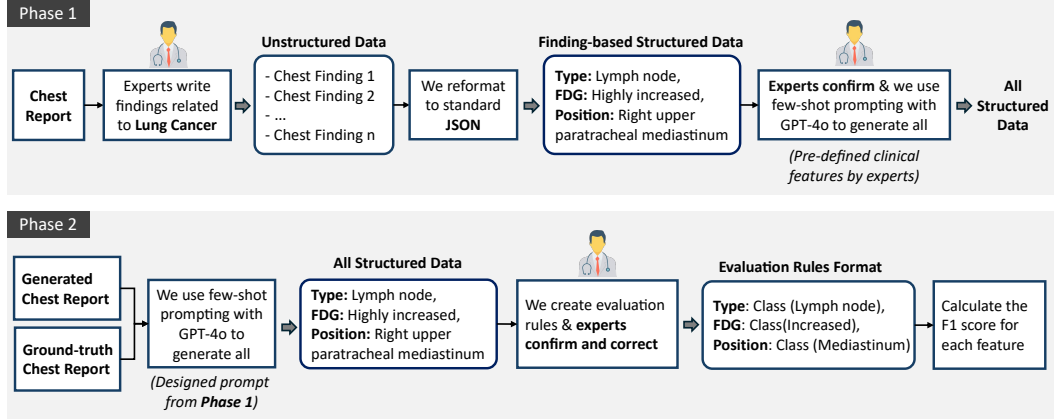


Figure 5: Clinical evaluation pipeline. Phase 1: Experts define structured clinical attributes from reports, which are validated and used to construct prompts for GPT-4o. Phase 2: GPT-4o extracts structured outputs from generated and ground-truth reports, which are mapped to clinical classes for F1-score evaluation.

```

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 6: 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 7 for an example.

118 used for this task is illustrated in Figures 6 and 7. The extracted information is subsequently mapped
 119 into categorical variables, which are validated by medical experts. The categories are defined as
 120 follows:

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

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

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 Types; F1-TP, which considers both lesion Types and Position; F1-TF, which evaluates lesion Types together with FDG uptake; and F1-TPF, which assesses all three attributes: Type, Position, and FDG uptake.

B.5 Task Evaluation by GPT-4o

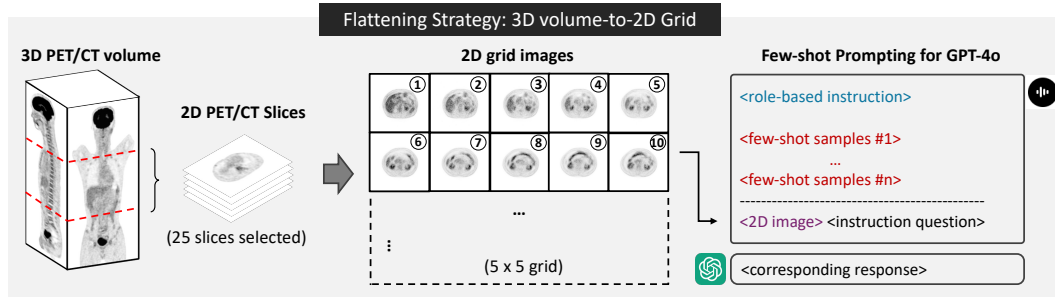


Figure 8: Visualization of the flattening strategy. Consecutive 2D slices from a 3D medical volume are arranged in numerical order (top-right corner) and concatenated into a 5x5 grid image to enable input into GPT-4o.

Due to the inability of GPT-4o to directly analyze 3D inputs, inspired by [23], 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 8. 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 9. 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.

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 9: 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 2. These findings reveal several key insights that are consistent with trends observed in the report generation task.

Table 2: Performance on VQA task. We define training configurations as: **O**-Original dataset, **G**-Report Generate dataset, **C**-Study Comparison 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 [21] | | | – | | 3.39 | 47.83 | 33.82 | 75.86 |
| | M3D [4] | | | – | | 0.03 | 11.80 | 9.66 | 59.87 |
| | RadFM [7] | | | – | | 0.04 | 11.71 | 12.24 | 61.93 |
| | GPT-4o* [14] | | | – | | 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 |

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

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

156 **Comparison between vision encoders:** Across all training settings, CT-ViT consistently outperforms
157 the Cosmos Tokenizer. This indicates that CT-ViT, which is specifically designed and pretrained on
158 3D medical imaging data, provides greater clinical relevance and effectiveness in improving VLM
159 performance compared to the Cosmos Tokenizer, which was pretrained on general-purpose datasets.

160 C.2 Report Generation and VQA Samples

161 We present examples of generated PET/CT reports and VQA interactions using the CT-ViT + Mistral-
162 7B combination. Figure 10 illustrates a sample from the report generation task, highlighting both
163 exact matches and discrepancies between predicted and ground truth reports for the chest and
164 abdomen–pelvis regions. Additionally, Figures 11 and 12 show examples of VQA conversations.

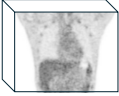
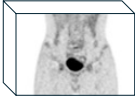
| Report Generation Task | |
|---|---|
| <Question>: <image> Please write a detailed medical report for this image. | |
| <PET/CT volume>  Chest volume | <Ground Truth> - VI: Hình ảnh bắt xạ theo đặc điểm sinh lý ở tim. Không phát hiện hình ảnh tràn dịch màng phổi hai bên và tràn dịch màng tim. Hình 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 10: 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**.

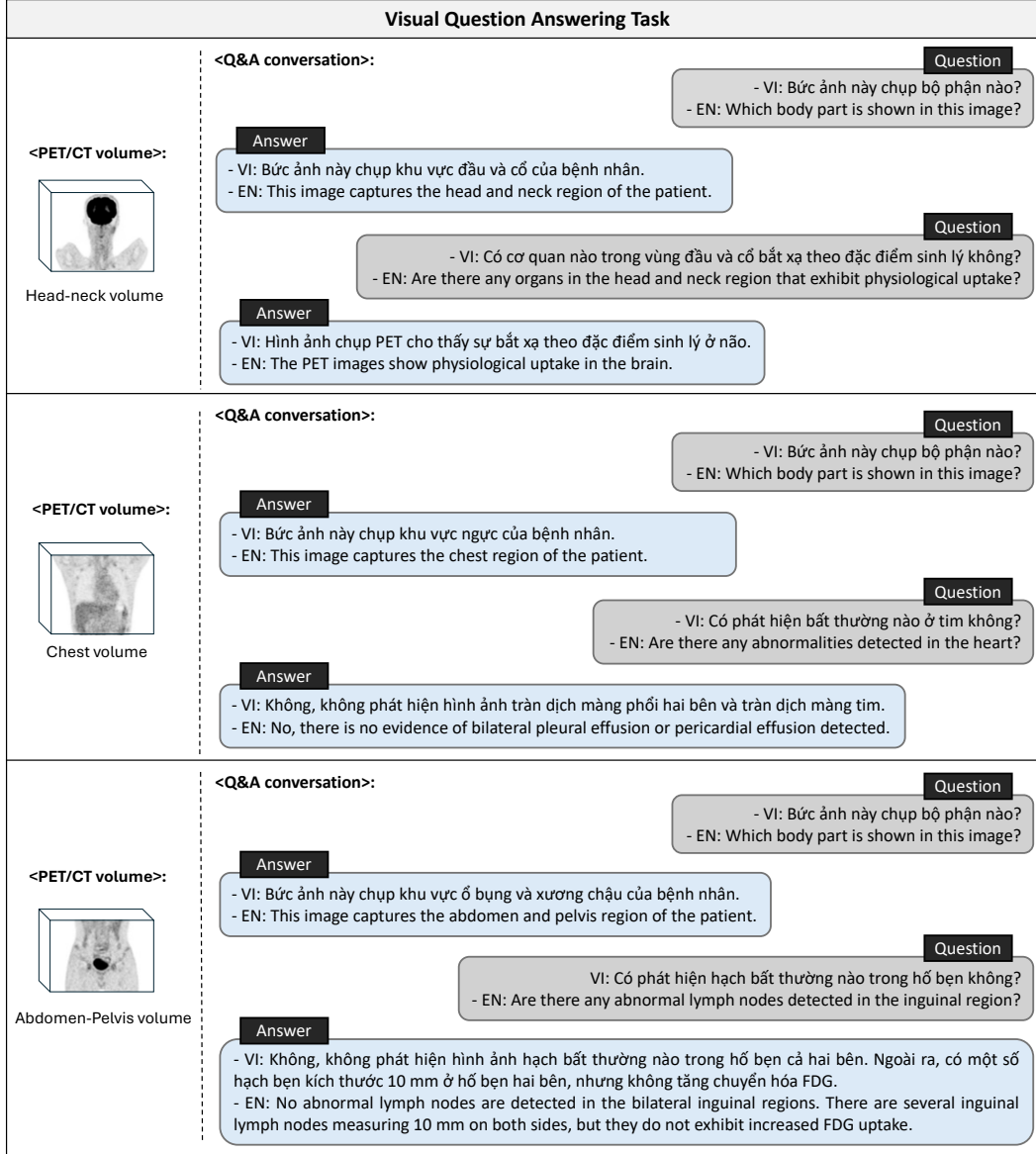


Figure 11: Short-form VQA interaction in Vietnamese (EN: translated) between a user and the CT-ViT + Mistral-7B model. The example illustrates concise factual queries and direct responses.

References

- [1] 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.
- [2] 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.
- [3] 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.

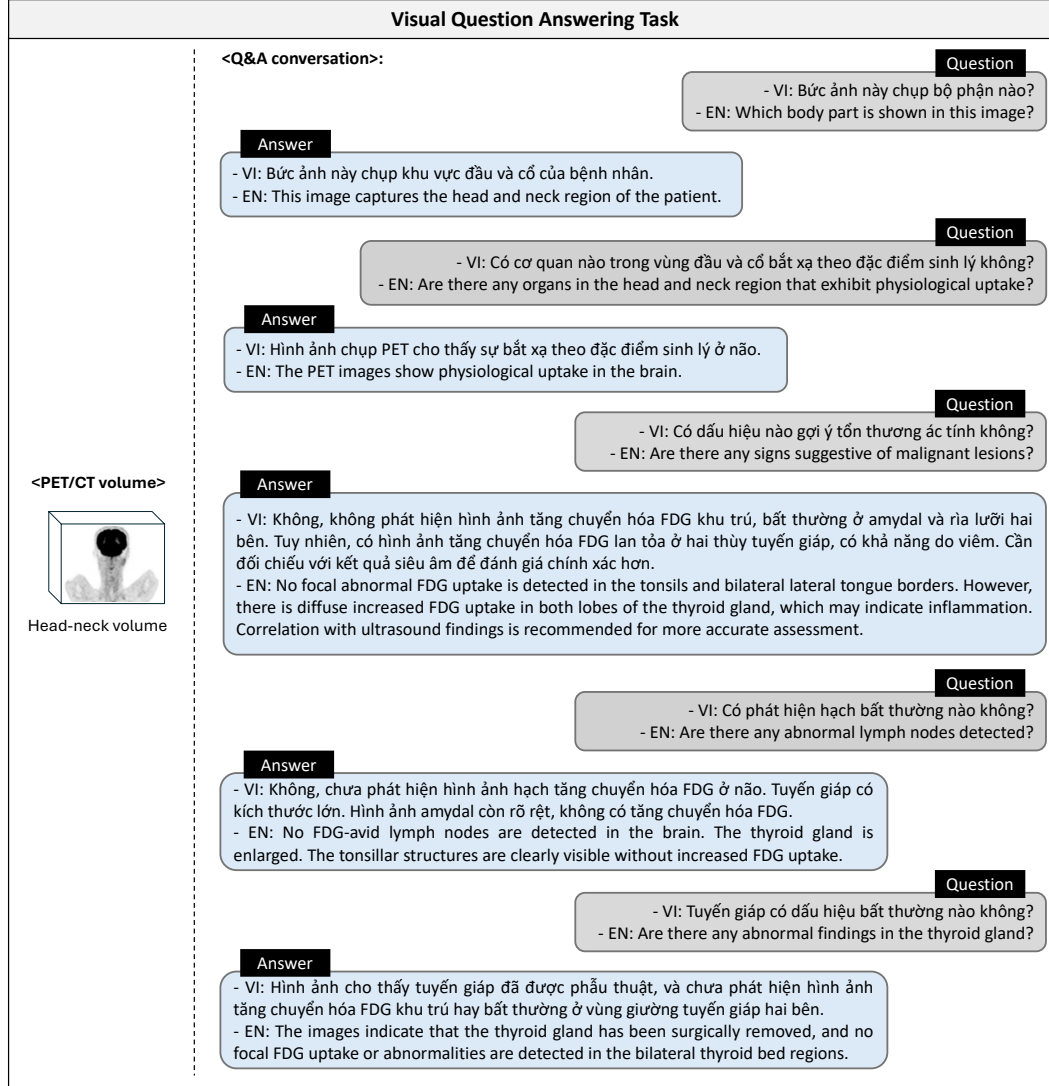


Figure 12: Long-form VQA interaction in Vietnamese (EN: translated) using the CT-ViT + Mistral-7B model. The conversation includes complex multi-sentence reasoning and detailed medical explanation.

- 176 [4] Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical
177 image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*,
178 2024.
- 179 [5] Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. CT2REP: Automated radiology report
180 generation for 3d medical imaging. In *Proceedings of the 2024 International Conference on*
181 *Medical Image Computing and Computer-Assisted Intervention*, pages 476–486. Springer, 2024.
- 182 [6] Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar
183 Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Cesar
184 Truys, et al. Merlin: A vision language foundation model for 3d computed tomography.
185 *Research Square*, pages rs–3, 2024.
- 186 [7] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist
187 foundation model for radiology by leveraging web-scale 2d&3d medical data, 2023.
- 188 [8] Omkar Chakradhar Thawakar, Abdelrahman M Shaker, Sahal Shaji Mullappilly, Hisham
189 Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Khan. Xraygpt:

- 190 Chest radiographs summarization using large medical vision-language models. In *Proceedings*
191 *of the 23rd Workshop on Biomedical Natural Language Processing*, pages 440–448, 2024.
- 192 [9] Shawn Xu, Lin Yang, Christopher Kelly, Marcin Sieniek, Timo Kohlberger, Martin Ma, Wei-
193 Hung Weng, Atilla Kiraly, Sahar Kazemzadeh, Zakkai Melamed, et al. Elixr: Towards a general
194 purpose x-ray artificial intelligence system through alignment of large language models and
195 radiology vision encoders. *arXiv preprint arXiv:2308.01317*, 2023.
- 196 [10] Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier,
197 Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes
198 Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv*
199 *preprint arXiv:2401.12208*, 2024.
- 200 [11] Junling Liu, Ziming Wang, Qichen Ye, Dading Chong, Peilin Zhou, and Yining Hua. Qilin-
201 med-vl: Towards chinese large vision-language model for general healthcare. *arXiv preprint*
202 *arXiv:2310.17956*, 2023.
- 203 [12] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen,
204 Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. Huatuogpt, towards taming language model to
205 be a doctor. *arXiv preprint arXiv:2305.15075*, 2023.
- 206 [13] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan
207 Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint*
208 *arXiv:2408.03326*, 2024.
- 209 [14] OpenAI. Gpt-4o: Openai’s multimodal model with vision, audio, and text capabilities. <https://openai.com/index/gpt-4o>, 2024. Accessed: 2025-04-30.
- 211 [15] Ibrahim Ethem Hamamci, Sezgin Er, Anjany Sekuboyina, Enis Simsar, Alperen Tezcan,
212 Ayse Gulnihan Simsek, Seval Nil Esirgun, Furkan Almas, Irem Doğan, Muhammed Furkan
213 Dasdelen, et al. Generatect: Text-conditional generation of 3d chest ct volumes. In *European*
214 *Conference on Computer Vision*, pages 126–143. Springer, 2024.
- 215 [16] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit
216 Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model
217 platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.
- 218 [17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,
219 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual
220 models from natural language supervision. In *International Conference on Machine Learning*,
221 pages 8748–8763, 2021.
- 222 [18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint*
223 *arXiv:1711.05101*, 2017.
- 224 [19] Dat Quoc Nguyen and Anh Tuan Nguyen. Phobert: Pre-trained language models for vietnamese.
225 *arXiv preprint arXiv:2003.00744*, 2020.
- 226 [20] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv*
227 *preprint arXiv:1608.03983*, 2016.
- 228 [21] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan
229 Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision
230 assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:
231 28541–28564, 2023.
- 232 [22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang,
233 Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Interna-*
234 *tional Conference on Learning Representations*, 1(2):3, 2022.
- 235 [23] Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi.
236 Open-vocabulary action localization with iterative visual prompting. *IEEE Access*, 2025.