Journal of Science and Technology – Smart Systems and Devices
Volume 36, Issue 2, May 2026, 137–144

Bridging Vision and Language in Medical Imaging: Towards Robust Automated Report Generation from PET/CT Scans

Thanh Trung Nguyen*
Department of Medical Equipment, 108 Military Central Hospital, Ha Noi, Vietnam
**Corresponding author email: trungntc10@benhvien108.vn*

Abstract

Vision-Language Models (VLMs) have achieved significant success in multimodal reasoning in general domains, yet their application to medical imaging remains limited, especially in specialized data domains such as PET/CT. In this study, we introduce Vietnamese Positron Emission Tomography - Vision-Language Model (ViPET-VLM), a novel pipeline specifically designed for medical report generation and visual question answering tasks on PET/CT data. ViPET-VLM integrates a fusion module to combine morphological information from CT with functional signals from PET, thereby forming a richer multimodal representation. To enhance clinical reliability, we propose a regularization mechanism with specialized loss functions that not only ensure diagnostic accuracy but also guide the model's attention to critical regions of interest in the images. ViPET-VLM was evaluated on a comprehensive, expert-validated PET/CT dataset and demonstrated marked improvements over current state-of-the-art methods in both report generation and medical question answering. The model shows potential for enhancing the accuracy and clinical applicability of VLMs for medical imaging.

Keywords: Medical report generation, PET/CT, vision-language models.

1. Introduction

In recent years, Vision-Language Models (VLMs) have become a core research direction in Artificial Intelligence due to their ability to learn unified representations from image and text data [1–5]. Trained on massive datasets with billions of image-caption pairs, these models achieve cross-modal alignment, enabling a direct connection between visual and linguistic semantics. This has unlocked potential for numerous downstream tasks such as image captioning [6, 7], visual question answering [4–7], automatic report generation [8, 9], as well as zero-shot image classification [1, 10].

Although general-purpose VLMs like CLIP [1], Flamingo [11], and GPT-4o [12] have demonstrated outstanding performance on various natural image benchmarks, their application in the medical domain remains challenging. The primary reason stems from the domain gap: medical images possess fundamentally different structures, textures, and purposes compared to everyday photographs [13, 14]. Furthermore, the accompanying textual descriptions in medicine are often highly specialized, context-rich, and require expert knowledge for interpretation [15–17].

Several recent works, such as MedCLIP [18] and MedFlamingo [19], have attempted to adapt general VLM architectures for medical data by retraining or fine-tuning them on specialized datasets. However, current approaches still face several limitations. Firstly, the range of modalities is narrow: most studies focus on chest X-rays [20], MRI and CT scans [21], or histopathology images [22], while functional modalities like PET, which are particularly crucial in oncology, cardiology, and neurology, are largely overlooked. Secondly, existing medical vision-language datasets are predominantly monolingual, mainly in English, leading to a severe lack of resources for low-resource languages such as Vietnamese. Lastly, the majority of available data consists only of short captions [23] or limited annotations [24], which is insufficient to reflect the complexity of real-world medical reports.

To address these gaps, we develop Vietnamese Positron Emission Tomography - Vision-Language Model (ViPET-VLM), a novel VLM pipeline specifically designed for PET/CT data. This model integrates a fusion module to synchronize structural information from CT and functional information from PET, while also leveraging a regularization mechanism with specialized loss components to both enhance clinical accuracy and guide the model's attention towards diagnostically significant regions of interest (ROIs).

The main contributions of this paper are as follows:

(i) We propose ViPET-VLM, the first model specifically designed for PET/CT data to concurrently exploit both structural (CT) and functional (PET) information.

p-ISSN 3093-3285
e-ISSN 3093-3315
https://doi.org/10.51316/jst.190.ssad.2026.36.2.15
Received: Oct 13, 2025; Revised: Jan 28, 2026;
Accepted: Jan 28, 2026; Online: Feb 24, 2026

(ii) ViPET-VLM incorporates a fusion module to unify multimodal signals, along with regularization mechanisms based on specialized loss functions to improve clinical accuracy and direct the model's focus onto critical ROIs.

(iii) We conduct experiments on a comprehensive PET/CT dataset with full clinical reports, demonstrating that ViPET-VLM achieves superior performance compared to existing state-of-the-art VLMs in both medical report generation and visual question answering tasks.

2. Related Works

2.1. Multimodal Datasets in the Medical Domain

Recent advancements in medical VLMs have been primarily driven by datasets that link medical images with their corresponding text annotations. Previous research has predominantly focused on modalities such as CT, MRI, and X-ray, providing captions or diagnostic reports [23, 24]. For 2D data, prominent examples include PMC-OA [23] and MIMIC-CXR [24]. Other datasets have expanded to 3D data to support volumetric modeling, such as M3D-Data [25], MedMD [26], and CT-RATE [27].

Recently, the ViMed-PET dataset [28], the first publicly available dataset comprising 3D whole-body PET/CT volumes, was released, opening a promising research avenue for deep learning models in multimodal biomedical image analysis. The emergence of this dataset not only facilitates the fair evaluation of new methods but also encourages the research community to develop more advanced architectures for learning representations from large-scale PET/CT data. In this study, we select ViMed-PET as the foundation for all our experiments to validate the effectiveness of our proposed method.

2.2. Vision-Language Models in the Medical Domain

Recent studies have developed numerous specialized vision-language models for medicine, with various designs tailored to leverage the unique characteristics of clinical data. MedCLIP [18] retains the CLIP-style dual- encoder architecture (image + text) but decouples images from reports to extend its scope beyond tightly paired data. It also replaces the standard InfoNCE loss with a medical knowledge-guided semantic matching loss to mitigate false negatives, thereby enhancing zero- shot/few-shot transfer performance. Meanwhile, Med- Flamingo [19] adheres to the Flamingo framework, combining a pre-trained vision encoder with gated perceiver-style cross-attention layers to interleave image tokens with a Large Language Model (LLM). The medical variant is further trained on image-text data from textbooks and articles, enabling generative visual question answering (generative VQA) in a few-shot setting. LLaVA-Med [29] adopts an adapter/projection approach, projecting visual features into the LLM's token space (e.g., Vicuna) and then performing visual instruction tuning. The supervision data is constructed from image-caption pairs from PubMed Central, combined with GPT-4-style synthetic instructions organized into a curriculum. For chest X-rays, CXR- LLaVA [30] extends the LLaVA architecture by fine- tuning on a large-scale dataset of chest X-rays with corresponding reports and Q&A pairs to generate finding-level free-text.

In summary, existing medical VLMs have demonstrated significant potential in leveraging image-text data, ranging from dual-encoder architectures (MedCLIP), interleaved cross-attention mechanisms (Med-Flamingo), to methods that project features into the LLM space for instruction tuning (LLaVA-Med, CXR-LLaVA, LLaVA-Rad), as well as designs that exploit temporal information (BioViL-T). However, the majority of these models still focus on X- ray or MRI data and have not been effectively extended to niche domains, particularly PET/CT imaging, where the combination of functional and structural signals is pivotal for clinical diagnosis. This gap underscores the necessity for specialized architectures aimed at richer multimodal representations and higher clinical reliability, thereby laying the groundwork for ViPET- VLM, our proposed vision-language model architecture.

3. Methodology

3.1. Component Architecture Selection

A typical vision-language model generally comprises two main components: a vision encoder and a text encoder. In this study, we utilize the 3D vision encoder CT-ViT [27], as it is the only publicly available Vision Transformer pre-trained on 3D CT imaging data. For the language component, we employ Mistral-7B [31], a language model recommended by state-of-the-art VLMs such as LLaVA-Med [29] and M3D [25]. These models are fine-tuned on biomedical instruction datasets, enhancing their ability to comprehend complex clinical texts and engage in sophisticated, clinically meaningful medical dialogue.

3.2. Fine-Tuning Flow

We utilize a curated PET/CT dataset to fine-tune the base component models through a structured procedure. As illustrated in Fig. 1, the ViPET-VLM pipeline consists of three stages: (1) Fine-tuning the 3D vision encoders to adapt to the 3D PET/CT data modality and learn cross-modal representations between the PET and CT domains via a fusion module, (2) Aligning vision-language attributes between images and text, and (3) Instruction-tuning the model using Low-Rank Adaptation (LoRA) [32] to optimize it for downstream multimodal clinical tasks.

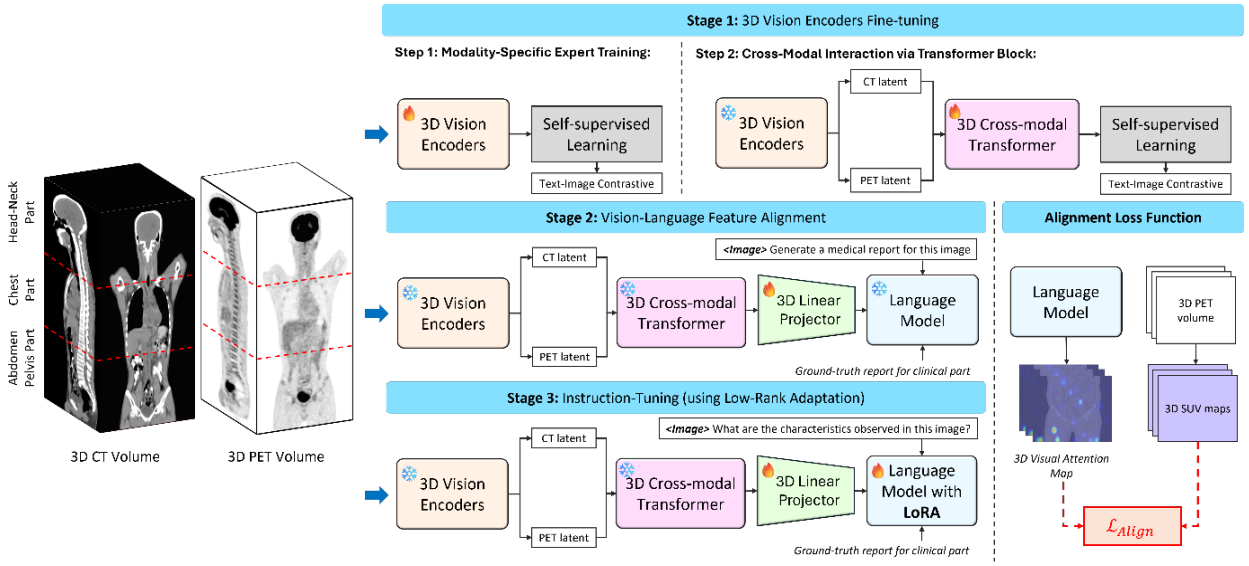

Fig. 1. Overview of the proposed ViPET-VLM method

Additionally, we propose an advanced method to improve the fine-tuning of the entire architecture by designing an alignment loss function. This loss function acts as a regularization mechanism, guiding the model's attention towards critical regions of interest in the images, specifically the high-SUV areas that require emphasis in PET images.

3.2.1. Fine-tuning the 3D vision encoders

We employ self-supervised learning (SSL), specifically CLIP-style fine-tuning, to pre-train the PET and CT encoders. This strategy enables the encoders to capture semantic information (aligned with text) from 3D medical data. This stage is divided into two steps: (1) CLIP-style fine-tuning of the PET and CT vision encoders, and (2) Learning cross-modal representations between the PET and CT data domains.

- Medical report separation

We propose using GPT-4o to partition the content of the full PET/CT report into two distinct descriptive segments for PET and CT information. Concretely, the prompt consists of (i) the original report text as input and (ii) an instruction that enforces a strict JSON-only output schema:

$$\{\ \text{"}PET\text{"}: \langle R_{PET} \rangle, \text{"}CT\text{"}: \langle R_{CT} \rangle\ \}$$

To minimize cross-contamination between modalities, we impose explicit content constraints in the instruction: *PET* is restricted to metabolic information only, including tracer uptake patterns and quantification (e.g., hyper-/hypo-metabolic foci, intensity and distribution, SUV/SUVmax, focal and diffuse uptake); *CT* is restricted to anatomical information only, including morphology and localization (e.g., organ/region, size measurements, shape, margins), while explicitly excluding uptake-related terms and SUV values.

- Fine-tuning specialized vision encoders

We fine-tune two independent 3D vision encoders: f_ϕ^{PET} for the PET image domain and f_θ^{CT} for the CT image domain. Correspondingly, we denote g_ϕ^{PET} and g_θ^{CT} as the text encoders for the PET and CT domains, respectively. The output visual representations and from the two 3D vision encoders are defined as:

$$v_{PET} = f_\phi^{PET}(x_{PET}), v_{CT} = f_\theta^{CT}(x_{CT}) \quad (1)$$

where x_{PET} and x_{CT} are the original 3D PET and CT images, respectively. Similarly, the output text representations are:

$$t_{PET} = g_\phi^{PET}(R_{PET}), t_{CT} = g_\theta^{CT}(R_{CT}) \quad (2)$$

Each vision encoder is fine-tuned using a CLIP-style loss function as follows:

$$\mathcal{L}_{PET} = -\frac{1}{N}\Sigma_{i=1}^{N} log \frac{exp\left(sim\left(v_{PET}^i, t_{PET}^i\right)/\tau\right)}{\sum_{j=1}^{N} exp\left(sim\left(v_{PET}^j, t_{PET}^j\right)/\tau\right)} \quad (3)$$

$$\mathcal{L}_{CT} = -\frac{1}{N}\Sigma_{i=1}^{N} log \frac{exp\left(sim\left(v_{CT}^i, t_{CT}^i\right)/\tau\right)}{\sum_{j=1}^{N} exp\left(sim\left(v_{CT}^j, t_{CT}^j\right)/\tau\right)} \quad (4)$$

where sim(·,·) is the cosine similarity, τ is a temperature parameter, and N is the minibatch size.

- Learning cross-modal representations between PET and CT domains

After fine-tuning the PET and CT vision encoders, the output PET image representations $V_{PET} = \{v^i_{PET}\}$ and CT image representations $V_{CT} = \{v^i_{CT}\}$ are combined through the cross-attention mechanism of a Transformer-based fusion module. This design is particularly well-motivated in PET/CT, where the two modalities are routinely co-registered in clinical

systems; thus, the attention operation is constrained to learn cross-modal associations under a shared spatial reference, preserving structural correspondence while enhancing semantic coupling. This combination is expressed by the following formulas:

$$Q = V_{PET}W_Q, K = V_{CT}W_K, V = V_{CT}W_V \quad (5)$$

$$\bar{Z} = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (6)$$

Here, we use PET features as queries to reflect the clinician-inspired reading workflow: experts often first identify suspicious metabolic uptake on PET and then localize and characterize the corresponding region on CT for anatomical delineation and differential interpretation. In this sense, PET acts as the "driver" signal that initiates cross-modal retrieval, while CT provides structured anatomical evidence as contextual support.

Similar to the standard Transformer architecture, we also incorporate residual connections, layer normalization, and a feed-forward network (FFN) to stabilize the training process:

$$Z' = LayerNorm(V_{PET} + \bar{Z}) \quad (7)$$

$$Z_{cross} = LayerNorm\big(Z' + FFN(Z')\big) \quad (8)$$

Finally, the fused representation Z_{cross} is aligned with the full PET/CT report using a CLIP-style loss function as follows:

$$\mathcal{L}_{cross} = -\frac{1}{N}\Sigma_{i=1}^{N} log\frac{exp(sim(z_{cross}^i, t_{full}^i)/\tau)}{\sum_{j=1}^{N} exp\big(sim\big(z_{cross}^j, t_{full}^j\big)/\tau\big)} \quad (9)$$

where $t_{full} = f_\psi^{full}(R)$ is the text representation encoded from the full PET/CT report.

3.2.2. Vision-language concept alignment

In this stage, we fine-tune a linear projection layer to map the visual feature representations into the embedding space of the language encoder. We use single-turn image-text pairs from a VQA dataset generated from the original dataset, where each sample consists of a PET/CT image and an instruction-based question. Example prompts include: "What are the key findings in this medical image?" or "Please write a detailed medical report for this medical image."

The training objective is to accurately reproduce the original text response given the image and prompt. During this process, the weights of the vision encoder and the language model are frozen, allowing only the weights of the linear projection layer to be updated. This design ensures an effective and stable alignment between the image and text embeddings, while also reducing the risk of overfitting and preserving the pre-trained representations.

3.2.3. Instruction tuning

In the final stage, we conduct fine-tuning of the entire VLM using the VQA dataset, which includes both single-turn question-answer pairs and multi-turn conversational interactions. During this process, the weights of the vision encoder are kept frozen to preserve the previously learned visual representations. Only the parameters of the vision-language projection layer and the language model are updated, applying the Low-Rank Adaptation (LoRA) [32] method for efficient fine-tuning.

This stage enhances the model's ability to understand and answer a diverse range of medical questions by integrating information from both image and text modalities. The inclusion of both simple and complex conversational formats also improves the model's robustness and operational stability in biomedical VQA tasks.

3.2.4. Alignment loss function

The language model component of ViPET-VLM is based on a Transformer decoder, which processes input through L decoder layers, each equipped with a multi-head attention module. At layer l, this module consists of H attention heads, where each head h (with $1 \le h \le H$) computes attention separately based on its corresponding attention map M_l^h. This attention mechanism models the relationships between image tokens and text tokens. To analyze the contribution of image tokens to text generation, we focus on the attention scores between image tokens and subsequent text tokens, referred to as the visual attention map $M_v^{l,h}$, which is a submatrix of the overall attention map M_l^h. To obtain an overview of the attention distribution across image tokens, we can compute the average visual attention map M_v by aggregating the attention across all heads and all layers:

$$M^v = \frac{1}{LH}\Sigma_{l=1}^{L}\Sigma_{h=1}^{H}M_{lh}^v \quad (10)$$

Separately, from the original PET x_{PET} image, we use a simple thresholding method to obtain a binary mask matrix S. The regions of interest correspond to the areas with a value of 1 in S, which also represent the high-SUV regions that require emphasis when cross-referenced with x_{PET}. From this, we design an alignment loss function $\mathcal{L}_{align}$ to guide the model's attention, via the average attention matrix M^v, towards the regions of interest defined by S as follows:

$$\mathcal{L}_{align} = \Sigma_{s\in S}\left(1 - \frac{\Sigma_{c\in S}M_c^v}{\Sigma_{c'=1}^{T}M_{c'}^v}\right)^2 \quad (11)$$

where s represents each reference region in S, c and c' are indices of the image tokens, and T is the total number of image tokens. Among all T image tokens, this loss

function encourages higher attention scores on the image tokens belonging to each region s, thereby creating an attentional constraint for the model on the important regions of the PET image. The final loss function of the model is formulated as shown below, with $\mathcal{L}_{align}$ acting as a regularization term:

$$\mathcal{L}_{final} = \mathcal{L}_{llm} + \lambda \mathcal{L}_{align} \quad (12)$$

where $\mathcal{L}_{llm}$ is the original loss function of the language component, and λ is a regularization hyperparameter to control the influence of $\mathcal{L}_{align}$ in the final loss.

4. Evaluation

4.1. Experimental Design

4.1.1. Dataset

- Data description

We use the ViMed-PET dataset [28], collected from a central general hospital in Vietnam. It comprises 2,757 whole-body PET/CT studies conducted over four years, with a total of 2,757 pairs of 3D CT–PET images and their corresponding full clinical reports. Each 3D scan contains approximately 250-500 CT and PET slices, covering the area from the head to the upper thigh, and includes various pathologies such as lung cancer, thyroid cancer, and other clinical conditions. The images are stored in DICOM format, accompanied by detailed data on age, gender, weight, tracer activity, and scan parameters, and were acquired using a GE Discovery 710/STE PET/CT system with CT-based attenuation correction. All personally identifiable information has been removed from the images and reports, including patient, physician, and institutional details. The reports were extracted and standardized into JSON format. To expand the number of samples and enhance the alignment quality between the two data domains, each PET/CT case was divided into three anatomical regions (head-neck, chest, abdomen-pelvis) with a 20-slice overlap to prevent information loss at the boundaries. This strategy generated 8,271 image-report pairs, enabling the model to learn region-specific clinical features more accurately and improving overall performance.

- Data processing

We constructed a specialized Visual Question Answering (VQA) dataset from ViMed-PET to serve the model's training and evaluation stages. The VQA dataset includes 27,855 single-turn question-answer samples and 8,271 multi-turn conversational dialogues generated using GPT-4o with a few-shot prompting strategy. This allows the model to be trained on contextual reasoning over PET/CT images.

4.1.2. Evaluation metrics

We evaluate the model's performance based on common Natural Language Processing (NLP) metrics, including BLEU-4, ROUGE, and BERTScore. Specifically, BLEU-4 is used to measure the precision of 4-gram sequences in the generated text. ROUGE-1 captures the coverage of unigrams, which is useful for summarization tasks. ROUGE-L assesses text fluency based on the longest common subsequence between the generated and reference texts. Meanwhile, BERTScore measures the semantic similarity between the output and ground-truth texts by leveraging contextual embeddings from a pre-trained BERT model.

4.1.3. Comparative experiments

We compare ViPET-VLM against various standard models, including LLaVA-Med [29], M3D [25], RadFM [26], and GPT-4o [11], on two tasks: PET/CT report generation and medical question answering on PET/CT.

4.2. Comparative Results with Other Models

The quantitative results comparing ViPET-VLM and other methods on the PET/CT report generation and medical VQA tasks are presented in Table 1 and 2, respectively.

Table 1. Quantitative comparison of ViPET-VLM and other methods on the PET/CT report generation task

Setup	**NLP Metrics (↑)**			
Model	**BLEU-4**	**ROUGE-1**	**ROUGE-L**	**BERT**
LLaVA-Med [29]	0.01	50.08	27.89	64.63
M3D [25]	0.04	41.01	23.53	67.21
RadFM [26]	0.06	54.23	28.33	69.49
GPT-4o* [11]	31.12	67.96	52.76	81.09
ViPET-VLM (w/o fusion & w/o reg)	53.30	77.79	68.60	88.35
ViPET-VLM (w/o fusion only)	<u>54.02</u>	<u>77.92</u>	<u>70.58</u>	88.36
ViPET-VLM (w/o reg only)	53.72	77.86	69.11	<u>88.44</u>
ViPET-VLM (Ours)	**55.70**	**78.63**	**72.04**	**89.74**
Diff.	**+3.1%**	**+0.9%**	**+2.1%**	**+1.5%**

*Note: The best results are in **bold** and the second best are <u>underlined</u>. ↑ indicates that higher is better. reg means regularization. *GPT-4o is evaluated under few-shot prompting.*

Table 2. Quantitative comparison of ViPET-VLM and other methods on the medical question answering task for PET/CT.

Setup	**NLP Metrics (↑)**			
Model	**BLEU-4**	**ROUGE-1**	**ROUGE-L**	**BERT**
GPT-4o* [11]	3.01	49.35	30.09	71.92
ViPET-VLM (w/o fusion & w/o reg)	31.14	65.61	51.22	82.50
ViPET-VLM (w/o fusion only)	<u>33.12</u>	<u>66.47</u>	51.18	<u>83.27</u>
ViPET-VLM (w/o reg only)	31.70	65.93	<u>51.24</u>	82.91
ViPET-VLM (Ours)	**33.41**	**67.35**	**54.02**	**84.01**
Diff.	**+0.9%**	**+1.3%**	**+5.4%**	**+0.9%**

*Note: The best results are in **bold** and the second best are <u>underlined</u>. ↑ indicates that higher is better. reg means regularization. *GPT-4o is evaluated under few-shot prompting.*

4.2.1. PET/CT report generation task

From Table 1, we observe that ViPET-VLM outperforms existing methods across all metrics in the PET/CT report generation task. Specifically, the model achieves a BLEU-4 score of 55.70%, a significant increase compared to GPT-4o (31.12%) and previous methods (<0.1 for LLaVA-Med, M3D, RadFM). Similarly, ViPET-VLM attains the highest ROUGE-1 (78.63%), ROUGE-L (72.04%), and BERTScore (89.74%), demonstrating a superior ability to capture both n-gram overlap and semantic content.

To isolate the contribution of multimodal fusion, we additionally evaluate ViPET-VLM in a PET-only setting (w/o fusion). Notably, even without fusion, the model already surpasses all baselines, achieving 53.30% in BLEU-4, 77.79% in ROUGE-1, 68.60% in ROUGE-L, 88.35% in BERTScore without regularization, and 54.02% in BLEU-4, 77.80% in ROUGE-1, 70.58% in ROUGE-L, 88.36% in BERTScore with regularization. More importantly, when comparing (w/o fusion & w/o reg) against (w/o reg only), adding the fusion module yields consistent improvements across all metrics: BLEU-4 increases from 53.30% to 53.72%, ROUGE-1 from 77.79% to 77.86%, ROUGE-L from 68.60% to 69.11%, and BERTScore from 88.35% to 88.44%. This demonstrates that multimodal fusion provides additional gains beyond a strong PET-only backbone by leveraging complementary structural cues from CT.

To isolate the contribution of the proposed regularization, we further compare ViPET-VLM without and with regularization under matched fusion settings. When fusion is removed (w/o fusion), adding regularization consistently improves generation quality: BLEU-4 increases from 53.30% to 54.02%, ROUGE-1 from 77.79% to 77.80%, ROUGE-L from 68.60% to 70.58%, and BERTScore from 88.35% to 88.36%.

Similarly, under the fusion-enabled configuration, regularization yields additional gains over the corresponding non-regularized variant (w/o reg only), further improving BLEU-4 from 53.72% to 55.70%, ROUGE-1 from 77.86% to 78.63%, ROUGE-L from 69.11% to 72.04%, and BERTScore from 88.44% to 89.74%. These results indicate that the regularization term contributes complementary benefits to fusion by promoting more clinically consistent and robust representations during fine-tuning.

When the full ViPET-VLM with PET and CT fusion and regularization term is used, performance improves across all metrics, with absolute gains of +3.1% in BLEU-4, +0.9% in ROUGE-1, +2.1% in ROUGE-L, and +1.5% in BERTScore compared to the second-best version. These results confirm that using both the multimodal integration of PET and CT and regularization technique provides complementary information, thereby enhancing the accuracy of PET/CT report generation.

4.2.2. Medical question answering on PET/CT task

As shown in Table 2, our ViPET-VLM surpasses the GPT-4o baseline on all metrics for the medical VQA task. When evaluating the contribution of multimodal data fusion, we find that integrating CT information and using proposed regularization improves the model's ability to comprehend content and generate answers. Overall, ViPET-VLM demonstrates improvements of +0.9% in BLEU-4, +1.3% in ROUGE-1, +5.4% in ROUGE-L, and +0.9% in BERTScore compared to the second-best ablation setup, confirming the benefits of multimodal fusion in the clinical VQA problem.

5. Conclusion

This study introduces ViPET-VLM, a vision-language model designed for report generation and question answering on PET/CT data, featuring a fusion module that combines CT and PET information. Evaluation on the ViMed-PET dataset shows that ViPET-VLM significantly improves performance compared to current methods, particularly in the semantic quality of the generated answers and reports. A current limitation is that the clinical reliability has not

been directly evaluated by physicians. In the future, we plan to collaborate with medical professionals to accurately measure clinical efficacy and extend the model to other medical imaging modalities.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, Learning transferable visual models from natural language supervision, in International Conference on Machine Learning, pp. 8748–8763, 2021.

[2] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang, The dawn of LMMs: Preliminary explorations with GPT-4V(ision), arXiv preprint arXiv:2309.17421, pp. 1–166, 2023.

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

[4] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, and C. Li, Llava-onevision: Easy visual task transfer, arXiv preprint arXiv:2408.03326, 2024.

[5] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, Qwen2.5-VL technical report, arXiv preprint arXiv:2502.13923, 2025.

[6] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, and T. Zhu, Qwen technical report, arXiv preprint arXiv:2309.16609, 2023.

[7] H. Liu, C. Li, Q. Wu, and Y. J. Lee, Visual instruction tuning, Advances in Neural Information Processing Systems, vol. 36, pp. 34892– 34916, 2023.

[8] S. Yan, W. K. Cheung, I. W. Tsang, K. Chiu, T. M. Tong, K. C. Cheung, and S. 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, pp. 14324–14333, 2024.

[9] T. T. Pham, N.-V. Ho, N.-T. Bui, T. Phan, P. Brijesh, D. Adjeroh, G. Doretto, A. Nguyen, C. C. Wu, H. Nguyen, and N. Le., 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, pp. 941–958, 2024.

[10] S. Javed, A. Mahmood, I. I. Ganapathi, F. A. Dharejo, N. Werghi, and M. 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, pp. 11450–11459, 2024.

[11] OpenAI, Gpt-4o: Openai's multimodal model with vision, audio, and text capabilities. 2024. Available: https://openai.com/index/gpt-4o

[12] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, A. Mensch, K. Milln, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, Flamingo: a visual language model for few-shot learning, arXiv preprint arXiv:2204.14198, 2022.

[13] H. Nori, Y. T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu, R. Luo, S. M. McKinney, R. O. Ness, H. Poon, T. Qin, N. Usuyama, C. White, and E. Horvitz, Can generalist foundation models outcompete special-purpose tuning? case study in medicine, arXiv preprint arXiv:2311.16452, 2023.

[14] J. H. Moon, H. Lee, W. Shin, Y.-H. Kim, and E. Choi, Multimodal understanding and generation for medical images and text via vision-language pre-training, IEEE Journal of Biomedical and Health Informatics, vol. 26, iss. 12, pp. 6070–6080, 2022. https://doi.org/10.1109/JBHI.2022.3207502

[15] M. Aljabri, M. AlAmir, M. AlGhamdi, M. Abdel-Mottaleb, F. Collado-Mesa, Towards a better understanding of annotation tools for medical imaging: A survey, Multimedia Tools and Applications, vol. 81, pp. 25877–25911, 2022.

[16] A. Davis, R. Souza, and J.-H. Lim, Knowledge-augmented language models interpreting structured chest x-ray findings, arXiv preprint arXiv:2505.01711, 2025.

[17] S. Yang, X. Wu, S. Ge, S. K. Zhou, and L. Xiao, Knowledge matters: Chest radiology report generation with general and specific knowledge, Medical Image Analysis, vol. 80, Art. no. 102510.

[18] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, Medclip: Contrastive learning from unpaired medical images and text, in Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, vol. 2022, p. 3876, 2022.

[19] M. Moor, Q. Huang, S. Wu, M. Yasunaga, C. Zakka, Y. Dalmia, E. P. Reis, P. Rajpurkar, and J. Leskovec, Med-flamingo: A multimodal medical few-shot learner, In Proceedings of Machine Learning for Health (ML4H) PMLR, pp. 353-367, 2023.

[20] P. Chambon, C. Bluethgen, J.-B. Delbrouck, R. Van der Sluijs, M. Połacin, J. M. Z. Chaves, T. M. Abraham, S. Purohit, C. P. Langlotz, and A. Chaudhari, Roentgen: Vision-language foundation model for chest x-ray generation, arXiv preprint arXiv:2211.12737, 2022.

[21] L. Xu, H. Sun, Z. Ni, H. Li, and S. Zhang, Medvilam: A multimodal large language model with advanced generalizability and explainability for medical data understanding and generation, arXiv preprint arXiv:2409.19684, 2024.

[22] Y. Xin, G. C. Ates, K. Gong, and W. Shao, Med3dvlm: An efficient vision-language model for 3D medical image analysis, arXiv preprint arXiv:2503.20047, 2025.

[23] W. Lin, Z. Zhao, X. Zhang, C. Wu, Y. Zhang, Y. Wang, and W. Xie, Pmc-clip: Contrastive language-image pre-training using biomedical documents, in International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 525–536, Springer, 2023.

[24] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, Mimic-cxr, a deidentified publicly available database of chest radiographs with free-text reports, Scientific data, vol. 6, iss. 1, p. 317, 2019.

[25] F. Bai, Y. Du, T. Huang, M. Q.-H. Meng, and B. Zhao, M3d: Advancing 3D medical image analysis with multi-modal large language models, arXiv preprint arXiv:2404.00578, 2024.

[26] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data, Nature Communications, vol. 16, 2025, Art. no. 7866.

[27] I. E. Hamamci, S. Er, and B. 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, pp. 476–486, Springer, 2024.

[28] Nguyen, H. T., Nguyen, D. T., Nguyen, T. M. D., Nguyen, T. T., Truong, T. N., Pham, H. H., Nguyen, P. L., Toward a vision-language foundation model for medical data: Multimodal dataset and benchmarks for Vietnamese PET/CT report generation, (2025) arXiv preprint arXiv:2509.24739.

[29] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, Llava-med: Training a large language-and-vision assistant for biomedicine in one day, Advances in Neural Information Processing Systems, vol. 36, pp. 28541–28564, 2023.

[30] S. Lee, J. Youn, H. Kim, M. Kim, and S. H. Yoon, CXR-LLAVA: a multimodal large language model for interpreting chest X-ray images, European Radiology, 35(7), pp.4374-4386, 2024.

[31] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, Mistral 7b, arXiv preprint arXiv:2310.06825, 2023.

[32] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, Lora: Low-rank adaptation of large language models, International Conference on Learning Representations, vol. 1, iss. 2, p. 3, 2022.