Region-Grounded Report Generation for 3D Medical Imaging: A Fine-Grained Dataset and Graph-Enhanced Framework

Cong Huy Nguyen[1,*], Son Dinh Nguyen[1,*], Guanlin Li[2], Tuan Dung Nguyen[1], Aditya Narayan Sankaran[2], Mai Huy Thong[3], Thanh Trung Nguyen[3], Mai Hong Son[3], Reza Farahbakhsh[2], Phi Le Nguyen[1,†], Noel Crespi[2,†]

[1]AI4LIFE, Hanoi University of Science and Technology, Vietnam
{huy.nc235504, son.dn225997, dung.nt232198m}@sis.hust.edu.vn, lenp@soict.hust.edu.vn
[2]SAMOVAR, Télécom SudParis, Institut Polytechnique de Paris, France
{guanlin_li, aditya-narayan.sankaran, reza.farahbakhsh, noel.crespi}@telecom-sudparis.eu
[3]108 Military Central Hospital, Vietnam
maihuythong27121995@gmail.com, trungntc10@benhvien108.vn, alex.hong.son@gmail.com

Abstract

Automated medical report generation for 3D PET/CT imaging is fundamentally challenged by the high-dimensional nature of volumetric data and a critical scarcity of annotated datasets, particularly for low-resource languages. Current "black-box" methods map whole volumes to reports, ignoring the clinical workflow of analyzing localized Regions of Interest (RoIs) to derive diagnostic conclusions. In this paper, we bridge this gap by introducing *VietPET-RoI*, the first large-scale 3D PET/CT dataset with fine-grained RoI annotation for a low-resource language, comprising 600 PET/CT samples and 1,960 manually annotated RoIs, paired with corresponding clinical reports. Furthermore, to demonstrate the utility of this dataset, we propose *HiRRA*, a novel framework that mimics the professional radiologist diagnostic workflow by employing graph-based relational modules to capture dependencies between RoI attributes. This approach shifts from global pattern matching toward localized clinical findings. Additionally, we introduce new clinical evaluation metrics, namely *RoI Coverage* and *RoI Quality Index*, that measure both RoI localization accuracy and attribute description fidelity using LLM-based extraction. Extensive evaluation demonstrates that our framework achieves SOTA performance, surpassing existing models by 19.7% in BLEU and 4.7% in ROUGE-L, while achieving a remarkable 45.8% improvement in clinical metrics, indicating enhanced clinical reliability and reduced hallucination. Our code and dataset are available on GitHub.

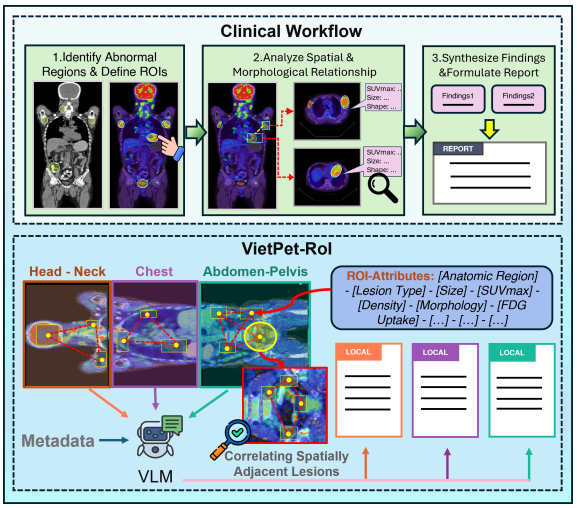

Figure 1: **Illustration of VietPET-RoI annotation.** Following doctors' conventional workflow, VietPET-RoI provides hierarchical annotations at both region-level and RoI-level with structured clinical attributes.

1 Introduction

Recent advances in Vision-Language Models (VLMs) have driven significant progress in healthcare AI, enabling the automated generation of clinical reports from medical images. Contemporary medical VLMs, such as LLaVA-Med (Li et al., 2023a), M3D-LaMed (Bai et al., 2024), and RadFM (Wu et al., 2025), have shown impressive capabilities in interpreting diagnostic imaging.

However, despite these general advancements, automated report generation for 3D PET/CT remains in its infancy. Current models exhibit suboptimal accuracy and are prone to significant hallucinations (Chen et al., 2024a), with state-of-the-art VLMs falling considerably short of real-world

*Equal contribution.
†Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 19725–19740
July 2-7, 2026

Table 1: **Comparison with existing PET/CT benchmarks.** VietPET-RoI uniquely provides both full reports and fine-grained RoI annotations (3D Boxes, Attributes).

Dataset	Year	Lang	Dataset Profile				Annotation Granularity				
			Size	Disease	Dim	Public	Report	Metadata	RoI BBox	RoI Attrs	Phys/Path
ViMed-PET (Nguyen et al., 2025)	2025	VN	2,757	Multi	3D	✓	✓	✓	✗	✗	✗
PETRG-Lym (Jiao et al., 2025)	2025	CN	824	Single	3D	✗	✓	✗	✗	✗	✗
PET2Rep (Zhang et al., 2025)	2025	CN	565	Multi	2D	✓	✓	✗	✗	✗	✗
AutoPET-RG (Jiao et al., 2025)	2025	CN	135	Single	3D	✓	✓	✗	✗	✗	✗
VietPET-RoI (Ours)	**2026**	**VN**	**600**[1]	**Multi**	**3D**	✓	✓	✓	✓	✓	✓

Lang: Language (VN: Vietnamese, CN: Chinese). **Metadata**: De-identified patient metadata. **RoI BBox**: Manual 3D Bounding Box annotation. **RoI Attrs**: Rich structured attributes (Density, Size, etc). **Phys/Path**: Physiological/Pathological uptake.

clinical requirements (Zhang et al., 2025), highlighting substantial challenges in 3D multimodal analysis. We posit that this limitation stems from a fundamental methodological divergence. Existing PET/CT report generation models predominantly rely on an end-to-end paradigm, attempting to map complex, high-dimensional whole-volume scans directly to a final text report (Messina et al., 2022). This "black-box" strategy ignores the intrinsic complexity of PET/CT data. In practice, radiologists do not interpret a 3D volume as a single monolithic input; instead, they systematically identify specific Regions of Interest (RoIs), evaluate their individual attributes, and analyze the spatial and physiological inter-relationships between these abnormalities (Waite et al., 2019). Only after this granular synthesis do they derive diagnostic conclusions and draft a formal report (shown in Figure 1). Consequently, the current end-to-end training paradigm lacks the clinical inductive bias, leading to reports that lack both precision and interpretability.

To address this challenge, two fundamental components are required: (i) a dataset with fine-grained RoI-level annotations to provide a basis for grounded learning (Xie et al., 2024; Boecking et al., 2022; de Castro et al., 2025), and (ii) a model architecture that replicates the hierarchical reasoning process of a medical expert (Zhang et al., 2024). However, existing PET/CT datasets and models (Bai et al., 2024; Nguyen et al., 2025) fail to meet these requirements, particularly for low-resource languages like Vietnamese.

In this paper, we bridge this research gap by introducing VietPET-RoI, the first 3D PET/CT dataset featuring fine-grained RoI-level grounding. Our dataset comprises 600 samples from 200 patients with 1,960 RoIs and reports in Vietnamese. Each RoI is manually labeled with comprehensive clinical attributes, providing a structured foundation for learning from segmentation to generation.

[1]600 region-level samples derived from 200 patients.

Furthermore, we propose *HiRRA*, a novel VLM architecture designed to mimic the professional diagnostic workflow. *HiRRA* comprises two core components: a dual-stream encoder and a graph-based relational module. The former preserves separate CT and PET features before fusion to ensure high-fidelity volumetric data; the latter models inter-RoI dependencies vital for diagnosing metastatic patterns and systemic disease (Hu et al., 2024; Seo et al., 2021). Experimental results demonstrate that HiRRA substantially outperforms existing baselines, achieving a 19.7% improvement in BLEU-4 over traditional end-to-end methods. To ensure a more rigorous evaluation of clinical utility, we also propose two clinical metrics: RoI Coverage and RoI Quality Index, which measure the fidelity of region-level attribute descriptions.

In summary, the primary contributions of our work include:

- First, we introduce *VietPET-RoI*, the first public RoI-grounded PET/CT dataset with 600 samples from 200 patients and 1,960 RoIs with comprehensive clinical annotations, addressing RoI-level supervision gap and medical AI scarcity for low-resource languages.
- Second, we propose *HiRRA*, a VLM emulating the diagnostic workflow. HiRRA integrates CT and PET information through hierarchical feature extraction, capturing global volumetric and localized RoI features, achieving state-of-the-art on linguistic and clinical metrics.
- Third, we design clinical metrics specifically tailored for medical report generation, directly assessing clinical factors such as *RoI Coverage* and *RoI Quality Index*.

2 Related work

2.1 Existing Medical 3D datasets

While large-scale PET/CT datasets like AutoPET (Gatidis et al., 2022) and RIDER (Muzi

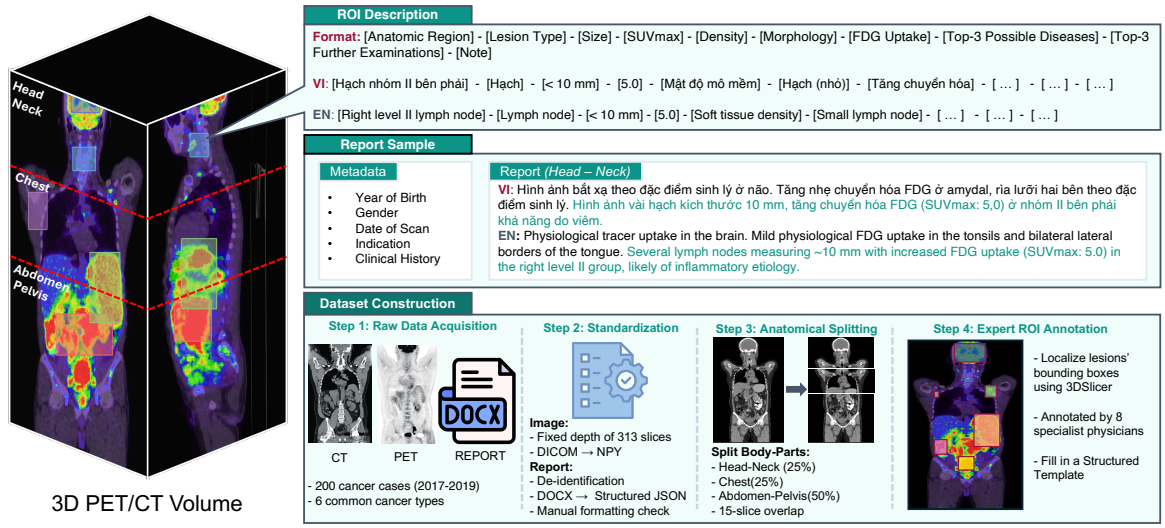

Figure 2: **Overview of the VietPET-RoI dataset.** The figure displays (top) the multimodal data samples including 3D PET/CT volumes, structured RoI descriptions, and clinical reports; and (bottom) the four-stage curation pipeline, spanning from raw data acquisition to expert-level annotation.

et al., 2015) support dense segmentation or lesion detection, they lack aligned clinical reports, limiting their utility for multimodal modeling. Conversely, recent volumetric report generation benchmarks, including ViMed-PET (Nguyen et al., 2025), PETRG-Lym (Jiao et al., 2025), and PET2Rep (Zhang et al., 2025), pair scans with diagnostic text but rely on coarse-grained supervision, omitting explicit lesion localization. This dichotomy highlights a critical gap: the lack of datasets combining fine-grained region grounding with diagnostic reporting to enable interpretable, clinically aligned modeling.

2.2 3D Report Generation VLMs

Recent advances in vision–language models (VLMs) have extended medical report generation to 3D volumetric imaging, enabling richer spatial reasoning over CT, MRI, and PET/CT scans. Representative models like M3D (Bai et al., 2024) and Med3DVLM (Xin et al., 2025) have integrated 3D encoders with large language models, achieving promising results when generating global diagnostic reports from full volumes. However, these end-to-end approaches rely on scan-level supervision, lacking explicit region-level grounding. This diverges from standard clinical practice, where radiologists follow a region-centric workflow - detecting, localizing, and characterizing abnormalities before report synthesis. Consequently, existing models fail to capture this intermediate reasoning, creating a gap between volumetric perception and clinically meaningful, structured generation.

3 VietPET-RoI: A RoI-Grounded Vietnamese 3D PET/CT Dataset

We detail the dataset's pipeline and clinical statistics, proving its suitability for 3D report generation.

3.1 Dataset Construction Process

The dataset was curated by oncology specialists at a leading public hospital in Vietnam, comprising 200 cancer cases spanning the six most common malignancies examined between 2017 and 2019. PET/CT scans were standardized to a unified format with 313 axial slices from head to upper thighs, then divided into three anatomical regions (head-neck, chest, abdomen-pelvis) with 15-slice overlap to yield 600 region-level image-report pairs (shown in Figure 2). Details on preprocessing procedures (DICOM conversion, de-identification, report standardization) are provided in Appendix A.

The key novelty of *VietPET-RoI* lies in fine-grained RoI annotations. Unlike existing PET/CT datasets that provide only global scan-report pairs, eight nuclear medicine physicians collaborated to localize and describe every abnormal finding or clinically relevant physiological structure mentioned in the reports. Using 3DSlicer (Pieper et al., 2004), physicians marked each RoI with a 3D bounding box and completed a structured template recording ten clinical attributes: anatomical location, lesion type, size, SUVmax, density, morphology, FDG uptake, top-3 differential diagnoses with follow-up examinations, and additional notes.

A total of 1,960 RoIs were annotated, with all samples rigorously verified by senior nuclear

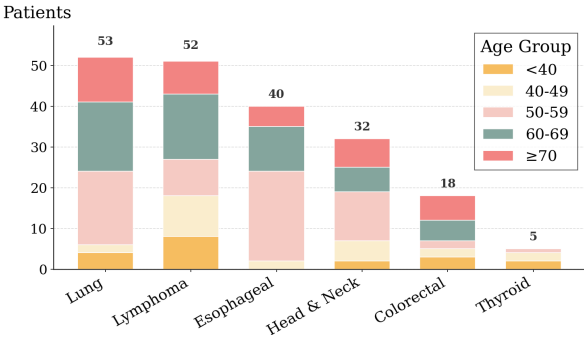

Figure 3: Data distribution across the six cancer types.

medicine physicians to ensure consistency. Any discrepancies were resolved through consensus discussion to guarantee high-quality ground truth. Unlike existing datasets providing only the final reports, our fine-grained RoI annotations enable models to learn direct associations between specific regions and corresponding clinical findings, rather than relying on ambiguous global image-text mappings. Moreover, this design replicates the clinical diagnostic workflow, facilitating multi-task learning including segmentation, description generation, and decision support.

3.2 Dataset Characteristics

VietPET-RoI provides a multimodal dataset comprising textual reports, 3D PET/CT volumes, and fine-grained RoI-description. In total, the dataset contains 600 region-level samples derived from 200 cancer patients covering six common and clinically important malignancies: lymphoma, head and neck cancer, lung cancer, esophageal cancer, thyroid cancer and colorectal cancer. The cohort was selected to include diverse sex, age, and body habitus based on patient metadata, reducing bias toward any specific demographic group. Figure 3 shows the age distribution across the six cancer types, while key dataset-level statistics are summarized in Table 2.

For each study, the CT volumes are stored as 3D arrays of size slices $\times$ 512 $\times$ 512, whereas the PET volumes are represented as slices $\times$ 256 $\times$ 256, following typical clinical in-plane resolutions and allowing straightforward voxel-wise fusion between modalities. After RoI annotation, we observe an average of 3.27 RoIs per body-region sample, with a minimum of 1 and a maximum of 23 RoIs. The abdomen_pelvis region is the most densely annotated, typically containing approximately 3.9 RoIs per sample, reflecting the prevalence of abdominal and pelvic lesions in the target cancer types. Approximately 42% of RoIs correspond to pathological findings, while about 58% capture physiological uptake patterns and serve as negative or contextual examples for the models. This configuration yields a compact but clinically rich dataset with dense region-level supervision, well suited for training and evaluating RoI-grounded PET/CT report generation and other multimodal medical AI tasks. The dataset will be publicly released for non-commercial research purposes under appropriate usage agreements upon paper acceptance.

Table 2: Key characteristics of the VietPET-RoI dataset.

Characteristic	**Value**
Patients	200
Region-level samples	600
Age (mean $\pm$ SD)	57.9 $\pm$ 12.2 years
Age range	13–81 years
CT volume size	313 $\times$ 512 $\times$ 512
PET volume size	313 $\times$ 256 $\times$ 256
Total RoIs	1,960
Avg RoIs/sample	3.27
Most annotated region	Abdomen-pelvis ($\approx$3.9/sample)
RoI composition	42% pathological / 58% physiological

4 HiRRA: A Hierarchical Region-Aware Framework for Report Autogeneration

Overview. To empirically validate the utility of the VietPET-RoI dataset, we propose HiRRA, a Hierarchical Region-aware for Report Autogeneration. Unlike conventional end-to-end VLMs that directly map global image features to textual reports, HiRRA is architected to emulate the professional diagnostic workflow, where physicians synthesize reports by integrating holistic volumetric scans with a granular analysis of specific RoIs. As illustrated in Figure 4, the HiRRA framework is composed of three primary modules designed to facilitate this hierarchical translation. The first component is the Dual Encoder, which independently processes anatomical CT and functional PET to preserve high-fidelity features prior to fusion. Subsequently, a Hierarchical Feature Extractor utilizes a bifurcated strategy: a Global Context block captures overarching volumetric characteristics, while a Local Context block extracts fine-grained features from annotated RoIs. Finally, an LLM-based Decoder integrates these multi-granularity features to generate clinical reports. By guiding attention through RoI visual tokens, this structured supervision ensures the model significantly outperforms traditional methods trained only on image-text pairs.

Multimodal Dual Encoder. Motivated by the clinical workflow where physicians overlay metabolic

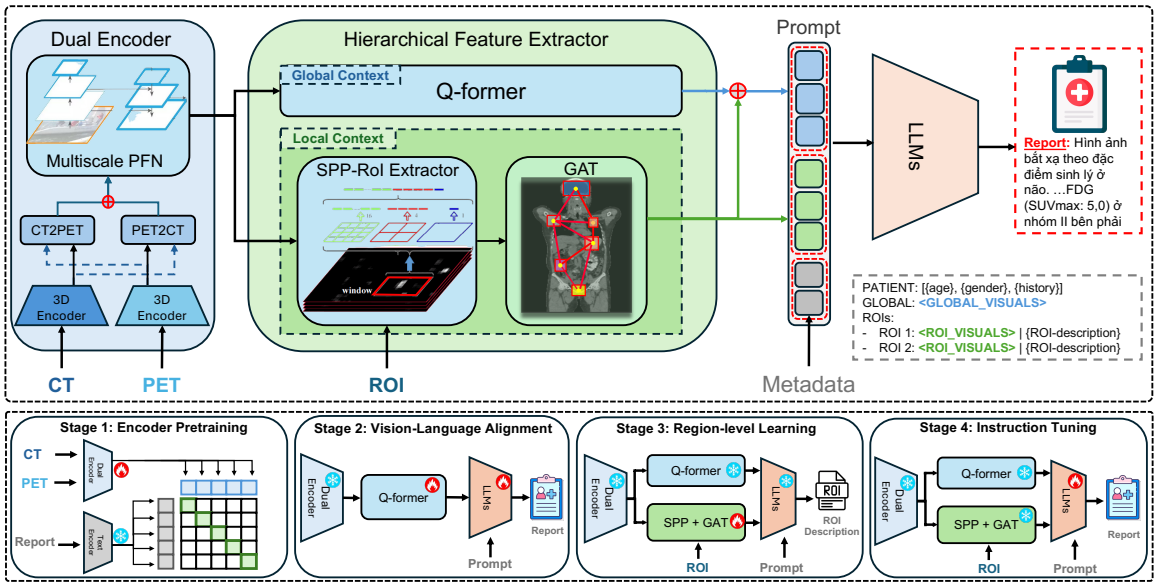

Figure 4: **The overall architecture of HiRRA.** The framework processes paired PET/CT volumes through a Dual Encoder and a Hierarchical Feature Extractor. The *Global Context* is captured via Q-former, while the *Local Context* is using SPP-RoI extraction and GATv2. Finally, the LLM generates the report using a semantic-injected prompt.

heatmaps (PET) onto anatomical scans (CT), we design a dual-stream architecture to extract modality-specific features before fusion. Given paired volumes $I_{CT}, I_{PET} \in \mathbb{R}^{B\times 1\times T\times H\times W}$, we employ two separate 3D encoders $\mathcal{E}_{CT}$ and $\mathcal{E}_{PET}$ based on CT-ViT (Hamamci et al., 2024b) to extract features $F_{CT}, F_{PET} \in \mathbb{R}^{B\times N\times D}$, where $N = T'\times H'\times W'$ and $D = 512$. Unlike single-encoder approaches that prematurely fuse CT and PET channels, our dual-stream design preserves modality-specific characteristics: CT captures anatomical structures (tissue boundaries, organ morphology) while PET quantifies metabolic activity (glucose uptake patterns) (Beyer et al., 2004).

To bridge the semantic gap, we apply bidirectional cross-attention (Li et al., 2023b):

$$\tilde{F}_{CT} = F_{CT} + \text{CA}(F_{CT}, F_{PET}), \quad (1)$$

$$\tilde{F}_{PET} = F_{PET} + \text{CA}(F_{PET}, F_{CT}), \quad (2)$$

where $\text{CA}(\cdot,\cdot)$ denotes cross-attention with query from the first argument and key-value from the second. The fused representation F_{visual} is obtained by averaging $\tilde{F}_{CT}$ and $\tilde{F}_{PET}$, followed by a Multi-scale Pyramid Feature Network (PFN) (Lin et al., 2017) to capture multi-resolution features.

Hierarchical Feature Extractor. To synthesize comprehensive visual representations, we extract features at two distinct levels of granularity: global and local context. The Global Context block captures the overarching environment from the input CT and PET volumes. This is operationalized using a Q-former (Li et al., 2023b) with query vectors, which compresses the high-dimensional volumetric data into a fixed-length latent representation.

For the Local RoI-Aware Context, we extract region-specific information to augment the LLM's inferential capabilities during report generation. Traditional methods typically process RoIs independently, thereby overlooking the critical clinical relationships between disparate lesions. However, diagnostic accuracy in PET/CT often hinges on these relationships: spatially proximate lesions may indicate local invasion, while distant lesions exhibiting similar metabolic patterns frequently suggest metastatic spread. To effectively capture these dependencies, we employ a Graph Neural Network to learn the spatio-morphological correlations between RoIs. Specifically, we construct a graph whose each node represents an individual RoI. Given N RoIs with features $\mathbf{h}_i \in \mathbb{R}^D$ and bounding boxes $\mathbf{b}_i \in \mathbb{R}^6$, we establish edges based on two criteria: spatial proximity (geometric distance d_{ij} between centroids $\mathbf{c}_i$ and $\mathbf{c}_j$) and morphological similarity (feature cosine similarity s_{ij}). An edge (i, j) is created if $d_{ij} < \tau_d$ or $s_{ij} > \tau_s$, enabling the graph to capture both adjacent lesions and distant metastases.

For each edge, we encode the relationship by constructing edge features:

$$\mathbf{e}_{ij} = \text{MLP}\big([\mathbf{h}_i \| \mathbf{h}_j \| \mathbf{g}^s_{ij} \| \mathbf{g}^m_{ij}]\big), \quad (3)$$

where $\mathbf{g}^s_{ij} = [d_{ij}, \mathbf{r}_{ij}, v_i/v_j]$ encodes spatial features (distance, relative direction, volume ratio), $\mathbf{g}^m_{ij} = [s_{ij}, \bar{I}_i, \bar{I}_j]$ encodes morphological features

(similarity and mean intensities $\bar{I}$ from CT/PET at RoI locations), and $\|$ denotes concatenation. We apply GATv2 (Brody et al., 2022) with its default attention mechanism for message passing, producing enhanced representations $\mathbf{h}'_i$ that incorporate contextual information from spatially adjacent and morphologically similar lesions.

Semantic-Injected LLM Decoder. Instead of relying solely on visual tokens, HiRRA employs a description-guided prompting strategy. The LLM decoder input consists of a composite prompt: (1) patient demographics and clinical history; (2) global visual tokens `<GLOBAL_VISUALS>`; and (3) structured RoI information `RoI` i: `<RoI_VISUALS> | {RoI-description}`, where `<RoI_VISUALS>` represents projected enhanced features $\mathbf{h}'_i$ and `{RoI-description}` contains structured clinical attributes (SUVmax, size, FDG).

Training Strategy. We implement a four-stage training paradigm, wherein tailored prompting schemes are designed for each phase to facilitate progressive alignment and learning.

Stage 1: Encoder Pretraining. We pretrain dual encoders on ViMed-PET (Nguyen et al., 2025) using CLIP-style contrastive learning, aligning CT-PET features with Vietnamese text reports.

Stage 2: Vision-Language Alignment. With encoders and LLM frozen, we train only the Q-former and projection layers. The prompt contains only patient metadata and `<GLOBAL_VISUALS>` tokens for image-report alignment.

Stage 3: Region-level Learning. We integrate local context modules (SPP-RoI, GATv2), freezing Stage 2 components. The prompt extends to full structure with `<RoI_VISUALS>` tokens and RoI description slots. Using bounding box supervision, the model generates RoI-level attribute descriptions.

Stage 4: Instruction Tuning. We fine-tune end-to-end via LoRA (Hu et al., 2021) ($r = 16$, $\alpha = 32$) with unfrozen RoI vision modules. Using the full prompt template, the model synthesizes RoI findings into complete Vietnamese reports.

5 Evaluation

5.1 Evaluation Metrics

To ensure a comprehensive assessment of both linguistic fluency and clinical validity, we employ a dual-evaluation strategy comprising standard linguistic metrics and a proposed clinical protocol.

Standard NLP Metrics. We report n-gram and embedding-based metrics. i.e., BLEU, ROUGE, BERTScore to evaluate lexical and semantic similarity. All scores are reported as percentages.

Proposed Clinical Evaluation Protocol. While the standard linguistic metrics effectively measure the text overlap, they often fail to capture the clinical correctness, anatomical precision, and diagnostic hierarchy essential for 3D PET/CT reporting. To this end, we introduce a structured protocol assessing both the correctness and semantic fidelity of the RoIs identified in the generated reports. To enable this, we first utilize *LangExtract* (Goel, 2025) to parse both ground truth and generated reports into structured objects defined by five key clinical fields: $E = \{\text{region, lesion, density, morphology, fdg\_uptake}\}$. Based on this structured representation, we define two clinical metrics as follows.

RoI Coverage - *Quantitative Identification*. We frame RoI evaluation as a bipartite matching problem. Specifically, we compute a pairwise similarity matrix between the set of predicted RoI text spans $\hat{\mathcal{R}}$ and ground-truth spans $\mathcal{R}$, utilizing cosine similarity within a BERT-based embedding space. Optimal one-to-one alignment is established via the Hungarian matching algorithm (Kuhn, 1955). Pairs exceeding similarity threshold τ are defined as True Positives (TP), forming the basis for Precision, Recall, and F1-score calculations.

RoI Quality Index - *Qualitative Fidelity*. For matched RoI pairs, we evaluate attribute-specific accuracy using the proposed RoI Quality Index (RoIQ). This metric is designed to enforce a clinical hierarchy, operationalizing the principle that accurate anatomical region and lesion type identification are fundamental prerequisites for a valid diagnostic description. To reflect this, we define S_{region} and S_{lesion} as the similarity scores for these critical attributes, while $\mathcal{A}_{\text{valid}}$ represents the set of additional non-empty attributes extracted from the clinical fields. The RoIQ is formulated as follows:

$$\text{RoIQ} = \sqrt{S_{\text{region}} \cdot S_{\text{lesion}}} \times \left(\frac{1}{|\mathcal{A}_{\text{valid}}|} \sum_{k \in \mathcal{A}_{\text{valid}}} S_k \right).$$

The first term, represented by the geometric mean of the core attributes, serves as a non-linear penalty for hallucinations in critical anatomical or pathological contexts. By structuring the metric in this manner, we ensure that high performance in secondary descriptors (e.g., size or morphology) cannot compensate for fundamental errors in localization or lesion categorization. Consequently, RoIQ

provides a more clinically grounded assessment of report quality than standard token-based overlaps, emphasizing diagnostic reliability.
Our metrics were informed by the clinical intuition of physicians from a hospital in Vietnam, and their medical correctness and clinical usefulness were subsequently validated by these clinicians.

5.2 Experimental Goals

To comprehensively evaluate the VietPET-RoI benchmark and the proposed HiRRA framework, we investigate the following research questions:
RQ1 - Performance on Low-Resource Languages: We benchmark representative VLMs on VietPET-RoI to assess their capability in handling native Vietnamese medical reports versus translated data in English versions.
RQ2 - Hallucination Mitigation and RoI Grounding: We evaluate the efficacy of the HiRRA pipeline compared to fine-tuned baselines in reducing hallucinations. This involves our new metric that measures both the quantitative recall of detected regions and the qualitative semantic accuracy of their descriptions.
RQ3 - Performance on Clinical Tasks: We assess model performance on clinical tasks, specifically differential diagnosis prediction and RoI description generation, to demonstrate dataset versatility.
RQ4 - Impact of Multimodal Fusion: We compare single-modality (CT or PET only) versus joint CT+PET fusion to assess the necessity of integrating anatomical and metabolic information for accurate RoI characterization.

5.3 Model selection

To establish a robust benchmark for *VietPET-RoI*, we evaluate representative state-of-the-art 2D and 3D vision-language models. For 2D modeling, we compare *InternVL3* (Chen et al., 2024b; OpenGVLab, 2025), a large-scale generalist foundation model, against *MedGemma* (Sellergren et al., 2025), a specialized medical backbone. For 3D volumetric reasoning, we incorporate four leading medical VLMs: *RadFM* (Wu et al., 2025) and *MedM-VL* (Shi et al., 2025) as general-purpose backbones with large-scale multimodal pretraining, and *Med3DVLM* (Xin et al., 2025) and *M3D-LaMed* (Bai et al., 2024) with architectures optimized for localization and report generation. While models like *CT2Rep* (Hamamci et al., 2024a) and *PETAR* (Maqbool et al., 2025) advance report generation, they have distinct structural limitations: CT2Rep is restricted to global, CT-only volume-to-report mapping, and PETAR requires pre-existing oracle lesion masks to generate isolated lesion captions. In contrast, our framework uniquely performs end-to-end RoI discovery, structured multi-attribute modeling, and explicit inter-RoI relational reasoning via GATv2 to synthesize coherent whole-body PET/CT reports

Table 3: **Benchmarking VLMs on VietPET-RoI.** Results are reported for Vietnamese (VI) and English (EN) inference. Best scores for VI are in **bold** and underline for EN. All metrics are reported as percentages (%).

Model	Lang	BLEU-4	R-1	R-L	BERT
InternVL3 (2D)	VI	0.25	**28.20**	**18.28**	64.05
	EN	0.94	16.64	11.05	81.50
MedGemma (2D)	VI	**0.82**	25.59	17.61	65.36
	EN	0.39	13.58	8.87	80.12
RadFM	VI	0.39	19.09	13.84	**83.33**
	EN	0.44	12.32	8.78	80.68
MedM-VL	VI	0.13	1.41	1.36	47.28
	EN	0.93	22.02	15.15	84.27
Med3DVLM	VI	0.17	1.31	1.24	62.67
	EN	0.25	12.54	9.44	67.14
M3D-LaMed	VI	0.15	1.18	1.08	62.92
	EN	0.45	13.42	9.96	67.44

5.4 Performance on Low-Resource Languages

To evaluate the capability of existing medical VLMs on low-resource languages as proposed in RQ1, we benchmark state-of-the-art 2D and 3D models on VietPET-RoI. Results in Table 3 reveal that all models perform extremely poorly on Vietnamese generation, with BLEU-4 scores near zero (MedGemma: 0.82, Med3D-VLM: 0.17, M3D-LaMed: 0.15). This collapse arises from a pronounced mismatch between existing model pretraining, which primarily covers English radiology text and 2D or 3D CT volumes, and the VietPET-RoI setting that requires reasoning over Vietnamese clinical language, 3D PET/CT volumes, and structured region-level clinical descriptions. In addition, VietPET-RoI requires models to describe metabolic activity (SUV_{max}), lesion morphology, and precise anatomical regions, making it a challenging benchmark that tests true PET/CT understanding rather than surface-level text generation.

5.5 Hallucination Mitigation and RoI Grounding

To address RQ2, we compare HiRRA against three top-performing 3D VLMs (MedM-VL (Shi et al., 2025), Med3DVLM (Xin et al., 2025), M3D-LaMed (Bai et al., 2024)) fully fine-tuned on VietPET-RoI. Additionally, we evaluate HiRRA

Table 4: **Quantitative benchmarking against state-of-the-art methods.** We evaluate the generation quality (NLP Metrics) and clinical alignment (RoI Clinical Metrics). *Correct* indicates the number of successfully grounded regions (out of 416). **Bold** denotes the best performance. The rows highlighted in gray show our model's performance and its relative improvement (↑) over the second-best baseline.

Method	Generation Quality (NLP)				Clinical RoI Alignment					
	BLEU-4	R-1	R-L	BERT	Correct	Total	Prec.	Rec.	F1	RoIQ
MedM-VL	31.69	67.11	50.00	91.92	62	416	12.88	19.12	14.16	23.24
Med3D-VLM	45.53	71.06	62.51	86.49	179	416	31.07	43.02	36.08	39.00
M3D-LaMed	44.30	74.14	64.39	85.90	193	416	34.90	46.39	39.83	36.31
HiRRA (No RoI)	52.48	77.75	66.51	95.13	187	416	32.29	44.95	37.58	33.89
HiRRA (Ours)	**62.80**	**80.40**	**69.66**	**95.79**	**223**	416	**35.17**	**53.60**	**42.47**	**56.86**
Improv. (Δ%)	+19.7%	+3.4%	+4.7%	+0.7%	+15.5%	-	+0.8%	+15.5%	+6.6%	**+45.8%**

Table 5: **Clinical Prediction and RoI Description Generation performance (RQ3)**. We report Recall (Rec.) and Precision(Prec.) for clinical tasks, and standard NLP metrics (BLEU-4, ROUGE-L) for description generation. Best results are highlighted in bold.

Model	Disease Pred.		Exam. Pred.		RoI Desc.	
	Prec.	Rec.	Prec.	Rec.	B-4	R-L
M3D-LaMed	67.4	73.2	21.6	21.3	36.0	69.3
Med3D-VLM	76.5	79.9	41.0	42.3	**62.3**	75.8
MedM-VL	**95.0**	**98.7**	**51.7**	**52.0**	50.3	**78.8**

under two configurations: *No RoI* (global context only) and *Full Pipeline* (with RoI features). Evaluation employs both traditional NLP metrics and our proposed clinical metrics: *RoI Coverage* and *RoI Quality Index*.

Results in Table 4 show that HiRRA Full Pipeline achieves state-of-the-art performance across all metrics. For generation quality, HiRRA attains BLEU-4 of 62.80 and BERTScore of 95.79, substantially outperforming the strongest baseline (Med3DVLM). More importantly, HiRRA demonstrates significant improvements in clinical alignment with 45.8% gain in RoIQ and 15.5% increase in RoI Recall, indicating superior hallucination reduction. Ablation comparison confirms the critical role of RoI: *HiRRA No RoI* maintains strong fluency (BERT 95.13) but achieves only 0.3389 in RoIQ, proving that global context alone is insufficient for clinical accuracy. This validates that RoI grounding is essential for bridging visual perception with accurate diagnostic reporting.

5.6 Performance on Clinical Tasks

We assess whether VietPET-RoI supports multitask learning by training models on two tasks: differential diagnosis and further examinations prediction, structured RoI description generation. Results are summarized in Table 5.

For clinical prediction, MedM-VL (Shi et al., 2025) achieves the highest Recall on disease prediction (98.7%), while examination prediction remains significantly lower (52.0%), indicating that follow-up recommendation requires more complex reasoning. For RoI description generation, models achieve strong performance with BLEU-4 up to 62.3% and ROUGE-L of 78.8%, demonstrating that VietPET-RoI provides effective fine-grained supervision for generation tasks. These results confirm that VietPET-RoI supports multitask learning, creating potential for a complete automated pipeline: from RoI segmentation to region-specific description generation, faithfully mirroring clinical workflows and paving the way for future diagnostic support systems.

5.7 Impact of Multimodal Fusion

Unlike prior approaches often limited to single-modality inputs (e.g., ViMed-PET (Nguyen et al., 2025)), Table 6 confirms the necessity of multimodal fusion. The joint CT+PET framework achieves superior performance (36.1 BLEU-4), significantly outperforming single-modality baselines by a margin of 2.6–4.9 points. This validates that integrating anatomical structure (CT) (Kinahan et al., 2003) with metabolic intensity (PET) (Townsend, 2008; Von Schulthess et al., 2009) is indispensable for accurate lesion description, effectively resolving the perceptual bottlenecks inherent in unimodal processing. Specifically, this cross-modal synergy enables the disambiguation of hypermetabolic signals, allowing the model to distinguish between pathological lesions and physiological uptake (e.g., in the brain or bladder) through anatomical referencing, which significantly enhances the report's clinical validity.

Table 6: **Impact of Multimodal Fusion.** Quantitative ablation demonstrating the superiority of the joint CT+PET framework over single-modality baselines.

Input Modality	BLEU-4	R-1	R-L	BERT
CT only	31.3	65.4	56.3	92.7
PET only	33.6	66.3	58.0	93.1
CT + PET	**36.1**	**69.3**	**59.4**	**94.0**

6 Conclusion

We addressed the scarcity of fine-grained PET/CT datasets by introducing a novel dataset that comprises 600 region-level samples with 1,960 comprehensively annotated RoIs, representing the first 3D PET/CT dataset with fine-grained RoI-level annotations. We also proposed HiRRA, a novel VLM architecture that integrates local RoI features and global context for report generation. Additionally, we introduced clinical evaluation metrics (RoI Coverage and RoI Quality Index) designed specifically for medical report generation assessment. Models trained on VietPET-RoI demonstrated substantial improvements, achieving 19.7% and 4.7% gains over the strongest baseline method (M3D-LaMed) in BLEU-4 and ROUGE-L, alongside 15.5% and 45.8% improvements in clinical metrics, indicating enhanced reliability and reduced hallucination.

Limitations

Despite its fine-grained RoI annotations, VietPET-RoI has several limitations. First, the dataset scale remains modest with 200 patients from a single institution, which may affect generalization and introduce potential demographic bias. Expanding to multiple medical centers would enhance diversity. Second, the RoI annotation process requires specialized medical expertise and is time-intensive (averaging 15-30 minutes per case), limiting rapid scaling. Third, due to the dataset's uniqueness, we lack established benchmarks for comprehensive evaluation, particularly for tasks such as automatic RoI segmentation. While we have established baseline methods for report generation, developing benchmarks for intermediate tasks remains an important research direction. Finally, the 3D architecture of HiRRA demands substantial computational resources, which may limit deployment in resource-constrained settings. We believe VietPET-RoI establishes an important foundation for region-aware PET/CT research and encourages the community to develop additional methods and benchmarks in this domain.

Ethical Considerations

This study was conducted in strict adherence to medical research ethics standards and the ACM Code of Ethics. The protocol was reviewed and approved by the Ethics Committee of the data-providing institution.

Informed Consent and Privacy: Due to the retrospective nature of the study, the Institutional Review Board (IRB) granted a formally validated waiver of informed consent, determining that the research poses minimal risk to subjects. All patient data underwent a rigorous de-identification process validated by institutional authorities to ensure complete anonymity. Personal Protected Information (PPI) was removed in compliance with international standards (e.g., HIPAA Safe Harbor). Consequently, there are no privacy risks associated with the public release of this dataset.

Intended Use and Fairness: VietPET-RoI is released exclusively for non-commercial research to support medical AI development. It is not a diagnostic tool and must not replace clinical judgment; any deployment requires further rigorous clinical validation. We acknowledge that data sourced from a single institution may contain inherent demographic biases. To mitigate potential harm, we advise future researchers to validate models on diverse external populations before clinical consideration.

The dataset will be publicly released for non-commercial research purposes under specified usage terms upon paper acceptance. Given the complete anonymization process, there are no privacy concerns associated with public release. All ethical considerations have been reviewed and validated by institutional authorities, ensuring compliance with medical research ethics standards.

Acknowledgements

This research is partially supported by Toray Industries (H.K.) Ltd. Vietnam. We also extend our deepest gratitude to the physicians and staff at the collaborating hospital for their expertise in data curation and RoI annotation. This research was made possible by the contributions of patients whose anonymized medical records form the basis of this dataset.

References

Fan Bai and 1 others. 2024. Advancing 3d medical image analysis with multi-modal large language models. *Preprint*. M3D-LaMed: Generalist 3D medical vision–language model.

Thomas Beyer, Gerald Antoch, Stefan Müller, Thomas Egelhof, Lutz S Freudenberg, Jörg Debatin, and Andreas Bockisch. 2004. Acquisition protocol considerations for combined pet/ct imaging. *Journal of Nuclear Medicine*, 45(1 suppl):25S–35S.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, and 1 others. 2022. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer.

Shaked Brody, Uri Alon, and Eran Yahav. 2022. How attentive are graph attention networks? *Preprint*, arXiv:2105.14491.

Kai Chen, Zhihong Chen, Yalun Gu, Wentao Wu, and Wanli Ouyang. 2024a. Towards systematic evaluation of hallucination for large vision language models in the medical context. *arXiv preprint arXiv:2408.08472*.

Z. Chen and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual–linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198.

Daniel Coelho de Castro, Aurelia Bustos, Shruthi Bannur, Stephanie L Hyland, Kenza Bouzid, Maria Teodora Wetscherek, Maria Dolores Sánchez-Valverde, Lara Jaques-Pérez, Lourdes Pérez-Rodríguez, Kenji Takeda, and 1 others. 2025. Padchest-gr: A bilingual chest x-ray dataset for grounded radiology report generation. *NEJM AI*, 2(7):AIdbp2401120.

D Dhouib, A Naït-Ali, C Olivier, and MS Naceur. 2021. Roi-based compression strategy of 3d mri brain datasets for wireless communications. *IRBM*, 42(3):146–153.

Sergios Gatidis, Tobias Hepp, Marcel Früh, Christian La Fougère, Konstantin Nikolaou, Christina Pfannenberg, Bernhard Schölkopf, Thomas Küstner, Clemens Cyran, and Daniel Rubin. 2022. A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. *Scientific Data*, 9(1):601.

Akshay Goel. 2025. LangExtract.

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. 2024a. Ct2rep: Automated radiology report generation for 3d medical imaging. *Preprint*, arXiv:2403.06801.

Ibrahim Ethem Hamamci, Sezgin Er, Anjany Sekuboyina, Enis Simsar, Alperen Tezcan, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Furkan Almas, Irem Dogan, Muhammed Furkan Dasdelen, Chinmay Prabhakar, Hadrien Reynaud, Sarthak Pati, Christian Bluethgen, Mehmet Kemal Ozdemir, and Bjoern Menze. 2024b. Generatect: Text-conditional generation of 3d chest ct volumes. *Preprint*, arXiv:2305.16037.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Young-jun Hu, Da-hye Kang, Jeong Won Hwang, Hwan-Ho Cho, Sang-Keun Lee, Mee-Hye Kim, Won-Jun Jung, and Kyoungjune Yang. 2024. Graph neural network model for prediction of non-small cell lung cancer lymph node metastasis using protein-protein interaction network and 18f-fdg pet/ct radiomics. *International Journal of Molecular Sciences*, 25(2):698.

Wenpei Jiao, Kun Shang, Hui Li, Ke Yan, Jiajin Zhang, Guangjie Yang, Lijuan Guo, Yan Wan, Xing Yang, Dakai Jin, and Zhaoheng Xie. 2025. Vision-language models for automated 3d pet/ct report generation. *arXiv preprint arXiv:2511.20145*.

Paul E Kinahan, Bruce H Hasegawa, and Thomas Beyer. 2003. X-ray-based attenuation correction for positron emission tomography/computed tomography scanners. *Seminars in Nuclear Medicine*, 33(3):166–179. Key kept as requested, original paper 2003.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. *Preprint*, arXiv:1612.03144.

Danyal Maqbool, Changhee Lee, Zachary Huemann, Samuel D. Church, Matthew E. Larson, Scott B. Perlman, Tomas A. Romero, Joshua D. Warner, Meghan Lubner, Xin Tie, Jameson Merkow, Junjie Hu, Steve Y. Cho, and Tyler J. Bradshaw. 2025. Petar: Localized findings generation with mask-aware vision-language modeling for pet automated reporting. *Preprint*, arXiv:2510.27680.

Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo Andía, Cristian Tejos, Claudia Prieto, and Daniel Capurro. 2022. A survey on deep learning and explainability for automatic report generation from medical images. *ACM Computing Surveys (CSUR)*, 54(10s):1–40.

Peter Muzi, Michelle Wanner, and Paul Kinahan. 2015. Data from rider lung pet-ct. *(No Title)*.

Huu Tien Nguyen, Dac Thai Nguyen, The Minh Duc Nguyen, Trung Thanh Nguyen, Thao Nguyen Truong, Huy Hieu Pham, Johan Barthelemy, Minh Quan Tran, Thanh Tam Nguyen, Quoc Viet Hung Nguyen, and 1 others. 2025. Toward a vision-language foundation model for medical data: Multimodal dataset and benchmarks for vietnamese pet/ct report generation. *arXiv preprint arXiv:2509.24739*.

OpenGVLab. 2025. Internvl3: Advanced multimodal large language models. https://internvl.github.io/blog/2025-04-11-InternVL-3.0/. Accessed 2025-12-01.

Andrés Ortiz, Juan M Górriz, Javier Ramírez, Francisco J Martinez-Murcia, and Alzheimer's Disease Neuroimaging Initiative. 2014. Automatic roi selection in structural brain mri using som 3d projection. *PloS one*, 9(4):e93851.

Steve Pieper, Michael Halle, and Ron Kikinis. 2004. 3d slicer. In *2004 2nd IEEE international symposium on biomedical imaging: nano to macro (IEEE Cat No. 04EX821)*, pages 632–635. IEEE.

A. Sellergren and 1 others. 2025. Medgemma: A collection of medical vision–language foundation models based on gemma 3. https://developers.google.com/health-ai-developer-foundations/medgemma. Google Health AI Developer Foundations, accessed 2025-12-01.

Hyeonsoo Seo, Colin Huang, Maxime Bassenne, Ran Xiao, and Lei Xing. 2021. Lymph node graph neural networks for cancer metastasis prediction. *arXiv preprint arXiv:2106.01711*.

Y. Shi and 1 others. 2025. Medm-vl: What makes a good medical lvlm? In *Medical Image Computing and Computer-Assisted Intervention – MICCAI*, Lecture Notes in Computer Science. Springer.

David W Townsend. 2008. Multimodality imaging of structure and function. *Physics in Medicine & Biology*, 53(4):R1.

Gustav K Von Schulthess, Hans C Steinert, and Thomas F Hany. 2009. Integrated pet/ct: current applications and future directions. *Radiology*, 251(3):708–736.

Stephen Waite, Arkadij Grigorian, Robert G Alexander, Stephen L Macknik, Marisa Carrasco, David J Heeger, and Susana Martinez-Conde. 2019. Analysis of perceptual expertise in radiology–current knowledge and a new perspective. *Frontiers in human neuroscience*, 13:213.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Towards generalist foundation model for radiology by leveraging web-scale 2d & 3d medical data. *Nature Communications*.

Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, and 1 others. 2024. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*.

Yu Xin, Gorkem Can Ates, Kuang Gong, and Wei Shao. 2025. Med3dvlm: An efficient vision–language model for 3d medical image analysis. *IEEE Journal of Biomedical and Health Informatics*. To appear.

Junsan Zhang, Ming Cheng, Qiaoqiao Cheng, Xiuxuan Shen, Yao Wan, Jie Zhu, and Mengxuan Liu. 2024. Hierarchical medical image report adversarial generation with hybrid discriminator. *Artificial Intelligence in Medicine*, 151:102846.

Yichi Zhang, Wenbo Zhang, Zehui Ling, Gang Feng, Sisi Peng, Deshu Chen, Yuchen Liu, Hongwei Zhang, Shuqi Wang, Lanlan Li, and 1 others. 2025. Pet2rep: Towards vision-language model-drived automated radiology report generation for positron emission tomography. *arXiv preprint arXiv:2508.04062*.

Supplementary Material

A Detailed Dataset Construction Process

A.1 Dataset Collection and Pre-processing

The dataset was curated by oncology specialists at a leading public hospital in Vietnam to ensure clinical accuracy and authenticity. After screening, 200 cancer cases spanning the six most common malignancies were selected. These cases were drawn from examinations performed between 2017 and 2019 and cover a diverse range of sex, age, and body habitus, supporting the development of models with improved generalization (see Table 2).

Clinical PET/CT at the site is acquired using several protocols (Beyer et al., 2004) that differ in image format and slice count. To reduce protocol-related variability, all studies were standardized to a single whole-body protocol with 313 axial slices from the head to the upper thighs. The raw PET/CT volumes were exported in DICOM format, containing rich metadata such as patient information, body weight, standardized uptake values (SUV), and other physiological attributes. The corresponding reports were collected as physician-typed Word documents (DOC), following a semi-structured template with fields including "Gender," "Examination Date," "Indication," "Clinical History," "Impression," and "Image Description.".

After collection, a series of preprocessing steps was applied to obtain model-ready inputs. For the textual reports, all patient-identifiable information was removed to satisfy ethical and privacy requirements. The de-identified reports were then converted from DOCX to a standardized JSON schema, enabling direct alignment with the corresponding images during training. Automatic extraction was followed by manual review to correct spelling or formatting inconsistencies and to ensure the reliability of key fields. For the 3D PET/CT data, the DICOM volumes were converted to NumPy array (NPY) format. To increase the number of samples and refine the supervision signal, each whole-body study (image plus report) was further divided into three anatomically meaningful regions: head–neck, chest, and abdomen–pelvis. Neighboring regions were defined with a 15-slice overlap to preserve continuity and avoid missing important findings at segment boundaries. Concretely, the head–neck region corresponds to the first 25% of slices in the scan; the chest region starts 25% below the last slice of the head–neck segment; and the abdomen–pelvis region covers the remaining slices from the chest down to the pelvis. These proportions were validated by clinical experts and empirically verified to ensure that each regional volume fully covers all relevant RoIs for that anatomical area. This regional segmentation increases the effective sample size and improves the alignment between visual content and textual descriptions, enabling more precise region-level learning and, in turn, better overall model performance.

Table 7: Schema of the region-of-interest (RoI) annotation template.

Field	Description
Anatomic region	Anatomical location of the finding (e.g., left supraclavicular, liver segment VI, right ovary).
Lesion type	Type of lesion or physiological structure; physiological findings are explicitly labeled as "physiological".
Size	In-plane lesion size in millimeters.
SUVmax	Maximum standardized uptake value of the lesion, when available.
Density	Qualitative CT density / attenuation (e.g., hypodense, soft-tissue density).
Morphology	Morphological appearance (e.g., rounded, irregular, thickened wall, nodule).
FDG uptake	Qualitative FDG metabolism (e.g., increased or non-increased uptake).
Top-3 possible diseases	Three most likely diagnoses, listed in descending order of likelihood.
Top-3 further examinations	Three recommended follow-up tests or diagnostic procedures, ordered by priority.
Physical region	Coarse body region (1 = head–neck, 2 = chest, 3 = abdomen–pelvis).
Note	Optional free-text comments if needed.

A.2 Region-of-Interest (RoI) Annotation

From the three-region split of each of the 200 cases, we obtained 600 image–report pairs for region-

level supervision. We then collaborated with eight physicians to annotate regions of interest (RoIs) and provide structured descriptions. Annotation was performed in 3DSlicer (Pieper et al., 2004): every abnormal finding or clinically relevant physiological structure mentioned in the report was localized on the corresponding PET/CT volume and marked with a 3D bounding box. For each RoI, annotators completed a compact structured template that records key clinical attributes, as summarized in Table 7. 3DSlicer (Pieper et al., 2004) exports the coordinates of each RoI together with its structured string description in this format, resulting in fine-grained, region-level supervision for every study. By linking spatial localization with rich clinical semantics, the annotations provide strong guidance for report generation models to focus on clinically important regions and help reduce irrelevant or hallucinated content in the generated reports (Dhouib et al., 2021; Ortiz et al., 2014).

B Detailed RoI Metrics

B.1 Ground Truth RoI Annotation Schema

To capture the semantic richness of PET/CT reports, our dataset utilizes a structured annotation schema for Regions of Interest (RoIs). Each ground truth RoI is represented as a structured string containing 11 distinct clinical attributes. The annotation format is defined as follows:

```
[Anatomic Region] - [Lesion Type] - [Size] - [SUVmax] - [Density] - [Morphology] - [FDG Uptake] - [Top-3 Differential Diagnoses] - [Top-3 Recommended Examinations] - [Physical Region ID] - [Clinical Note]
```

Example Instance:

```
[Cecum] - [Focal hypermetabolism] - [Unclear] - [12.3] - [Soft tissue density] - [Focal] - [Very intense hypermetabolism] - [Colon cancer (cecum), Inflammatory bowel disease (Crohn's disease), Appendicitis/Abscess] - [Colonoscopy and biopsy, Abdominal MRI/CT, Blood tests] - [3] - [Very intense focal FDG uptake (SUVmax 12.3) in the cecum. Highly suggestive of colon cancer...]
```

B.2 Information Extraction Framework

To evaluate generation quality, we employ *LangExtract* (Goel, 2025), an LLM-based extraction framework designed to parse unstructured generated reports into structured RoI objects. Specifically, we utilize Gemini-2.5-Pro as the backbone LLM. For every predicted sentence in a report, the framework extracts the information into the following structured JSON format:

```
{
  "extraction_text": "...", // The original sentence text
  "anatomic_region": "...",
  "lesion_type": "...",
  "density": "...",
  "morphology": "...",
  "fdg_uptake": "..."
}
```

While the ground-truth RoI schema contains 10 clinical attributes and 1 data management identifier, the primary objective of our clinical metrics (RoI Coverage and RoIQ) is to assess the correctness and semantic fidelity of the localized lesions. After consultation with clinical partners, we determined that only the five extracted fields above directly reflect this capability, as they characterize the observable properties of a lesion. The remaining six attributes were excluded for the following methodological reasons:

- **Data Management Artifacts:** The *physical region code* is a preprocessing identifier used solely to split full-body scans into anatomical segments during dataset construction, rather than a clinically observable feature.
- **Unstructured Context:** *Additional notes* consist of free-text commentary derived from the complete clinical report. These contain auxiliary observations extending beyond the scope of individual RoIs and cannot be systematically evaluated through field-level matching.
- **Downstream Diagnostic Reasoning:** *Top-3 possible diseases* and *top-3 further examinations* are synthesized from the entire report context. They exhibit high inter-annotator variability and are downstream from RoI identification-they depend on correct lesion characterization but do not define it.
- **Numerical Regression Constraints:** *Size* and *SUVmax* are quantitative measurements requiring precise numerical computation from raw NumPy imaging arrays. Evaluating these fields would conflate the model's spatial identification ability with numerical regression, which current VLMs cannot reliably perform from visual tokens alone.

Therefore, these five attributes isolate and directly measure the core capability our metrics are

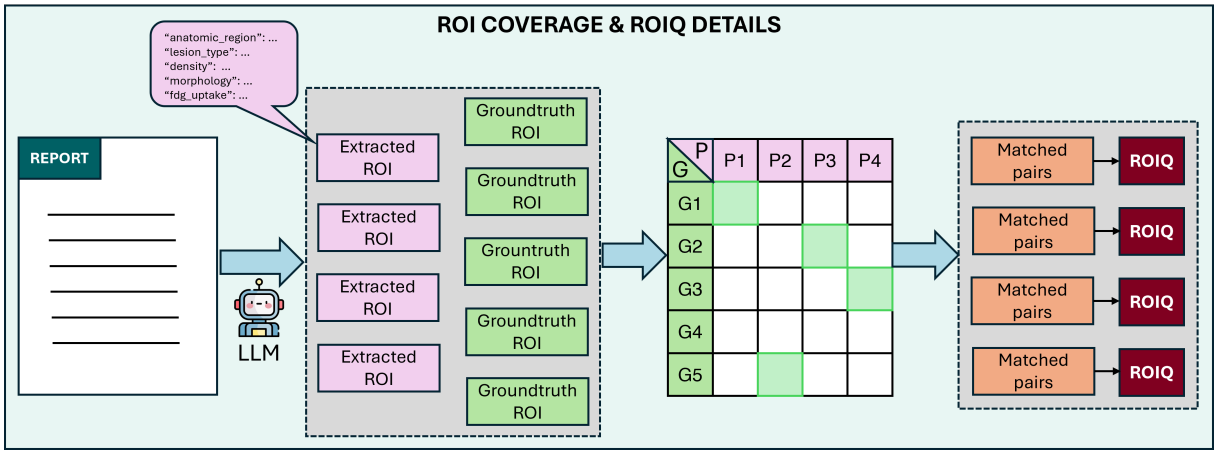

Figure 5: **Overview of our proposed clinical evaluation protocol.** We utilize an LLM-based framework to extract structured clinical attributes from reports. RoI Coverage is quantified by aligning predicted and ground-truth RoIs via embedding-based Hungarian matching. For aligned pairs, the RoI Quality Index (RoIQ) measures semantic fidelity, strictly enforcing anatomical and lesion-type correctness.

designed to assess: whether the model correctly identifies and semantically characterizes lesions as observable entities.

B.3 Quantitative Evaluation: RoI Coverage

We introduce **RoI Coverage** to evaluate lesion localization as a set-based detection problem. We compare the set of Ground Truth RoIs (G) against the set of Predicted RoIs (P) extracted by LangExtract.

B.3.1 Embedding and Similarity Calculation

For each RoI, we isolate five comparable text fields: *anatomic region, lesion type, density, morphology,* and *fdg uptake*. We utilize a clinical embedding model to convert the text value of each field into a vector representation. For a predicted RoI $p_i \in P$ and a ground truth RoI $g_j \in G$, we calculate the similarity score $S(p_i, g_j)$ based on the cosine similarities of their constituent fields (excluding `extraction_text`).

B.3.2 Optimal Matching (Hungarian Algorithm)

We construct a similarity matrix $M \in \mathbb{R}^{|P|\times|G|}$ where each entry $M_{i,j}$ represents the similarity between p_i and g_j. To resolve the alignment, we apply the Hungarian Algorithm (Kuhn, 1955) (Linear Sum Assignment) to find the optimal set of pairs that maximizes the total similarity score. A pair (p_i, g_j) is considered a valid match (True Positive) only if their similarity score exceeds a predefined threshold τ:

$$\text{Match}(p_i, g_j) = \mathbb{1}[M_{i,j} \geq \tau]$$

B.3.3 Classification Metrics

Based on valid matches, we quantify **RoI Coverage** using True Positives (TP), False Positives (FP), and False Negatives (FN):

- TP: Successfully matched ground truth RoIs.
- $FP = |P| - TP$: Predicted RoIs that did not match any ground truth (hallucinations).
- $FN = |G| - TP$: Ground truth RoIs that were not detected.

Standard classification metrics are then computed as:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}$$
$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

B.4 Qualitative Evaluation: RoI Quality Index (RoIQ)

For every successfully matched pair of RoIs, we introduce the **RoI Quality Index (RoIQ)** to assess the semantic accuracy of the generated details. This metric prioritizes the correct identification of critical attributes (region and lesion type) before evaluating descriptive attributes (density, morphology, FDG uptake).

Let $S_{\text{region}}, S_{\text{lesion}}, S_{\text{density}}, S_{\text{morphology}}, S_{\text{uptake}}$ denote the similarity scores for the five fields. We define the set of secondary attributes as $\mathcal{A} = \{S_{\text{density}}, S_{\text{morphology}}, S_{\text{uptake}}\}$. Since descriptive attributes may be absent in the ground truth, we compute the mean score only over the subset

of non-empty attributes, $\mathcal{A}_{\text{valid}} \subseteq \mathcal{A}$. The RoIQ is defined as:

$$\text{RoIQ} = \sqrt{S_{\text{region}} \cdot S_{\text{lesion}}} \times \left(\frac{1}{|\mathcal{A}_{\text{valid}}|} \sum_{s \in \mathcal{A}_{\text{valid}}} s \right)$$

Note: If a descriptive attribute is missing in the ground truth, it is excluded from $|\mathcal{A}_{\text{valid}}|$ to prevent penalizing the model for not generating non-existent features.

B.5 Sensitivity Analysis of Similarity Threshold (τ)

We conducted an experiment to evaluate the sensitivity of the similarity threshold τ across representative values on a held-out development set of 60 samples (240 RoI pairs) via grid search over $\tau \in [0.5, 0.95]$ with a step size of 0.05.

According to the experimental results presented in Table 8, decreasing τ to 0.55 increases Recall (0.502) but substantially degrades Match Quality ($RoIQ = 0.501$), indicating that the metric becomes overly permissive and accepts semantically distinct pairs as valid matches. Conversely, increasing τ to 0.80 causes a significant drop in both Recall (0.337) and F1 (0.295), penalizing clinically valid semantic variations such as synonymous radiological expressions.

The selected $\tau = 0.70$ achieves the best trade-off between comprehensive lesion coverage and semantic precision, and is therefore adopted as our default threshold.

Table 8: Sensitivity analysis of the similarity threshold (τ) on the validation set (60 samples).

Threshold (τ)	F1	Recall	Precision	RoIQ
0.55	0.435	0.502	0.368	0.501
0.70 (Selected)	**0.371**	**0.422**	**0.332**	**0.573**
0.80	0.295	0.337	0.264	0.525

C Implementation Details

We implemented our framework using PyTorch and the HuggingFace `transformers` library. All experiments were conducted on a single NVIDIA A100 (80GB) GPU. We provide the detailed training hyperparameters and optimization settings in Table 9.

C.1 Progressive Training Strategy

Our model employs a four-stage progressive training curriculum to ensure stable convergence and prevent catastrophic forgetting. Starting from vision encoder pretraining on large-scale medical data, we sequentially align global visual features, integrate region-specific information, and perform end-to-end finetuning. This staged approach allows each component to specialize in its designated task before joint optimization, significantly improving training stability compared to end-to-end training from scratch.

In the final end-to-end finetuning stage (Stage 4), all model components are jointly optimized with carefully tuned learning rates to balance exploration and exploitation. We adopt Low-Rank Adaptation (LoRA) (Hu et al., 2021) for parameter-efficient finetuning of the large language model (Qwen3-8B), applying rank-32 adapters with $\alpha = 64$ and dropout rate of 0.1 to the query, key, value, and output projection layers. This reduces trainable parameters by 99.7% while maintaining model expressiveness. To prevent overfitting, we employ early stopping with a patience of 5 epochs, monitoring validation loss with a minimum improvement threshold of 0.001.

C.2 Training Hyperparameters

Learning Rate Strategy. We employ a conservative learning rate strategy to prevent catastrophic forgetting of pretrained knowledge. Vision components (encoder, feature extractors) use an order-of-magnitude lower learning rate (1×10^{-6}) compared to LoRA adapters (5×10^{-6}), allowing the language model to adapt to multimodal inputs while preserving visual representations learned in earlier stages. A cosine annealing scheduler with 10% linear warmup steps gradually reduces learning rates, facilitating smooth convergence.

Regularization. To mitigate overfitting on the relatively small medical dataset, we apply multiple regularization techniques: (1) weight decay of 0.02 on all parameters except biases and layer normalization weights; (2) gradient clipping with maximum norm of 1.0 to stabilize training; and (3) dropout of 0.1 in LoRA adapters. These techniques collectively prevent the model from memorizing training samples while maintaining generalization capability.

Table 9: Training hyperparameters for end-to-end finetuning (Stage 4). Earlier stages use similar configurations with component-specific learning rates detailed in the supplementary materials.

Hyperparameter	**Value**
Optimization	
Optimizer	AdamW
Learning Rate (LoRA)	5×10^{-6}
Learning Rate (Vision)	1×10^{-6}
LR Scheduler	Cosine Annealing
Warmup Ratio	10% of total steps
Weight Decay	0.02
Adam β_1, β_2	$(0.9, 0.999)$
Adam ϵ	1×10^{-8}
Regularization	
Max Gradient Norm	1.0
Dropout (LoRA)	0.1
Batch Configuration	
Batch Size per GPU	1
Effective Batch Size	2
Total Epochs	10
Hardware & Efficiency	
Mixed Precision	BF16
GPU	1 × NVIDIA A100 (80GB)
Number of Workers	8

C.3 Loss Function

Our model is optimized using the standard autoregressive language modeling objective with cross-entropy loss:

$$\mathcal{L}_{\text{LM}} = -\frac{1}{T}\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, \mathbf{x}) \qquad (4)$$

where $\mathbf{x} = [\mathbf{x}_{\text{demo}}, \mathbf{x}_{\text{global}}, \mathbf{x}_{\text{RoI}}]$ represents the multimodal input comprising patient demographics, global visual context ($M = 32$ query vectors), and RoI-specific features (graph-enhanced visual embeddings), y_t denotes the target token at position t, and T is the target sequence length. The loss is computed only on target tokens, with prompt tokens masked by setting their labels to -100. During training, we apply teacher forcing where ground-truth tokens are used as input for predicting subsequent tokens, enabling efficient parallel computation across the sequence dimension.