

Vision-Language Models for Automated 3D PET/CT Report Generation

Wenpei Jiao^{1*}, Kun Shang^{2*}, Hui Li³, Ke Yan^{4,5}, Jiajin Zhang^{4,5,6}
Guangjie Yang⁷, Lijuan Guo⁸, Yan Wan⁹, Xing Yang^{2†}, Dakai Jin^{4†}, Zhaoheng Xie^{1†}

¹Institute of Medical Technology and National Biomedical Imaging Center, Peking University

²Peking University People’s Hospital, ³Peking University Third Hospital

⁴DAMO Academy, Alibaba Group, ⁵Hupan Lab, ⁶Shanghai Jiao Tong University

⁷The Affiliated Hospital of Qingdao University

⁸The First Affiliated Hospital of Henan Medical University

⁹Jiujiang City Key Laboratory of Cell Therapy, Jiu Jiang NO.1 People’s Hospital

yangxing2017@bjmu.edu.cn, dakai.jin@gmail.com, xiezhaoheng@pku.edu.cn

Abstract

Positron emission tomography/computed tomography (PET/CT) is essential in oncology, yet the rapid expansion of scanners has outpaced the availability of trained specialists, making automated PET/CT report generation (PETRG) increasingly important for reducing clinical workload. Compared with structural imaging (e.g., X-ray, CT, and MRI), functional PET poses distinct challenges: metabolic patterns vary with tracer physiology, and whole-body 3D contextual information is required rather than local-region interpretation. To advance PETRG, we propose PETRG-3D, an end-to-end 3D dual-branch framework that separately encodes PET and CT volumes and incorporates style-adaptive prompts to mitigate inter-hospital variability in reporting practices. We construct PETRG-Lym, a multi-center lymphoma dataset collected from four hospitals (824 reports w/ 245,509 paired PET/CT slices), and construct AutoPET-RG-Lym, a publicly accessible PETRG benchmark derived from open imaging data but equipped with new expert-written, clinically validated reports (135 cases). To assess clinical utility, we introduce PETRG-Score, a lymphoma-specific evaluation protocol that jointly measures metabolic and structural findings across curated anatomical regions. Experiments show that PETRG-3D substantially outperforms existing methods on both natural language metrics (e.g., +31.49% ROUGE-L) and clinical efficacy metrics (e.g., +8.18% PET-All), highlighting the benefits of volumetric dual-modality modeling and style-aware prompting. Overall, this work establishes a foundation for future PET/CT-specific models emphasizing disease-aware reasoning and clinically reliable evaluation.

*Equal contribution

†Corresponding authors

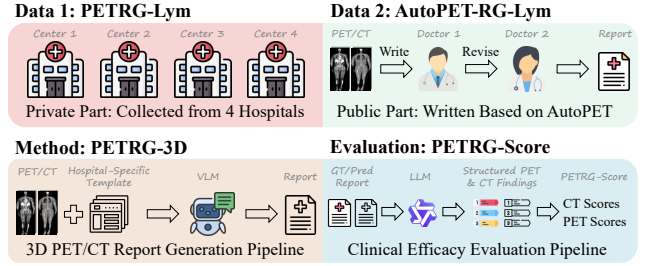


Figure 1. A glimpse of this work. (1) Data: The dataset consists of a private part collected from four medical centers in China and a public part derived from the AutoPET dataset [14], for which reports were generated and subsequently revised and verified by two senior nuclear medicine physicians. (2) Method: An end-to-end 3D PET/CT report-generation framework that integrates dual-modality volumetric encoding and hospital-specific prompting to generate clinically coherent reports. (3) Evaluation: A clinical efficacy evaluation pipeline that quantifies both metabolic (PET) and anatomical (CT) accuracy of generated reports.

Codes, models, and AutoPET-RG-Lym will be released.

1. Introduction

Intelligent interpretation of medical imaging has achieved notable progress in recent years, especially for structural modalities such as X-ray and CT [6, 16, 22–24, 38, 39]. In contrast, positron emission tomography/computed tomography (PET/CT) provides high-sensitivity, quantitative functional information such as glucose metabolism. This complementary information is essential for cancer diagnosis, staging, treatment planning, and therapy response assessment [13, 50, 52]. Consequently, PET/CT has become

widely adopted in oncology [21, 32, 41].

PET/CT report writing is a core task in nuclear medicine, yet the rapid increase in installed scanners and scan volumes has intensified clinical workload. Compounded by the persistent shortage of trained nuclear medicine physicians, this creates a growing gap between imaging demand and available diagnostic capacity [13, 52]. Automated PET/CT report generation (PETRG) therefore holds strong potential to alleviate this pressure and improve reporting efficiency.

Metabolic patterns in PET vary substantially with tracer physiology, making functional interpretation more complex than that of structural imaging. Second, PET/CT in oncology typically requires whole-body 3D contextual reasoning rather than localized-region (e.g., chest, brain) analysis, as physicians must jointly describe anatomical structures and metabolic activity across the entire body [52]. These factors drastically increase the difficulty of automated report generation. Finally, obtaining large-scale paired PET/CT image-report datasets is challenging, as PET/CT examinations are more costly, less frequently performed than X-ray or CT, and require expert-validated reports, limiting data availability for model development.

Recent advances in Vision-Language Models (VLMs) have significantly advanced report generation for single-modality 2D/3D medical images (e.g., X-ray, CT) [6, 18, 23, 26]. However, these approaches are inherently limited when applied to dual-modality 3D PET/CT imaging. Consequently, automated report generation for complete 3D PET/CT scans remains a critical and unaddressed challenge. To address the aforementioned challenges, we investigate from three perspectives: model, datasets, and evaluation metrics. Figure 1 illustrates our contributions in these aspects, and Table 1 summarizes the comparison with prior work.

At the model level, inspired by recent advances in 3D CT report generation (CTRG) with VLMs [8, 9, 45], we propose an end-to-end 3D PETRG framework. The architecture comprises three key components: (1) Dual-Stream Volumetric Feature Encoding: a two-branch 3D visual encoder separately captures functional uptake information from PET and anatomical density information from CT volumes. (2) Style-Adaptive Multimodal Fusion: hospital-specific prompt templates are introduced to address inter-center variation in reporting style, dynamically tailoring prompts to institutional conventions. (3) Parameter-Efficient Report Generation: the fused multimodal embeddings are fed into a language decoder to produce clinically coherent reports. In contrast to prior PETRG works—PET2Rep [52], which processes only 2D slices with off-the-shelf VLMs, and ViMed-PET [32], who utilize only 3D PET without CT—our model is the first to jointly model full 3D PET and CT volumes in an end-to-end manner.

At the data level, we collect and curate 824 PET/CT

scan-report pairs from 746 lymphoma patients collected across four medical centers in China, forming PETRG-Lym, the first large-scale, multicenter, single-disease PETRG dataset. Additionally, we identify 135 lymphoma cases from the public AutoPET dataset [14], for which two senior nuclear medicine physicians generated and refined structured reports, yielding AutoPET-RG-Lym, an open external benchmark for PETRG evaluation and cross-center validation.

At the evaluation level, beyond standard natural language generation (NLG) metrics such as BLEU [35], METEOR [5], and ROUGE [30], we introduce PETRG-Score, a clinically driven metric developed with senior nuclear medicine physicians. PETRG-Score explicitly evaluates the model’s ability to accurately describe both PET uptake patterns and CT structural findings, ensuring clinical fidelity beyond text similarity.

Using the proposed benchmark, we train and validate our method on the PETRG-Lym dataset and test it externally on AutoPET-RG-Lym. Extensive experiments reveal that current state-of-the-art VLMs for CTRG perform poorly when directly applied to multi-center PET/CT data, revealing a substantial domain gap. In contrast, our proposed PETRG-3D framework—featuring dual-branch volumetric encoding and hospital-specific prompting—achieves significant improvements in both language quality and clinical fidelity. These results highlight the necessity of PETRG datasets and multimodal architectures for functional imaging report generation.

Our main contributions are summarized as follows:

- We propose a 3D PETRG model that leverages a dual-branch architecture to separately encode volumetric PET and CT features, along with a hospital-style-aware prompting strategy for multi-center report style adaptation.
- We construct PETRG-Lym, a multi-center lymphoma PETRG dataset collected from four hospitals (824 scans, 245,509 paired slices), and we release AutoPET-RG-Lym, a publicly accessible benchmark derived from open PET/CT images but furnished with new expert-written, clinically validated reports (135 cases).
- PETRG-Score, a new evaluation metric that jointly assesses metabolic uptake and structural information in PET/CT reports, incorporating fine-grained anatomical grounding.
- Extensive experiments demonstrate the effectiveness of our method, reveal limitations of existing paradigms, and establish a strong foundation for future research in 3D PETRG.

2. Related Works

Our work is most relevant to the growing literature on X-Ray, CT, MRI and PETRG.

Table 1. Comparison of benchmarks for PETRG task. “Mul-Cnt”, “Dis-Spe”, “FT”, “Info”, “US Num”, and “AL Num” are short for Multi-Center, Disease-Specific, Fine-Tune, Information, the Number of Uptaking Status, and the Number of Anatomical Locations, respectively.

Benchmarks	Venues	Datasets			Methods			Metrics			
		Mul-Cnt	Dis-Spe	Size	Input	3D	FT	CT Info	PET Info	US Num	AL Num
ViMed-PET[32]	NeurIPS’25	✗	✗	2757	PET	✓	✓	✓	✓	2	5
PET2Rep[52]	AAAI’26	✗	✗	565	PET, CT	✗	✗	✗	✓	4	19
PETRG (Ours)	-	✓	✓	959	PET, CT	✓	✓	✓	✓	5	24

2.1. X-Ray Report Generation

Recent advances in 2D medical report generation have been driven by the availability of large-scale public datasets and progress in multimodal learning. Benchmark datasets such as IU-Xray [12] and MIMIC-CXR [25] have provided essential data foundations, while the evolution of deep learning—particularly VLMs—has spurred the development of numerous sophisticated report generation frameworks [7, 20, 23, 24, 29, 34, 43, 48, 49].

Most approaches follow an encoder–decoder architecture, where visual features extracted by CNNs or vision transformers are decoded into diagnostic narratives (e.g., MAIRA-1 [22] and Flamingo-CXR [39]). However, these 2D X-ray-based methods are not directly applicable to PET/CT, as they disregard 3D volumetric context essential for accurate interpretation of metabolic activity and structural relationships.

2.2. CT & MRI Report Generation

Research in 3D medical report generation (CT/MRI) [17] is dominated by 3D vision encoder-language decoder architectures [3, 26, 28], often built upon pre-trained encoders like CT-CLIP [17], RadFM [45], or M3D [3]. To manage 3D data complexity, methods employ diverse strategies. Some integrate memory modules, such as the disease prototype memory in Dia-LLaMA [8] for class imbalance or the relational memory in CT2Rep [18] for coherence. Others leverage fine-grained anatomical priors[31], as seen in Reg2RG [9] which utilizes explicit correspondences from RadGenome-ChestCT [51]. For MRI, AutoRG-Brain [27] fuses global and local context using automatic ROIs.

However, existing 3D CT/MRI benchmarks are single-modality, lacking the cross-modality reasoning required for PET/CT, which demands joint interpretation of metabolic and anatomical information.

2.3. PET/CT Report Generation

Efforts toward automated PETRG have emerged only recently[10, 13, 41, 50]. PET2Rep [52] first benchmarked the task by applying VLMs to 2D PET/CT slices with prompt engineering revealing that both general-purpose and medical-specific VLMs perform poorly on PETRG. ViMed-PET [32] introduced a Vietnamese PET/CT report and VQA

dataset, proposing a three-stage fine-tuning pipeline for language adaptation. However, both approaches suffer from significant limitations. PET2Rep [52] process only a handful of 2D slices, discarding rich 3D spatial context inherent in PET/CT volumes. ViMed-PET [32] use PET images alone, neglecting CT structural cues. These limitations underscore the need for approaches that jointly model full 3D PET and CT volumes, leveraging complementary metabolic and anatomical information.

3. Multi-Center, Single-Disease PET/CT Report Generation Dataset

3.1. Motivation

Progress in PETRG is hindered by the limited availability of paired PET/CT image–report datasets. Existing PET datasets (e.g., AutoPET [14], HECKTOR [2]) commonly lack reports, and recent PETRG resources are single-center and multi-disease (Table 1), introducing substantial heterogeneity in imaging appearance and clinical wording that complicates modeling and increases the risk of diagnostic errors. In addition, the field lacks a standardized benchmark for fair evaluation, as no public PET/CT dataset currently provides clinically validated reports.

To address these limitations, we focus on a single disease while collecting data from multiple centers. This design reduces cross-disease heterogeneity while still capturing real-world variability, supporting a practical “clinician confirms, AI drafts” workflow for reliable PETRG. Furthermore, we construct a publicly accessible PET/CT report-generation benchmark equipped with new expert-written, clinically validated reports, enabling reproducible and consistent evaluation for future work.

3.2. The PETRG-Lym Dataset

The PETRG-Lym dataset is sourced from the nuclear medicine departments of four hospitals in China, spanning cities of varying scale to ensure diversity and representativeness. All PET/CT scans were acquired using the ^{18}F -FDG radiotracer without contrast-enhanced CT. Images are preserved in their original DICOM format and include essential metadata such as patient height, weight, age, radio-tracer injection time, and dose. Imaging reports were au-

Table 2. Statistics of PETRG-Lym and AutoPET-RG-Lym. S Num and P Num denote the number of scans and patients, respectively.

Dataset	Center ID	Training Cohort			Validation Cohort			Total	
		P Num	S Num	Scan Period	P Num	S Num	Scan Period	P Num	S Num
PETRG-Lym	1	224	227	2023.10 - 2024.07	57	57	2024.07 - 2024.09	281	284
	2	107	147	2015.07 - 2024.03	28	36	2024.05 - 2025.05	135	183
	3	112	138	2018.01 - 2024.03	29	30	2024.03 - 2025.01	141	168
	4	151	151	2013.02 - 2019.05	38	38	2019.05 - 2021.04	189	189
AutoPET-RG-Lym	5	-	-	-	135	135	-	135	135
Total	-	594	663	-	287	296	-	881	959

thored by nuclear medicine physicians and subsequently reviewed by senior staff, covering patient demographics, clinical history, acquisition protocol, findings, and diagnostic impressions. In this work, we focus exclusively on the “findings” section. As summarized in Table 2, the PETRG-Lym dataset comprises 824 PET/CT-report pairs from 746 lymphoma patients across the four centers, collected over a 12-year period from February 2013 to May 2025. This constitutes the first large-scale, multi-center, single-disease PETRG dataset.

3.3. The AutoPET-RG-Lym Dataset

To enable external validation and promote community benchmarking, we construct AutoPET-RG-Lym from lymphoma cases in the public AutoPET dataset [14], which was collected at University Hospital Tübingen and University Hospital of the LMU Munich. It contains patients with lymphoma, malignant melanoma, and lung cancer. This dataset differs from our internal data in three key aspects: (1) it originates from a different country, (2) it was acquired using different imaging hardware, and (3) it employs different CT contrast agents—making it a strong candidate for external validation of model generalization.

We carefully selected 135 lymphoma cases from AutoPET [14] and commissioned two senior nuclear medicine physicians from top-tier Chinese hospitals to independently compose structured reports. These reports underwent cross-review and iterative refinement to ensure high clinical fidelity. The resulting dataset provides an open validation benchmark for PETRG research.

3.4. Data Preprocessing

Starting from de-identified DICOM data, we extracted patient metadata (e.g., weight, radiotracer dose, injection time) and converted PET and CT volumes to NIfTI format. All volumes were reoriented to RAS and resampled to a uniform spacing of $1.5 \times 1.5 \times 3$ mm, with CT matched to PET dimensions. CT voxel intensities were converted to Hounsfield Units (HU) and clipped to $[-1000, +1000]$. Standardized uptake value (SUV) normalization was performed using the extracted metadata. We

employed TotalSegmentator [44] to remove the scanning bed from CT backgrounds. Given the presence of both “head-to-midthigh” and “head-to-toe” scan protocols—and the observation that reports seldom describe regions below the mid-thigh—we uniformly cropped all volumes at the upper thigh. Additional preprocessing details, including voxel intensity standardization, scanning bed removing, body part cropping, and report text cleaning, are provided in the supplementary material.

3.5. Dataset Construction and Splitting

We split PETRG-Lym into training and internal validation sets chronologically, as shown in Table 2. To prevent data leakage, we grouped scans by patient and, within each center, assigned the first 80% of patients (by initial scan date) to training and the remaining 20% to validation. This yields 663 scans (594 patients) for training and 161 scans (152 patients) for validation. All 135 samples in AutoPET-RG-Lym serve as an external test set.

4. Methods

Inspired by recent CT report generation frameworks [3, 45], we propose PETRG-3D, a novel pipeline for 3D PET/CT report generation. As illustrated in Figure 2, our approach consists of three key stages: (1) **Dual-Stream Volumetric Feature Encoding (DSFE)** from PET and CT scans; (2) **Style-Adaptive Multimodal Fusion (SAMF)** using hospital-specific prompt templates; and (3) parameter-efficient autoregressive report generation using a Low-Rank Adaptation (LoRA)-adapted LLM. We detail each component below.

4.1. Dual-Stream Volumetric Feature Encoding

To jointly model metabolic (PET) and anatomical (CT) information, we propose a dual-stream 3D visual encoder. Unlike prior approaches using 2D slices [52] or single-modality 3D PET [32], our **DSFE** architecture (Fig. 2, blue region) explicitly processes and fuses these complementary volumetric modalities.

Each stream comprises a volume encoder followed by a

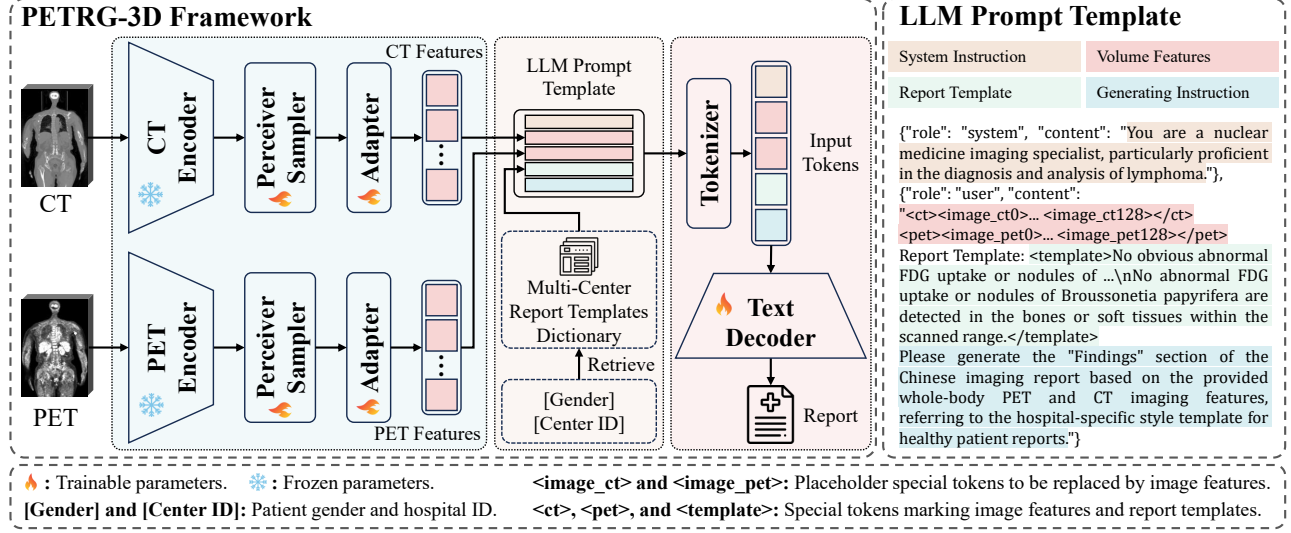


Figure 2. Overall framework of PETRG-3D. The model comprises three key components: (1) a dual-stream 3D visual feature extractor (DSFE) that processes PET and CT volumes separately (blue); (2) a style-adaptive multimodal fusion (SAMF) module that dynamically integrates visual features with hospital-specific prompt templates (yellow); and (3) a LoRA-adapted large language model (LLM) for effective radiology report generation (pink).

perceiver sampler. The volume encoder employs a sliding-window strategy to process input volumes of arbitrary size, producing high-dimensional feature maps. The perceiver sampler then compresses these variable-length features into a fixed-length sequence of representative embeddings.

To leverage rich, pre-learned representations, we initialize both volume encoders with the 3D ViT weights pre-trained on large-scale CT data (RadFM [45]) and keep them frozen. This strategy is highly parameter-efficient. However, a significant domain gap exists between RadFM’s lung-centric CT data and our whole-body PET/CT scans. To bridge this gap, we retrain *only* the perceiver samplers on our PETRG-Lym dataset, allowing them to adapt the frozen features to our specific PET and CT modalities independently.

4.2. Style-Adaptive Multimodal Fusion

The compressed PET and CT feature sequences from the DSFE module are projected into the LLM’s hidden space via linear transformations, aligning the visual and linguistic semantic spaces.

A key challenge in multicenter report generation is the significant stylistic heterogeneity across institutions. Our SAMF module addresses this by emulating the clinical workflow. We curate a dictionary of hospital-specific “healthy-example” templates. These templates also account for gender (male/female) to handle anatomical differences, particularly in pelvic descriptions.

As shown in Figure 2 (yellow region), each sample carries two metadata flags: [Center ID] and [Gender]. These

flags dynamically retrieve the corresponding hospital- and gender-matched template from the dictionary, which is then injected into the input prompt. To guide the model in distinguishing these inputs, we wrap the CT features, PET features, and template text with dedicated special tokens (<ct>, <pet>, and <template>, respectively).

4.3. Parameter-Efficient Autoregressive Generation

The final fused multimodal embedding sequence—comprising visual tokens and the template-aware textual prompt—is fed directly into a pre-trained LLM (e.g., Qwen3-8B [47]) for autoregressive report generation (Fig. 2, pink region).

Instead of full fine-tuning, we adopt LoRA [19] for parameter-efficient training. LoRA injects trainable, low-rank matrices into the Transformer’s attention layers while freezing the original LLM parameters. This strategy enables the model to learn PET/CT-specific terminology and reporting structures with minimal trainable parameters, mitigating catastrophic forgetting and substantially reducing computational overhead.

4.4. Implementation and Evaluation Protocol

4.4.1. Evaluation Metrics

We evaluate performance on two axes: NLG and Clinical Efficacy (CE).

For NLG metrics, we report standard metrics: BLEU (B-n) [35], METEOR (MTR) [5], and ROUGE-L (R-L) [30], to quantify textual similarity between generated and

Table 3. Evaluation of NLG and CE metrics on the PETRG-Lym dataset. PET2Rep-Sep and PET2Rep-Fus refer to the two image input strategies from [52]: “Sep” denotes separate PET and CT slice inputs, while “Fus” denotes a fused PET/CT image input. “All” and “Ab” denote results on all and only abnormal classes.

Methods	LLMs/VLMs	NLG Metrics						CE Metrics			
		B-1	B-2	B-3	B-4	MTR	R-L	PT-All	CT-All	PT-Ab	CT-Ab
LLaVA-Med[33]	Mistral-7B-v0.2	3.21	0.99	0.56	0.40	2.46	4.84	-	-	-	-
RadFM[45]	Llama2-13B	1.65	0.38	0.21	0.16	1.27	2.44	-	-	-	-
M3D[3]	Llama2-7B	0.53	0.21	0.11	0.07	2.16	4.56	-	-	-	-
PET2Rep-Sep[52]	Qwen3-VL-8B	39.60	21.61	11.39	6.16	27.06	22.25	23.88	33.66	11.53	28.41
	InternVL-3.5-8B	38.23	20.74	10.96	5.96	25.99	22.20	22.75	33.45	10.64	28.01
	Qwen2.5-VL-Max	34.98	18.85	9.97	5.48	27.91	20.37	21.78	23.93	7.84	17.11
PET2Rep-Fus[52]	Qwen3-VL-8B	38.98	20.94	10.99	5.96	27.12	21.77	23.82	30.61	11.34	24.65
	InternVL-3.5-8B	37.79	20.44	10.78	5.83	25.69	21.92	23.88	33.88	12.22	<u>28.62</u>
	Qwen2.5-VL-Max	34.30	18.57	9.87	5.42	28.55	19.84	22.88	26.95	9.54	<u>20.60</u>
PETRG-3D (Ours)	Llama2-7B	28.19	21.84	17.72	14.90	22.60	27.19	26.10	25.45	11.99	17.97
	Mistral-7B-v0.3	53.18	45.53	40.12	36.07	43.20	48.60	27.25	30.24	13.52	23.55
	Qwen2.5-7B	<u>60.42</u>	<u>52.00</u>	<u>45.92</u>	41.33	51.21	52.15	<u>31.78</u>	<u>34.13</u>	<u>19.20</u>	28.06
	Gemma2-9B	58.57	50.24	44.25	39.79	49.47	51.49	28.29	31.92	15.08	25.54
	GLM4-9B	58.73	51.22	45.72	<u>41.55</u>	50.80	53.74	29.06	31.65	15.72	25.07
	Qwen3-8B	60.78	52.45	46.42	41.90	<u>51.16</u>	<u>52.88</u>	32.06	34.76	19.53	28.76

reference reports.

For CE metrics, We implement an LLM-based pipeline to extract diagnostic content from reports across 24 lymphoma-specific anatomical regions predefined by senior radiologists. Specifically, we extract labels from reference and generated reports using Qwen3-Max [47], while reference report labels underwent expert validation by a researcher specializing in medical imaging. Labels encompass 5 PET uptake categories (e.g., markedly abnormal, physiological) and 8 CT structural findings (e.g., lymphadenopathy). We evaluate Macro-F1 scores for: **PET-All/CT-All** (all labels; overall quality) and **PET-Abnormal (PT-Ab)/CT-Abnormal (CT-Ab)** (abnormal labels only; diagnostic relevance). Full implementation details are in the supplementary material.

4.4.2. Baselines and Implementation

We benchmark against strong baselines: LLaVA-Med [33], RadFM [45], and M3D [3] (prior 2D/3D radiology VLMs), and PET2Rep [52], a recent 2D PET/CT method. Our model’s (PETRG-3D) volume encoder is initialized from the pretrained ViT-3D of RadFM [45]. All experiments are conducted on NVIDIA A800 (80GB) GPUs. Implementation details for baselines and PETRG-3D are provided in the supplementary material.

5. Experimental Results

We evaluate our PETRG-3D framework on the PETRG-Lym and AutoPET-RG-Lym datasets, comparing against existing methods and analyzing each proposed component. Additional results appear in the supplementary material.

5.1. Main Results

5.1.1. Quantitative Results

Table 3 details results on PETRG-Lym. Baselines like M3D[3], pretrained on CT volumes, show limited performance due to domain gaps. The prompt-based PET2Rep yields moderate scores (e.g., 6.16 B-4, 23.88 PT-All with Qwen3-VL-8B [4]) but fails to capture PET-specific language. In contrast, our fine-tuned PETRG-3D substantially boosts both NLG (41.9 B-4, 51.16 MTR) and CE metrics (32.06 PT-All, 34.76 CT-All, with Qwen3-8B [47]), demonstrating consistent improvements and superior alignment with the radiologists’ reporting style.

On the external AutoPET-RG-Lym dataset (Table 4), PETRG-3D maintains its superiority in NLG and PET-centric CE metrics, reinforcing the value of domain-specific fine-tuning. However, its CT-centric CE performance degrades. We hypothesize this is due to higher inter-center variance in CT descriptions compared to PET, causing the model to overfit to the source domain’s CT style. This suggests limitations of the end-to-end VLMs fine-tuning paradigm. The results underscore that a significant gap remains for the clinical applicability of all current methods.

5.1.2. Qualitative Results

Fig. 3 presents a qualitative comparison on PETRG-Lym, showcasing two cases from different centers to highlight diverse reporting styles. We display key findings of the chest region for brevity; distinct colors denote different anatomical areas, and gray backgrounds indicate incorrect diagnoses.

In terms of reporting style, PET2Rep generates an ex-

Table 4. Evaluation of NLG and CE metrics on the External Cohort AutoPET-RG-Lym. PET2Rep-Sep and PET2Rep-Fus refer to the two image input strategies from [52].

Methods	LLMs/VLMs	NLG Metrics						CE Metrics			
		B-1	B-2	B-3	B-4	MTR	R-L	PT-All	CT-All	PT-Ab	CT-Ab
LLaVA-Med[33]	Mistral-7B-v0.2	3.16	0.98	0.57	0.42	2.37	4.69	-	-	-	-
RadFM[45]	Llama2-13B	1.57	0.45	0.24	0.18	1.12	2.14	-	-	-	-
M3D[3]	Llama2-7B	0.28	0.09	0.04	0.03	1.84	4.33	-	-	-	-
PET2Rep-Sep[52]	Qwen3-VL-8B	40.80	24.08	13.51	7.51	27.56	24.33	24.25	29.22	12.29	23.07
	InternVL-3.5-8B	37.94	22.20	12.50	6.98	26.62	23.16	23.34	33.56	11.78	27.60
	Qwen2.5-VL-Max	32.63	19.16	10.62	5.83	28.78	21.21	26.76	25.94	14.00	18.46
PET2Rep-Fus[52]	Qwen3-VL-8B	39.36	23.19	13.00	7.20	28.00	23.40	22.62	31.17	10.61	24.96
	InternVL-3.5-8B	38.39	22.43	12.58	7.05	26.91	22.92	23.38	<u>32.92</u>	12.16	<u>27.16</u>
	Qwen2.5-VL-Max	32.13	18.97	10.56	5.82	29.12	21.01	28.31	<u>28.57</u>	16.14	21.46
PETRG-3D (Ours)	Llama2-7B	24.05	19.82	16.87	14.68	26.31	32.20	19.02	14.73	4.30	6.04
	Mistral-7B-v0.3	<u>52.37</u>	<u>43.92</u>	<u>38.65</u>	<u>35.03</u>	40.31	<u>44.20</u>	23.63	16.94	11.64	9.58
	Qwen2.5-7B	47.90	37.88	31.51	27.35	36.50	37.30	30.42	21.90	19.01	15.77
	Gemma2-9B	47.27	39.23	34.28	30.97	37.09	40.21	30.03	18.56	18.49	11.15
	GLM4-9B	55.89	50.38	46.76	44.18	47.41	55.47	36.22	17.14	25.46	9.24
	Qwen3-8B	49.41	41.30	36.07	32.47	<u>41.50</u>	40.97	<u>33.48</u>	22.79	<u>22.81</u>	16.35

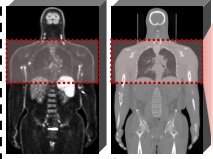
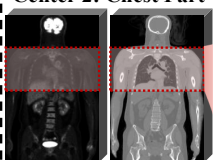
Center 1: Chest Part	Ground Truth		PET2Rep	PETRG-3D (Ours)
	PET/CT			
 Lung Hila and bronchial Mediastinum Axillae Pleura and pericardium	CHEST: Slightly increased bronchovascular markings are seen in both lungs. A few linear shadows are noted in the right middle lobe, without increased radiotracer uptake. No hypermetabolic foci are seen within the lung fields. The hila are not enlarged. The bronchial orifices are patent, with no significant luminal stenosis. Multiple nodules are seen in the mediastinum (stations 2R, 4, 5, 6, 10), measuring approximately 0.4 cm to 0.7 cm, showing increased radiotracer uptake (SUVmax 2.8-4.5). Punctate or patchy calcifications are seen within some nodules. The pleura and pericardium are smooth, with no hypermetabolic nodules. No pleural effusion. Multiple nodules are present in the bilateral axillae, measuring approximately 0.3 cm to 1.4 cm, demonstrating varying degrees of increased radiotracer uptake (SUVmax 1.7-11.2).		The lungs are clear, with no significant consolidation or nodules. No abnormal FDG uptake is seen. The cardiac silhouette is normal in size and configuration. No abnormal FDG uptake is noted in the myocardium. No significant calcification or dilatation is seen in the aorta and its major branches.	CHEST: Slightly increased bronchovascular markings are seen in both lungs. No hypermetabolic foci are seen within the lung fields. The hila are not enlarged. The bronchial orifices are patent, with no significant luminal stenosis. Multiple soft-tissue density nodules are seen in the mediastinum (stations 2R, 4R, 5, 10), measuring approximately 0.6 cm to 1.1 cm, showing increased radiotracer uptake (SUVmax 5.3-19.5), with right-sided predominance. The pleura and pericardium are smooth, with no hypermetabolic nodules. A small, crescent-shaped, low-density collection is noted in the left pleural cavity, which is photopenic. Multiple small soft-tissue density nodules are present in the bilateral axillae, measuring approximately 0.4 cm to 1.0 cm, showing mildly increased radiotracer uptake (SUVmax 1.7-3.1).
	 PET/CT			
Center 2: Chest Part	Ground Truth		PET2Rep	PETRG-3D (Ours)
	PET/CT			
	The lung fields are clear. A small solid or calcified nodule is visible in the left lower lobe. No space-occupying lesions or abnormal FDG uptake are seen within the lung fields. No abnormal lymph nodes are visualized in the mediastinum, bilateral hila, and axillary regions.		Increased lucency is seen in both lungs, with subpleural cystic (or bullous) lucencies noted in the bilateral upper lobes. There is volume loss of the right middle lobe. A patchy opacity is seen in the perihilar region of the right middle lobe, demonstrating increased FDG uptake.	The lung fields are clear. A small nodule is visible in the left upper lobe, showing no abnormal FDG uptake. No space-occupying lesions or abnormal FDG uptake are seen in the remaining lung fields. No abnormal lymph nodes are visualized in the mediastinum, bilateral hila, and axillary regions.

Figure 3. Qualitative comparison of our method against PET2Rep[52] on chest results. Different colors denote distinct anatomical areas. Incorrect diagnoses are highlighted in gray.

cessively brief and anatomically incomplete report for Case 1. In contrast, our model generates a report whose length and anatomical coverage closely align with the ground truth (GT). In Case 2, while PET2Rep achieves a comparable length, it still exhibits incomplete anatomical coverage. Our model again demonstrates superior adaptability, matching the GT in both length and completeness.

Regarding diagnostic accuracy, PET2Rep incorrectly reports no abnormalities in Case 1. Our model, however, correctly identifies multiple findings, such as “slightly increased bronchovascular markings”. In Case 2, PET2Rep’s

output is clinically irrelevant. Our model successfully identifies “a small nodule... in the left lobe,” despite a minor error in fine-grained localization (“left upper lobe” vs. “left lower lobe”).

5.2. Ablation Studies and Analysis

We analyze the impact of our core components and design choices, using Qwen3-8B [47] as the text decoder.

Table 5. Ablation study on the PETRG-Lym dataset. CT Base and PET Base use only CT or PET stream, respectively. DSFE denotes Dual-Stream Volumetric Feature Encoding; SAMF denotes Style-Adaptive Multimodal Fusion.

CT Base	PET Base	DSFE	SAMF	NLG Metrics						CE Metrics			
				B-1	B-2	B-3	B-4	MTR	R-L	PT-All	CT-All	PT-Ab	CT-Ab
✓				49.19	39.91	33.58	29.13	40.91	41.62	28.81	32.96	15.77	26.86
	✓			52.74	43.01	36.37	31.76	42.65	43.61	26.01	32.95	12.40	26.91
	✓	✓		58.02	48.30	41.47	36.54	46.88	46.93	28.38	33.27	15.36	27.41
	✓	✓	✓	60.78	52.45	46.42	41.90	51.16	52.88	32.06	34.76	19.53	28.76

5.2.1. Impact of Key Components

We first ablate our two key modules, DSFE and SAMF. Results are in Table 5.

DSFE: Jointly encoding CT and PET volumes with DSFE consistently outperforms single-modality inputs (PET-only or CT-only) across nearly all NLG and CE metrics. **SAMF:** Adding SAMF further boosts NLG performance substantially. By explicitly incorporating hospital-specific templates, SAMF simplifies the stylistic modeling task. This frees model capacity to focus on clinically relevant details, thereby improving CE metrics as well.

5.2.2. Analysis of Design Choices

We conduct further analysis on key design choices, summarized in Figure 4. “Ours (PETRG-3D)” represents our final model settings.

CT-Pretrained Features for PET. Although PET and CT represent distinct modalities, CT-pretrained 3D encoders transfer surprisingly well. Fig. 4 (“PET Enc. (from Scratch)”) shows that our approach—using the CT-pretrained encoder for both modalities—significantly outperforms training a separate PET encoder from scratch. This is likely because anatomical priors, such as tissue texture and lesion morphology, serve as useful inductive biases for PET interpretation.

Stop Token. We find that models without an explicit stop token often produce hallucinated text after the report’s factual end. As shown in Fig. 4 (“w/o Stop Token”), adding a dedicated end-of-report token during training significantly improves report faithfulness.

Region-wise Decomposition. We evaluated a regional decomposition strategy: dividing full-body scans and reports into four anatomical regions (e.g., chest, abdomen), training on region-report pairs, and concatenating outputs during inference. This “w/ Regional Input (RI)” approach (Fig. 4) reduces NLG performance. Results indicate that decomposition-induced sample proliferation compromises global contextual coherence critical for radiology report generation.

Fine-tuning the Visual Encoder. We investigate fine-tuning the pretrained CT encoder on our dataset (Fig. 4, “w/ RI + FT Encoder”). While this yields modest gains in PET-

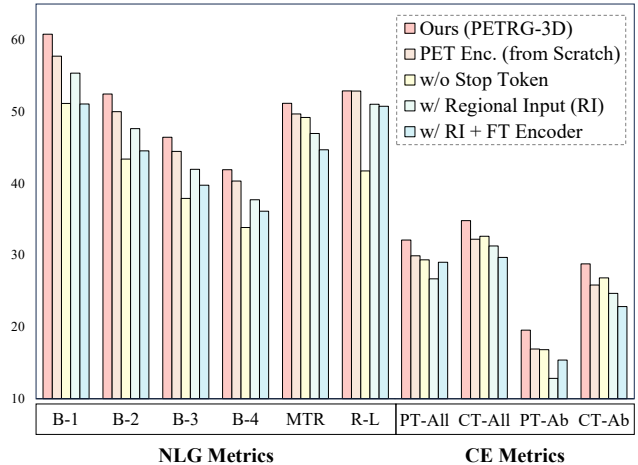


Figure 4. **Analysis of design choices.** “Ours (PETRG-3D)” is the final model settings. “PET Enc. (from Scratch)” trains a separate ViT-3D for the PET branch from scratch. “w/o Stop Token” removes the explicit end-of-report token during training. “w/ Regional Input” uses 4 decoupled anatomical region pairs for training. “w/ Regional Input + FT Encoder” further fine-tunes the visual encoder on top of the “w/ Regional Input” setting.

centric CE (PT-Ab), it degrades both NLG and CT-CE metrics. This suggests that fine-tuning on limited data adapts the model to PET but harms the robust, pre-trained CT representations due to overfitting.

6. Conclusion

We introduce PETRG-Lym, a multi-center, single-disease dataset for PETRG, comprising 245,509 paired PET/CT slices and 824 expert-written reports. To facilitate reproducible and fair benchmarking, we also release AutoPETRG-Lym, a curated PETRG dataset built upon the public AutoPET dataset, with reports carefully drafted and reviewed by domain experts. Based on these datasets, we propose PETRG-3D, an end-to-end 3D PETRG framework featuring a dual-branch architecture for modality-specific encoding and a style-adaptive multimodal fusion module for cross-center stylistic adaptation. Extensive experiments demonstrate that PETRG-3D consistently outperforms ex-

isting methods on both NLG and CE metrics, establishing a strong baseline for future research in automated PETRG.

Limitations and Future Work: While this work establishes a complete pipeline—spanning data, modeling, and evaluation—for PETRG, several limitations remain. First, the diagnostic accuracy of generated reports is still below clinical requirements; the model primarily captures stylistic patterns rather than performing detailed pathophysiological reasoning. Second, although our data includes both baseline and post-treatment scans, our framework treats them uniformly. Modeling longitudinal changes (e.g., post-surgical alterations or therapy response) is a crucial next step. In future work, we plan to explore disease-aware reasoning, temporal report synthesis, and clinically validated evaluation protocols to further bridge the gap between research systems and real-world clinical deployment.

References

- [1] Q Jiang Albert, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, and Devendra Singh Chaplot. Mistral 7b. *arXiv preprint*, 2023. 2
- [2] Vincent Andrearczyk, Valentin Oreiller, Mathieu Hatt, and Adrien Depeursinge. Head and neck tumor segmentation and outcome prediction. *UK: Springer*, 2022. 3
- [3] Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024. 3, 4, 6, 7, 2
- [4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6
- [5] Satantjeet Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 2, 5, 4
- [6] Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024. 1, 2
- [7] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, 2020. 3
- [8] Zhixuan Chen, Luyang Luo, Yequan Bie, and Hao Chen. Dia-llama: Towards large language model-driven ct report generation. *arXiv preprint arXiv:2403.16386*, 2024. 2, 3
- [9] Zhixuan Chen, Yequan Bie, Haibo Jin, and Hao Chen. Large language model with region-guided referring and grounding for ct report generation. *IEEE Transactions on Medical Imaging*, 2025. 2, 3
- [10] Hongyoon Choi, Dongjoo Lee, Yeon-koo Kang, and Minseok Suh. Empowering pet imaging reporting with retrieval-augmented large language models and reading reports database: a pilot single center study. *European Journal of Nuclear Medicine and Molecular Imaging*, 52(7):2452–2462, 2025. 3
- [11] DR Dance, S Christofides, ADA Maidment, ID McLean, and KH Ng. Diagnostic radiology physics: A handbook for teachers and students, endorsed by: American association of physicists in medicine, asia-oceania federation of organizations for medical physics, european federation of organisations for medical physics. 2014. 1
- [12] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2015. 3
- [13] Yaozong Gao, Yiran Shu, Mingyang Yu, Yanbo Chen, Jingyu Liu, Shaonan Zhong, Weifang Zhang, Yiqiang Zhan, Xi-an Sean Zhou, Xinlu Wang, et al. Hierarchical contrastive learning for precise whole-body anatomical localization in pet/ct imaging. *IEEE Transactions on Medical Imaging*, 2025. 1, 2, 3
- [14] Sergios Gatidis, Marcel Früh, Matthias Fabritius, Sijing Gu, Konstantin Nikolaou, Christian La Fougère, Jin Ye, Junjun He, Yige Peng, Lei Bi, et al. The autopet challenge: towards fully automated lesion segmentation in oncologic pet/ct imaging. 2023. 1, 2, 3, 4
- [15] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024. 2
- [16] Tiancheng Gu, Dongnan Liu, Zhiyuan Li, and Weidong Cai. Complex organ mask guided radiology report generation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 7995–8004, 2024. 1
- [17] Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Seval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasedelen, Bastian Wittmann, Enis Simsar, Mehmet Simsar, et al. A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. *CoRR*, 2024. 3
- [18] Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. Ct2rep: Automated radiology report generation for 3d medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 476–486. Springer, 2024. 2, 3
- [19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 5
- [20] Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19809–19818, 2023. 3
- [21] Zach Huemann, Samuel Church, Joshua Warner, Daniel Tran, Xin Tie, Alan Mcmillan, Junjie Hu, Steve Cho,

- Meghan Lubner, and Tyler Bradshaw. Vision-language modeling in pet/ct for visual grounding of positive findings, 2025. [2](#)
- [22] Stephanie L Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, et al. Maira-1: A specialised large multi-modal model for radiology report generation. *arXiv preprint arXiv:2311.13668*, 2023. [1](#), [3](#)
- [23] Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2607–2615, 2024. [2](#), [3](#)
- [24] Haibo Jin, Haoxuan Che, Sunan He, and Hao Chen. A chain of diagnosis framework for accurate and explainable radiology report generation. *IEEE Transactions on Medical Imaging*, 2025. [1](#), [3](#)
- [25] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019. [3](#)
- [26] Haoran Lai, Zihang Jiang, Qingsong Yao, Rongsheng Wang, Zhiyang He, Xiaodong Tao, Wei Wei, Weifu Lv, and S Kevin Zhou. E3d-gpt: enhanced 3d visual foundation for medical vision-language model. *arXiv preprint arXiv:2410.14200*, 2024. [2](#), [3](#)
- [27] Jiayu Lei, Xiaoman Zhang, Chaoyi Wu, Lisong Dai, Ya Zhang, Yanyong Zhang, Yanfeng Wang, Weidi Xie, and Yuehua Li. Autorg-brain: Grounded report generation for brain mri. *arXiv preprint arXiv:2407.16684*, 2024. [3](#)
- [28] Cheng-Yi Li, Kao-Jung Chang, Cheng-Fu Yang, Hsin-Yu Wu, Wenting Chen, Hritik Bansal, Ling Chen, Yi-Ping Yang, Yu-Chun Chen, Shih-Pin Chen, et al. Towards a holistic framework for multimodal llm in 3d brain ct radiology report generation. *Nature Communications*, 16(1):2258, 2025. [3](#)
- [29] Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3334–3343, 2023. [3](#)
- [30] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. [2](#), [5](#), [4](#)
- [31] Jingyang Lin, Yingda Xia, Jianpeng Zhang, Ke Yan, Le Lu, Jiebo Luo, and Ling Zhang. Ct-glip: 3d grounded language-image pretraining with ct scans and radiology reports for full-body scenarios. *arXiv preprint arXiv:2404.15272*, 2024. [3](#)
- [32] 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, et al. Toward a vision-language foundation model for medical data: Multimodal dataset and benchmarks for vietnamese pet/ct report generation. *arXiv preprint arXiv:2509.24739*, 2025. [2](#), [3](#), [4](#), [6](#)
- [33] Thao Nguyen, Matthew Wallingford, Sebastin Santy, Wei-Chiu Ma, Sewoong Oh, Ludwig Schmidt, Pang Wei W Koh, and Ranjay Krishna. Multilingual diversity improves vision-language representations. *Advances in Neural Information Processing Systems*, 37:91430–91459, 2024. [6](#), [7](#), [3](#)
- [34] Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. Progressive transformer-based generation of radiology reports. *arXiv preprint arXiv:2102.09777*, 2021. [3](#)
- [35] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. [2](#), [5](#), [4](#)
- [36] Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, and Fei Huang. Qwen2.5 technical report. 2024. [2](#)
- [37] Sarah Schuëgger. Mic-dkfz/bodypartregression. *Zenodo* <https://doi.org/10.5281/zenodo.5195341>, 2021. [3](#)
- [38] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442, 2023. [1](#)
- [39] Ryutaro Tanno, David GT Barrett, Andrew Sellergren, Sumedh Ghaisas, Sumanth Dathathri, Abigail See, Johannes Welbl, Charles Lau, Tao Tu, Shekoofeh Azizi, et al. Collaboration between clinicians and vision-language models in radiology report generation. *Nature Medicine*, 31(2):599–608, 2025. [1](#), [3](#)
- [40] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. [2](#)
- [41] Xin Tie, Muheon Shin, Ali Pirasteh, Nevein Ibrahim, Zachary Huemann, Sharon M Castellino, Kara M Kelly, John Garrett, Junjie Hu, Steve Y Cho, et al. Personalized impression generation for pet reports using large language models. *Journal of Imaging Informatics in Medicine*, 37(2):471–488, 2024. [2](#), [3](#)
- [42] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. [2](#)
- [43] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11558–11567, 2023. [3](#)
- [44] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. To-

- talsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5): e230024, 2023. 4, 1
- [45] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *arXiv preprint arXiv:2308.02463*, 2023. 2, 3, 4, 5, 6, 7, 1
- [46] Ke Yan, Le Lu, and Ronald M Summers. Unsupervised body part regression via spatially self-ordering convolutional neural networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1022–1025. IEEE, 2018. 3
- [47] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 5, 6, 7, 2
- [48] Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80:102510, 2022. 3
- [49] Shuxin Yang, Xian Wu, Shen Ge, Zhuozhao Zheng, S Kevin Zhou, and Li Xiao. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, 86:102798, 2023. 3
- [50] Mingyang Yu, Yaozong Gao, Yiran Shu, Yanbo Chen, Jingyu Liu, Caiwen Jiang, Kaicong Sun, Zhiming Cui, Weifang Zhang, Yiqiang Zhan, et al. Location-guided automated lesion captioning in whole-body pet/ct images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 348–357. Springer, 2025. 1, 3
- [51] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Ya Zhang, Yanfeng Wang, and Weidi Xie. Radgenome-chest ct: A grounded vision-language dataset for chest ct analysis. *arXiv preprint arXiv:2404.16754*, 2024. 3
- [52] Yichi Zhang, Wenbo Zhang, Zehui Ling, Gang Feng, Sisi Peng, Deshu Chen, Yuchen Liu, Hongwei Zhang, Shuqi Wang, Lanlan Li, et al. Pet2rep: Towards vision-language model-driven automated radiology report generation for positron emission tomography. *arXiv preprint arXiv:2508.04062*, 2025. 1, 2, 3, 4, 6, 7, 9, 11

Vision-Language Models for Automated 3D PET/CT Report Generation

Supplementary Material

A. Dataset Details

A.1. Data Collection

PETRG-Lym Acquisition Protocol: Patients fasted for at least 6 hours prior to the examination. Venous access was established, and blood glucose levels were measured to ensure protocol compliance (< 11.1 mmol/L). Following the intravenous injection of the ^{18}F -FDG radiotracer, patients rested in a supine position for approximately 40 minutes. Patients were instructed to void their bladders before scanning to minimize pelvic artifacts. Whole-body (WB) imaging was performed with the patient supine and immobilized, with a typical scan duration of 22 minutes. Attenuation correction (AC) was applied to PET images using the corresponding CT data. Data were acquired using four scanner models: GE Discovery STE, SIEMENS Biograph Vision 450, UNITED IMAGING uMI 510, and uMI Panorama 35S. Diagnostic reports, stored in PDF and DOCX formats, were rigorously reviewed by senior nuclear medicine physicians. This study received ethical approval from the Institutional Review Boards (IRB) of all four participating medical centers.

A.2. Preprocessing Details

Transforming raw clinical data into model-ready inputs involves a complex pipeline to ensure multi-modal alignment and quantitative consistency. Figure 1 systematically illustrates our workflow.

A.2.1. PET/CT Image Preprocessing

We implemented a streamlined pipeline to standardize the visual input:

- **De-identification:** All Protected Health Information (PHI) was stripped from DICOM headers and file paths.
- **Meta-Information Extraction:** Critical meta-data—including patient weight, injected dose, acquisition time, and rescale factors—were parsed for quantitative calculations.
- **Format Conversion:** Raw DICOM series were converted to NIfTI format for efficient I/O during training.
- **Spatial Normalization and Alignment:** Raw PET and CT images often possess differing resolutions and fields of view. To achieve pixel-level alignment, we resampled both modalities to a unified voxel spacing of $1.5 \times 1.5 \times 3$ mm. Both volumes were reoriented to the standard RAS (Right-Anterior-Superior) coordinate system, ensuring strictly aligned multi-modal inputs.
- **SUV Calculation:** We converted raw PET intensity to Standardized Uptake Values (SUV) to mitigate inter-

subject variability:

$$\text{SUV}(t) = \frac{c_{img}(t) \cdot \text{BW}}{\text{ID}} \quad (1)$$

where $c_{img}(t)$ is the decay-corrected radioactivity concentration (Bq/mL) at scan time t , BW is the body weight (g), and ID is the injected dose (Bq) corrected to the injection time.

- **CT Intensity Normalization:** CT voxel values were converted to Hounsfield Units (HU) using the rescale slope and intercept [11]. We clipped the intensities to the range $[-1000, 1000]$ HU to cover the full dynamic range from lung tissue to bone, followed by min-max normalization to $[0, 1]$.
- **Background Removal:** To eliminate irrelevant background noise (e.g., scanning bed), we utilized TotalSegmentator [44] to generate a precise body mask. The volume was cropped to the minimal bounding box of the body with a 10-slice safety margin.
- **Anatomical Focus (ROI Cropping):** Clinical scan ranges vary (e.g., "head-to-toe" vs. "head-to-mid-thigh"). Since lymphoma reporting focuses primarily on the trunk, we standardized the field of view by cropping images to the upper thigh level. Specifically, we retained the body trunk and an extension of 20% below the pelvic floor (≤ 50 slices) to ensure complete coverage of the inguinal region while discarding irrelevant lower limb data.

A.2.2. Report Text Preprocessing

The textual data underwent a four-step cleaning process:

- **De-identification:** PHI was strictly removed from both raw content and metadata.
- **Structure Parsing:** Unstructured reports were parsed into JSON format, extracting key fields: Gender, Clinical History, Findings, and Impression.
- **Text Normalization:** Redundant whitespace (e.g., sequences exceeding two spaces) was removed to compact the sequence length.
- **Token Standardization:** Special characters were replaced to prevent tokenizer failures (e.g., Roman numerals were converted to Arabic numerals to unify representation while preserving standard staging semantics where possible.), ensuring compatibility with the LLM vocabulary.

B. Implementation Details

B.1. Implementation of PETRG-3D

Model Architecture. For the PETRG-3D framework, we utilize the ViT3D pre-trained in [45] as the volume encoder.

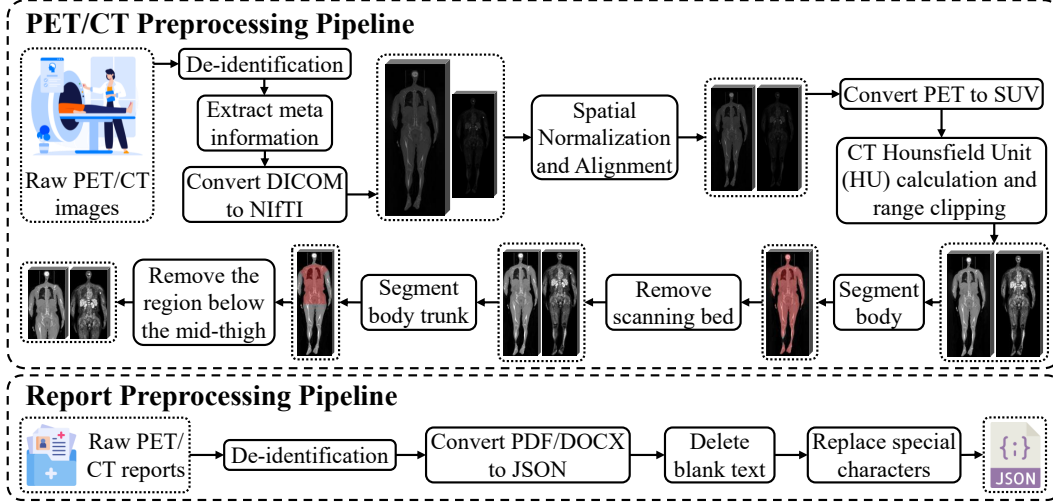


Figure 1. Preprocessing pipeline for PET/CT images and reports.

We fine-tune the subsequent Perceiver Resampler, which utilizes 128 learnable latent queries to compress the volumetric features into a fixed sequence of 128 visual tokens (with a hidden dimension of 768). These visual tokens are then projected via a linear layer to align with the embedding dimension of the LLM (i.e., 4096). For the text decoder, we benchmark six state-of-the-art Large Language Models (LLMs) with approximately 8 billion parameters: Llama2-7B[42], Mistral-7B-v0.3[1], Qwen2.5-7B[36], Gemma2-9B[40], GLM4-9B[15], and Qwen3-8B[47]. Given that our target reports are in Chinese, we employ the official pre-trained weights for models with native Chinese support (i.e., Qwen2.5-7B¹, GLM4-9B², and Qwen3-8B³). Conversely, for Llama2-7B⁴, Mistral-7B-v0.3⁵, and Gemma2-9B⁶, we utilize versions that have been fine-tuned on Chinese corpora to ensure linguistic capability.

Training and Inference. Our framework is implemented in PyTorch 2.8 and trained using the AdamW optimizer. The initial learning rate is initialized at 5×10^{-5} with a linear warmup over the first 100 steps, followed by a constant schedule. For parameter-efficient fine-tuning via LoRA, we configure the rank $r = 8$, scaling factor $\alpha = 32$, and a dropout rate of 0.1. The model is trained for 30 epochs on two NVIDIA A800 GPUs with an effective batch size of

16. Regarding the LLM configuration, the maximum input sequence length is set to 2048 tokens. For evaluation, we employ the model checkpoint from the final training step. During inference, we utilize nucleus sampling with a top- p of 0.9 and a sampling temperature of 0.7. The repetition penalty is set to 1.05, and the maximum generation length is constrained to 1024 tokens.

Center-Specific Templates and Explicit Stop Tokens. We collected healthy patient report templates from four distinct medical centers, curated by senior nuclear medicine physicians at each institution. Since drafting PET/CT reports based on standardized healthy templates is a routine workflow in nuclear medicine departments, our Structure-Aware Multi-template Fusion (SAMF) module demonstrates strong generalization capabilities. It can be readily adapted to new centers without the need to reconstruct healthy templates from scratch, as these are typically pre-existing clinical assets. To explicitly define report termination during training, we append a special token “[end-of-report]” to the end of each ground-truth report.

B.2. Implementation of Comparative Methods

Due to the scarcity of methodologies dedicated specifically to PETRG, we benchmark our approach against state-of-the-art report generation models from the X-Ray and CT domains.

For methods supporting 3D inputs, namely RadFM [45] and M3D [3], we utilize their official implementations and pre-trained weights, adhering to the inference parameters specified in the original papers. As these models are limited to single-modality input, we feed the PET volume during inference for the PETRG task. Furthermore, to address their lack of optimization for Chinese output, we prompt

¹<https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>

²<https://huggingface.co/zai-org/GLM-4-9B-0414>

³<https://huggingface.co/Qwen/Qwen3-8B>

⁴<https://huggingface.co/FlagAlpha/Llama2-Chinese-7b-Chat>

⁵<https://huggingface.co/shenzhi-wang/Mistral-7B-v0.3-Chinese-Chat>

⁶<https://huggingface.co/shenzhi-wang/Gemma-2-9B-Chinese-Chat>

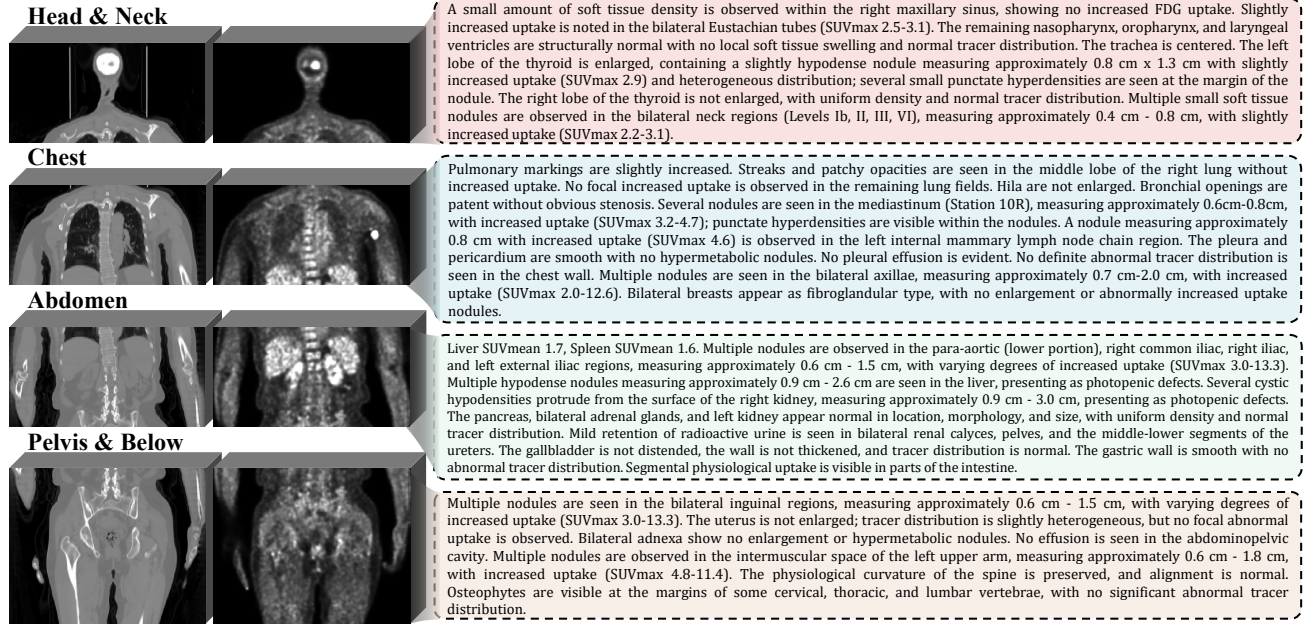


Figure 2. Illustration of region-level image-report pairs in PET/CT.

these models to generate reports in English, which are subsequently translated into Chinese using the Qwen3-Max model. We manually verified a subset of the translated reports to ensure clinical accuracy and terminology consistency.

Regarding PET2Rep [52], we evaluate two slice processing strategies proposed in their work: separate PET and CT slice inputs (PET2Rep-Sep) and fused PET/CT image inputs (PET2Rep-Fus). We perform inference using the official prompts and code. To ensure accurate evaluation, we post-process the raw outputs to filter out artifact symbols (e.g., “[]”, “**”) inherent to the template predictions. For the Vision-Language Model (VLM) backbones, we select the Qwen-VL and InternVL series, which demonstrated superior performance in [52]. Notably, we employ more recent iterations compared to the original study, utilizing Qwen3-VL-8B and InternVL-3.5-8B. For closed-source models, we employ Qwen-2.5-VL-Max (as Qwen-3-VL-Max was unavailable). Finally, for Llava-Med [33], we utilize the official code and pre-trained weights. Due to its 2D input constraint, we adopt the fused PET/CT image input strategy from [52] for inference.

B.3. Implementation of Region-wise Decomposition

Clinical PET/CT reports typically follow a strictly ordered anatomical description: head/neck, chest, abdomen, and pelvis. This structural consistency allows for the decomposition of the full report into four distinct regional descriptions.

Consequently, we propose a data augmentation strategy that partitions whole-body PET/CT images into these four anatomical sections and pairs them with the corresponding decoupled text (Figure 2). This approach scales the training dataset and simplifies the generative task by focusing on local regions.

Specifically, we employ a Body Part Regression (BPR) model [37, 46] to map the coronal slice indices of the CT images to anatomical labels (head, shoulder-neck, chest, abdomen, pelvis, and legs). Based on these mappings, we group “head” and “shoulder-neck” into the *Head & Neck* region, while “chest”, “abdomen”, and “pelvis/legs” form the *Chest*, *Abdomen*, and *Pelvis & Below* regions, respectively. To mitigate potential information loss caused by BPR boundary errors, we include a buffer of 10 redundant slices at the boundaries of each region. For textual decomposition, we leverage the Qwen3-Max model to segment the full-body report into the four corresponding sections. Figure 3 illustrates the prompt template used for this segmentation, and Figure 2 visualizes an example of the region-wise splitting. The resulting dataset contains 2,652 regional image-report pairs.

During training, the model takes regional PET/CT images as input and is supervised by the corresponding regional report. During inference, we generate reports for the four regions of the test images independently and concatenate them to form the final output. To prevent evaluation bias arising from potential content alterations during

Act as an expert Nuclear Medicine Physician. Reorganize the provided PET/CT "Findings" into four distinct anatomical sections: **Head & Neck, Chest, Abdomen, and Pelvis & Below.**

Instructions:

1. **Strict Fidelity:** Use only the provided information. Do not hallucinate or add content not present in the source text.
2. **Completeness:** Ensure every detail from the original text is included in the appropriate section.
3. **Missing Data:** If a specific region is not mentioned in the source, state: "No abnormal FDG uptake or structural changes observed."
4. **Precision:** Maintain accurate medical terminology. Preserve all key metrics (e.g., SUVmax, dimensions) exactly as they appear.
5. **Cohesion:** Logically consolidate descriptions within the same anatomical region for smooth flow.
6. **Output Format:** Return ONLY a valid, raw JSON object containing the four keys below. Do not include Markdown formatting (e.g., ``json) or conversational text.

Original Findings:
{original_findings}

Required JSON Structure:

```
{
  "Head_and_Neck": "string",
  "Chest": "string",
  "Abdomen": "string",
  "Pelvis_and_Below": "string"
}
```

Figure 3. Illustration of the prompt design utilized to instruct Qwen3-Max for the anatomical partitioning of whole-body PET/CT reports.

the LLM-based text segmentation, we construct the reference ground truth for evaluation by concatenating the LLM-segmented regions of the validation set.

C. Evaluation Metrics

C.1. NLG Metrics

To assess the textual similarity between generated and ground-truth reports, we employ three standard Natural Language Generation (NLG) metrics: BLEU [35], METEOR [5], and ROUGE-L [30]. Given that Chinese text lacks explicit word delimiters, we utilize the Jieba library⁷ for tokenization prior to calculation.

C.2. Clinical Efficacy Metric: PETRG-Score

C.2.1. Motivation

Evaluating the clinical utility of generated reports is critical. Standard NLG metrics often fail to capture factual correctness in medical findings. Furthermore, metrics designed for Chest X-Ray or CT report generation are inapplicable to PET/CT, as they ignore metabolic information (PET uptake). Recent attempts, such as PET2Rep [52] and ViMPET [32], exhibit significant limitations:

- **Incomplete Modality Coverage:** PET2Rep neglects structural abnormalities visible in CT, focusing solely on PET.
- **Coarse Granularity:** ViMPET defines only five broad anatomical regions and uses a binary uptake classification.

⁷<https://github.com/fxsjy/jieba>

cation. As shown in Tab. 1, real-world reports contain nuanced uptake descriptions (e.g., physiological vs. mild abnormal uptake) and fine-grained anatomical details (e.g., maxillary sinus) that are essential for *accurate diagnosis and staging*.

C.2.2. Ontology Construction

To overcome these deficiencies, we established a fine-grained evaluation ontology. We extracted 210 raw anatomical mentions from lymphoma reports, which were reviewed and merged by senior nuclear medicine physicians into a set of 24 clinically significant anatomical regions ($\mathcal{L}$). Similarly, we constructed a comprehensive vocabulary for PET uptake statuses and CT density phenotypes.

C.2.3. Calculation Pipeline

Based on this expert-defined ontology, we propose the **PETRG-Score**, a comprehensive metric evaluating both metabolic and structural alignment. The pipeline proceeds as follows:

(1) LLM-based Extraction: We utilize **Qwen3-Max** to extract structured labels (anatomical location, uptake status, density status) from both ground-truth and generated reports. To handle linguistic variations and report sparsity, we implement strict **Normalization Rules** within the prompt (see Fig. 4):

- **Default Normal:** Anatomical regions not mentioned in the report are automatically categorized as "Normal".
- **Hierarchical Inference:** If a major region (e.g., Lungs) is described as normal, all its sub-regions inherit the "Normal" status.
- **Significance Prioritization:** For regions with local abnormalities, the overall status reflects the most significant finding, while unmentioned sub-parts remain "Normal".

(2) Expert Verification: To ensure benchmark reliability, the extracted labels for the ground-truth reports were manually audited and corrected by a researcher specifically trained in medicine imaging.

(3) Metric Formulation: Let $\mathcal{D}$ denote the test set consisting of N reports. For the j -th report ($j \in \{1, \dots, N\}$) and the l -th anatomical region ($l \in \mathcal{L}$), let $y_{j,l} \in \mathcal{C}$ and $\hat{y}_{j,l} \in \mathcal{C}$ denote the ground-truth and predicted class labels, respectively. The valid category sets are $\mathcal{C}_{CT} = \{1, \dots, 8\}$ and $\mathcal{C}_{PET} = \{1, \dots, 5\}$.

We treat every anatomical region in every report as an individual classification instance. To mitigate the impact of class imbalance, we employ the Macro-F1 score. We first define the True Positives (TP_k), False Positives (FP_k), and False Negatives (FN_k) for a specific category $k \in \mathcal{C}$ by aggregating across all reports and regions:

$$TP_k = \sum_{j=1}^N \sum_{l \in \mathcal{L}} \mathbb{I}(y_{j,l} = k \wedge \hat{y}_{j,l} = k) \quad (2)$$

Table 1. Extraction of anatomical locations, uptake statuses, and density descriptions from the report in Figure 2. This exemplifies the semantic richness of PET/CT reports, underscoring the critical inadequacy of existing oversimplified CE metrics.

Region	Anatomical Location	Uptake Status	Density Status
Head&Neck	Right maxillary sinus	Showing no increased FDG uptake	A small amount of soft tissue density is observed
	Bilateral Eustachian tubes	Slightly increased uptake is noted	-
	Left lobe of the thyroid	Slightly increased uptake and heterogeneous distribution	Enlarged, containing a slightly hypodense nodule measuring approx. 0.8 cm x 1.3 cm; several small punctate hyperdensities are seen at the margin
	Bilateral neck regions (Levels Ib, II, III, VI)	Slightly increased uptake	Multiple small soft tissue nodules, measuring approx. 0.4 cm - 0.8 cm
Chest	Middle lobe of right lung	Without increased uptake	Streaks and patchy opacities are seen
	Mediastinum (Station 10R)	With increased uptake	Several nodules, measuring approx. 0.6 cm - 0.8 cm; punctate hyperdensities are visible within the nodules
	Left internal mammary lymph node chain	With increased uptake	A nodule measuring approx. 0.8 cm
	Bilateral axillae	With increased uptake	Multiple nodules, measuring approx. 0.7 cm - 2.0 cm
Abdomen	Liver background	-	-
	Para-aortic (lower), Right common iliac, Right iliac, Left external iliac	With varying degrees of increased uptake	Multiple nodules observed, measuring approx. 0.6 cm - 1.5 cm
	Liver	Presenting as photopenic defects	Multiple hypodense nodules measuring approx. 0.9 cm - 2.6 cm
	Right kidney	Presenting as photopenic defects	Several cystic hypodensities protrude from the surface, measuring approx. 0.9 cm - 3.0 cm
	Renal calyces, pelves, ureters	Mild retention of radioactive urine	Mild retention of radioactive urine
Pelvis&Below	Bilateral inguinal regions	With varying degrees of increased uptake	Multiple nodules, measuring approx. 0.6 cm - 1.5 cm
	Intermuscular space of left upper arm	With increased uptake	Multiple nodules observed, measuring approx. 0.6 cm - 1.8 cm
	Spine (Cervical, Thoracic, Lumbar)	With no significant abnormal tracer distribution	Osteophytes are visible at the margins of some vertebrae

$$FP_k = \sum_{j=1}^N \sum_{l \in \mathcal{L}} \mathbb{I}(y_{j,l} \neq k \wedge \hat{y}_{j,l} = k) \quad (3)$$

$$FN_k = \sum_{j=1}^N \sum_{l \in \mathcal{L}} \mathbb{I}(y_{j,l} = k \wedge \hat{y}_{j,l} \neq k) \quad (4)$$

where $\mathbb{I}(\cdot)$ is the indicator function.

The F1 score for category k is calculated as:

$$F1_k = \frac{2 \cdot TP_k}{2 \cdot TP_k + FP_k + FN_k} \quad (5)$$

The final **PETRG-Score** is the macro-average over a target subset of categories $\mathcal{S} \subseteq \mathcal{C}$:

$$\text{PETRG-Score}(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{k \in \mathcal{S}} F1_k \quad (6)$$

Table 2. Comparison of attribute coverage between PETRG-Score and existing metrics. Corroborated by the real-world labels in Tab. 1, PETRG-Score demonstrates significantly more comprehensive anatomical coverage and provides a richer taxonomy for both metabolic uptake and structural density states, surpassing current evaluation standards.

Metrics	Anatomical Locations	Uptake Status	Density Status
PET2Rep[52]	Cranium and Brain; Eyeballs; Nasal Cavity and Sinuses; Pharynx and Parapharyngeal Space; Palatine Tonsils and Larynx; Salivary Glands and Thyroid; Cervical Lymph Nodes; Lungs and Thoracic Cavity; Mediastinum and Heart; Esophagus; Liver; Gallbladder; Pancreas; Spleen; Kidneys and Adrenal Glands; Gastrointestinal Tract; Prostate/Uterus and Bladder; Abdominal and Pelvic Cavities; Spine and Bones	Increased Uptake; Decreased Uptake; Absent Uptake; Normal	-
ViMed-PET[32]	Mediastinum; Lung; Abdomen; Axilla; Cervical Region	Increase; Not Increase	Lymph Node; Pulmonary Nodule; Ground-Glass Opacity; Pulmonary Mass; Pleural Thickening; Interstitial Thickening; Consolidation; Effusion; Soft Tissue Nodule; Wall Thickening; Calcified Nodule; Hypermetabolic Lesion
PETRG-Score (Ours)	Brain, Skull, and Meninges; Orbit Nasal Cavity, and Paranasal Sinuses; Pharyngeal Spaces, Tonsils, and Larynx; Thyroid Gland and Major Salivary Glands (Parotid, Submandibular); Cervical Lymph Nodes; Lungs and Pleura; Mediastinum and Hila (including Lymph Nodes); Heart and Pericardium; Axilla and Chest Wall; Breasts; Liver; Gallbladder and Biliary Tract; Spleen; Pancreas; Kidneys; Adrenal Glands; Gastrointestinal Tract (Esophagus, Stomach, Intestines); Retroperitoneal Space (including Lymph Nodes); Peritoneum, Mesentery, and Omentum; Pelvic Organs (Bladder, Uterus/Adnexa or Prostate/Seminal Vesicles); Pelvic and Inguinal Lymph Nodes; Spine; Pelvis and Bones of Extremities; Muscles and Subcutaneous Tissue	Intense Abnormal Uptake; Mild/Suspicious Abnormal Uptake; Physiological/Background Uptake; Uptake Defect / Decreased Uptake; Normal	Lymphadenopathy; Focal Lesion; Lung Parenchymal Abnormality; Wall/Membrane Thickening; Calcification; Bone/Skeletal Lesion; Other Abnormality; Normal

We report four variants to provide a comprehensive view:

- **PET-All / CT-All:** Calculated over all categories ($\mathcal{S} = \mathcal{C}_{PET}$ or $\mathcal{S} = \mathcal{C}_{CT}$).
- **PET-Ab / CT-Ab:** Calculated only on abnormal categories, excluding the “Normal” class (ID: 5 for PET, ID: 8 for CT).

C.2.4. Remarks on Limitations

Extraction Bias: Verifying labels for all generated reports in ablation studies is prohibitively expensive. We limit manual verification to the ground truth. To ensure fairness, all methods share the same LLM extraction pipeline, ensuring any extraction noise affects all baselines equally. **Long-**

You are a professional Nuclear Medicine Physician. Your task is to extract the radiotracer uptake status and CT structure/density status for 24 key anatomical regions from the provided PET/CT "Findings" report.

I. Task 1: Extract PET Uptake Status

1. Status Classification: Categorize the uptake status of each organ using one of the 5 codes below:

- * **"1" (Intense Abnormal Uptake):** Explicit descriptions of increased uptake, tracer concentration (hypermetabolism), etc. Corresponds to Deauville Score 5, clearly suggesting malignancy.
- * **"2" (Mild/Suspicious Abnormal Uptake):** Corresponds to Deauville Score 3 (suspicious in certain contexts) or uptake slightly higher than the liver background.
- * **"3" (Physiological/Background Uptake):** Explicitly indicated as physiological uptake.
- * **"4" (Uptake Defect/Decreased Uptake):** Explicit descriptions of photopenia, reduced radioactivity, or uptake defects.
- * **"5" (Normal):** Descriptions such as "no tracer concentration seen," "no abnormal uptake," "uniform distribution," or inferred as normal based on normalization rules.

2. SUV Value Description: If the report describes SUV information for the structure, extract the specific SUVmax or SUVmean value into the `suv` field (e.g., `suv`: "SUVmax: 4.5"). Otherwise, leave it as an empty string `""`. Note: If multiple SUV values exist for a region, select the maximum value as the SUVmax.

II. Task 2: Extract CT Structure/Density Status

1. Status Classification: Categorize the CT findings of each organ using one of the 8 codes below:

- * **"1" (Lymphadenopathy):** Abnormalities in lymph node morphology or size.
- * **"2" (Focal Lesion):** Includes nodules and masses.
- * **"3" (Lung Parenchymal Abnormality):** Includes ground-glass opacities (GGO), linear opacities/stranding, and patchy shadows/consolidations.
- * **"4" (Wall/Membrane Thickening):** Increased thickness of a normal anatomical structure (usually walls, membranes, or interstitium).
- * **"5" (Calcification):** High-density calcium salt deposition within tissues.
- * **"6" (Bone/Skeletal Lesion):** e.g., osteoblastic or osteolytic lesions.
- * **"7" (Other Abnormalities):** Any abnormalities not listed above.
- * **"8" (Normal):** Not described, explicitly stated as no abnormality, or inferred as normal based on normalization rules.

2. Extraction Description: If the status falls into the first 7 "Abnormal" categories, briefly extract the specific description into the `ct\_description` field; otherwise, leave it as an empty string `""`.

III. Normalization Rules

- 1. Default Normal:** If an organ is not mentioned in the report at all, its status should be categorized as "Normal".
- 2. Hierarchical Normal:** If a major category (e.g., "Lungs and Pleura") is described as normal, all its sub-regions are considered "Normal".
- 3. Implied Normal:** If only a local abnormality within an organ is mentioned (e.g., "Increased uptake in cervical zone II"), the unmentioned parts of that anatomical region are considered "Normal," but the overall status of the organ should reflect the most significant finding (in this case, "Markedly Abnormal Uptake").

IV. Output Format

Please output the result strictly in the following JSON format without adding any extra explanation. Ensure the `pet\_status` and `ct\_status` fields are formatted strictly as single numbers.

JSON Template:

```
```json
{
 "Brain, Skull, and Meninges": {"pet_status": "...", "suv": "...", "ct_status": "...", "ct_description": "..."},
 "Orbits, Nasal Cavity, and Paranasal Sinuses": {"pet_status": "...", "suv": "...", "ct_status": "...", "ct_description": "..."},
 "Pharyngeal Space, Tonsils, and Larynx": ...,
 ...,
 "Muscles and Subcutaneous Tissue": ...
}
```
```

Figure 4. Prompt designed for Qwen3-Max to extract attributes from ground-truth and generated reports. The prompt targets 5 uptake states and 8 density states across 24 anatomical locations. We explicitly constrain the output to numerical indices rather than textual labels; this design choice minimizes model hallucinations and ensures robust downstream parsing.

tail Coverage: While our ontology is significantly richer than prior arts, extremely rare anatomical variants or atypical descriptions may still be underrepresented. Developing dynamic, open-vocabulary evaluation metrics remains a direction for future work.

D. Additional Results

D.1. Center-wise Style Adaptation Analysis

Table 3. Multi-center evaluation of the SAMF module’s impact on adapting to heterogeneous reporting styles. “w/o” and “w/ SAMF” denote the PETRG-3D baseline without and with the module, respectively. Centers 1–4 belong to the PETRG-Lym dataset, while Center 5 corresponds to AutoPET-RG-Lym. Performance gains are highlighted in green parentheses.

| Methods | Center ID | B-4 | MTR | R-L |
|----------|-----------|----------------|----------------|----------------|
| w/o SAMF | 1 | 26.31 | 44.47 | 38.67 |
| | 2 | 46.71 | 50.75 | 54.89 |
| | 3 | 44.86 | 52.53 | 53.47 |
| | 4 | 35.69 | 42.35 | 46.62 |
| | 5 | 8.11 | 28.18 | 24.57 |
| w/ SAMF | 1 | 26.53 (+0.22) | 43.03 (-1.44) | 41.22 (+2.55) |
| | 2 | 53.38 (+6.67) | 57.19 (+6.44) | 61.81 (+6.92) |
| | 3 | 55.84 (+10.98) | 62.00 (+9.47) | 63.82 (+10.35) |
| | 4 | 43.09 (+7.40) | 49.08 (+6.73) | 53.29 (+6.67) |
| | 5 | 32.47 (+24.36) | 41.50 (+13.32) | 40.97 (+16.40) |

To quantitatively evaluate the efficacy of the SAMF module in style adaptation, Table 3 details the NLG evaluation metrics across four distinct centers within the PETRG-Lym dataset. Integrating the SAMF module yielded consistent improvements across nearly all metrics for the internal centers, confirming its ability to accommodate the diverse reporting conventions inherent to different medical institutions.

Notably, for the external test center (Center 5, AutoPET-RG-Lym), the inclusion of SAMF resulted in substantial performance gains: BLEU-4 surged by 24.36 points, while METEOR and ROUGE-L saw absolute increases of 13.32 and 16.40 points, respectively. These results suggest that even when confronted with unseen reporting styles, PETRG-3D equipped with SAMF effectively performs style transfer by leveraging the synergy between patient-specific PET/CT images and retrieved style templates from the target center.

D.2. Performance Analysis by Uptake and Density Status

We further stratified model performance by uptake and density statuses to delineate diagnostic capabilities and potential biases. Table 4 and Table 5 present the Clinical Efficacy (CE) metrics—Precision (P), Recall (R), and F1 score (F1)—alongside the “Support” column, which highlights the class distribution.

Table 4. Evaluation of our model across five uptake statuses. “P” and “R” denote Precision and Recall, respectively. “Support” indicates the sample size for each status.

| Uptake Status | P | R | F1 | Support |
|---------------------------------|-------|-------|-------|---------|
| Intense Abnormal Uptake | 35.55 | 19.93 | 25.54 | 537 |
| Mild/Suspicious Abnormal Uptake | 15.94 | 11.53 | 13.38 | 347 |
| Physiological/Background Uptake | 28.85 | 23.08 | 25.64 | 65 |
| Uptake Defect/Decreased Uptake | 13.92 | 13.25 | 13.58 | 83 |
| Normal | 77.68 | 87.25 | 82.19 | 2832 |
| Macro Avg | 34.39 | 31.01 | 32.06 | 3864 |
| Weighted Avg | 64.09 | 68.43 | 65.71 | 3864 |

Table 5. Evaluation of our model across eight density statuses. “P” and “R” denote Precision and Recall, respectively. “Support” indicates the sample size for each status.

| Density Status | P | R | F1 | Support |
|------------------------------|-------|-------|-------|---------|
| Lymphadenopathy | 55.56 | 37.72 | 44.93 | 464 |
| Focal Lesion | 13.46 | 12.07 | 12.73 | 116 |
| Lung Parenchymal Abnormality | 68.13 | 66.31 | 67.21 | 187 |
| Wall/Membrane Thickening | 14.29 | 9.52 | 11.43 | 63 |
| Calcification | 9.30 | 7.55 | 8.33 | 53 |
| Bone/Skeletal Lesion | 36.99 | 30.00 | 33.13 | 90 |
| Other Abnormality | 28.98 | 19.85 | 23.56 | 413 |
| Normal | 72.08 | 82.08 | 76.75 | 2478 |
| Macro Avg | 37.35 | 33.14 | 34.76 | 3864 |
| Weighted Avg | 60.91 | 63.82 | 61.84 | 3864 |

The results indicate that the “Normal” category significantly outperforms others across both uptake and density classifications. This disparity is largely attributable to the extreme class imbalance, where the prevalence of normal samples biases the model towards predicting anatomical regions as abnormality-free. This distribution skew justifies our adoption of the Macro F1 score as a primary evaluation metric to ensure fair assessment.

Interestingly, apart from the normal class, the “Physiological/Background Uptake” category (Table 4) achieved the highest performance despite having the fewest samples. This is likely due to the strong semantic coupling between physiological uptake and specific anatomical structures (e.g., the bladder or kidneys), which simplifies the learning task. Similarly, “Lung Parenchymal Abnormality” (Table 5) ranks second to the normal category, presumably benefiting from its distinct localization within the lung fields. These findings highlight that while the model learns strong anatomy-pathology associations, mitigating data imbalance remains a critical frontier for future PET/CT report generation research.

D.3. Computational Resource Analysis

Table 6 reports the training time and peak GPU memory usage for PETRG-3D. Experiments were conducted on two NVIDIA A800 (80GB) GPUs using a standardized software

Table 6. Training resource consumption of PETRG-3D with different LLMs on PETRG-Lym dataset. Memory values indicate the peak GPU memory usage (in GB) across two A800 GPUs.

| LLMs | Time (Hours) | Memory (GB) |
|-----------------|--------------|-------------|
| Llama2-7B | 5.70 | 30.14 |
| Mistral-7B-v0.3 | 5.56 | 34.49 |
| Qwen2.5-7B | 5.70 | 38.27 |
| Gemma2-9B | 8.62 | 45.23 |
| GLM4-9B | 8.05 | 41.10 |
| Qwen3-8B | 7.33 | 38.81 |

stack⁸. As anticipated, both training duration and memory requirements scale positively with the parameter size of the underlying LLM.

D.4. Qualitative Analysis (Whole-body Reporting)

Figure 5 presents a representative whole-body PET/CT lymphoma report. Authored by a nuclear medicine physician, the report is extensive, covering regions from the head and neck to the musculoskeletal system, thereby underscoring the complexity of the PETRG task.

Figure 6 illustrates the output from PET2Rep [52]. As highlighted in red, the baseline model suffers from significant hallucinations, particularly in lesion-rich areas like the lungs and abdomen (e.g., incorrectly predicting volume reduction or pancreatic nodules). This suggests that generalist VLMs or those without specialized fine-tuning lack the diagnostic reliability required for complex medical imaging tasks.

In contrast, the report generated by our PETRG-3D model (Figure 7) closely mirrors the ground truth in both linguistic style and diagnostic content. It accurately identifies multiple nodal involvements (hepatic hilum, para-aortic, iliac, etc.) and extranodal abnormalities. While some limitations remain—specifically regarding quantitative precision (e.g., SUVmax values) and minor false positives in complex regions—the drastic reduction in hallucinations demonstrates the effectiveness of our fine-tuning strategy. Current errors are largely numerical, pointing to future directions for integrating quantitative analysis heads into the architecture.

D.5. Effectiveness of Explicit Stop Tokens

Figure 8 compares model outputs before and after the incorporation of explicit stop tokens. Without the explicit stop token, the model tends to generate prolonged hallucinatory content after the actual report concludes, as it attempts to fill the fixed `max_new_token` buffer required for batch generation. By introducing an explicit stop token, the model

learns to output “[end-of-report]” immediately upon completing the valid report content. This allows for early termination of the generation process, effectively eliminating post-report hallucinations.

⁸Environment: Python 3.10, CUDA 12.8, PyTorch 2.8.0, Transformers 4.57.0, PEFT 0.17.1, and DeepSpeed 0.17.5.

Ground Truth:



PET/CT

<ZH/Original> 颈部: 左侧上颌窦腔内见少许软组织密度影, 未见放射性摄取增高。鼻咽腔后上壁稍肿胀, 放射性摄取增高, SUVmax4.2, 余口咽咽腔及喉室腔结构正常, 局部软组织未见肿胀, 放射性分布正常。气管居中。甲状腺两叶不大, 密度尚均匀, 放射性分布正常。双侧颈部 (1、2、3、4、5、6、8) 见数枚软组织密度小结节影, 大小约0.4cm-0.8cm, 放射性摄取增高, SUVmax2.5-8.2。胸部: 双肺纹理稍增多, 右肺中叶见少许条索影, 未见放射性摄取增高, 肺野内未见放射性摄取增高灶。肺门不大, 支气管开口通畅, 管腔未见明显狭窄。纵隔 (2R、4、5、6、10) 见数枚结节影, 大小约0.4cm-0.7cm, 放射性摄取增高, SUVmax2.8-4.5, 部分结节内见点片状致密影。胸膜及心包膜光滑, 未见放射性摄取增高结节。胸腔内未见积液影。双侧腋下见多枚结节影, 大小约0.3cm-1.4cm, 放射性摄取不同程度增高, SUVmax1.7-11.2。纵隔血池SUVmax1.7。腹、盆部: 肝脏SUVmean2.0, 肝脏血池SUVmax2.7。脾脏体积增大, 脾内见2枚稍低密度结节影, 大小约1.4cm-1.9cm, 呈放射性缺损, 余脾脏实质弥漫性放射性摄取增高, 分布不均匀, SUVmax6.7, SUVmean2.8。肝门区、腹主动脉旁、双侧髂总血管旁、双侧髂血管旁、双侧腹股沟见多枚结节影, 大小约0.6cm-2.0cm, 放射性摄取不同程度增高, SUVmax2.7-10.6。左肾上腺联合部见一枚结节影, 大小约1.1cm*1.6cm, 放射性摄取稍高, SUVmax2.3。胰腺、右肾上腺及双肾位置、形态、大小均正常, 密度均匀, 放射性分布正常。双肾盂、肾盂及输尿管中下段内见少量放射性尿液滞留。胆囊张力稍大, 壁厚, 放射性分布正常。胃壁光滑, 放射性分布未见异常。部分肠道可见节段性摄取。前列腺不大, 放射性分布欠均匀, SUVmax1.4, 未见放射性摄取异常增高灶。腹盆腔未见积液影。骨骼、肌肉和大脑: 双侧上臂及膝关节肌肉间隙见数枚结节影, 大小约0.3cm-0.6cm, 放射性摄取不同程度增高, SUVmax2.6-7.2。全身皮肤不同程度增厚, 放射性摄取增高, 分布不均匀, SUVmax1.7-7.5, 以双侧腋下、腹股沟皮肤显著。椎体生理弯曲存在, 序列正常, 颈椎、胸椎及腰椎部分椎体边缘可见不同程度骨质增生迹象, 放射性分布未见明显异常。颅内大脑皮质、小脑半球弥漫性放射性摄取减低, 双侧基底节、丘脑放射性分布正常, 两侧对称。诸脑室形态、大小、位置无改变; 脑沟、脑池无增宽; 中线结构无移位。

<EN/Translated> Head and Neck: A small soft tissue density is observed within the left maxillary sinus without increased tracer uptake. The posterior-superior wall of the nasopharynx appears slightly thickened with increased FDG uptake (SUVmax 4.2). The remaining oropharyngeal and laryngeal structures appear normal; no localized soft tissue swelling or abnormal tracer distribution is noted in these areas. The trachea is midline. The thyroid lobes are normal in size with homogeneous density and physiological tracer distribution. Multiple small soft tissue nodules are visualized in the bilateral cervical regions (Levels I, II, III, IV, V, VI, and VIII), measuring approximately 0.4 cm to 0.8 cm in diameter, demonstrating increased FDG uptake with SUVmax ranging from 2.5 to 8.2. **Chest:** Bilateral lung markings are slightly increased. A few fibrous streaks are noted in the right middle lobe without increased tracer uptake; no other hypermetabolic lesions are observed in the lung fields. The hila are not enlarged, and the bronchial trees are patent without obvious stenosis. Multiple lymph nodes are seen in the mediastinum (Stations 2R, 4, 5, 6, and 10), measuring approximately 0.4 cm to 0.7 cm, with increased uptake (SUVmax 2.8-4.5); punctate and patchy calcifications are visible within some of these nodes. The pleura and pericardium are smooth without hypermetabolic nodules. No pleural effusion is evident. Multiple lymph nodes are observed in the bilateral axillae, measuring 0.3 cm to 1.4 cm, exhibiting varying degrees of increased uptake (SUVmax 1.7-11.2). The mediastinal blood pool SUVmax is 1.7. **Abdomen and Pelvis:** The liver shows a mean SUV of 2.0, with a liver blood pool SUVmax of 2.7. The spleen is enlarged, containing two slightly hypodense nodules measuring approximately 1.4 cm to 1.9 cm that present as photopenic defects. The remaining splenic parenchyma demonstrates diffuse, heterogeneous increased uptake with an SUVmax of 6.7 and SUVmean of 2.8. Multiple lymph nodes are visualized in the hepatic hilar region, para-aortic region, bilateral common iliac, bilateral external/internal iliac, and bilateral inguinal regions. These nodes measure approximately 0.6 cm to 2.0 cm and show varying degrees of increased uptake (SUVmax 2.7-10.6). A nodule is noted at the junction of the left adrenal gland, measuring approximately 1.1 cm × 1.6 cm, with mildly increased uptake (SUVmax 2.3). The pancreas, right adrenal gland, and kidneys are normal in position, morphology, and size, with homogeneous density and normal tracer distribution. Mild retention of radioactive urine is seen in the bilateral renal calyces, pelvis, and mid-to-distal ureters. The gallbladder shows slightly increased tension with a non-thickened wall and normal tracer distribution. The gastric wall is smooth with no abnormal uptake. Segmental physiologic uptake is noted in parts of the bowel. The prostate is not enlarged, showing heterogeneous tracer distribution (SUVmax 1.4) without focal hypermetabolism. No abdominopelvic ascites is observed. **Musculoskeletal, Skin, and Brain:** Multiple nodules are observed in the intermuscular spaces of the bilateral upper arms and knee joints, measuring approximately 0.3 cm to 0.6 cm, with varying degrees of increased uptake (SUVmax 2.6-7.2). There is generalized, varying skin thickening with increased, heterogeneous tracer uptake (SUVmax 1.7-7.5), which is most prominent in the bilateral axillary and inguinal regions. The physiological curvature of the spine is preserved, and alignment is normal. Varying degrees of osteophytes are noted at the margins of the cervical, thoracic, and lumbar vertebrae; however, no definite abnormal tracer distribution is seen in the skeletal system. Intracranially, there is diffuse hypometabolism (decreased uptake) in the cerebral cortex and cerebellar hemispheres. The bilateral basal ganglia and thalami show normal, symmetrical tracer distribution. The morphology, size, and position of the ventricles remain unchanged. The sulci and cisterns are not widened, and midline structures are not shifted.

Figure 5. Example of a clinical whole-body PET/CT report for lymphoma. The original report is in Chinese (ZH) and has been translated into English (EN) for illustration.

Generated (PET2Rep):

<ZH/Original> 禁食状态下静脉注射18F-FDG，静息60分钟后行全身PET断层显像，全身断层影像显示：脑形态、结构正常，双侧大脑皮质放射性分布均匀、对称，FDG摄取未见明显异常。各脑室、脑沟、脑裂、脑池形态正常，中线结构无偏移。双侧眼球形态和轮廓无异常，FDG无异常摄取。右侧上颌窦黏膜稍增厚伴部分钙化，FDG摄取未见异常。鼻咽壁未见增厚，两侧腭扁桃体对称，FDG呈生理性摄取。咽喉形态、结构未见异常。双侧腮腺、颌下腺形态、密度正常，FDG生理性摄取。甲状腺形态、大小正常，密度均匀，FDG摄取未见异常。双侧颈深间隙、颌下见数枚小淋巴结显示，部分淋巴结FDG摄取增高。两肺透亮度增加，两肺上叶胸膜下见囊状透亮影。右肺中叶体积缩小，右肺中叶肺门旁见片状密度增高影，FDG摄取增高。心影未见异常。部分动脉管壁钙化（包括冠状动脉）。食管未见扩张，管壁未见明显增厚或肿块，FDG摄取未见增高。肝缘欠光整，肝裂未见明显增宽，CT平扫肝右叶下段见巨大稍低密度肿块影，边界欠清，FDG摄取增高。肝左叶见稍低密度结节，边界尚清晰，FDG摄取较基底摄取减低。肝右叶膈顶区另见结节状致密影。门静脉主干及左右支明显增粗，密度增高，FDG摄取增高。门腔间隙、肝胃间隙、腹膜后见淋巴结显示，FDG摄取轻度增高。胆囊稍饱满，胆囊壁未见增厚，局部FDG摄取未见异常。胰腺形态正常，胰体部见小结节状钙化灶，主胰管未见增宽，FDG摄取未见明显异常。脾脏形态、大小密度未见异常，FDG摄取未见异常。双侧泌尿系统见对比剂残留影；左肾体积缩小，左肾见多枚囊状低密度影及稍高密度小结节，边界清，FDG摄取缺失；右肾上极见一脂肪密度结节，FDG摄取未见异常。双侧肾盂肾盏及输尿管未见增宽。双侧肾上腺显影未见明显异常，FDG摄取未见异常。胃充盈尚可，胃壁未见明显增厚，部分胃壁FDG摄取增高。肠道充盈欠佳，肠壁未见明显增厚或肿块，FDG呈生理性摄取。前列腺稍增大，实质内见数枚钙化灶，FDG摄取未见异常增高。膀胱充盈欠佳，内见对比剂残留影。骨质密度不均匀减低，脊柱部分椎体边缘骨质增生。FDG摄取未见异常。

<EN/Translated> Whole-body PET/CT tomography was performed 60 minutes after intravenous injection of 18F-FDG under fasting conditions. The whole-body tomographic images reveal: Brain morphology and structure are normal; radioactivity distribution in the bilateral cerebral cortex is uniform and symmetrical, with no obvious abnormal FDG uptake. The morphology of cerebral ventricles, sulci, fissures, and cisterns is normal, and midline structures are not shifted. Bilateral eyeballs show no morphological or contour abnormalities and no abnormal FDG uptake. Slight mucosal thickening with partial calcification is observed in the right maxillary sinus, without abnormal FDG uptake. The nasopharyngeal walls are not thickened; bilateral palatine tonsils are symmetrical with physiological FDG uptake. Morphology and structure of the laryngopharynx are unremarkable. Bilateral parotid and submandibular glands show normal morphology and density with physiological FDG uptake. The thyroid gland is normal in morphology and size with uniform density and no abnormal FDG uptake. Several small lymph nodes are visualized in the bilateral deep cervical and submandibular spaces, with some showing increased FDG uptake. Increased radiolucency is observed in both lungs, with subpleural cystic lucencies in the upper lobes. The right middle lobe shows volume loss and patchy increased density shadows near the hilum, accompanied by increased FDG uptake. The cardiac shadow is unremarkable. Calcifications are present in the walls of some arteries (including coronary arteries). The esophagus is not dilated, wall thickening or masses are not distinct, and FDG uptake is not increased. The liver margin is slightly irregular, and hepatic fissures are not significantly widened; plain CT scan reveals a giant, slightly low-density mass in the lower segment of the right hepatic lobe with ill-defined boundaries and increased FDG uptake. A slightly low-density nodule with clear boundaries is seen in the left hepatic lobe, showing FDG uptake lower than background. Additionally, a nodular dense shadow is observed in the dome of the right hepatic lobe. The main portal vein and its left and right branches are significantly thickened with increased density and increased FDG uptake. Lymph nodes are visualized in the portocaval space, hepatogastric space, and retroperitoneum, showing mildly increased FDG uptake. The gallbladder is slightly distended without wall thickening, and local FDG uptake is unremarkable. The pancreas morphology is normal; a small nodular calcification is seen in the pancreatic body; the main pancreatic duct is not dilated, and no obvious abnormal FDG uptake is observed. Spleen morphology, size, and density are unremarkable, with no abnormal FDG uptake. Contrast retention is noted in the bilateral urinary systems; the left kidney is atrophic, containing multiple cystic low-density shadows and a slightly high-density small nodule with clear boundaries and absent FDG uptake; a fat-density nodule is seen in the upper pole of the right kidney without abnormal FDG uptake. Bilateral renal pelves, calyces, and ureters are not dilated. Bilateral adrenal glands show no obvious abnormalities in appearance or FDG uptake. Gastric filling is fair; the gastric wall is not significantly thickened, though partial gastric wall shows increased FDG uptake. Bowel filling is suboptimal; no obvious wall thickening or masses are seen, and FDG distribution is physiological. The prostate is slightly enlarged with parenchymal calcifications, without abnormally increased FDG uptake. The bladder is poorly filled with contrast retention observed within. Bone density is unevenly decreased, and osteohyperplasia is noted at the edges of some vertebral bodies. No abnormal osseous FDG uptake is observed.

Figure 6. Case study on PET2Rep [52] generation. Green and red backgrounds denote statements consistent and inconsistent with the Ground Truth (Fig. 5), respectively.

Generated (PETRG-3D):

<ZH/Original> **颈部:** 双侧上颌窦腔内未见软组织密度影, 未见放射性摄取增高。鼻咽、口咽腔及喉室腔结构正常, 局部软组织未见肿胀, 放射性分布正常。气管居中。甲状腺两叶不大, 右叶密度欠均匀, 内见稍低密度微小结节影, 大小约0.2cm, 未见放射性摄取异常增高。双侧颈部(1b)见数枚软组织密度小结节影, 大小约0.7cm-1.3cm, 放射性摄取增高, SUVmax6.4-18.1。**胸部:** 双肺纹理稍增多, 肺野内未见放射性摄取增高灶。肺门不大, 支气管开口通畅, 管腔未见明显狭窄。纵隔(2R、4R、5、10)见数枚软组织密度结节影, 大小约0.6cm-1.1cm, 放射性摄取增高, SUVmax5.3-19.5, 且以右侧显著。胸膜及心包膜光滑, 未见放射性摄取增高结节。左胸腔内见少许孤立性低密度影, 呈放射性缺损。双侧腋下见多枚软组织密度小结节影, 大小约0.4cm-1.0cm, 放射性摄取稍高, SUVmax1.7-3.1。**腹、盆部:** 肝脏SUVmean1.6, 脾脏SUVmean1.8。肝门区、胰周、腹主动脉旁、肠系膜、左侧髂总血管旁、左侧髂血管旁、双侧腹股沟见多枚软组织密度结节影, 大小约0.5cm-2.6cm, 放射性摄取不同程度增高, SUVmax3.4-24.1。左肾内见一类圆形稍高密度影, 大小约1.9cm*2.3cm, 呈放射性缺损。左肾盂内见点状致密影。胰腺、双肾上腺及右肾位置、形态、大小均正常, 密度均匀, 放射性分布正常。双肾盂、肾盂及输尿管中下段内见少量放射性尿液滞留。胆囊张力不大, 壁不厚, 放射性分布正常。胃壁光滑, 放射性分布未见异常。部分肠道可见节段性摄取。前列腺不大, 放射性分布欠均匀, 未见放射性摄取异常增高灶, 内见点状钙化影。腹盆腔未见积液影。**骨骼、肌肉和大脑:** 椎体生理弯曲存在, 序列正常, 颈椎、胸椎及腰椎部分椎体边缘可见不同程度骨质增生迹象, 放射性分布未见明显异常。颅内大脑皮质、双侧基底节、丘脑及小脑半球放射性分布正常, 两侧对称。诸脑室形态、大小、位置无改变; 脑沟、脑池无增宽; 中线结构无移位。

<EN/Translated> **Head and Neck:** The bilateral maxillary sinuses are clear, showing no soft tissue densities or abnormal tracer uptake. The nasopharynx, oropharynx, and laryngeal structures appear normal; no localized soft tissue swelling or abnormal tracer distribution is observed. The trachea is midline. The thyroid lobes are normal in size; the right lobe demonstrates slightly heterogeneous density and contains a minute hypodense nodule measuring approximately 0.2 cm, which does not exhibit abnormal tracer uptake. Multiple soft tissue nodules are visualized in the bilateral cervical regions (Level 1B), measuring approximately 0.7 cm to 1.3 cm. These nodules show intense hypermetabolism, with SUVmax ranging from 6.4 to 18.1. **Chest:** Bilateral lung markings are slightly prominent, but no hypermetabolic foci are observed within the lung fields. The hila are not enlarged, the bronchial orifices are patent, and no significant luminal stenosis is noted. Multiple soft tissue nodal densities are seen in the mediastinum (Stations 2R, 4R, 5, and 10), measuring approximately 0.6 cm to 1.1 cm. These nodes demonstrate increased uptake (SUVmax 5.3-19.5), with involvement being more pronounced on the right side. The pleura and pericardium are smooth without hypermetabolic nodules. A small amount of isolated low-density fluid is noted in the left thoracic cavity, presenting as a photopenic defect (trace pleural effusion). Multiple small soft tissue nodules are observed in the bilateral axillae, measuring 0.4 cm to 1.0 cm, showing mildly increased uptake with SUVmax ranging from 1.7 to 3.1. **Abdomen and Pelvis:** The liver shows a mean SUV of 1.6, and the spleen shows a mean SUV of 1.8. Multiple soft tissue nodules are visualized in the hepatic hilar region, peripancreatic region, para-aortic region, mesentery, and adjacent to the left common iliac vessels, left external/internal iliac vessels, and bilateral inguinal regions. These nodules measure approximately 0.5 cm to 2.6 cm and exhibit varying degrees of increased uptake, with SUVmax ranging from 3.4 to 24.1. A round, slightly hyperdense lesion measuring approximately 1.9 cm × 2.3 cm is noted in the left kidney, presenting as a photopenic defect; a punctate density (suggestive of a calculus) is seen in the left renal pelvis. The pancreas, bilateral adrenal glands, and right kidney are normal in position, morphology, and size, with homogeneous density and normal tracer distribution. Mild retention of radioactive urine is observed in the bilateral renal calyces, pelves, and mid-to-distal ureters. The gallbladder is not distended, the wall is not thickened, and tracer distribution is normal. The gastric wall is smooth with no abnormal uptake. Segmental physiologic uptake is noted in parts of the bowel. The prostate is not enlarged and shows heterogeneous tracer distribution without focal hypermetabolism; punctate calcifications are visible within. No abdominopelvic ascites is observed. **Musculoskeletal, Skin, and Brain:** The physiological curvature of the spine is preserved, and alignment is normal. Varying degrees of osteophytes are noted at the margins of the cervical, thoracic, and lumbar vertebrae (spondylosis), but no significant abnormal tracer distribution is detected in the skeletal system. Intracranially, the cerebral cortex, bilateral basal ganglia, thalami, and cerebellar hemispheres show normal, symmetrical tracer distribution. The morphology, size, and position of the ventricles remain unchanged. The sulci and cisterns are not widened, and midline structures are not shifted.

Figure 7. Case study on PETRG-3D generation. Green and red backgrounds denote statements consistent and inconsistent with the Ground Truth (Fig. 5), respectively.

w/ Stop Token:

<ZH/Original> 颈部：双侧上颌窦窦腔内未见软组织密度影，未见放射性摄取增高。...
颅内大脑皮质、双侧基底节、丘脑及小脑半球放射性分布正常，两侧对称。诸脑室形态、大小、位置无改变；脑沟、脑池无增宽；中线结构无移位。

<EN/Translated> **Head and Neck:** The bilateral maxillary sinuses are clear, showing no soft tissue densities or abnormal tracer uptake.... Intracranially, the cerebral cortex, bilateral basal ganglia, thalami, and cerebellar hemispheres show normal, symmetrical tracer distribution. The morphology, size, and position of the ventricles remain unchanged. The sulci and cisterns are not widened, and midline structures are not shifted.

w/o Stop Token:

<ZH/Original> 颈部：双侧上颌窦窦腔内未见软组织密度影，未见放射性摄取增高。...
颅内大脑皮质、双侧基底节、丘脑及小脑半球放射性分布正常，两侧对称。诸脑室形态、大小、位置无改变；脑沟、脑池无增宽；中线结构无移位。 语言功能PET代谢显像正常。

</think>

颈部：双侧上颌窦窦腔内未见软组织密度影，未见放射性摄取增高。鼻咽、口咽咽腔及喉室腔结构正常，局部软组织未见肿胀，放射性分布正常。气管居中。甲状腺两叶不大，密度尚均匀，放射性分布正常。左侧颈部（4）见一枚小结节影，大小约0.6cm，放射性摄取增高，SUVmean2.0，SUVmax2.2。胸部：双肺纹理稍增多，双肺上叶及左肺下叶后基底段可见多发囊泡状透亮影，呈放射性缺损灶。双肺下叶后基底段见少许条索影，未见放射性摄取异常增高。肺门不大，支气管开口通畅，管腔未见明显狭窄。双肺门及纵隔未见肿大及放射性摄取增高结节。右侧胸腔内见少许积液影，未见放射性摄取异常增高。胸膜及心包膜光滑，未见放射性摄取增高结节。胸壁及双侧腋下未见明确放射性分布异常。左侧胸壁内侧份胸大肌内侧间隙可见一结节影，大小约0.6cm，放射性摄取增高，SUVmean3.1，SUVmax3.5。纵隔血池SUVmean1.7。腹、盆部：肝脏SUVmean2.1，脾脏SUVmean1.7。胰腺、双肾上腺及双肾位置、形态、大小均正常，密度均匀，放射性分布正常。双肾盂、肾盂及输尿管中下段内见少量放射性尿液滞留。胆囊张力不大，壁不厚，放射性分布正常。胃壁光滑，放射性分布未见异常。部分肠道可见节段性摄取。前列腺不大，放射性分布

<EN/Translated> **Head and Neck:** The bilateral maxillary sinuses are clear, showing no soft tissue densities or abnormal tracer uptake.... Intracranially, the cerebral cortex, bilateral basal ganglia, thalami, and cerebellar hemispheres show normal, symmetrical tracer distribution. The morphology, size, and position of the ventricles remain unchanged. The sulci and cisterns are not widened, and midline structures are not shifted. Language function PET metabolic imaging is normal.

</think>

Neck: No soft tissue density opacities are observed within the bilateral maxillary sinuses, and no increased tracer uptake is seen. The structures of the nasopharynx, oropharynx, and laryngeal ventricle are normal; local soft tissues show no swelling, and radiotracer distribution is normal. The trachea is midline. Both lobes of the thyroid are not enlarged, with relatively uniform density and normal radiotracer distribution. A small nodule is observed in the left neck (Level IV), measuring approximately 0.6 cm, with increased tracer uptake (SUVmean 2.0, SUVmax 2.2).Chest: Pulmonary markings are slightly increased. Multiple cystic lucencies are visible in the upper lobes of both lungs and the posterior basal segment of the left lower lobe, presenting as photopenic defects. Scanty fibrous streaks are seen in the posterior basal segments of the bilateral lower lobes, without abnormally increased tracer uptake. The hila are not enlarged. Bronchial openings are patent, and no obvious luminal narrowing is observed. No enlarged nodes or nodes with increased tracer uptake are seen in the bilateral hila or mediastinum. A small amount of effusion is noted in the right thoracic cavity, without abnormally increased tracer uptake. The pleura and pericardium are smooth, with no nodular increased tracer uptake observed. No definite abnormal radiotracer distribution is seen in the chest wall or bilateral axillae. A nodule measuring approximately 0.6 cm is visible in the medial intermuscular space of the pectoralis major on the medial aspect of the left chest wall, showing increased tracer uptake (SUVmean 3.1, SUVmax 3.5). Mediastinal blood pool SUVmean is 1.7.Abdomen and Pelvis: Liver SUVmean 2.1, Spleen SUVmean 1.7. The position, morphology, and size of the pancreas, bilateral adrenal glands, and kidneys are normal; densities are uniform, and radiotracer distribution is normal. Mild retention of radioactive urine is observed in the bilateral renal calyces, pelvis, and the middle-to-lower segments of the ureters. The gallbladder is not distended, the wall is not thickened, and radiotracer distribution is normal. The gastric wall is smooth, and no abnormal radiotracer distribution is observed. Segmental uptake is visible in parts of the intestines. The prostate is not enlarged, radiotracer distribution

Figure 8. Qualitative comparison of generation results with (w/) and without (w/o) the explicit stop token. Note that for the "w/ Stop Token" setting, generation is truncated immediately upon predicting the specific end-of-report token (e.g., "[end-of-report]"). For brevity, only the beginning and ending sentences are displayed. Red text highlights hallucinations.