This CVPR Findings paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

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

Wenpei Jiao[1], Ke Yan[2,3], Jiajin Zhang[2,3,4], Dakai Jin[2†], Zhaoheng Xie[1†]

[1]Institute of Medical Technology and National Biomedical Imaging Center, Peking University
[2]DAMO Academy, Alibaba Group, [3]Hupan Lab, [4]Shanghai Jiao Tong University

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. Codes, models, and AutoPET-RG-Lym are released at: https://github.com/Jorwnpay/PETRG-3D.

1. Introduction

Intelligent interpretation of medical imaging has achieved notable progress in recent years, especially for structural

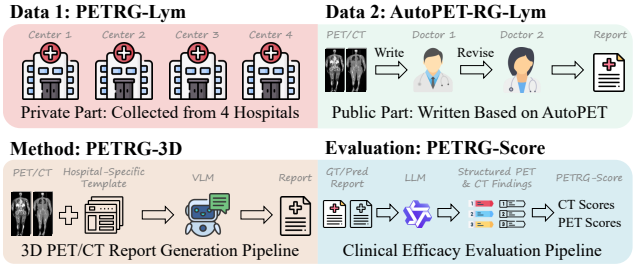


Figure 1. A glimpse of this work. (1) Data: The dataset consists of a private part collected from four medical centers in China and a public part derived from the AutoPET dataset [12], for which reports were generated and subsequently revised and verified by two senior nuclear medicine physicians. (2) 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.

modalities such as X-ray and CT [5, 13, 19–21, 33, 34]. 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 [11, 42, 44]. Consequently, PET/CT has become widely adopted in oncology [18, 29, 35].

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[11, 44]. 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

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[29]	NeurIPS'25	✗	✗	2757	PET	✓	✓	✓	✓	2	5
PET2Rep[44]	AAAI'26	✗	✗	565	PET, CT	✗	✗	✗	✓	4	19
PETRG (Ours)	-	✓	✓	959	PET, CT	✓	✓	✓	✓	5	24

rather than localized-region (e.g., chest, brain) analysis, as physicians must jointly describe anatomical structures and metabolic activity across the entire body [44]. 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) [5, 15, 20, 23]. 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 [7, 8, 38], 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 [44], which processes only 2D slices with off-the-shelf VLMs, and ViMed-PET [29], 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 [12], 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 [32], METEOR [4], and ROUGE [27], 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 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.

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 [10] and MIMIC-CXR [22] have provided essential data foundations, while the evolution of deep learning—particularly VLMs—has spurred the development of numerous sophisticated report generation frameworks [6, 17, 20, 21, 26, 31, 36, 40, 41].

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 [19] and Flamingo-CXR [34]). 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) [14] is dominated by 3D vision encoder-language decoder architectures [2, 23, 25], often built upon pre-trained encoders like CT-CLIP [14], RadFM [38], or M3D [2]. To manage 3D data complexity, methods employ diverse strategies. Some integrate memory modules, such as the disease prototype memory in Dia-LLaMA [7] for class imbalance or the relational memory in CT2Rep [15] for coherence. Others leverage fine-grained anatomical priors[28], as seen in Reg2RG [8] which utilizes explicit correspondences from RadGenome-ChestCT [43]. For MRI, AutoRG-Brain [24] 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[9, 11, 35, 42]. PET2Rep [44] 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 [29] 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 [44] process only a handful of 2D slices, discarding rich 3D spatial context inherent in PET/CT volumes. ViMed-PET [29] 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 PETRG Datasets

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 [12], HECKTOR [1]) 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, radiotracer injection time, and dose. Imaging reports were authored 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 [12], which was collected at University Hospital Tübingen and University

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

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 [12] 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 [37] 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 [2, 38], 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 [44] or single-modality 3D PET [29], our **DSFE** architecture (Fig. 2, blue region) explicitly processes and fuses these complementary volumetric modalities.

Each stream comprises a volume encoder followed by a 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 pretrained on large-scale CT data (RadFM [38]) 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.

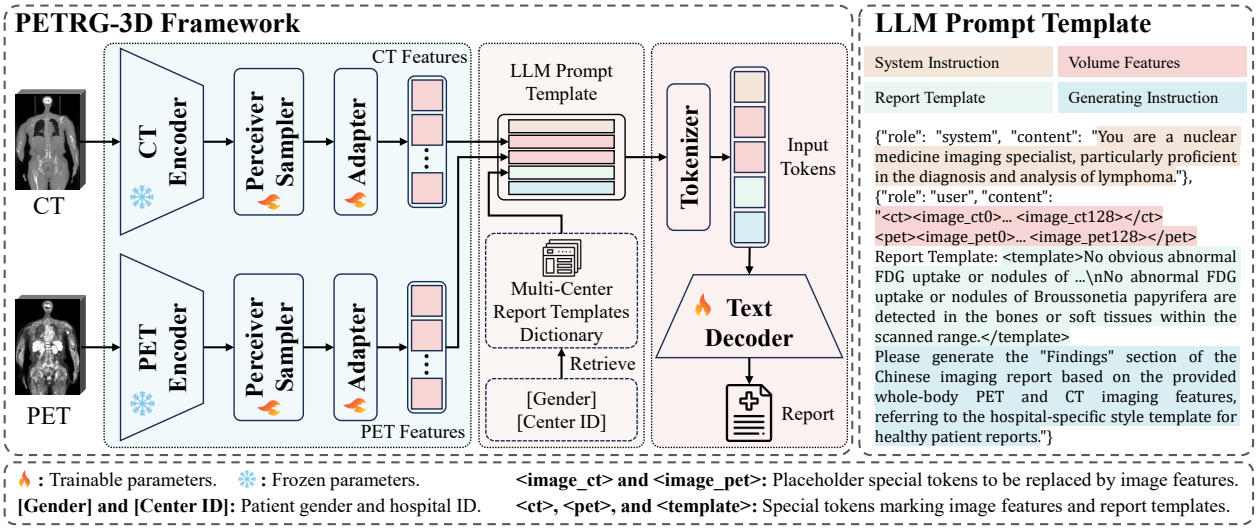


Figure 2. Overall framework of PETRG-3D. The model comprises three key components: (1) a dual-stream 3D visual feature extractor (DSFE) that processes PET and CT volumes separately *(blue)*; (2) a style-adaptive multimodal fusion (SAMF) module that dynamically integrates visual features with hospital-specific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

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 [39]) for autoregressive report generation (Fig. 2, pink region).

Instead of full fine-tuning, we adopt LoRA [16] 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) [32], METEOR (MTR) [4], and ROUGE-L (R-L) [27], to quantify textual similarity between generated and 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 [39], 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.

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 [44]: "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. †: adaptation and fine-tuned on PETRG-Lym.

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[30]	Mistral-7B-v0.2	3.21	0.99	0.56	0.40	2.46	4.84	-	-	-	-
RadFM[38]	Llama2-13B	1.65	0.38	0.21	0.16	1.27	2.44	-	-	-	-
M3D[2]	Llama2-7B	0.53	0.21	0.11	0.07	2.16	4.56	-	-	-	-
PET2Rep-Sep[44]	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[44]	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	28.62
	Qwen2.5-VL-Max	34.30	18.57	9.87	5.42	28.55	19.84	22.88	26.95	9.54	20.60
CT2Rep[15]†	Transformer	44.02	29.86	21.16	15.95	33.79	26.33	23.24	28.25	9.15	21.50
RadFM[38]†	Qwen3-8B	40.26	27.86	20.52	16.23	32.66	28.06	26.62	32.24	13.74	26.34
Dia-LLaMA[7]†	Qwen3-8B	41.13	31.02	24.51	20.28	33.57	32.82	24.19	34.16	9.78	27.85
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	60.42	52.00	45.92	41.33	**51.21**	52.15	31.78	34.13	19.20	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	41.55	50.80	**53.74**	29.06	31.65	15.72	25.07
	Qwen3-8B	**60.78**	**52.45**	**46.42**	**41.90**	51.16	52.88	**32.06**	**34.76**	**19.53**	**28.76**

4.4.2. Baselines and Implementation

We benchmark against strong baselines: LLaVA-Med [30], RadFM [38], and M3D [2] (prior 2D/3D radiology VLMs), and PET2Rep [44], a recent 2D PET/CT method. Our model's (PETRG-3D) volume encoder is initialized from the pretrained ViT-3D of RadFM [38]. 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[2], 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 [3]) 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 [39])), 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 excessively 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

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 [44]. †: adaptation and fine-tuned on PETRG-Lym.

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[30]	Mistral-7B-v0.2	3.16	0.98	0.57	0.42	2.37	4.69	-	-	-	-
RadFM[38]	Llama2-13B	1.57	0.45	0.24	0.18	1.12	2.14	-	-	-	-
M3D[2]	Llama2-7B	0.28	0.09	0.04	0.03	1.84	4.33	-	-	-	-
PET2Rep-Sep[44]	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[44]	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	32.96	12.16	27.16
	Qwen2.5-VL-Max	32.13	18.97	10.56	5.82	29.12	21.01	28.31	28.57	16.14	21.46
CT2Rep[15]†	Transformer	37.08	21.80	12.56	7.16	28.91	21.84	20.98	27.03	6.88	19.68
RadFM[38]†	Qwen3-8B	34.90	20.54	12.40	7.57	27.48	22.90	27.35	30.78	14.58	23.88
Dia-LLaMA[7]†	Qwen3-8B	41.64	27.51	18.14	11.90	32.54	27.91	26.50	31.72	14.10	25.39
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	52.37	43.92	38.65	35.03	40.31	44.20	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	41.50	40.97	33.48	22.79	22.81	16.35

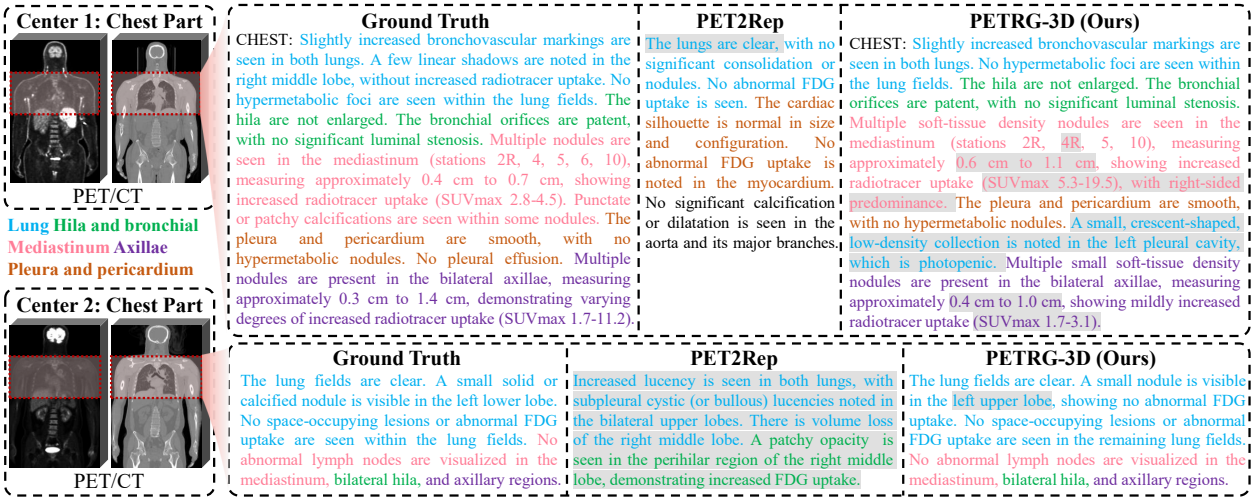


Figure 3. Qualitative comparison of our method against PET2Rep[44] on chest results. Different colors denote distinct anatomical areas. Incorrect diagnoses are highlighted in gray.

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 [39] as the text decoder.

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

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**

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-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.

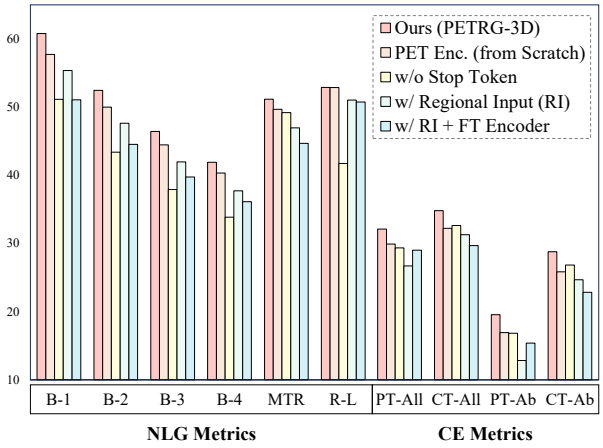


Figure 4. **Analysis of design choices.** "Ours (PETRG-3D)" is the final model settings. "PET Enc. (from Scratch)" trains a separate ViT-3D for the PET branch from scratch. "w/o Stop Token" removes the explicit end-of-report token during training. "w/ Regional Input" uses 4 decoupled anatomical region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

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 AutoPET-RG-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 existing 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.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (62394311, 62394310), Beijing Natural Science Foundation (Z210008), National Biomedical Imaging Facility Grant, and the startup funds of Peking University Health Science Center.

References

[1] Vincent Andrearczyk, Valentin Oreiller, Mathieu Hatt, and Adrien Depeursinge. Head and neck tumor segmentation and outcome prediction. *UK: Springer*, 2022. 3

[2] 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

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6

[4] Satanjeev 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

[5] 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

[6] 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

[7] 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, 6, 7

[8] 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

[9] 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

[10] 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

[11] Yaozong Gao, Yiran Shu, Mingyang Yu, Yanbo Chen, Jingyu Liu, Shaonan Zhong, Weifang Zhang, Yiqiang Zhan, Xiang 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, 3

[12] 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

[13] 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

[14] Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, 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

[15] 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, 6, 7

[16] 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

[17] 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

[18] 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. 1

[19] 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 multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668*, 2023. 1, 3

[20] 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

[21] 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

[22] 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

[23] 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

[24] 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

[25] 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

[26] 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

[27] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 2, 5

[28] 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

[29] 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. 1, 2, 3, 4

[30] 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

[31] 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

[32] 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

[33] 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

[34] 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

[35] 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. 1, 3

[36] 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

[37] 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. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5): e230024, 2023. 4

[38] 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, 6, 7

[39] 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

[40] 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

[41] 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

[42] 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

[43] 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

[44] 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-drived automated radiology report generation for positron emission tomography. *arXiv preprint arXiv:2508.04062*, 2025. 1, 2, 3, 4, 6, 7