PET2Rep: Towards Vision-Language Model-Drived Automated Radiology Report Generation for Positron Emission Tomography

Yichi Zhang[1,2,*], Wenbo Zhang[1,2,*], Zehui Ling[1,2,*] Gang Feng[3], Sisi Peng[3], Deshu Chen[1,2], Yuchen Liu[1,2], Hongwei Zhang[1,2], Shuqi Wang[1], Lanlan Li [1], Limei Han[1,2], Yuan Cheng[1,2,†], Zixin Hu[1,2,†], Yuan Qi[1,2,†], Le Xue[1,2,†]

[1] Fudan University, Shanghai, China
[2] Shanghai Academy of Artificial Intelligence for Science, Shanghai, China
[3] Shanghai Universal Medical Imaging Diagnostic Center, Shanghai, China

Abstract

Positron emission tomography (PET) is a cornerstone of modern oncologic and neurologic imaging, distinguished by its unique ability to illuminate dynamic metabolic processes that transcend the anatomical focus of traditional imaging technologies. Radiology reports are essential for clinical decision making, yet their manual creation is labor-intensive and time-consuming. Recent advancements of vision-language models (VLMs) have shown strong potential in medical applications, presenting a promising avenue for automating report generation. However, existing applications of VLMs in the medical domain have predominantly focused on structural imaging modalities, while the unique characteristics of molecular PET imaging have largely been overlooked. To bridge the gap, we introduce PET2Rep, a large-scale comprehensive benchmark for evaluation of general and medical VLMs for radiology report generation for PET images. PET2Rep stands out as the first dedicated dataset for PET report generation with metabolic information, uniquely capturing whole-body image-report pairs that cover dozens of organs to fill the critical gap in existing benchmarks and mirror real-world clinical comprehensiveness. In addition to widely recognized natural language generation metrics, we introduce a series of clinical efficacy metrics to evaluate the quality of radiotracer uptake pattern description in key organs in generated reports. We conduct a head-to-head comparison of 30 cutting-edge general-purpose and medical-specialized VLMs. The results show that the current state-of-the-art VLMs perform poorly on PET report generation task, falling considerably short of fulfilling practical needs. Moreover, we identify several key insufficiency that need to be addressed to advance the development in medical applications. We believe PET2Rep will serve as a platform for the development and application of VLMs for PET imaging, accelerating the development of trustworthy reporting tools that can genuinely alleviate radiologist burden and enhance patient care.

Project — https://github.com/YichiZhang98/PET2Rep

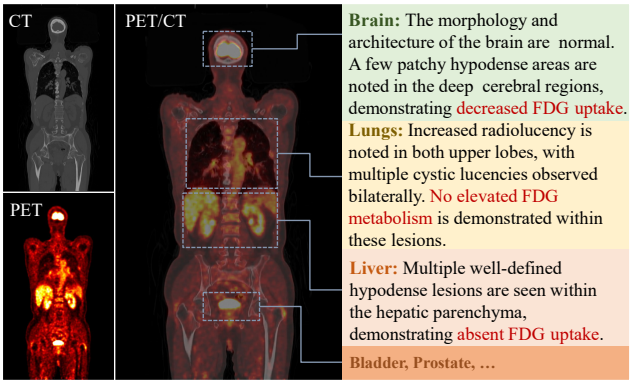

Figure 1: An overview of PET2Rep benchmark. Each case contains whole-body PET/CT images with radiology report.

Introduction

Radiology plays a crucial role in modern healthcare, enabling clinicians to visualize internal structures of patients and make informed decisions for diagnosis and treatment. Positron Emission Tomography (PET) stands as a cornerstone in contemporary oncological and neurological imaging, offering unparalleled insights into dynamic metabolic processes. Unlike imaging modalities like X-ray and CT which primarily focus on information of anatomical structures, PET excels at visualizing metabolic information of physiological functions. By tracking the distribution of radioactive tracers, PET can detect early signs of disease progression, monitor treatment response, and guide personalized therapy plans (Peng et al. 2023; Xue et al. 2024). This functional imaging capability has revolutionized the diagnosis and management of various conditions (Schwenck et al. 2023). In the clinical workflow, radiology reports play a pivotal role in translating imaging into actionable information for healthcare providers. These reports summarize the radiologist's interpretation of the images, highlight key observations, and provide recommendations for further evaluation or treatment (Pang, Li, and Zhao 2023). However, the process of manually creating these reports is inherently labor-intensive and time-consuming, often burdening radiologists

*These authors contributed equally.

with a significant administrative workload. This not only limits their capacity to handle a growing volume of imaging studies, but also introduces potential delays in patient care (Ashraf et al. 2023).

The recent surge in vision-language models (VLMs) has attracted interest from the medical community, where these models hold great potential to transform various aspects of clinical practice like automatic medical report generation (Zhang et al. 2024). Leveraging the power of large-scale pretraining, VLMs can analyze medical images and generate corresponding textual descriptions, effectively bridging the gap between visual data and clinical language. However, existing applications of VLMs in the medical domain have predominantly focused on structural imaging modalities (Liu et al. 2024; Hamamci et al. 2024; Zhu et al. 2025b), while the unique characteristics and clinical value of PET imaging have largely been overlooked in the current research landscape. As the analysis of PET images poses unique challenges due to the need to integrate functional and anatomical information and specialized knowledge required to interpret tracer uptake patterns (Coleman et al. 2010; Matsubara et al. 2022), it is worth rethinking that *How Far are VLMs from Effective Radiology Report Generation for Positron Emission Tomography Imaging?*

To answer this question, we introduce PET2Rep, a comprehensive benchmark for the evaluation of radiology report generation for PET imaging. Compared with existing medical benchmarks, the key advantages of PET2Rep can be concluded in the following three aspects.

The First PET/CT Report Dataset. PET2Rep is the first dataset dedicated to PET/CT report generation. Unlike other modalities like X-ray and CT which primarily focus on anatomical structures, PET operates at the molecular level, enabling the assessment of metabolic information. This unique feature allows for early disease identification, often before anatomical changes are visible on other imaging modalities (Gatidis et al. 2024). PET2Rep is a large-scale multi-modal dataset of 565 cases with paired PET, CT and corresponding radiology reports. Given the high cost of PET/CT scans and the need for specialized expertise in report writing, there is currently no relevant dataset available, which highlights the importance of PET2Rep in advancing research in this field.

Whole-Body Imaging with Radiology Reports. Existing medical imaging benchmarks are often limited to specific anatomical domains. For instance, chest X-ray report generation primarily address thoracic pathologies (Liu et al. 2024), while those for CT reports concentrate on the volume and morphology of organs and lesions in the chest (Hamamci et al. 2024) or abdominal regions (Bassi et al. 2025). In contrast, PET2Rep encompasses a much broader anatomical scope, with images ranging from the head and neck to the proximal limbs. Consequently, its corresponding reports provide detailed evaluations of dozens of organs body-wide, demanding a more extensive scope of medical knowledge for accurate interpretation, as shown in Figure.1. This holistic approach more closely simulates real-world oncology practice, where radiologists conduct comprehensive assessments rather than focusing on isolated areas.

Data Collection from Clinical Scenarios. Many existing medical multimodal benchmarks are developed from public imaging archives (Sepehri et al. 2024; Chen et al. 2024). These frameworks often generate tasks that probe for superficial understanding of the image, such as identifying the imaging modality or naming marked organs, rather than complex clinical reasoning (Ye et al. 2024; Zhou et al. 2025). Such scenarios test for basic medical knowledge and differ significantly from the complex demands of a real clinical workflow. In contrast, PET2Rep is collected from real clinical scenarios and incorporates data directly from the clinical setting, ensuring that the benchmark authentically reflects the challenges radiologists encounter in their daily work. This ensures the authenticity and clinical relevance of the PET2Rep benchmark while minimizing the risk of data leakage, thereby reflecting the generalization performance of VLMs in real-world clinical scenarios.

To make a comprehensive evaluation of the performance of VLMs, we establish a standardized evaluation pipeline for PET/CT radiology report generation. We formulate a prompting framework incorporating essential elements including imaging modality specifications and clinical objectives and design a structured report template aligned with radiological training protocols. This approach ensures faithful translation of image-derived information into formatted reports that maintain consistency with expert-generated radiological reports. We conduct a comprehensive evaluation state-of-the-art models, including 19 general purpose and 11 medical-specific VLMs on PET2Rep benchmark. The experimental results show that current cutting-edge VLMs exhibit suboptimal performance on the task, falling considerably short of fulfilling real-world requirements. Furthermore, our analysis reveals several critical limitations that must be tackled to drive progress in clinical applications.

Related Works

Positron Emission Tomography

Positron Emission Tomography (PET) is a clinical imaging technique that reveals ongoing metabolic processes in the body by detecting gamma photons generated from positron annihilation after injecting radioactive tracers. As the most widely used tracer, fluorodeoxyglucose (FDG) assesses local glucose uptake to evaluate organ metabolism and detect tumor metastasis, enabling monitoring of treatment progress (Ren et al. 2019). Clinically, PET is primarily used for early tumor screening for cancer detection (Gatidis et al. 2024; Peng et al. 2023), organ metabolic function assessment (Xue et al. 2024; Zhang et al. 2025), and treatment monitoring (van der Geest et al. 2021). The unique metabolic imaging capability provides indispensable insights for disease diagnosis and treatment optimization in clinical applications.

Vision-Language Models

Vision-Language Models (VLMs) have emerged as a transformative development in artificial intelligence, effectively bridging the gap between visual perception and natural language understanding (Zhang et al. 2024). The swift progress

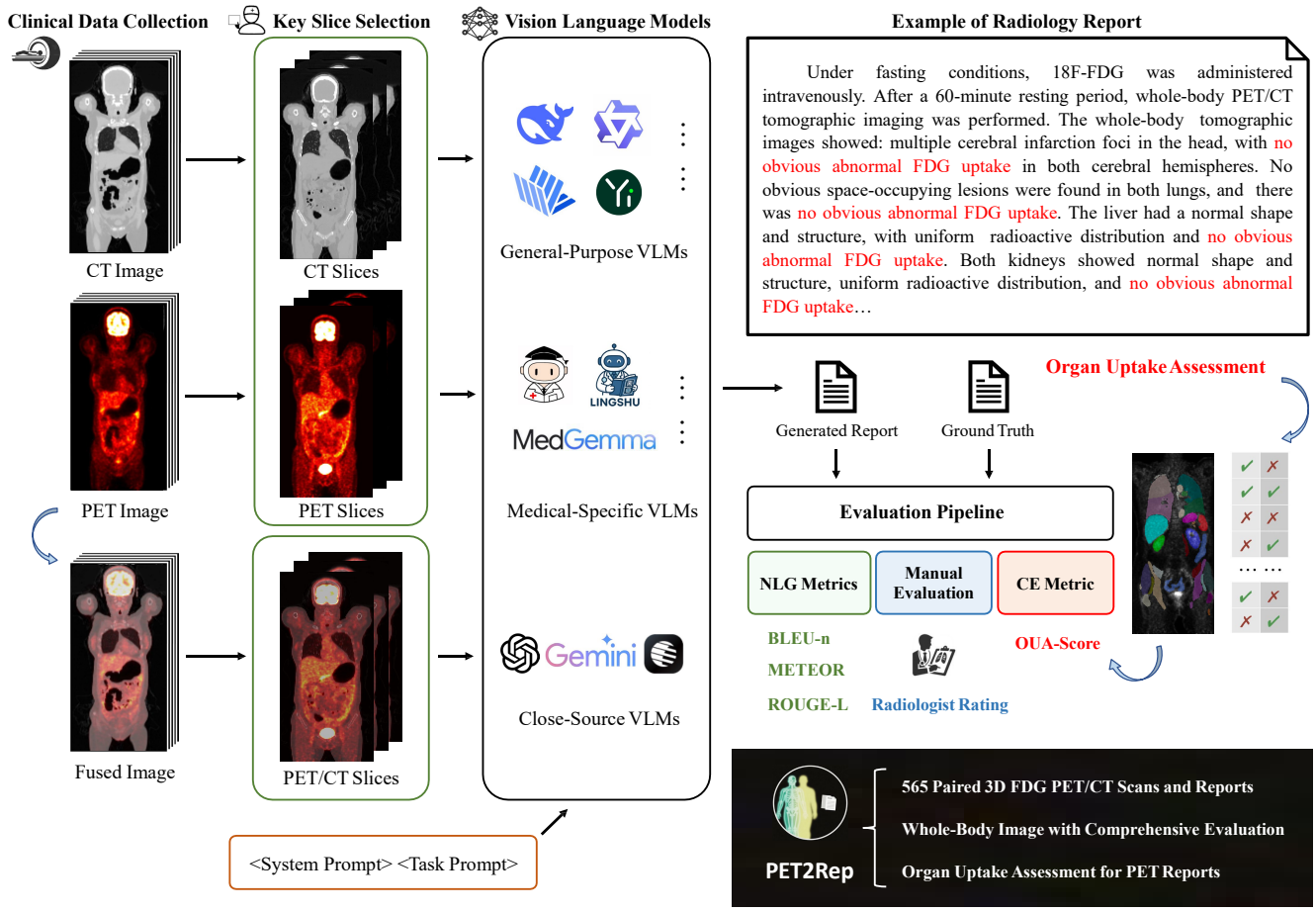

Figure 2: Pipeline of the PET2Rep benchmark for evaluation of VLM-based PET radiology report generation. First, PET/CT images are analyzed by VLMs with a designed prompt format to include necessary information such as image modality, clinical task, and designed report template based on radiologist training guidelines. Then the generated reports are evaluated against the ground-truth reports with widely recognized natural language generation (NLG) metrics and a novel clinical efficacy (CE) metric for PET imaging. We further conduct manual scoring by radiologists for more comprehensive evaluation.

in VLM development is largely attributed to innovative pre-training strategies and architectural designs, which have demonstrated remarkable capabilities across a wide array of tasks such as visual question answering and image captioning (Chen et al. 2022; Feng et al. 2025; Lin et al. 2025). Beyond general-purpose vision tasks, VLMs are making significant inroads into specialized fields like medical image analysis (Peng et al. 2025). VLMs can generate diagnostic reports, answer clinical questions, and highlight regions of interest, offering substantial support to healthcare professionals and promising to enhance the efficacy and accuracy of medical diagnoses (Zhang, Shen, and Jiao 2024; Jiang et al. 2024; Lin, Xu, and Qin 2025).

Radiology Report Generation

The core task of radiology report generation is to transform medical imaging information into accurate and standardized textual reports. Early studies primarily focused on training encoder-decoder architectures for report generation, where the image features are extracted by an encoder and then fed into a decoder to predict the corresponding report (Jing, Xie, and Xing 2018). Given the complexity and the inherent variability in radiological findings, several approaches utilize confidential guidance or attention mechanism to enhance the adaptability (Song et al. 2022; Wang et al. 2024). Due to the impressive performance in a variety of downstream tasks (Zhang et al. 2024), there has been a surge of investigating VLMs for radiology report generation (Hamamci, Er, and Menze 2024; Chen et al. 2025).

PET2Rep Benchmark

We introduce PET2Rep, a comprehensive benchmark designed to evaluate the performance of VLMs for generating radiology reports from PET images. PET2Rep is the first PET/CT dataset with paired structured radiology reports. A key distinction of PET2Rep is its data sourcing. Unlike

benchmarks that rely on data aggregated from online repositories, all data in our work were meticulously collected from actual clinical scenarios. This approach guarantees the authenticity and clinical relevance of the benchmark. Furthermore, by sourcing directly from clinical settings, we mitigate the risk of data leakage, ensuring that PET2Rep accurately reflects the complexity and diversity of real-world radiological practice. The comprehensive workflow of the PET2Rep benchmark is detailed in Figure. 2 and will be elaborated upon in the subsequent sections.

Dataset Construction

We collect 565 cases of 3D whole-body FDG PET/CT imaging from one local medical center, which is the most widely used PET tracer in oncology. As a non-specific tracer, FDG can be used for whole-body imaging to reflect tissue glucose metabolism, which makes the imaging useful in assessing the systemic distribution and metastasis of tumors. Structured radiology reports are constructed based on radiologist-designed templates, which play a pivotal role in assisting physicians to interpret whole-body PET/CT scans in a standardized and organized manner, thereby enhancing clarity and supporting clinical decision-making. The report provides a detailed and objective description of the findings from the PET/CT images in a systematic, head-to-toe sequence, ensuring that no anatomical region is overlooked. It enumerates all detected abnormalities without offering interpretative conclusions, serving as a factual foundation for subsequent clinical assessment. More detailed information and examples of the dataset are shown in the Appendix.

Data Preprocessing

To ensure the accuracy and reliability of multi-modal image analysis, rigorous data pre-processing is indispensable for bringing all imaging modalities into a consistent and interpretable format. In PET/CT imaging, a critical preprocessing step involves resampling the CT images to match the lower spatial resolution of the PET images. This coregistration process aligns both modalities to a common matrix size, ensuring voxel-wise correspondence across datasets. Following resampling, the CT intensities are standardized using z-score normalization to reduce inter-scan variability. Additionally, normalization of PET data is performed by converting the raw radioactivity counts into Standardized Uptake Values (SUV) a widely adopted quantitative metric in PET imaging that accounts for factors such as the injected radiotracer dose and the patient's body weight (Lucignani, Paganelli, and Bombardieri 2004). To emulate the clinical diagnostic workflow, we implement a fusion process that integrates PET and CT scans. This approach combines the functional information from PET with the anatomical detail provided by CT, reflecting the manner in which radiologists interpret these modalities in clinical practice. The resulting composite image enables visualization of metabolic activity within its precise anatomical context. Such integration is essential for accurately localizing regions of abnormal radiotracer uptake and facilitating a comprehensive assessment of the patient's condition.

Key Slice Selection

Given that the original PET/CT images are three-dimensional, while most existing VLMs are designed for 2D images, it becomes necessary to select out representative 2D slices from the 3D imaging. In this study, we select the coronal plane as the view for slice sampling, following clinical conventions in which radiologists commonly utilize this view for comprehensive head-to-toe assessments. The coronal plane offers an optimal perspective, capturing the global anatomical context and encompassing the majority of key organs. By analyzing multiple coronal slices, VLMs can effectively capture the salient information embedded within the full 3D scan. Building upon this design, we further emulate the diagnostic process of radiologists and design two strategies for report generation as described below.

Input Separate PET and CT Images. In this strategy, we maintain the distinction between functional and anatomical information by providing the model with two distinct, parallel inputs. For each anatomical location of interest, we extract a corresponding pair of 2D slices: a grayscale slice from the CT volume for structural context, and a pseudo-colored slice from the PET volume to highlight metabolic activity. Specifically, we identify three key locations for analysis, resulting in a total input of six images for the VLMs (a PET/CT pair for each location). This dual-input approach compels the model to learn the complex correlations between anatomical structure and functional uptake, mirroring the cognitive process of a radiologist integrating two different sets of images.

Input Fused PET/CT Images. This strategy involves pre-integrating the multimodal information into a single image before presenting it to the model. For each selected location, we generate a fused image by superimposing the pseudo-colored PET slice directly onto its corresponding grayscale CT slice. In our implementation, we create these fused images for three key locations, providing the model with a total of three input PET/CT images. Each fused image presents an immediate composite view, in which metabolic hotspots are precisely localized within their anatomical context. This method simulates the final visualization that radiologists use for diagnosis. By supplying the model with pre-fused inputs, we eliminate the need for it to learn the fusion process, enabling it to focus directly on understanding the integrated functional and structural patterns.

Experimental Setup

In this study, we evaluate a range of VLMs encompassing both medical-specific and general-purpose models including open-source options and those accessible via proprietary APIs. The weights of open-source models were sourced from respective official Hugging Face repositories. To guide the models in generating radiology-style reports, we design a standardized prompt format specifying the imaging modality, clinical task, and a report template derived from radiologist training guidelines. This ensures that image interpretations are expressed in a format consistent with manually authored radiological reports. Our evaluation are conducted in a zero-shot setting, which serves as a stringent test of

the models' generalization ability, revealing how well they can handle complex medical imaging tasks without any task-specific fine-tuning. All tests were executed using NVIDIA A100 GPUs with 80GB of memory.

Evaluated Models

To comprehensively assess the performance of VLMs on the PET2Rep benchmark, we conducted a systematic evaluation of diverse state-of-the-art VLMs as follows.

General-Purpose VLMs. General-purpose VLMs are trained on large and diverse datasets to perform a wide spectrum of multimodal tasks. Their broad training enables strong visual understanding and reasoning capabilities, with versatility and scalability. We select following general-purpose VLMs for evaluation, including Qwen2.5-VL series (Bai et al. 2025), InternVL3 series (Zhu et al. 2025a), Yi-VL series (Young et al. 2024), LLaVA-V1.5 (Liu et al. 2023a), OmniLMM (Yu et al. 2024), VisualGLM (Du et al. 2022) and Deepseek-VL2 (Wu et al. 2024) .

Medical-Specific VLMs. In contrast to general-purpose models, medical-specific VLMs are tailored for clinical applications, emphasizing domain adaptation and integration of specialized medical knowledge. Trained on curated medical datasets and aligned with diagnostic workflows, these models prioritize accuracy and reliability in healthcare settings. We select following medical-specific VLMs for evaluation, including LLaVA-Med (Li et al. 2023), Med-Flamingo (Moor et al. 2023), Qilin-Med-VL (Liu et al. 2023b), RadFM (Wu et al. 2023), MedDr (He et al. 2024), HuatuoGPT-Vision (Chen et al. 2024), MedVLM-R1 (Pan et al. 2025), and latest MedGemma series (Sellergren et al. 2025) and Lingshu series (Xu et al. 2025).

Closed-Source VLMs. Closed-source VLMs are developed and maintained by enterprises with inaccessible source code, typically provided to users via APIs for integration into applications. We select following closed-source VLMs for evaluation, including Gemini 2.5 Pro (Comanici et al. 2025), GPT-4o (Hurst et al. 2024), Moonshot-v1 (MoonshotAI 2025) and Qwen-VL-Max (Bai et al. 2023).

Evaluation Metrics

To assess the performance of VLMs in radiology report generation, we compare the generated reports against the ground-truth reports using the following aspects.

Natural Language Generation (NLG) Metrics. In line with existing research, we adopt widely recognized NLG metrics, including BLEU-n (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), and ROUGE-L (Lin 2004). Specifically, BLEU-n evaluates n-gram overlap between generated and reference reports, ROUGE-L measures alignment via the longest common subsequences, and METEOR accounts for synonyms and paraphrases to capture semantic similarity.

PET Clinical Efficacy (CE) Metrics. NLG metrics primarily focus on word and sentence similarity while neglecting diagnostic accuracy. Reports with opposite diagnostic conclusions may achieve similar NLG scores. Conversely, results with correct uptake assessments but inconsistent formatting in the report text might receive lower NLG scores.

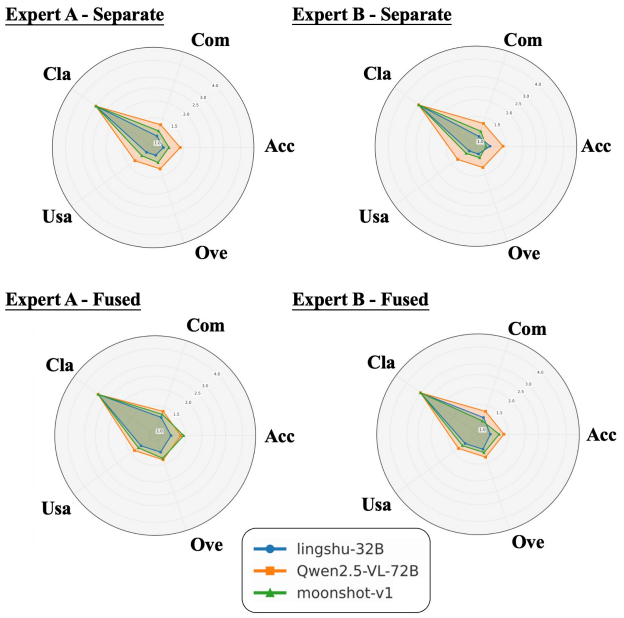

Figure 3: Performance comparison of three VLMs under different task settings for manual evaluation by two radiologists rated across five dimensions, including Medical Accuracy (Acc), Key Findings Completeness (Com), Expression Clarity (Cla), Clinical Usability (Usa) and Overall Rating (Ove).

Existing studies have explored the proposal of clinical efficacy metrics by utilizing text classifiers to extract abnormality labels for CT report evaluation (Hamamci et al. 2024). However, these methods are not applicable to PET reports. To assess the clinical efficacy of PET reports, we introduce a series of CE metrics to evaluate descriptions regarding radiotracer uptake patterns in key organs within generated PET reports. Given that the whole-body PET imaging data used in our study, we extract the assessment of uptake levels corresponding to each key organ from the report text and compare these assessments with the corresponding entries in the ground truth reports. For each key organ, we define four states of radiotracer uptake: *Increased Uptake*, *Decreased Uptake*, *Absent Uptake*, and *Normal*. Given the clinical focus on anomaly detection, we categorize the first three states into three distinct positive classes, with *Normal* serving as the negative class. Our evaluation method involves independently calculating the precision, recall, and F1-score for each of the three positive classes. The final CE metrics are the macro-average across these three positive classes. The implementation details are elaborated in the Appendix. Compared to NLG metrics, CE metrics shift the evaluation from text-matching problem to multi-label classification assessment that more closely aligns with clinical diagnosis.

Results and Analysis

Table. 1 summarizes the performance of the evaluated VLMs, with the last column providing an overall reference score as the average of all metrics. After reviewing the evaluation results, we have drawn following conclusions.

Model (year/month)	**NLG Metrics**			**CE Metrics**			**Overall**
	BL-4	**MTR**	**RG-L**	**Pre**	**Rec**	**F1**	**(%)**
Template Baseline	0.3150(0.0482)	0.1475(0.0141)	0.5110(0.0319)	0.2282(0.0179)	0.2220(0.0106)	0.2249(0.0123)	27.5
General-Purpose VLMs							
Qwen2.5-VL-7B (25/1)	0.3050(0.0476)	0.1407(0.0198)	0.5075(0.0340)	0.2233(0.0236)	0.1974(0.0083)	0.2094(0.0132)	26.4
	0.3057(0.0467)	0.1390(0.0186)	0.5088(0.0320)	0.2284(0.0227)	0.2023(0.0075)	0.2144(0.0121)	26.6
Qwen2.5-VL-32B (25/1)	0.1777(0.0421)	0.0063(0.0110)	0.4165(0.0516)	0.3402(0.0781)	0.0418(0.0127)	0.0743(0.0214)	17.6
	0.1851(0.0408)	0.0063(0.0111)	0.4295(0.0486)	0.2728(0.0447)	0.0308(0.0047)	0.0554(0.0082)	16.3
Qwen2.5-VL-72B (25/1)	0.2223(0.0585)	0.0655(0.0172)	0.4234(0.0588)	0.2474(0.0513)	0.0295(0.0024)	0.0527(0.0043)	17.3
	0.2273(0.0584)	0.0645(0.0171)	0.4306(0.0594)	0.2917(0.0328)	0.0393(0.0049)	0.0693(0.0084)	18.7
InternVL3-8B (25/4)	0.2439(0.0627)	0.0606(0.0443)	0.4739(0.0630)	0.2425(0.0151)	0.2107(0.0114)	0.2254(0.0119)	24.3
	0.2509(0.0529)	0.0641(0.0463)	0.4845(0.0566)	0.2333(0.0153)	0.2099(0.0074)	0.2208(0.0087)	24.4
InternVL3-14B (25/4)	0.2513(0.0684)	0.0472(0.0528)	0.4835(0.0910)	0.2366(0.0206)	0.2057(0.0095)	0.2199(0.0129)	24.1
	0.2495(0.0671)	0.0532(0.0506)	0.4813(0.0904)	0.2322(0.0196)	0.1982(0.0099)	0.2137(0.0131)	23.8
InternVL3-38B (25/4)	0.1377(0.0924)	0.0775(0.0483)	0.4371(0.1199)	0.2711(0.0203)	0.2072(0.0141)	0.2344(0.0127)	22.8
	0.1446(0.0855)	0.0825(0.0480)	0.4618(0.0913)	0.2674(0.0258)	0.2435(0.0298)	0.2546(0.0278)	24.2
InternVL3-78B (25/4)	0.3090(0.0525)	0.1233(0.0359)	0.4997(0.0401)	0.2355(0.0255)	0.0520(0.0119)	0.0850(0.0157)	21.7
	0.3090(0.0518)	0.1262(0.0318)	0.5008(0.0397)	0.2369(0.0492)	0.0748(0.0083)	0.1132(0.0122)	22.7
Yi-VL-6B (24/1)	0.0065(0.0316)	0.0002(0.0056)	0.0479(0.0709)	0.1144(0.0430)	0.0061(0.0020)	0.0115(0.0038)	3.1
	0.0374(0.0733)	0.0029(0.0165)	0.1156(0.1432)	0.1519(0.0261)	0.0260(0.0033)	0.0444(0.0055)	6.3
Yi-VL-34B (24/1)	0.2610(0.1071)	0.0848(0.0664)	0.4439(0.1420)	0.2305(0.0159)	0.1869(0.0079)	0.2063(0.0098)	23.6
	0.2854(0.0809)	0.0898(0.0645)	0.4779(0.0950)	0.2303(0.0211)	0.2038(0.0072)	0.2160(0.0116)	25.1
LLaVa-V1.5-7B (23/9)	0.1198(0.0508)	0.0126(0.0515)	0.3043(0.0639)	0.2044(0.0287)	0.1022(0.0091)	0.1306(0.0121)	14.6
	0.0328(0.0141)	0.0056(0.0369)	0.1717(0.0283)	0.2460(0.0764)	0.0337(0.0093)	0.0592(0.0163)	9.2
OmniLMM-12B(24/4)	0.0412(0.0627)	0.0075(0.0232)	0.1339(0.1324)	0.1789(0.0330)	0.0173(0.0027)	0.0316(0.0050)	6.8
	0.0397(0.0614)	0.0067(0.0238)	0.1293(0.1336)	0.2095(0.0393)	0.0180(0.0040)	0.0331(0.0071)	7.3
VisualGLM-6B 23/5	0.0361(0.0519)	0.0182(0.0517)	0.1338(0.1214)	0.0662(0.0710)	0.0002(0.0002)	0.0004(0.0004)	4.3
	0.0306(0.0492)	0.0208(0.0588)	0.1173(0.1157)	0.3404(0.1494)	0.0014(0.0006)	0.0029(0.0012)	8.6
DeepSeek-VL2 (24/12)	0.2697(0.0675)	0.0939(0.0976)	0.4875(0.0536)	0.2170(0.0137)	0.1532(0.0076)	0.1795(0.0081)	23.4
	0.2817(0.0637)	0.1054(0.0974)	0.4936(0.0476)	0.2198(0.0269)	0.1571(0.0135)	0.1831(0.0176)	24.0
Medical-Specific VLMs							
MedDr(24/4)	0.2667(0.1012)	0.1564(0.0434)	0.4571(0.1168)	0.2270(0.0245)	0.1820(0.0201)	0.2020(0.0215)	24.9
	0.2801(0.0874)	0.1536(0.0389)	0.4742(0.0951)	0.2397(0.0275)	0.2113(0.0084)	0.2243(0.0138)	26.4
HuatuoGPT-Vision (24/6)	0.1384(0.0865)	0.0000(0.0000)	0.3399(0.1112)	0.1692(0.0232)	0.0814(0.0186)	0.1097(0.0207)	14.0
	0.2573(0.0546)	0.0743(0.0278)	0.4834(0.0577)	0.2183(0.0200)	0.1620(0.0148)	0.1859(0.0164)	23.0
MedVLM-R1 (25/2)	0.1602(0.1112)	0.0006(0.0097)	0.3472(0.1742)	0.2246(0.0285)	0.1019(0.0117)	0.1399(0.0150)	16.2
	0.1708(0.1294)	0.0003(0.0070)	0.3358(0.1840)	0.2321(0.0324)	0.1204(0.0077)	0.1583(0.0110)	17.0
MedGemma-4B (25/7)	0.3015(0.0517)	0.1215(0.0384)	0.5077(0.0385)	0.2276(0.0185)	0.2260(0.0113)	0.2266(0.0129)	26.8
	0.2874(0.0773)	0.1207(0.0339)	0.4875(0.0786)	0.2362(0.0162)	0.2245(0.0091)	0.2301(0.0103)	26.4
MedGemma-27B (25/7)	0.2185(0.0552)	0.0297(0.0157)	0.4390(0.0696)	0.2300(0.0375)	0.0391(0.0079)	0.0667(0.0130)	17.1
	0.2251(0.0574)	0.0309(0.0153)	0.4521(0.0781)	0.2853(0.0435)	0.0846(0.0141)	0.1304(0.0201)	20.1
Lingshu-7B (25/6)	0.2848(0.0855)	0.1079(0.0727)	0.4793(0.0933)	0.2281(0.0162)	0.1970(0.0100)	0.2112(0.0097)	25.1
	0.2775(0.0945)	0.1030(0.0748)	0.4700(0.1119)	0.2273(0.0220)	0.1942(0.0106)	0.2093(0.0138)	24.7
Lingshu-32B (25/6)	0.3050(0.0650)	0.1554(0.0520)	0.4999(0.0604)	0.2250(0.0178)	0.2035(0.0079)	0.2135(0.0109)	26.7
	0.2987(0.0713)	0.1531(0.0597)	0.4939(0.0733)	0.2328(0.0151)	0.2071(0.0071)	0.2191(0.0091)	26.8
Closed-Source VLMs							
Gemini 2.5 Pro (25/6)	0.1535(0.0411)	0.0186(0.0199)	0.3987(0.0438)	0.1705(0.0299)	0.0215(0.0056)	0.0381(0.0092)	13.4
	0.1536(0.0420)	0.0199(0.0201)	0.4025(0.0477)	0.2394(0.0571)	0.0311(0.0066)	0.0550(0.0115)	15.0
GPT-4o (24/5)	0.2023(0.0422)	0.0287(0.0160)	0.4023(0.0421)	0.3375(0.0891)	0.0527(0.0110)	0.0910(0.0185)	18.6
	0.2134(0.0425)	0.0318(0.0132)	0.4168(0.0412)	0.2540(0.0450)	0.0728(0.0085)	0.1130(0.0135)	18.5
Moonshot-v1 (25/1)	0.3064(0.0496)	0.1261(0.0301)	0.5157(0.0339)	0.2603(0.0220)	0.1457(0.0096)	0.1866(0.0117)	25.7
	0.2923(0.0464)	0.1055(0.0302)	0.5142(0.0316)	0.2327(0.0232)	0.1803(0.0132)	0.2030(0.0160)	25.5
Qwen-VL-Max (25/1)	0.2315(0.0375)	0.0269(0.0035)	0.4462(0.0377)	0.2764(0.0313)	0.1897(0.0139)	0.2248(0.0183)	23.3
	0.2479(0.0406)	0.0265(0.0047)	0.4649(0.0409)	0.2844(0.0220)	0.1802(0.0092)	0.2204(0.0113)	23.7

Table 1: Evaluation of general-purpose and medical-specific VLMs on PET2Rep benchmark. Evaluation results presented in gray and white represent the results of separate PET and CT images and fused PET/CT images, respectively.

Overall Ineffectiveness or Even Failure. All evaluated VLMs exhibit limited effectiveness in generating structured radiology reports. Alarmingly, most models fail to surpass even a simple template baseline. The requirement to produce comprehensive, whole-body structured reports presents a substantial challenge for existing VLMs. Many models are unable to consistently adhere to the prescribed report structure, occasionally generating disorganized, unusable, or even empty outputs, which yield near-zero scores across evaluation metrics and are thus omitted from the results table. Even when models attempt to follow the template, they often include irrelevant information or omit mandatory sections, underscoring their difficulty in capturing the core task requirements. This pattern suggests that many VLMs are overfitted to narrow training distributions, typically focused on specific tasks such as single-organ interpretation or generic image captioning, rather than holistic clinical reporting. As a result, they struggle to generalize to clinical applications, where accuracy, completeness, and structural consistency are essential.

State-of-the-Art Models Merely Match the Baseline. Although the most advanced models, such as the Lingshu and MedGemma series, outperform other VLMs, their performance remains only marginally comparable to the baseline. This underwhelming result indicates that even state-of-the-art VLMs are far from ready for practical application in clinical workflow. While these models can generate coherent text with high NLG metrics, they frequently omit critical clinical details, such as subtle tracer uptake abnormalities, leading to low CE metrics. Manual review by radiologists further confirms that the outputs of these models are largely unusable. The accurate interpretation of tracer uptake patterns, combined with the extensive medical knowledge required for comprehensive whole-body assessment, remains a major challenge, highlighting the gap between general language proficiency and specialized clinical expertise. Further manual evaluation by two radiologists in Figure. 3 demonstrate that the outputs of state-of-the-art models are also mostly unusable. The nuanced interpretation of tracer uptake patterns and the broad medical knowledge required for whole-body assessment remain significant challenges, highlighting a critical gap between general language proficiency and specialized clinical expertise.

Larger Model Does Not Necessarily Translate to Better Performance. Our evaluation reveals an intriguing phenomenon that within the same model series, larger-scale models do not consistently outperform their smaller counterparts. In some cases, larger models appear to overlook task requirements, generating irrelevant or fabricated details such as patient names and ages, which negatively impact their evaluation performance. This observation suggests that the inferior performance of larger models may not stem from model scaling itself, but rather from insufficient exposure to domain-specific data and task-oriented training. Therefore, for specialized and highly structured tasks like PET report generation, architectural innovation and targeted fine-tuning may play a more critical role.

Further details regarding experimental results analysis and case studies are presented in the Appendix.

Discussion and Conclusion

In this work, we present PET2Rep, the first comprehensive benchmark specifically designed for evaluating radiology report generation in PET imaging, addressing a critical gap between existing research and clinical application. The benchmark consists of 565 whole-body PET/CT image-report cases, representing a significant advancement in this domain. Another key innovation is the introduction of a series of clinical efficacy metrics to evaluate the quality of radiotracer uptake pattern description in key organs in generated reports PET reports, which is a decisive factor in clinical decision-making, as omissions in critical findings can alter therapeutic pathways.

Our experimental results clearly reveal the critical limitations of current VLMs. Despite their reported success on various multimodal medical benchmarks, all models fail to surpass even a simple template baseline in PET2Rep, with some models generating disorganized or structurally non-compliant reports. These findings underscore the need for fundamental advancements in clinically grounded evaluation frameworks and rigorous alignment with real-world reporting standards to achieve genuine clinical applicability. Many existing benchmark tasks assess the capabilities of VLMs through visual question answering, which primarily reflects superficial image understanding and falls short of the deep clinical reasoning required for diagnosis and treatment. Moreover, existing clinical report generation datasets are largely confined to localized anatomical structures, overlooking the integration of structural and functional information necessary for comprehensive whole-body evaluation. In this context, PET2Rep serves as an expert-informed and clinically aligned benchmark that helps bridge this gap, providing a foundation for exploring the potential of large models toward more generalizable medical intelligence and facilitating progress in domain-specific model development.

While PET2Rep represents a significant step forward, several limitations should be acknowledged. At present, our evaluations are limited to 2D slices, which do not fully capture the three-dimensional spatial relationships and volumetric information critical for comprehensive image interpretation (Zhang et al. 2022). Moreover, clinically important quantitative indicators, such as standardized uptake values (SUVs) and lesion volume measurements, are not yet incorporated into the current evaluation framework. To address these limitations, we plan to expand the benchmark to support full 3D PET/CT evaluations, enabling more complete spatial and volumetric analysis (Xue et al. 2025). Key quantitative measures, including SUVs and lesion volumes, will be reintegrated to enhance the benchmark's clinical validity. In addition, while the current version supports only Chinese reports, future iterations will extend to multilingual evaluation, improving generalizability and facilitating broader clinical adoption of VLMs across diverse healthcare systems (Qiu et al. 2024). These planned enhancements will significantly improve the benchmark's clinical relevance and utility for developing more robust report generation systems.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 82394432 and 92249302), and the Shanghai Municipal Science and Technology Major Project (Grant No. 2023SHZDZX02). The computations in this research were performed using the CFFF platform of Fudan University.

References

Ashraf, N.; Tahir, M. J.; Saeed, A.; Ghosheh, M. J.; Alsheikh, T.; Ahmed, A.; Lee, K. Y.; and Yousaf, Z. 2023. Incidence and factors associated with burnout in radiologists: A systematic review. *European Journal of Radiology Open*, 11: 100530.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Banerjee, S.; and Lavie, A. 2005. 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*, 65–72.

Bassi, P. R.; Yavuz, M. C.; Wang, K.; Chen, X.; Li, W.; Decherchi, S.; Cavalli, A.; Yang, Y.; Yuille, A.; and Zhou, Z. 2025. RadGPT: Constructing 3D Image-Text Tumor Datasets. *arXiv preprint arXiv:2501.04678*.

Chen, J.; Gui, C.; Ouyang, R.; Gao, A.; Chen, S.; Chen, G. H.; Wang, X.; Zhang, R.; Cai, Z.; Ji, K.; et al. 2024. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*.

Chen, J.; Guo, H.; Yi, K.; Li, B.; and Elhoseiny, M. 2022. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18030–18040.

Chen, Z.; Bie, Y.; Jin, H.; and Chen, H. 2025. Large language model with region-guided referring and grounding for ct report generation. *IEEE Transactions on Medical Imaging*.

Coleman, R. E.; Hillner, B. E.; Shields, A. F.; Duan, F.; Merlino, D. A.; Hanna, L. G.; Stine, S. H.; and Siegel, B. A. 2010. PET and PET/CT reports: observations from the National Oncologic PET Registry. *Journal of Nuclear Medicine*, 51(1): 158–163.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; and Tang, J. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 320–335.

Feng, Y.; Liu, Y.; Yang, S.; Cai, W.; Zhang, J.; Zhan, Q.; Huang, Z.; Yan, H.; Wan, Q.; Liu, C.; et al. 2025. Vision-language model for object detection and segmentation: A review and evaluation. *arXiv preprint arXiv:2504.09480*.

Gatidis, S.; Früh, M.; Fabritius, M. P.; Gu, S.; Nikolaou, K.; Fougère, C. L.; Ye, J.; He, J.; Peng, Y.; Bi, L.; et al. 2024. Results from the autoPET challenge on fully automated lesion segmentation in oncologic PET/CT imaging. *Nature Machine Intelligence*, 1–10.

Hamamci, I. E.; Er, S.; and Menze, B. 2024. Ct2rep: Automated radiology report generation for 3d medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 476–486. Springer.

Hamamci, I. E.; Er, S.; Wang, C.; Almas, F.; Simsek, A. G.; Esirgun, S. N.; Doga, I.; Durugol, O. F.; Dai, W.; Xu, M.; et al. 2024. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834*.

He, S.; Nie, Y.; Chen, Z.; Cai, Z.; Wang, H.; Yang, S.; and Chen, H. 2024. Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. *arXiv e-prints*, arXiv–2404.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jiang, Y.; Omiye, J. A.; Zakka, C.; Moor, M.; Gui, H.; Alipour, S.; Mousavi, S. S.; Chen, J. H.; Rajpurkar, P.; and Daneshjou, R. 2024. Evaluating general vision-language models for clinical medicine. *medRxiv*, 2024–04.

Jing, B.; Xie, P.; and Xing, E. 2018. On the Automatic Generation of Medical Imaging Reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2577–2586.

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

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Lin, H.; Hong, D.; Ge, S.; Luo, C.; Jiang, K.; Jin, H.; and Wen, C. 2025. Rs-moe: A vision-language model with mixture of experts for remote sensing image captioning and visual question answering. *IEEE Transactions on Geoscience and Remote Sensing*.

Lin, H.; Xu, C.; and Qin, J. 2025. Taming Vision-Language Models for Medical Image Analysis: A Comprehensive Review. *arXiv preprint arXiv:2506.18378*.

Liu, B.; Zou, K.; Zhan, L.; Lu, Z.; Dong, X.; Chen, Y.; Xie, C.; Cao, J.; Wu, X.-M.; and Fu, H. 2024. GEMeX:

A Large-Scale, Groundable, and Explainable Medical VQA Benchmark for Chest X-ray Diagnosis. *arXiv preprint arXiv:2411.16778*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, J.; Wang, Z.; Ye, Q.; Chong, D.; Zhou, P.; and Hua, Y. 2023b. Qilin-med-vl: Towards chinese large vision-language model for general healthcare. *arXiv preprint arXiv:2310.17956*.

Lucignani, G.; Paganelli, G.; and Bombardieri, E. 2004. The use of standardized uptake values for assessing FDG uptake with PET in oncology: a clinical perspective. *Nuclear medicine communications*, 25(7): 651–656.

Matsubara, K.; Ibaraki, M.; Nemoto, M.; Watabe, H.; and Kimura, Y. 2022. A review on AI in PET imaging. *Annals of Nuclear Medicine*, 36(2): 133–143.

MoonshotAI. 2025. Moonshot model site. https://www.moonshot.cn/.

Moor, M.; Huang, Q.; Wu, S.; Yasunaga, M.; Dalmia, Y.; Leskovec, J.; Zakka, C.; Reis, E. P.; and Rajpurkar, P. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, 353–367. PMLR.

Pan, J.; Liu, C.; Wu, J.; Liu, F.; Zhu, J.; Li, H. B.; Chen, C.; Ouyang, C.; and Rueckert, D. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*.

Pang, T.; Li, P.; and Zhao, L. 2023. A survey on automatic generation of medical imaging reports based on deep learning. *BioMedical Engineering OnLine*, 22(1): 48.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Peng, C.; Zhang, K.; Lyu, M.; Liu, H.; Sun, L.; and Wu, Y. 2025. Scaling Up Biomedical Vision-Language Models: Fine-Tuning, Instruction Tuning, and Multi-Modal Learning. *arXiv preprint arXiv:2505.17436*.

Peng, L.; Liao, Y.; Zhou, R.; Zhong, Y.; Jiang, H.; Wang, J.; Fu, Y.; Xue, L.; Zhang, X.; Sun, M.; et al. 2023. [18F] FDG PET/MRI combined with chest HRCT in early cancer detection: a retrospective study of 3020 asymptomatic subjects. *European Journal of Nuclear Medicine and Molecular Imaging*, 50(12): 3723–3734.

Qiu, P.; Wu, C.; Zhang, X.; Lin, W.; Wang, H.; Zhang, Y.; Wang, Y.; and Xie, W. 2024. Towards building multilingual language model for medicine. *Nature Communications*, 15(1): 8384.

Ren, S.; Laub, P.; Lu, Y.; Naganawa, M.; and Carson, R. E. 2019. Atlas-based multiorgan segmentation for dynamic abdominal PET. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 4(1): 50–62.

Schwenck, J.; Sonanini, D.; Cotton, J. M.; Rammensee, H.-G.; la Fougère, C.; Zender, L.; and Pichler, B. J. 2023. Advances in PET imaging of cancer. *Nature Reviews Cancer*, 23(7): 474–490.

Sellergren, A.; Kazemzadeh, S.; Jaroensri, T.; Kiraly, A.; Traverse, M.; Kohlberger, T.; Xu, S.; Jamil, F.; Hughes, C.; Lau, C.; et al. 2025. MedGemma Technical Report. *arXiv preprint arXiv:2507.05201*.

Sepehri, M. S.; Fabian, Z.; Soltanolkotabi, M.; and Soltanolkotabi, M. 2024. MediConfusion: Can you trust your AI radiologist? Probing the reliability of multimodal medical foundation models. *arXiv preprint arXiv:2409.15477*.

Song, X.; Zhang, X.; Ji, J.; Liu, Y.; and Wei, P. 2022. Cross-modal contrastive attention model for medical report generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, 2388–2397.

van der Geest, K. S.; Treglia, G.; Glaudemans, A. W.; Brouwer, E.; Sandovici, M.; Jamar, F.; Gheysens, O.; and Slart, R. H. 2021. Diagnostic value of [18F] FDG-PET/CT for treatment monitoring in large vessel vasculitis: a systematic review and meta-analysis. *European journal of nuclear medicine and molecular imaging*, 48(12): 3886–3902.

Wang, Y.; Lin, Z.; Xu, Z.; Dong, H.; Luo, J.; Tian, J.; Shi, Z.; Huang, L.; Zhang, Y.; Fan, J.; et al. 2024. Trust it or not: Confidence-guided automatic radiology report generation. *Neurocomputing*, 578: 127374.

Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. *arXiv preprint arXiv:2308.02463*.

Wu, Z.; Chen, X.; Pan, Z.; Liu, X.; Liu, W.; Dai, D.; Gao, H.; Ma, Y.; Wu, C.; Wang, B.; et al. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Xu, W.; Chan, H. P.; Li, L.; Aljunied, M.; Yuan, R.; Wang, J.; Xiao, C.; Chen, G.; Liu, C.; Li, Z.; et al. 2025. Lingshu: A Generalist Foundation Model for Unified Multimodal Medical Understanding and Reasoning. *arXiv preprint arXiv:2506.07044*.

Xue, L.; Feng, G.; Zhang, W.; Zhang, Y.; Li, L.; Wang, S.; Peng, L.; Peng, S.; and Gao, X. 2025. PETWB-REP: A Multi-Cancer Whole-Body FDG PET/CT and Radiology Report Dataset for Medical Imaging Research. *arXiv preprint arXiv:2511.03194*.

Xue, L.; Fu, Y.; Gao, X.; Feng, G.; Qian, S.; Wei, L.; Li, L.; Zhuo, C.; Zhang, H.; and Tian, M. 2024. [18F] FDG PET integrated with structural MRI for accurate brain age prediction. *European Journal of Nuclear Medicine and Molecular Imaging*, 51(12): 3617–3629.

Ye, J.; Wang, G.; Li, Y.; Deng, Z.; Li, W.; Li, T.; Duan, H.; Huang, Z.; Su, Y.; Wang, B.; et al. 2024. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37: 94327–94427.

Young, A.; Chen, B.; Li, C.; Huang, C.; Zhang, G.; Zhang, G.; Wang, G.; Li, H.; Zhu, J.; Chen, J.; et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Yu, T.; Zhang, H.; Yao, Y.; Dang, Y.; Chen, D.; Lu, X.; Cui, G.; He, T.; Liu, Z.; Chua, T.-S.; et al. 2024. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, Y.; Liao, Q.; Ding, L.; and Zhang, J. 2022. Bridging 2D and 3D Segmentation Networks for Computation-Efficient Volumetric Medical Image Segmentation: An Empirical Study of 2.5 D Solutions. *Computerized Medical Imaging and Graphics*, 102088.

Zhang, Y.; Shen, Z.; and Jiao, R. 2024. Segment anything model for medical image segmentation: Current applications and future directions. *Computers in Biology and Medicine*, 108238.

Zhang, Y.; Xue, L.; Zhang, W.; Li, L.; Liu, Y.; Jiang, C.; Cheng, Y.; and Qi, Y. 2025. Seganypet: Universal prompt-able segmentation from positron emission tomography images. *arXiv preprint arXiv:2502.14351*.

Zhou, T.; Xu, Y.; Zhu, Y.; Xiao, C.; Bian, H.; Wei, L.; and Zhang, X. 2025. DrVD-Bench: Do Vision-Language Models Reason Like Human Doctors in Medical Image Diagnosis? *arXiv preprint arXiv:2505.24173*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025a. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Zhu, Q.; Hou, B.; Mathai, T. S.; Mukherjee, P.; Jin, Q.; Chen, X.; Wang, Z.; Cheng, R.; Summers, R. M.; and Lu, Z. 2025b. How well do multimodal LLMs interpret CT scans? An auto-evaluation framework for analyses. *Journal of Biomedical Informatics*, 168: 104864.