PETAR: Localized Findings Generation with Mask-Aware Vision-Language Modeling for PET Automated Reporting

Danyal Maqbool[1,2], Changhee Lee[2], Zachary Huemann[2], Samuel D. Church[1,2], Matthew E. Larson[2], Scott B. Perlman[2], Tomas A. Romero[2], Joshua D. Warner[2], Meghan Lubner[2], Xin Tie[2], Jameson Merkow[3], Junjie Hu[1], Steve Y. Cho[2], Tyler J. Bradshaw[2]
[1]University of Wisconsin–Madison Department of Computer Sciences, Madison, WI, USA
[2]University of Wisconsin–Madison Department Radiology, Madison, WI, USA
[3]Microsoft, Redmond, WA, USA

Abstract

Generating automated reports for 3D positron emission tomography (PET) is an important and challenging task in medical imaging. PET plays a vital role in oncology, but automating report generation is difficult due to the complexity of whole-body 3D volumes, the wide range of potential clinical findings, and the limited availability of annotated datasets. To address these challenges, we introduce PETARSeg-11K, the first large-scale, publicly available dataset that provides lesion-level correspondence between 3D PET/CT volumes and free-text radiological findings. It comprises 11,356 lesion descriptions paired with 3D segmentations. Second, we propose PETAR-4B, a 3D vision-language model designed for mask-aware, spatially grounded PET/CT reporting. PETAR-4B jointly encodes PET, CT, and 3D lesion segmentation masks, using a 3D focal prompt to capture fine-grained details of lesions that normally comprise less than 0.1% of the volume. Evaluations using automated metrics show PETAR-4B substantially outperforming all 2D and 3D baselines. A human study involving five physicians—the first of its kind for automated PET reporting—confirms the model's clinical utility and establishes correlations between automated metrics and expert judgment. This work provides a foundational dataset and a novel architecture, advancing 3D medical vision-language understanding in PET.

1. Introduction

Vision-language models (VLMs) have demonstrated considerable potential in automating radiology report generation, with promising applications for accelerating clinical workflows and mitigating radiologist burnout [19]. However, current research has predominantly focused on 2D imaging modalities—such as chest X-rays or individual computed tomography (CT) slices, while 3D modalities, such as CT, magnetic resonance (MRI), and positron emission tomography (PET) present greater technical challenges [46]. Of these, PET is severely underrepresented, despite its critical and expanding role in oncology for diagnosis, staging, and treatment response assessment. PET scans provide unique functional insights into metabolic activity, enabling early disease detection and treatment monitoring.

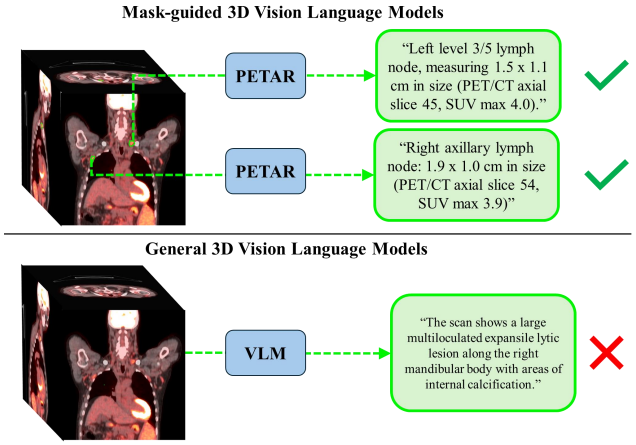


Figure 1. Overview of mask-guided PET/CT report generation. By incorporating lesion-level masks, PETAR produces anatomically fine-grained findings grounded in the 3D volume. In contrast, general 3D models perform global encoding without fine-grained anatomical correlation, hence they generate vague or clinically incomplete descriptions.

PET presents several unique challenges that existing VLMs are poorly equipped to handle. First, clinical PET reports require fine-grained, lesion-level descriptions—not just global study summaries—which allows physicians to track individual findings over time. These descriptions are complex, as they often specify a lesion's anatomical site,

sub-site, laterality, morphology, and metabolic activity, leading to a vast number of possible combinations (see Figure 1). This, together with the tendency to scan the entire body in a PET exam, makes PET reports arguably the longest in radiology, at times 3-fold longer than CT reports [38]. This combination of report length and descriptive granularity creates a significant challenge of fine-grained visual grounding. Second, these clinically relevant lesions can be numerous, very small (averaging less than 0.1% of the total image volume), and spatially dispersed, making them challenging for standard vision encoders. Existing 3D VLMs, such as CT2Rep [16], M3D [2], and Merlin [5] typically rely on global feature extraction and down-sampling, a process that can erase these fine-grained details [9]. Furthermore, existing medical 3D VLMs have been predominantly trained on CT images (anatomical imaging) rather than PET images (molecular/metabolic imaging) and are not designed to jointly process dual-modality PET/CT volumes. Compounding this, a critical barrier is the lack of any large-scale, public dataset that provides the necessary 3D lesion-level segmentations aligned with free-text radiological findings. Figure 1 illustrates our task and how conventional 3D medical VLMs fail to capture the specificity of lesion findings. Without accurate, lesion-specific findings, these models offer limited utility and reliability.

Our work addresses these issues by making two foundational contributions: a dataset and a model. Our dataset provides a direct connection between spatially localized lesion information and its corresponding natural language description, enabling precise alignment between imaging features and clinical language. Our model takes PET/CT scans and corresponding lesion segmentations as input to generate findings, leveraging the spatial cues in the masks to improve both specificity and interpretability. The segmentation masks can be obtained from advanced segmentation models [15] or FDA-cleared single-click PET lesion segmentation tools such as PetEdge [43]. Our contributions are as follows:

- We introduce **PETARSeg-11K**, a large-scale dataset of PET/CT studies with explicit lesion-level alignment between image space and report text. The dataset comprises **11,356 lesion descriptions** paired with corresponding 3D lesion segmentations, collected from over **5,000 whole-body PET/CT exams**. Each lesion is associated with a referring expression extracted from clinical radiology reports using a hybrid pipeline that couples anatomical cues with LLM-based parsing and normalization. To our knowledge, this is the first dataset to link localized PET/CT lesion masks with free-text descriptions.
- We propose **PETAR-4B**, a 3D mask-aware VLM designed for PET/CT findings generation. PETAR-4B jointly encodes PET, CT, and 3D lesion masks, enabling the model to condition language generation on both global disease context (e.g., disease distribution) and fine-grained lesion attributes (e.g., metabolic activity).
- We present a comprehensive evaluation of PETAR-4B and baseline 2D and 3D VLMs on lesion-level PET/CT findings generation. We employ automated evaluation metrics and human evaluations to quantify clinical usefulness. We benchmarked these metrics against expert physician preferences to understand which metrics are most reliable for PET findings generation.

2. Related work

2.1. Medical vision language models (VLMs)

Medical VLMs encode multimodal inputs, such as medical images, into a shared embedding space, and leverage large language models to generate clinically meaningful outputs. MAIRA [4, 21] generates detailed radiology reports from chest X-rays through aligned vision–language encoders, while LLaVA-Med [24] adapts general-domain VLMs to biomedical applications via curriculum learning on figure–caption pairs. However, these methods are limited to the 2D plane, discarding critical spatial information inherent in 3D medical volumes.

To extend multimodal reasoning into volumetric contexts, M3D [2] leverages a large-scale multimodal dataset for 3D segmentation and report generation. Merlin [5] aligns 3D CT data with both structured and unstructured clinical information within a unified embedding space, and CT2Rep [16] introduces a causal 3D feature extractor for automated chest CT report generation. RadFM [45], trained on both 2D and 3D data, seeks to build a generalist model capable of handling diverse radiologic tasks across modalities. A crucial gap within these works is that they are *mask-agnostic*, limiting their ability to localize and describe lesions or regions with fine-grained precision.

The datasets underpinning these models primarily rely on CT and X-ray modalities, with CT dominating recent 3D efforts. Several PET/CT datasets exist but each lacks key elements needed for lesion-level captioning. As summarized in Table 1, most public PET/CT datasets provide only images or masks, but lack the accompanying text findings needed for lesion-level captioning. ViMed-PET [30] and Pet2Rep [55] include textual findings but lack grounded lesion masks. Together, these datasets capture only parts of the task, underscoring the need for a resource that combines findings and masks for training PET/CT VLMs.

2.2. Fine-grained captioning

Advances have been made toward fine-grained visual grounding in 2D images. Early methods, such as Kosmos [34], Shikra [8], and GPT4ROI [52] incorporate explicit spatial cues—such as bounding box coordinates—into textual prompts to direct visual attention. More recent approaches have introduced flexible region-level interactions, allowing

segmentation masks or point-based cues to guide the model's focus and improve spatial grounding, such as AlphaClip, and ViP-LLaVA [6, 18, 37, 51]. To further improve fine-grained analysis, Describe Anything Model [26] introduces a zoomed in *focal prompt* as input. This zoomed in input further improves fine-grained analysis capabilities by incorporating both global and local information.

In the medical domain, Reg2RG [9] employs these ideas by introducing a mask input channel to generate more detailed, region-aware findings from CT scans. However, it is designed to use large, organ-level masks to generate region-level descriptions (e.g., entire lung). Its scope is also limited to single-modality chest CT and is not designed for the fine-grained abnormality-focused task of lesion-level captioning. Our work builds upon this direction by introducing a 3D, mask-aware architecture capable of jointly processing dual-modality PET and CT scans. Crucially, it uses mask-guided focal prompting to enable localized, clinically grounded captioning of small, discrete lesions across the body, a task that is anatomically and technically distinct.

3. PETARSeg-11K dataset construction

3.1. Preliminaries on PET/CT reports

PET reports describe how radiotracers are distributed in the body in a PET scan. This is usually complemented by a CT or MRI scan for anatomical reference. Within each body region, reports list the lesions or regions with abnormal radiotracer uptake and include measurements such as standardized uptake value (SUV_{max}), slice numbers, and qualitative descriptions. Because lesions are usually tracked across multiple scans, this information helps future readers identify the lesions of concern.

3.2. Dataset construction pipeline

This research was conducted under a protocol approved by our Institutional Review Board (IRB), which granted a waiver of informed consent. All data was fully anonymized.

Table 1. Comparison of publicly available PET/CT datasets

Dataset	Total scans	Findings	Grounded Masks	Whole body
RIDER Lung PET-CT [28]	244	×	×	×
Head-Neck PET-CT [40]	298	×	×	×
Lung-PET-CT-Dx [25]	251	×	×	×
FDG-PET-CT-Lesions [15]	1014	×	✓	✓
SPADe [13]	1614	×	×	✓
SegAnyPet [54]	5,731	×	✓	✓
Pet2Rep [55]	565	✓	×	✓
ViMed-PET [30]	2757	✓	×	✓
PETARSeg-11K (ours)	5126	✓	✓	✓

We collected a set of 33,000 PET/CT scans from patient hospital records. Then, we leveraged the pipeline developed by Huemann *et al*. [20] to create grounded segmentations for metabolic lesions. The method uses an ensemble

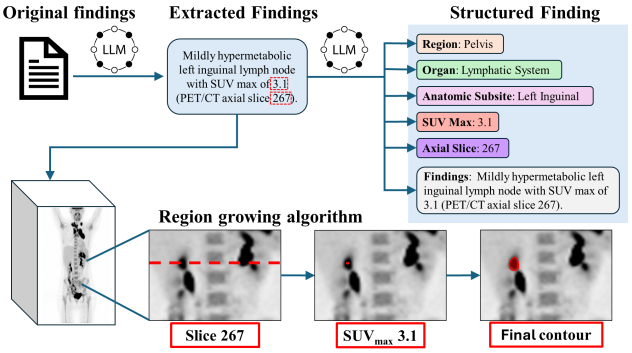


Figure 2. Overview of the PETARSeg-11K data pipeline. LLMs extract key lesion attributes (e.g., SUV_{max} and slice number) from radiology reports, which guide a region-growing algorithm to localize and refine PET lesions. The final findings are structured into standardized fields linking text to 3D lesion masks.

of language models to filter out irrelevant sentences, disambiguate references to prior studies, and accurately extract the SUV_{max} value and corresponding slice number. Specifically, we used an ensemble of Mistral-7B-Instruct and Mixtral-8×7B-Instruct. Lesion masks were generated using an iterative thresholding algorithm [22] applied to the PET volume. The algorithm first applies a threshold based on the reported SUV_{max}, producing an initial set of candidate regions. Connected components are identified, and the component whose SUV_{max} matches the reported SUV_{max} (within ± 0.1) and intersects the reported axial slice is selected. From the matching SUVmax and slice, we start at the max pixel and iteratively grow the contour until it stabilizes. The resulting dataset contains 11,356 lesion descriptions across 5,126 unique exams, covering multiple radiotracers including ^{18}F-fluorodeoxyglucose (FDG), ^{68}Ga-(DOTA-(Tyr^3)-octreotate) (DOTATATE), ^{18}F-fluciclovine, and ^{18}F-DCFPyL. The data was resampled to $3\,\mathrm{mm}$ resolution with dimensions $192 \times 192 \times 352$. Physician evaluations of a sample of this dataset recorded a 98% accuracy in contour locations [20].

Using a strong LLM, Qwen3-30B-A3B [48], each description is then formatted into a structured schema. An illustration of a complete sample is provided in Figure 2. This representation explicitly grounds spatial and anatomical references.

To strengthen the model's understanding of global anatomical structure, we curated a complementary pretraining dataset. Using TotalSegmentator [42], we automatically segmented the CT volumes of our PET/CT dataset, generating roughly 100,000 labeled segmentations spanning its 117 predefined anatomical classes. This dataset provides comprehensive anatomical coverage, and encourages model alignment between metabolic activity in PET scans and the structural anatomies in the CT scan.

Our dataset is the first publicly available multimodal PET/CT dataset with lesion-level granularity. As we highlight in Table 1, existing PET/CT datasets are typically limited in scope, focusing on specific organs, lacking text, lesion-level findings, or missing segmentation masks. In contrast, our dataset, which we term **PETARSeg-11K**, provides the first large-scale, whole-body PET/CT collection with aligned findings, 3D lesion masks, and textual descriptions.

4. PETAR model design

4.1. Mask aware inputs

Given a 3D PET scan $\boldsymbol{P} \in \mathbb{R}^{D\times W\times H}$, a CT scan $\boldsymbol{C} \in \mathbb{R}^{D\times W\times H}$, and a binary mask $\boldsymbol{M} \in \{0,1\}^{D\times W\times H}$ indicating the region of interest, the objective is to generate detailed diagnostic findings y focused on the masked region. Let f_θ denote our model parameterized by θ. The generation is:

$$y = f_\theta(\boldsymbol{P}, \boldsymbol{C}, \boldsymbol{M}). \tag{1}$$

4.2. Focal prompt

A key bottleneck in lesion-level understanding is that lesions are often a very small part of a volume, and so standard approaches lead to a risk of information loss during common global processing steps, such as resizing, cropping, or downsampling. For instance, in our dataset, the average binary mask $\boldsymbol{M}$, occupies less than 0.1% of the total scan volume, and the sizes range from 0.01% to 2%. To address this, we extended the ideas of the Describe Anything Model [26] to create a 3D focal prompt input to provide a localized, high-resolution view of the lesion.

We first center the crop on the mask $\boldsymbol{M}$ and extract a cubic subvolume covering the region of interest. To enhance robustness and avoid overfitting to fixed spatial positions, we apply random spatial perturbations to both the cube center and its side length, limiting deviations to 20% of the original mask center and cube size. The crop is adjusted such that the mask remains fully contained within the subcube, ensuring complete lesion visibility in the focal prompt.

Specifically, we denote the lesion mask center as $c \in \mathbb{R}^3$ and the desired crop size as r. We apply the small random perturbations to both the center and size as follows:

$$\tilde{c} = c + \triangle c, \quad \tilde{r} = r + \triangle r, \tag{2}$$

$$\triangle c,\ \triangle r \overset{\text{i.i.d.}}{\sim} \mathcal{U}(-0.2r, 0.2r), \tag{3}$$

where $\triangle c$ denotes a random translation offset, and $\triangle r$ represents a random variation in crop size.

The focal PET, CT, and mask crops are then extracted as:

$$\boldsymbol{F}_P, \boldsymbol{F}_C, \boldsymbol{F}_M = \text{Crop}(\boldsymbol{P}, \boldsymbol{C}, \boldsymbol{M}; \tilde{c}, \tilde{r}), \tag{4}$$

where $\boldsymbol{F}_P, \boldsymbol{F}_C, \boldsymbol{F}_M \in \mathbb{R}^{\tilde{r}\times\tilde{r}\times\tilde{r}}$ represent the PET, CT, and mask focal crops, and $\text{Crop}(\cdot)$ denotes the operation that extracts the 3D subvolume. These focal crop inputs can be seen in Figure 3.

4.3. Shared visual encoding

For a more efficient architecture, we adopted a shared 3D vision transformer [12] to encode both PET and CT volumes. PET scans on their own contain limited structural information about the body, and are complemented by CT scans for anatomical context. Unlike prior work, we propose to jointly process them to capture both pieces of information. Specifically, to prepare the sequential inputs for the shared transformer, the PET and CT volumes are split into non-overlapping 3D patches of size (s_D, s_W, s_H), which are then flattened into a sequence of K patches and linearly projected into token embeddings of dimension d.

$$\boldsymbol{Z}_P = \text{PatchEmbed}(\boldsymbol{P}) \in \mathbb{R}^{K\times d} \tag{5}$$

$$\boldsymbol{Z}_C = \text{PatchEmbed}(\boldsymbol{C}) \in \mathbb{R}^{K\times d} \tag{6}$$

where $\boldsymbol{Z}_C$ and $\boldsymbol{Z}_P$ are the patch embeddings. The mask embeddings are produced using a different set of parameters.

$$\boldsymbol{Z}_M = \text{PatchEmbed}(\boldsymbol{M}) \in \mathbb{R}^{K\times d} \tag{7}$$

The 3D vision transformer, denoted by $\mathcal{T}$, encodes the embeddings while incorporating the lesion mask into the PET scan via element-wise additive conditioning:

$$\mathbf{X}_{\text{PET}} = \mathcal{T}\big(\mathbf{Z}_P + \mathbf{Z}_M\big), \quad \mathbf{X}_{\text{CT}} = \mathcal{T}\big(\mathbf{Z}_C\big). \tag{8}$$

We then fuse the CT and mask-aware PET token embeddings to obtain the global features $\boldsymbol{X}$ by concatenating them along their embedding dimension:

$$\mathbf{X} = \text{Concat}(\mathbf{X}_{\text{PET}}, \mathbf{X}_{\text{CT}}) \in \mathbb{R}^{K\times d_v}. \tag{9}$$

The same process from Eq. (5)-(9) is applied to the focal crops $(\boldsymbol{F}_P, \boldsymbol{F}_C, \boldsymbol{F}_M)$, producing the focal embeddings $\tilde{\mathbf{X}}$. The global and focal features are then combined via element-wise addition and reduced via spatial pooling [2]. These visual tokens are then linearly projected into the language model's text embedding space:

$$\mathbf{T} = \big(\mathbf{X} + \tilde{\mathbf{X}}\big) \in \mathbb{R}^{K\times d_v}. \tag{10}$$

$$\mathbf{V} = \text{Proj}(\text{SpatialPooler}(\mathbf{T})). \tag{11}$$

Finally, the projected tokens $\mathbf{V}$ and a lesion-description query q are passed to the language model decoder to generate the localized findings y:

$$y = \text{LLMDecoder}(q, \mathbf{V}). \tag{12}$$

4.4. Training stages

We employ a three-stage training pipeline, where each stage optimizes the same objective in Eq. (13) while updating either partial or full model parameters. The entire procedure is first applied to our curated TotalSegmentator (TS)

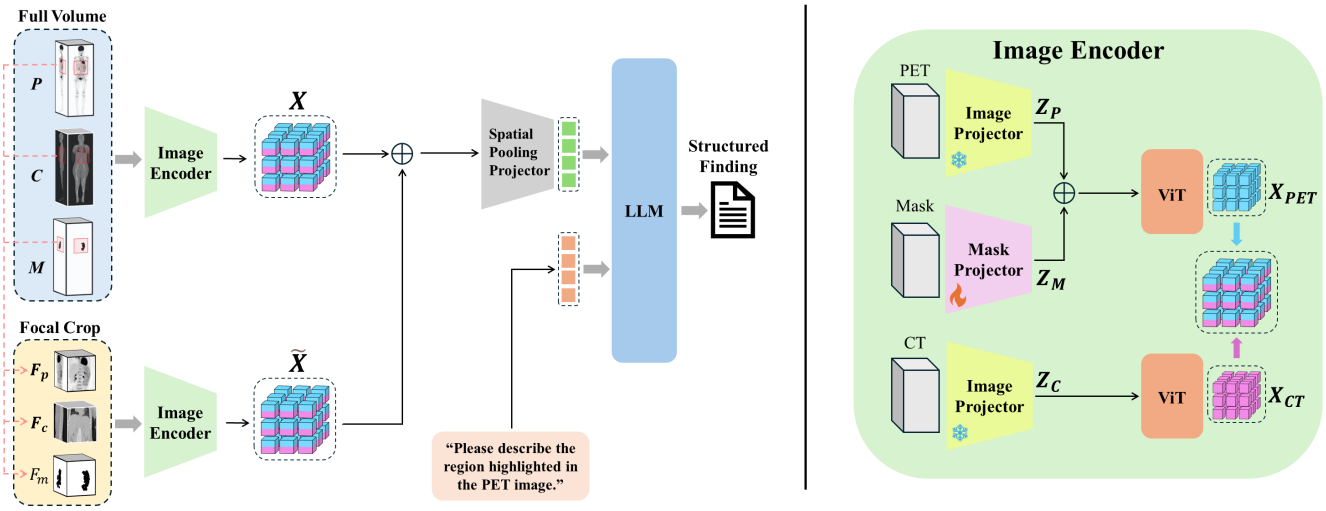


Figure 3. **Overview of the proposed framework.** The **left panel** illustrates the overall architecture, which integrates PET, CT, and lesion mask inputs through 3D convolutional image projectors and an M3D-CLIP backbone. The resulting multi-modal visual tokens are fused and spatially pooled before being passed to a Phi3-4B language model that generates clinically grounded text descriptions conditioned on visual features and textual prompts. The **right panel** details the *Image Encoder* design, where modality-specific projectors (for PET, CT, and mask inputs) map each input into a shared latent space. These embeddings are subsequently processed by a ViT encoder to produce modality-aligned visual tokens for downstream fusion.

dataset for pretraining, where each sample is formatted as a question–answer pair: *Question:* "What is the region highlighted by the mask?" *Answer:* "The highlighted region is the ⟨Region Class⟩". The same three-stage pipeline is then repeated on PETARSeg-11K.

Stage 1: Mask embedding alignment. We train only the projection head that maps PET/CT visual features into the language model's embedding space in Eq. (11). To ensure the learning focuses solely on this mapping, the mask-embedding weights are initialized to zero, and both the PET/CT encoder and the language model remain frozen.

Stage 2: Projector alignment. Next, we train the mask embedding module in Eq. (7) while freezing all other components. As shown in Figure 3 (right), the mask weights are trained to encode binary masks in a way that aligns with the underlying 3D anatomical and metabolic signals. This encourages the vision encoder to build meaningful correspondences between mask shapes and anatomical context.

Stage 3: Full finetuning. Finally, we unfreeze the full architecture and perform end-to-end finetuning. In this stage, PET/CT features, mask-conditioned representations, and the language model are jointly optimized to generate accurate region-aware descriptions from the input mask.

4.5. Training objective

Given a training dataset $\mathcal{D}$ of size D, each example consists of PET–CT–mask visual triplets, a lesion-description query q, and a target token sequence $y = (y_1, \ldots, y_N)$ of length N. The model is trained to predict each token autoregressively, conditioned on the visual features and all previously generated tokens. Formally, the objective is the autoregressive negative log-likelihood:

$$\mathcal{L}(\mathcal{D}, \theta) = -\sum_{(\mathbf{V},q,y)\sim\mathcal{D}} \sum_{i=1}^{N} \log p_\theta(y_i \mid \mathbf{V}, q, y_{<i}), \quad (13)$$

where $\mathbf{V}$ denotes the visual tokens encoded from the PET–CT–mask encoder in Eq. (11), and $y_{<i}$ denotes all preceding tokens in the sequence. This is used for all stages of training.

5. Experiments and analysis

5.1. Experimental setup

Implementation. We used the vision encoder and language model from M3D [2], a widely used medical VLM trained on large 3D medical image–text datasets. The vision encoder is a Vision Transformer with a 3D convolutional projection layer, and the language model is Phi3-4B [1]. All experiments were done on 2 NVIDIA L40S GPUs. All stages were trained for 10 epochs, taking about 20 hours total. A complete list of hyperparameters are provided in the supplementary material.

Test set. As PETARSeg-11k is the first dataset of its kind for PET scans, evaluation was performed on a held-out set of 1175 test samples. We also manually curated a test set of size 32 from the autoPET [15] dataset for our reader evaluations. This dataset does not have corresponding reports and therefore automatic evaluation metrics do not apply.

Table 2. Comparison of selected models on the PETARSeg-11k test set. The "(*finetuned*)" indicates the model was trained on PETARSeg-11k.

Model / Metric	BLEU	ROUGE-L	METEOR	CIDEr	BART	BERT	RaTE	GREEN
2D VLMs								
MedGemma-4B	0.124	0.352	0.358	0.027	-4.92	0.690	0.540	0.011
HuatuoGPT-Vision-7B	0.130	0.350	0.357	0.024	-4.85	0.695	0.584	0.015
InternVL3-8B	0.137	0.355	0.356	0.035	-4.72	0.694	0.614	0.030
Qwen3-VL-8B	0.375	0.343	0.364	0.051	-4.74	0.690	0.612	0.022
QoQ-Med	0.364	0.317	0.368	0.025	-4.64	0.680	0.505	0.006
ViP-LLaVA	0.064	0.301	0.308	0.009	-5.96	0.651	0.503	0.006
Qwen3-VL-8B (*finetuned*)	0.484	0.443	0.501	0.086	-4.40	0.750	0.608	0.060
MedGemma-4B (*finetuned*)	0.495	0.454	0.510	0.119	-4.36	0.754	0.613	0.086
3D VLMs								
Med3DVLM	0.004	0.080	0.066	0.005	-5.70	0.511	0.285	0.002
M3D	0.327	0.276	0.323	0.013	-5.32	0.634	0.505	0.000
M3D-RAD	0.343	0.300	0.340	0.016	-5.23	0.658	0.518	0.003
Reg2RG	0.044	0.060	0.108	0.001	-5.54	0.518	0.363	0.002
Reg2RG (*finetuned*)	0.478	0.416	0.487	0.055	-4.58	0.732	0.532	0.031
M3D-RAD (*finetuned*)	0.485	0.446	0.501	0.132	-4.34	0.750	0.627	0.071
PETAR-4B (*Ours*)	**0.535**	**0.524**	**0.560**	**0.457**	**-4.00**	**0.795**	**0.713**	**0.257**

Table 3. Ablation study of our model showing the effect of each component on multiple evaluation metrics. TS=TotalSegmentator pretraining.

Mask	CT	Focal	TS	BLEU	ROUGE	METEOR	CIDEr	BERTScore	BARTScore	RaTEScore	GREEN
×	×	×	×	0.485	0.446	0.501	0.132	0.750	-4.34	0.627	0.071
✓	×	×	×	0.480	0.445	0.498	0.137	0.748	-4.33	0.626	0.088
×	✓	×	×	0.477	0.436	0.499	0.134	0.746	-4.31	0.622	0.060
×	×	✓	×	0.528	0.518	0.550	0.397	0.787	-4.05	0.698	0.226
✓	×	✓	×	0.525	0.518	0.550	0.381	0.787	-4.05	0.700	0.232
×	✓	✓	×	0.517	0.507	0.540	0.428	0.784	-4.08	0.587	0.234
✓	✓	✓	×	0.521	0.519	0.551	0.439	0.787	-4.05	0.613	0.239
✓	✓	✓	✓	**0.535**	**0.524**	**0.560**	**0.457**	**0.795**	**-4.00**	**0.713**	**0.257**

Compared models. We compared a variety of 2D and 3D VLMs, including both standard and prompt-aware variants. For 2D models, we used strong general-purpose and medical-specific baselines: InternVL3-8B [57], Qwen3-VL-8B-instruct [48], HuatuoGPT-Vision-7B [7], MedGemma-4B [36], and QoQ-Med [11]. The 2D prompt-aware model ViP-LLaVA [6] adds visual grounding with mask-conditioned prompting. For 3D models, we evaluated M3D [2], Med3DVLM [47] and M3D-RAD [14], along with the 3D prompt-aware Reg2RG [9], which incorporates mask-guided conditioning. We finetuned representative 2D and 3D models on our dataset for fair comparison and validation. For 2D finetuning, we followed a similar strategy to [55] and used 3 representative slices of the PET scan from the sagittal, coronal, and axial views to approximate 3D context. During inference stage for 2D models, these 3 slices were provided and the model was queried to produce findings for these images. During inference for 3D models, the entire volume was provided.

Metrics We evaluated model predictions using 3 categories of metrics. N-gram overlap metrics (BLEU [32], ROUGE [27], METEOR [3], CIDEr [41]), which measure precision, recall, and consensus-weighted word overlap; semantic metrics (BERTScore [53], BARTScore [50]), which use contextual embeddings to measure semantic similarity; and language model–based metrics (RaTEScore [56], GREEN [31]), which use large language models to assess clinical correctness, factual consistency, and reasoning.

5.2. Main results

We found that current VLMs, primarily trained on CT and radiography data, are limited in their ability to interpret PET data, as shown in Table 2. This indicates that the domain shift introduced by PET images and reporting styles is substantial. 3D VLMs like M3D-RAD and Med3DVLM showed limited captioning ability, with GREEN scores of just 0.002–0.03 and BERTScores below 0.66. Thus, our proposed dataset offers an important resource for bridging the performance of VLMs from CT to PET/CT.

Finetuning existing vision–language models on our PET dataset yielded consistent performance gains, particularly in clinically grounded metrics such as GREEN and RaTE, which assess factual and interpretive accuracy. For instance, Qwen3-VL-8B improves GREEN from 0.022 to 0.066 after

fine-tuning, and M3D-RAD rises from 0.003 to 0.071. In contrast, our PETAR-4B model—designed with mask-aware conditioning and focal prompting—achieves substantial absolute improvements across all linguistic, semantic, and clinical metrics. As shown in Table 2, PETAR-4B surpassed the strongest 2D model (MedGemma) in BLEU-4 (0.535 vs 0.495), ROUGE-L (0.524 vs 0.454), and METEOR (0.560 vs 0.510). It also exceeded the top 3D model M3D-RAD (finetuned) on BERTScore (0.795 vs 0.750) and RaTEScore (0.713 vs 0.627), while achieving a higher GREEN score (0.257 vs 0.071). These results indicate that beyond data exposure, the architectural choices in PETAR-4B—particularly the use of mask-guided feature conditioning and region-focused prompting—are central to achieving clinically meaningful improvements in PET report generation.

Qualitative results are presented in Figure 4. These results illustrate the difficulty of the task, both in accurately describing the location and visual features of a lesion, as well as correctly interpreting it (e.g., "likely represents physiological activity"). The figure includes comparisons between PETAR and M3D-RAD before and after fine-tuning. In the zero-shot setting, M3D-RAD hallucinated irrelevant anatomy (e.g., describing the "maxillary ridge" instead of the correct paratracheal node). After fine-tuning on PETARSeg-11K (Fig. 4B), M3D-RAD showed improvement but still frequently misidentifies regions. For example, labeling uptake in the left inguinal node as the "left proximal femur." In contrast, PETAR consistently aligned textual descriptions with real visual features and anatomical context, demonstrating superior grounding and clinical interpretability. The strong quantitative gains in CIDEr (+0.325) and GREEN (+0.186) further validate this qualitative advantage, indicating that PETAR-4B not only captures descriptive meaning but also generates clinically meaningful descriptions. MedGemma, shows similar performance trends as M3D-RAD.

5.3. Human evaluation

Though automated metrics are widely used to evaluate medical report generation, it is unclear how well they reflect what clinicians actually value. Many metrics focus on word overlap rather than semantic meaning, so they may penalize valid paraphrases or reward text that sounds correct but is factually wrong. Even newer, language model–based metrics such as GREEN, which aim to capture meaning, can still overrate text that seems plausible but introduces unsupported details.

To assess our model's performance from an expert perspective, we conducted a human evaluation with five board-certified nuclear medicine physicians. They reviewed 116 pairs of real (i.e., physician-generated) and PETAR-generated lesion descriptions, blinded to the source of the descriptions, and rated each on a standardized 5-point scale for anatomical accuracy, interpretation correctness, and clinical usefulness. We excluded errors related to quantitative measurements (e.g., lesion size) since those values are sometimes hallucinated by the model, but can be directly derived from the input masks, so they are not the focus of our evaluation. They also evaluated 32 instances of the autoPET dataset for testing on unseen distributions. To the best of our knowledge, this is the first human study of PET report generation by vision-language models.

Table 4. Spearman's correlation between automated metrics and human evaluation scores. Higher correlation indicates stronger alignment with expert judgment.

Metric	**Spearman's ρ**
GREEN	0.592
RaTEScore	0.550
BERTScore	0.511
ROUGE	0.471
METEOR	0.438
CIDEr	0.421
BLEU	0.214
BARTScore	0.168

Table 5 shows that PETAR-4B produces clinically useful findings, with anatomical, interpretation, and utility scores all in the 3.7–3.9 range compared to human scores of 4.3–4.4. Furthermore, the physicians preferred, or considered equal, the model generated findings in $\approx 60\%$ of cases (69/116). PETAR-4B achieved similarly strong scores on the external AutoPET dataset, indicating consistent performance for out-of-distribution data.

To explore which evaluation metrics best align with human judgment for PET report generation, we used the ratings from the reader study and analyzed how existing automated metrics correlate with human scores. Results are shown in Table 4. Metrics that capture semantic and contextual similarity showed markedly stronger correlation with radiologist preferences compared to traditional n-gram–based scores like BLEU. This finding underscores the need for evaluation standards that better reflect clinical reasoning rather than surface-level text similarity. GREEN provided the best indicator of PET report quality (ρ=0.59).

5.4. Ablation studies

To evaluate the contribution of key design components in PETAR-4B, we performed ablation experiments on the mask input, CT modality, focal prompt, and TotalSegmentator pretraining (Table 3). Removing any of PETAR's modules led to a measurable decrease in performance across all metrics. The mask input provided crucial spatial grounding, substantially improving lesion localization and factual accuracy. Adding the CT modality enhanced anatomical coherence and contextual understanding, which translates to higher BERTScore and GREEN. The focal prompt had the strongest impact over-

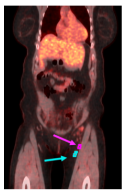
PET/CT Fusion	Ground Truth	PETAR	M3D-RAD (*baseline*)	M3D-RAD (*fine-tuned*)	MedGemma (*baseline*)	MedGemma (*fine-tuned*)
	•6.3 x 3.5 cm large nodal conglomerate involving left levels IIA, IIB and III measuring SUV max of 14.7	•The most FDG avid lymph node conglomerate is located in left level 2/3, measuring approximately 5.0 x 4.0 cm (max SUV 15.8).	•Sure, it is characterized by inflammation of the Synovium of the first metatarsophalangeal joint with subcutaneous gas.	•There is an intensely FDG avid left level 3/4 lymph node with an SUV Max of 10.1 and a size of 1.2 x 1.1 cm	•Uptake in the left lobe of the thyroid gland	•Intensely FDG avid lytic lesion in the C6 vertebra with SUV max of 14.7 (PET/CT axial slice 54).
	•Asymmetric intense uptake in the right vocal cord measuring SUV max of 5.9, presumably represents top normal physiologic uptake.	•There is a focus of moderate hypermetabolic activity fusing to the right posterior laryngeal wall (SUV max 6.0) that most likely represents physiologic activity.	•The report describes a lytic lesion with a soft tissue component arising from the posteromedial aspect of the distal femoral metaphysis, measuring 5.5 x 3.5 x 5 cm, with a 15 mm bony septa.	•Left supraclavicular lymph node, 0.8 cm, SUV max 3.8	•Uptake is present in the right upper lobe.	•There is a mildly enlarged left level 5 lymph node measuring 1.2 x 0.9 cm with moderate FDG uptake (PET/CT axial slice 86, SUV max 6.3).
	•Mildly hypermetabolic left inguinal lymph node with SUV max of 3.1	•For example, a left inguinal lymph node measures 1.2 cm with SUV max 3.0	•Key anatomical feature: Major muscle group on the right side, aiding in hip abduction and rotation.	•Left proximal femur without clear correlate on noncontrast CT images, SUV max 4.0.	•Uptake in the right upper lobe of the lung.	•Intensely FDG avid lytic lesion in the left anterior seventh rib with SUV max of 4.6 (PET/CT axial slice 144).
	•Mildly hypermetabolic left inguinal lymph node with SUV max of 2.5	•Bilateral inguinal lymph nodes with mild FDG uptake are nonspecific (SUV max 2.6 for bilateral inguinal lymph nodes	•AI-guided biopsy of the lesion in the left iliac bone	•A focus of moderate FDG uptake fusing to a subcutaneous soft tissue nodule/stranding within the right anterior abdominal wall (SUV max 3.8)	•Uptake is present in the right lobe of the liver.	•Intensely FDG avid left inguinal/pelvic lymph node with SUV max of 4.7 (PET/CT axial slice 263).
	•Examples include right upper paratracheal 2.1 x 1.6 cm node which measures SUV max of 12.4	•Intensely FDG avid right paratracheal lymph node with SUV max of 13.5	•The scan reveals a large infiltrative osseous mass involving the left maxillary ridge, with extension into the left orbit.	•Arteriovenous malformation (AVM)	•There is increased uptake in the right upper lobe of the lung.	•Left axillary lymph node measuring 1.4 x 1.1 cm, SUV max 8.7, axial slice 101.
	•Right lower paraesophageal node measuring SUV max of 6.5	•There is a focus of moderately increased FDG uptake within the distal esophagus (SUV max 7.0	•A large soft tissue mass lesion is seen involving the root of the mesentery, with surrounding fatty stranding, enlarged regional lymph nodes, and multiple metastatic lung nodules.	•Large soft tissue mass lesion is seen involving the root of the mesentery, with fatty stranding, enlarged regional lymph nodes, and multiple metastatic lung nodules.	•There is a focal area of increased uptake in the right lobe of the liver.	•Intensely FDG avid focus within the T9 vertebral body with SUV max of 5.6 (PET/CT axial slice 156).

Figure 4. A comparison between different models. PETAR consistently produces anatomically correct descriptions. Prior to fine-tuning, both MedGemma and M3D-RAD produce highly inaccurate results. After fine-tuning, the models still tend to make localisation errors. Anatomical descriptors are underlined for ease of comparison (red=incorrect, green=correct). Note: quantitative measurements (lesion size, SUVmax) are hallucinated by the models but can be easily replaced with directly measured values using the input lesion masks.

all, sharpening attention on fine-grained details that might otherwise be lost. TotalSegmentator pretraining also provided a measurable boost in performance. By combining all components, PETAR-4B achieved the best performance, indicating PET/CT reasoning benefits from the synergy between volumetric grounding and lesion-focused prompting.

5.5. Limitations and future work

A feature of our method is the requirement of mask inputs to achieve the best performance. Currently, these would need to come from physicians. However, this could easily fit into a typical physician reading workflow, where each lesion could be contoured by a physician using single-mouse-click segmentation methods (MIM's PETEdge+, which already has wide clinical adoption), and then the report findings would be drafted using PETAR. However, a fully automated pipeline is possible, wherein state-of-the-art lesion segmentation algorithms first identify and segment potential findings, which serve as inputs to PETAR. Substantial work has already been published on automated PET lesion detection and segmentation [35, 39, 49]. In future work, we will explore ways to build this pipeline using such models, and how to best convert lesion descriptions into templated PET reports.

Table 5. Human evaluation of **PETAR-4B** on an internal test set and an external AutoPET dataset. Five blinded nuclear medicine physicians compared PETAR findings with original reports.

Metric	Internal Reader Study		AutoPET (External)
	Human	PETAR-4B	PETAR-4B
Anatomical score	4.4 ± 0.9	3.9 ± 1.2	3.8 ± 1.1
Interpretation score	4.4 ± 0.8	3.9 ± 1.1	4.0 ± 1.0
Utility score	4.3 ± 0.9	3.7 ± 1.1	3.8 ± 1.1
PETAR preferred (# cases)	15		–
Preference tied (# cases)	54		–
PETAR not preferred (# cases)	47		–

6. Conclusion

We introduced PETAR, a 3D mask-aware vision–language framework for PET/CT report generation, along with PETARSeg-11k, the first large-scale dataset linking lesion-level PET/CT findings to descriptive clinical text. By combining metabolic (PET) and structural (CT) cues through mask-guided volumetric encoding and focal prompting, our model captures fine-grained details of small, spatially-dispersed lesions often lost by other models. PETAR sets a

new state-of-the-art performance across linguistic, semantic, and clinically grounded metrics. Our evaluations, including a human study with five board-certified physicians, confirmed the model's ability to produce accurate and clinically useful findings that show strong alignment with expert judgment.

7. Acknowledgments

Research reported in this publication was supported by the National Institute Of Biomedical Imaging And Bioengineering of the National Institutes of Health under Award Number R01EB033782. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. 5

[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. 2, 4, 5, 6

[3] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures*, pages 65–72, 2005. 6

[4] Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Fabian Falck, Ozan Oktay, Anja Thieme, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024. 2

[5] Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Cesar Truyts, et al. Merlin: A vision language foundation model for 3d computed tomography. *Research Square*, pages rs–3, 2024. 2

[6] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In *CVPR*, pages 12914–12923, 2024. 3, 6

[7] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. Huatuogpt-vision: Towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*, 2024. 6

[8] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 2

[9] Z. Chen, Y. Bie, H. Jin, and H. Chen. Large language model with region-guided referring and grounding for CT report generation. *IEEE Trans. Med. Imaging*, 44(8):3139–3150, 2025. 2, 3, 6

[10] The MONAI Consortium. MONAI: Medical open network for AI. https://monai.io, 2020. Version: 1.5.0. 1

[11] Wei Dai, Peilin Chen, Chanakya Ekbote, Paul Pu Liang, et al. Qoq-med: Building multimodal clinical foundation models with domain-aware grpo training. *arXiv preprint arXiv:2506.00711*, 2025. 6

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 4

[13] Sabri Eyuboglu, Geoffrey Angus, Bhavik N. Patel, Anuj Pareek, Guido Davidzon, Jin Long, Jared Dunnmon, and Matthew P. Lungren. Multi-task weak supervision enables anatomically-resolved abnormality detection in whole-body fdg-pet/ct. *Nature Communications*, 12(1):1880, 2021. 3

[14] Xiaotang Gai, Jiaxiang Liu, Yichen Li, Zijie Meng, Jian Wu, and Zuozhu Liu. 3d-rad: A comprehensive 3d radiology med-vqa dataset with multi-temporal analysis and diverse diagnostic tasks. *arXiv preprint arXiv:2506.11147*, 2025. Version v2 revised on 27 Oct 2025. 6

[15] Stefanie Gatidis and Tobias Kuestner. A whole-body fdg-pet/ct dataset with manually annotated tumor lesions (fdg-pet-

ct-lesions). *The Cancer Imaging Archive*, 2022. [Dataset]. 2, 3, 5, 1

[16] Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. Ct2rep: Automated radiology report generation for 3d medical imaging. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2024. 2

[17] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 2

[18] Hang Hua, Qing Liu, Lingzhi Zhang, Jing Shi, Zhifei Zhang, Yilin Wang, Jianming Zhang, and Jiebo Luo. Finecaption: Compositional image captioning focusing on wherever you want at any granularity, 2024. 3

[19] Shih-Cheng Huang, Kevin Lee, Kimberly Dodelzon, et al. Efficiency and quality of generative ai–assisted radiograph reporting. *JAMA Netw. Open*, 8(10):e2334943, 2025. 1

[20] Zachary Huemann, Samuel Church, Joshua D. Warner, Daniel Tran, Xin Tie, Alan B McMillan, Junjie Hu, Steve Y. Cho, Meghan Lubner, and Tyler J. Bradshaw. Vision-language modeling in pet/ct for visual grounding of positive findings, 2025. 3

[21] Stephanie L. Hyland, Shruthi Bannur, Kenza Bouzid, Daniel Coelho de Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, Noel Codella, Matthew P. Lungren, Maria Teodora Wetscherek, Ozan Oktay, and Javier Alvarez-Valle. Maira-1: A specialised large multimodal model for radiology report generation. Technical Report MSR-TR-2023-47, Microsoft Research, 2023. 2

[22] Walter Jentzen. An improved iterative thresholding method to delineate PET volumes using the delineation-averaged signal instead of the enclosed maximum signal. *J. Nucl. Med. Technol.*, 43(1):28–35, 2015. Epub 2015 Feb 5. 3

[23] Dhiraj Kalamkar, Dheevatsa Mudigere, Naveen Mellempudi, Dipankar Das, Kunal Banerjee, Sasikanth Avancha, Dharma Teja Vooturi, Nataraj Jammalamadaka, Jianyu Huang, Hector Yuen, Jiyan Yang, Jongsoo Park, Alexander Heinecke, Evangelos Georganas, Sudarshan Srinivasan, Abhisek Kundu, Misha Smelyanskiy, Bharat Kaul, and Pradeep Dubey. A study of bfloat16 for deep learning training, 2019. 2

[24] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *NeurIPS*, 2023. 2

[25] P. Li, S. Wang, T. Li, J. Lu, Y. HuangFu, and D. Wang. A large-scale ct and pet/ct dataset for lung cancer diagnosis (lung-pet-ct-dx) [data set]. The Cancer Imaging Archive (TCIA), 2020. 3

[26] Long Lian, Yifan Ding, Yunhao Ge, Sifei Liu, Hanzi Mao, Boyi Li, Marco Pavone, Ming-Yu Liu, Trevor Darrell, Adam Yala, and Yin Cui. Describe anything: Detailed localized image and video captioning. In *ICCV*, 2025. 3, 4

[27] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. Technical report, ACL Workshop on Text Summarization, 2004. 6

[28] P. Muzi, M. Wanner, and P. Kinahan. Data from rider lung pet-ct [data set]. The Cancer Imaging Archive (TCIA), 2015. 3

[29] National Library of Medicine. Nlm-scrubber: A hipaa compliant clinical text de-identification tool, 2025. Version current as of access date. 1

[30] 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, Quynh Anh Chau, Hong Son Mai, Thanh Trung Nguyen, and Phi Le Nguyen. Toward a vision-language foundation model for medical data: Multimodal dataset and benchmarks for vietnamese pet/ct report generation, 2025. 2, 3

[31] Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and Jean-Benoit Delbrouck. GREEN: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 374–390, Miami, Florida, USA, 2024. Association for Computational Linguistics. 6

[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. 6

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 8024–8035, 2019. 1

[34] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world, 2023. 2

[35] Maximilian Rokuss, Yannick Kirchhoff, Seval Akbal, Balint Kovacs, Saikat Roy, Constantin Ulrich, Tassilo Wald, Lukas T. Rotkopf, Heinz-Peter Schlemmer, and Klaus Maier-Hein. Lesionlocator: Zero-shot universal tumor segmentation and tracking in 3d whole-body imaging. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 30872–30885, 2025. 8

[36] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas

Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical report, 2025. 6

[37] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. In *CVPR*, pages 13019–13029, 2024. 3

[38] Xin Tie, Muheon Shin, Ali Pirasteh, Nevein Ibrahim, Zachary Huemann, Sharon M. Castellino, Kara M. Kelly, John Garrett, Junjie Hu, Steve Y. Cho, and Tyler J. Bradshaw. Personalized impression generation for pet reports using large language models. *Journal of Imaging Informatics in Medicine*, 37(2): 471–488, 2024. 2

[39] Xin Tie, Muheon Shin, Changhee Lee, Scott B. Perlman, Zachary Huemann, Amy J. Weisman, Sharon M. Castellino, Kara M. Kelly, Kathleen M. McCarten, Adina L. Alazraki, Junjie Hu, Steve Y. Cho, and Tyler J. Bradshaw. Automatic quantification of serial pet/ct images for pediatric hodgkin lymphoma using a longitudinally aware segmentation network. *Radiology: Artificial Intelligence*, 7(3):e240229, 2025. 8

[40] Martin Vallières, Emily Kay-Rivest, Léo Jean Perrin, Xavier Liem, Christophe Furstoss, Nader Khaouam, Phuc Félix Nguyen-Tan, Chang-Shu Wang, and Khalil Sultanem. Data from head-neck-pet-ct [data set]. The Cancer Imaging Archive (TCIA), 2017. 3

[41] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proc. IEEE CVPR*, pages 4566–4575, 2015. 6

[42] Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel T. Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5), 2023. 3

[43] Maria Werner-Wasik, Adam D. Nelson, Woo Choi, et al. What is the best way to contour lung tumors on pet scans? multiobserver validation of a gradient-based method using a nsclc digital pet phantom. *Int. J. Radiat. Oncol. Biol. Phys.*, 82(3):1164–1171, 2012. 2

[44] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, 2020. Association for Computational Linguistics. 1

[45] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *arXiv preprint arXiv:2308.02463*, 2023. 2

[46] Jing Wu, Yuli Wang, Zhusi Zhong, Weihua Liao, Natalia Trayanova, Zhicheng Jiao, and Harrison X. Bai. Vision-language foundation model for 3d medical imaging. *npj Artif. Intell.*, 1(17), 2025. Review. 1

[47] Yu Xin, Gorkem Can Ates, Kuang Gong, and Wei Shao. Med3dvlm: An efficient vision-language model for 3d medical image analysis. *IEEE Journal of Biomedical and Health Informatics*, PP(99):1–14, 2025. 6

[48] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 3, 6

[49] Fereshteh Yousefirizi, Ivan S. Klyuzhin, Joo Hyun O, Sara Harsini, Xin Tie, Isaac Shiri, Muheon Shin, Changhee Lee, Steve Y. Cho, Tyler J. Bradshaw, Habib Zaidi, François Bénard, Laurie H. Sehn, Kerry J. Savage, Christian Steidl, Carlos F. Uribe, and Arman Rahmim. Tmtv-net: fully automated total metabolic tumor volume segmentation in lymphoma pet/ct images – a multi-center generalizability analysis. *Eur. J. Nucl. Med. Mol. Imaging*, 51(7):1937–1954, 2024. 8

[50] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. In *NeurIPS*, 2021. 6

[51] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *CVPR*, pages 28202–28211, 2024. 3

[52] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. In *European Conference on Computer Vision*, pages 52–70. Springer, 2025. 2

[53] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *ICLR*, 2020. 6

[54] Yichi Zhang, Le Xue, Wenbo Zhang, Lanlan Li, Yuchen Liu, Chen Jiang, Yuan Cheng, and Yuan Qi. Seganypet: Universal promptable segmentation from positron emission tomography images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21107–21116, 2025. 3

[55] Yichi Zhang, Wenbo Zhang, Zehui Ling, Gang Feng, Sisi Peng, Deshu Chen, Yuchen Liu, Hongwei Zhang, Shuqi Wang, Lanlan Li, Limei Han, Yuan Cheng, Zixin Hu, Yuan Qi, and Le Xue. PET2Rep: Towards vision–language model–driven automated radiology report generation for positron

emission tomography. *arXiv preprint arXiv:2508.04062*, 2025. 2, 3, 6

[56] Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Ratescore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15004–15019, 2024. 6

[57] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 6

PETAR: Localized Findings Generation with Mask-Aware Vision-Language Modeling for PET Automated Reporting

Supplementary Material

Table 6. Cohort summary statistics.

Summary Statistics	Value (%)
Average Age (years)	62.2
Male	53.2
Female	46.9

Table 7. Indication frequency with low-prevalence categories grouped into "Others" (percent only).

Indication	Percent (%)
Restaging	39.7
Initial staging	29.6
Treatment response assessment	13.8
Metastatic workup	5.2
Suspected recurrence	3.2
Others	8.5

Table 8. Cancer type frequency with rare categories grouped into "Others" (percent only).

Cancer Type	Percent (%)
Lymphoma	22.8
Lung	22.5
Head and neck	11.5
Breast	7.6
Esophageal	4.4
Melanoma	4.2
Gynecologic (ovarian/cervical/endometrial)	3.6
Colorectal	3.5
Neuroendocrine tumor	3.3
Gynecologic cancer	3.1
Prostate	2.7
Others	20.8

8. Dataset information

8.1. Collection

PET/CT images were collected from patient scans conducted between 2010-2013, using scanners from GE Healthcare, at the University of Wisconsin-Madison hospital. Of the 5126 unique exams, 4798 are FDG, 178 are DOTATATE, 106 are [18F]Fluciclovine, and 44 are [18F]DCFPyL. Images were originally formatted in DICOM and converted into NIfTI. PET images were converted to units of SUV. Reports were formatted in csv. DICOM images were deidentified using Clinical Trial Processor, and reports were deidentified using NLM Scrubber [29].

The autoPET dataset is already deidentified, processed, and made available through the works of Gatidis *et al*. [15]. Note that the CT images for the AutoPET dataset were primarily contrast-enhanced CTs (i.e., iodine contrast), whereas our internal dataset consisted primarily of non-contrast-enhanced CTs.

8.2. Statistics

We collected statistics for our dataset regarding age, sex, and disease type. This is summarized in Table 6, Table 7, and Table 8.

8.3. Formatting

We used the following prompt for the LLM to structure the data in the format highlighted our dataset construction section.

"Given the following PET/CT finding:

{Findings}

Extract and format the information in this exact structure:

Region: Broad body section (e.g., Head, Neck, Chest, Abdomen, Pelvis etc.)

Organ: The nearest major organ or system involved (e.g., liver, pancreas, kidney, lung, bone etc.)

Anatomic Subsite: Specific substructure or precise location described (e.g., peripancreatic, left adrenal, posterior to left kidney, upper right gluteal etc.)

SUV Max: Numerical SUV max value mentioned

Axial Slice: Axial slice number mentioned

Findings: Copy the original sentence exactly as written Rules: If any field is not explicitly stated, write "N/A". Do not add interpretations beyond the sentence. Enclose the final output strictly between <extract> and </extract> tags."

9. Further implementation details

9.1. Training environment

We use the MONAI library [10] for 3D volumetric data processing, and PyTorch [33] together with the HuggingFace `transformers` and `accelerate` libraries [44] for model creation, optimization, and distributed training. All

Table 9. Hyperparameters used for LoRA-based fine-tuning of PETAR-4B

Hyperparameter	Value
Mixed precision	bf16
Model max length	512
Training epochs	5
Per-device batch size	2
Gradient accumulation	1
Learning rate	5e-5
Weight decay	0.0
Warmup ratio	0.03
LR scheduler	cosine
Gradient checkpointing	False
Evaluation strategy	steps
Evaluation steps	0.2
Eval accumulation steps	1
Dataloader workers	8
Pin memory	True

Table 10. Comparison of interpretation options for patient 1.

Patient ID	ROI ID	Preference	Description
Patient 1	Mask1	Option 1	There is a small, intensely FDG avid focus in the right iliac fossa with SUV max of [#] (PET/CT axial slice [#]) that is not seen on CT.
Patient 1	Mask1	Option 2	Intensely radiotracer avid nodular focus ([#] x [#] cm) in close proximity to the surgical clips, measuring SUV max of [#] correlates with the recently CT characterized nodularity and is most suspicious for recurrent disease.

experiments are launched using `accelerate launch` and leverage LoRA-based parameter-efficient fine-tuning for language models [17]. Mixed-precision training was enabled using `bfloat16` [23].

9.2. Hyperparameters

We fine-tuned the combination of our modified image encoder and the language component of M3D [2] using LoRA adapters applied to all linear layers in the language backbone. Evaluation and checkpointing are performed at fixed step intervals. All hyperparameters used during training are listed in Table 9.

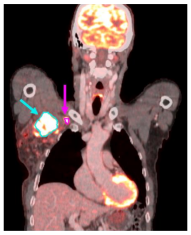
PET/CT Fusion	PETAR
	•There is an intensely hypermetabolic right supraclavicular lymph node (slice [#], SUV max [#]) that measures [#] x [#] cm. •There is a [#] x [#] cm left subpectoral lymph node with intense FDG uptake (PET/CT axial slice [#], SUV max [#]).
	•There is a new right axillary lymph node with intense FDG uptake (PET/CT axial slice [#], SUV max [#]). •Intensely FDG avid right supraclavicular nodal conglomerate/lymph mass with SUV max of [#] (PET/CT axial slice [#]).

Figure 5. PETAR-4B predictions for examples from the autoPET dataset.

Table 11. Anatomical accuracy rubric.

Score	Description
1	Completely incorrect. Wrong organ/system, wrong side, or irrelevant location.
2	Mostly incorrect. Correct general region but wrong substructure.
3	Partially correct. Correct organ/region but poor extent or boundary mismatch.
4	Nearly correct. Correct location with minor boundary issues.
5	Completely correct. Precise location and extent.

10. AutoPET examples

In Figure 5, we show examples of findings produced by PETAR-4B on the autoPET data.

11. Human evaluation details

11.1. Evaluation dataset preparations

We provided physicians with DICOM PET/CT images, with lesion contour overlays, in MIM Encore software. In the provided text, we replaced all numerics to [#] as these are hallucinated and can easily be extracted from the input contour. For every pair of ground truth and prediction, we shuffled the order for all examples to mitigate any bias in scoring. An example of what the physicians saw for scoring is shown in Table 10.

11.2. Guidance on scoring

We reproduce the scoring guidelines provided to the physicians below in Table 11, Table 12, Table 13. As the autoPET

Table 12. Interpretation accuracy rubric.

Score	Description
1	Completely incorrect; misleading or dangerous.
2	Mostly incorrect; plausible but clinically misleading.
3	Partially correct; right lesion but wrong qualifier.
4	Nearly correct; accurate meaning but missing detail.
5	Completely correct; matches ground truth interpretation.

Table 13. Overall utility rubric.

Score	Description
1	No utility; unusable or harmful.
2	Minimal utility; major edits required.
3	Moderate utility; usable after moderate edits.
4	High utility; only minor edits required.
5	Fully useful; clinically usable as written.

dataset does not have corresponding reports, we do not report preference scores or automated metrics as these require a reference ground truth.