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

Supplementary Material

Table 1. Cohort summary statistics.

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

Table 2. 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 3. 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

1. Dataset information

1.1. Collection

PET/CT images were collected from patient scans conducted between 2010-2013 using scanners from GE Healthcare. 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 [6].

The autoPET dataset is already deidentified, processed, and made available through the works of Gatidis *et al*. [3]. 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.

1.2. Statistics

We collected statistics for our dataset regarding age, sex, and disease type. This is summarized in Table 1, Table 2, and Table 3.

1.3. 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.

Begin:"

2. Further implementation details

2.1. Training environment

We use the MONAI library [2] for 3D volumetric data processing, and PyTorch [7] together with the HuggingFace `transformers` and `accelerate` libraries [8] for model creation, optimization, and distributed training. All

Table 4. 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 5. 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 [4]. Mixed-precision training was enabled using `bfloat16` [5].

2.2. Hyperparameters

We fine-tuned the combination of our modified image encoder and the language component of M3D [1] 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 4.

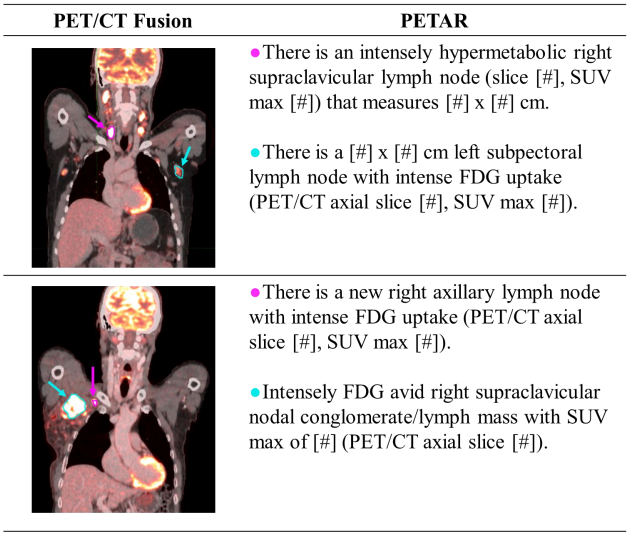

Figure 1. PETAR-4B predictions for examples from the autoPET dataset.

Table 6. 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.

3. AutoPET examples

In Figure 1, we show examples of findings produced by PETAR-4B on the autoPET data.

4. Human evaluation details

4.1. 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 5.

4.2. Guidance on scoring

We reproduce the scoring guidelines provided to the physicians below in Table 6, Table 7, Table 8.

Table 7. 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 8. 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.

References

[1] 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

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

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

[4] 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 *International Conference on Learning Representations*, 2022. 2

[5] 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. *arXiv preprint arXiv:1905.12322*, 2019. 2

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

[7] 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

[8] 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