Semantically Consistent Whole-Body PET/CT Report Generation with Local Lesion-Level Guidance

Mingyang Yu[1], Yaozong Gao[2], Yiran Shu[1,2], Jingyu Liu[2], Jiaming Wu[5], Kaicong Sun[1], Shuwei Bai[1], Yanbo Chen[2], Xiang Zhou[2], Yiqiang Zhan[2], Weifang Zhang[3], Shaonan Zhong[4], Xinlu Wang[4], Meixin Zhao[3(✉)], and Dinggang Shen[1,2,6(✉)]

[1] School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai 201210, China
Dinggang.Shen@gmail.com
[2] Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China
[3] Department of Nuclear Medicine, Peking University Third Hospital, Haidian District Beijing 100191, China
zhaomeixin.student@sina.com
[4] Department of Nuclear Medicine, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou 510000, China
[5] Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong SAR, China
[6] Shanghai Clinical Research and Trial Center, Shanghai 201210, China

Abstract. Automated report generation for whole-body PET/CT imaging can substantially reduce radiologists' workload and improve clinical efficiency. Existing report generation approaches typically maps the entire volumetric scan to a complete report. However, clinically meaningful findings in whole-body PET/CT are sparse and localized. Such global end-to-end modeling tends to omit lesion-specific semantic cues, limiting complete descriptions of lesions. To address this challenge, we propose a whole-body PET/CT report generation framework guided by local lesion-level cues. Our method is built upon three key strategies. First, we propose an aggregation module that compresses variable lesion candidates in PET/CT, enabling many-to-one lesion–description alignment that reflects real clinical reporting practice. Second, we propose local entity alignment and global relation alignment during pretraining to enforce cross-modal semantic consistency between report and image. Third, we use a two-stage report generation training pipeline, namely pretraining on structured reports and fine-tuning on free-text radiology reports, to reduce template bias. We have collected **3,015** whole-body PET/CT cases from **three** medical centers, including **120,549** annotated local lesions. Experimental results demonstrate that the proposed framework improves performance under clinically reliable evaluations of generated reports.

Keywords: Whole-body PET/CT · Report generation · Vision–language pretraining.

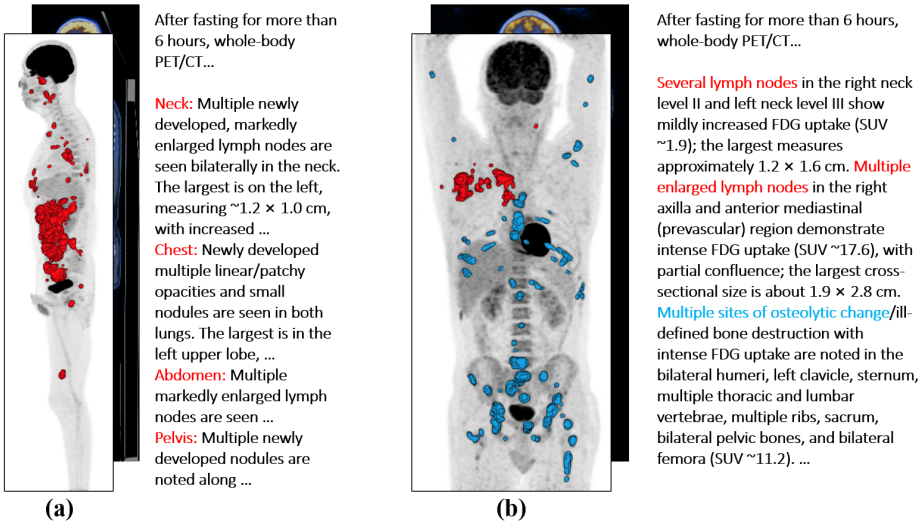

Fig. 1. Radiologists' reporting style in whole-body PET/CT often follows a many-to-one pattern, where multiple lesions are summarized in a single statement. Such summarization is commonly organized by anatomical regions (left, a), lesion types (right, b), or a hybrid of both.

1 Introduction

Automated report generation [16,12] has been extensively studied in recent years. However, most existing efforts focus on specific body parts, such as chest radiographs [10] and breast ultrasound [2], with limited anatomical field-of-view coverage, and the report content is largely driven by localized findings. In contrast, whole-body PET/CT spans diverse anatomical regions across the entire body, making report generation substantially more challenging due to more complex volumetric input and sparse localized abnormalities. Existing works [4,9] attempt to interpret whole-body PET/CT volumes using large models. However, this strategy is misaligned with clinical focus and computationally impractical. Moreover, the scarcity of large-scale paired whole-body PET/CT report datasets further limits effective training [9].

In contrast, studies [8,14] have explored mask-guided, local lesion-level captioning to produce more clinically useful lesion descriptions. In particular, PETAR [8] focuses on local finding generation with a relatively direct lesion–description correspondence, while PETRG [4] directly models the whole PET/CT volume for report generation. These studies demonstrate the potential of lesion-aware or full-scan PET/CT reporting, but local captioning does not fully address complete report generation, and whole-volume modeling may dilute small clinically relevant lesions within the large whole-body background. And other studies [1,13] show that structured reports can provide more efficient supervision and evaluation than free-text. However, transferring these ideas to whole-body, full-scan

PET/CT report generation remains challenging. First, whole-body PET/CT scans often contain multiple metastatic lesions, so the report rarely follows a one-to-one image–text mapping. Instead, it frequently exhibits a many-to-one lesion–description pattern, where multiple lesions are summarized in a single statement. In practice (Fig. 1), such summarization is often organized by lesion types (e.g., bone lesions), anatomical regions (e.g., lymph nodes or soft-tissue masses), or a combination of both, reflecting reporting styles and clinical indications. Second, such report generation models rely heavily on a visual encoder with strong semantic understanding. Prior research often adopts CLIP-style [11] vision–language pretraining for image–text alignment. However, these mechanisms are typically designed for one-to-one matching and do not explicitly model the many-to-one alignment common in whole-body PET/CT reports. Finally, free-text reports, especially those from single-center cohorts, often contain substantial template-driven content. During early training, models may overfit these frequent patterns, which can hinder learning sparse yet clinically critical lesion-specific cues.

To address these challenges, we propose a local lesion-level guidance framework to generate whole-body PET/CT reports with guidance from specific local lesions. Specifically, our contributions are threefold. First, a SetTransformer-based aggregation module [5] is used to compress variable lesion candidates into a fixed-length entity set, enabling many-to-one lesion-description alignment that reflects real clinical reporting practice. Second, CLIP-style pretraining is built on one-to-one image–text matching and treats all other pairs in the minibatch as negatives, implicitly assuming that different reports (and different images) are unrelated. However, PET/CT images and reports are less diverse than natural images and texts and often share substantial anatomy- and template-driven content, making this fully discriminative objective suboptimal for PET/CT image–text alignment. To address this, we introduce local entity alignment and global relation alignment during pretraining. Local entity alignment preserves paired image-text alignment, while global relation alignment replaces the all-negatives assumption and encourages the image embedding space to align with the text embedding space. Third, to mitigate overfitting to template-driven contents, we adopt a two-stage report generation training pipeline, first training on structured positive findings to emphasize clinically salient cues, and then fine-tuning on free-text reports to reduce template bias.

2 Method

Our method extends the LLaVA [7] framework. As shown in Fig. 2b, the model takes a varying number of lesion-guided PET/CT volume patches, paired with lesion attributes (e.g., SUVmax and lesion size), as input to generate the final findings report. Specifically, given the grouped 3D lesion patch candidates (e.g., by anatomical region), patches within each group are first encoded into lesion features and then aggregated per group into a fixed-length set of visual entities. Subsequently, the resulting entity sets are projected into visual entity

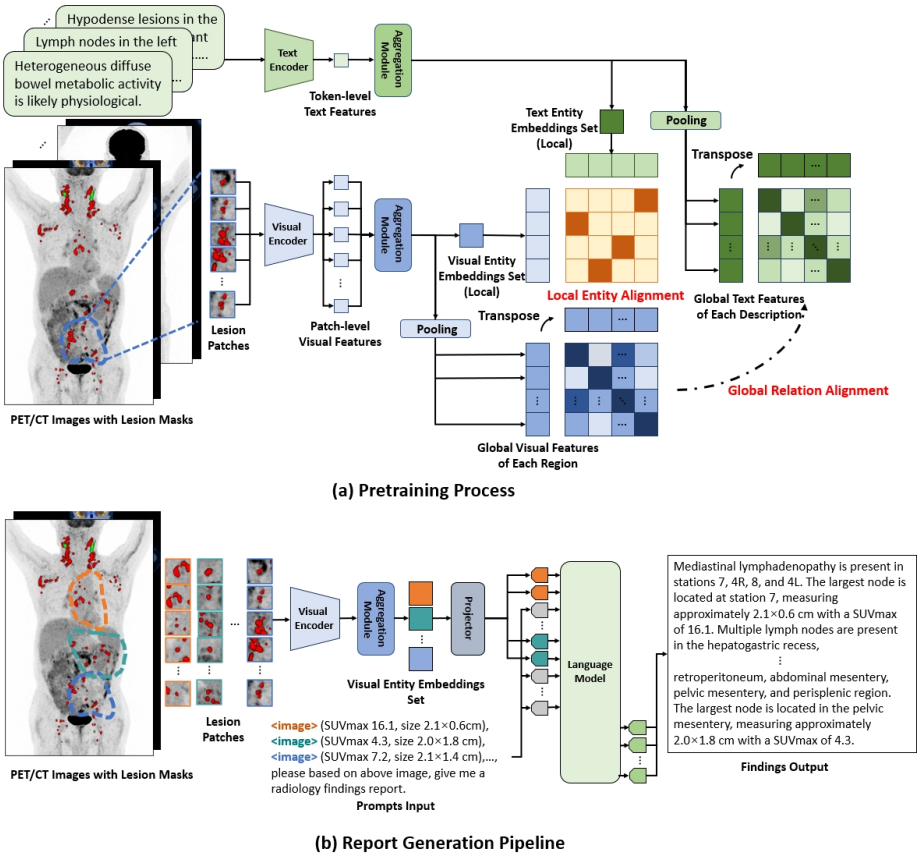

Fig. 2. Overview of the pretraining process and report generation pipeline. The pretraining pipeline (a, top) uses local entity alignment and global relation alignment to enforce semantic consistency both cross-modally (image–text) and within each modality. The model pipeline (b, bottom) shows how local lesion cues are encoded and aggregated, then combined with prompt embeddings to generate the final whole-body PET/CT report. All lesion volume patches are cropped as $64 \times 64 \times 64$ voxels with an isotropic spacing of 1.5 mm.

embeddings, fused with lesion-attribute prompt embeddings, and then fed to the language model to generate the whole-body PET/CT findings report.

2.1 SetTransformer-based Aggregation Module

As illustrated in Fig. 1, whole-body PET/CT studies often contain a large and highly varying number of lesion ROIs, particularly in multifocal disease. In some patients, the scan may include hundreds of tracer-avid metastatic lesions. Directly feeding all lesion features into the LLM is inefficient and largely redundant, since many ROIs share similar uptake patterns and overlapping findings within

the same anatomical region. In clinical practice, the corresponding report is typically concise, summarizing many lesion ROIs into a statement by lesion type or region, which naturally forms a many-to-one lesion-description relationship.

To better accommodate the many-to-one lesion–description relationship in whole-body PET/CT report generation, we introduce a SetTransformer-based aggregation module [5] into our framework. The SetTransformer uses attention-based set encoding to compress a variable-sized ROI set into a fixed number K of visual entities, and it is order-insensitive, making it naturally suitable for tasks that only care about lesion content rather than the input order. Accordingly, if all lesion ROIs in a case can be organized into n groups (e.g., by anatomical region), corresponding to n key summarization statements, then the ROIs can be compressed into $n \times K$ visual entities. This compact representation is designed to retain sufficient information for report generation and reflects the many-to-one nature of clinical reporting practice.

2.2 Local Entity Alignment and Global Relation Alignment

CLIP-style pretraining is widely used in vision–language learning. It uses a one-hot contrastive target. For each image, only the paired text is treated as positive, while all other texts are treated as negatives (and vice versa). This assumption is improper for whole-body PET/CT because PET/CT images and reports are less diverse than natural counterparts, and they often share substantial anatomy- and template-driven content, making the discriminative objective suboptimal for PET/CT image–text alignment.

To better support PET/CT alignment and our many-to-one lesion–description setting, we introduce two complementary objectives during the pretraining of the vision encoder and the aggregation module: local entity alignment and global relation alignment (Fig. 2a). For local entity alignment, we retain the paired positive supervision and model many-to-one correspondence by compressing paired lesion volumetric patches and their descriptions into fixed-length entity sets of size K. This design is motivated by the fact that a single report statement may mention multiple localized concepts, e.g., "Several lymph nodes are present in the right level II and left level III cervical nodal stations." which contains both "right neck level II" and "left neck level III". We then align K visual entities with K textual entities via one-to-one matching, so that each textual entity corresponds to a specific visual entity. For global relation alignment, we replace the absolute "all negatives" assumption with cross-sample relationships. We pool each entity set into a single embedding, build pair-to-pair similarity matrices in both modalities, and minimize their discrepancy so that visual representations align with the textual semantic space. The total loss is formulated as $\mathcal{L} = \mathcal{L}_{\text{align}} + \lambda_{\text{r}}\mathcal{L}_{\text{rel}}$, where $\lambda_{\text{r}} = 1/B$ balances the batch-level global relation loss with the local entity alignment loss; in our experiments, $B = 32$.

Local Entity Alignment Given a mini-batch of B paired samples, we represent the b-th pair by K ℓ_2-normalized visual entities $E_b = \{e_{b,i}\}_{i=1}^{K}$ and K

ℓ_2-normalized textual entities $T_b = \{t_{b,j}\}_{j=1}^{K}$, where $b \in \{1, \ldots, B\}$. We compute the similarity matrix $S_b \in \mathbb{R}^{K \times K}$, obtain a one-to-one assignment π_b via Hungarian matching, and apply a temperature-scaled cross-entropy loss:

$$\mathcal{L}_{\text{align}} = -\frac{1}{BK} \sum_{b=1}^{B} \sum_{i=1}^{K} \log \frac{\exp\big(S_b(i, \pi_b(i))/\tau\big)}{\sum_{j=1}^{K} \exp\big(S_b(i,j)/\tau\big)}, \tag{1}$$

$$S_b(i,j) = e_{b,i}^{\top} t_{b,j}, \quad \pi_b = \text{Hungarian}(-S_b). \tag{2}$$

Global Relation Alignment To align cross-sample semantic geometry across modalities, we first form sample-level embeddings by mean pooling the K entities: $u_b = \frac{1}{K}\sum_{i=1}^{K} e_{b,i}$ and $v_b = \frac{1}{K}\sum_{j=1}^{K} t_{b,j}$. We stack them as $U = [u_1; \ldots; u_B]$ and $V = [v_1; \ldots; v_B]$, compute relation distributions with temperature τ, and minimize a symmetric KL divergence:

$$P_I = \text{softmax}\big(UU^{\top}/\tau\big)\,, P_T = \text{softmax}\big(VV^{\top}/\tau\big)\,, \tag{3}$$

$$\mathcal{L}_{\text{rel}} = \frac{1}{2}\Big(\text{KL}(P_T \| P_I) + \text{KL}(P_I \| P_T)\Big). \tag{4}$$

2.3 Two-stage Report Generation Training Pipeline

Free-text whole-body PET/CT reports often include many repeated template sentences (e.g., "No abnormal radiotracer uptake ..."). These sentences can dominate training, so the model learns common but low-information wording instead of rare yet clinically important abnormalities.

To address this issue, we employ a two-stage training pipeline. In Stage one, we train on structured reports that contain **positive lesions only**, covering **471** anatomical locations and **61** lesion-type categories, to keep the model focus on lesion-relevant cues and avoid learning template bias and normal findings. In Stage two, we then finetune this model on full free-text reports that include both positive and negative statements to learn realistic narrative style and institution-specific reporting conventions. These structured reports in Stage one are derived from structured labels rather than written as free text, i.e., they are constructed from each study via an automated lesion-centric pipeline. Specifically, we first apply a lesion segmentation model to extract 3D lesion volume patches. Then we use a set of lesion classification models to assign each lesion category (e.g., anatomical location, lesion type) together with computed quantitative attributes (e.g., SUVmax and lesion size). All results are manually reviewed by radiologists and mapped to corresponding sentence in free-text report to ensure consistency between structured report and free-text report.

3 Experiments and Results

3.1 Dataset

Our dataset is collected from three medical centers and including 3,015 whole-body PET/CT studies. We have identified 120,549 abnormal regions. Among

which, 83,395 lesions are aligned with report-described content. The dataset contains **2,408** FDG studies and **607** PSMA studies. We use 2,811 studies for training and 204 studies for testing. We obtain lesion regions using an nnU-Net-based [3] lesion segmentation model. Each lesion has structured attributes, including anatomical location, CT findings, lesion size, and SUVmax. The structured labels cover 471 anatomical locations and 61 CT-finding categories. We normalize PET and CT images to $[0, 1]$ and apply standard data augmentation such as random rotation and scaling.

3.2 Training Details

Training was conducted on 8 NVIDIA H20 GPUs. During the pretraining stage, the CNN-based visual encoder was implemented as a 3D ResNet-50, and the Transformer-based visual encoder was implemented as a ViT-3D. The visual encoders were initialized with Kaiming initialization. Clinically similar lesion candidates were grouped by anatomical region or lesion category before SetTransformer aggregation. We set the number of SetTransformer output entities to $K = 8$ to balance information preservation and computational efficiency. The temperature was set to $\tau = 0.1$. The batch size was $B = 32$, and $\lambda_{\mathrm{r}} = 1/32$. The model was optimized using AdamW with a learning rate of 1×10^{-4}. During report generation fine-tuning, the model was initialized with the pretrained vision and text encoders. The vision encoder was frozen, and the language model was fine-tuned using LoRA adapters.

3.3 Comparison Study

We evaluate our method on our two-stage report generation training curriculum: 1) structured reports and 2) free-text reports. Structured reports contain positive lesion findings only, enabling a focused assessment of lesion-level semantic fidelity without interference from template-driven content. Free-text reports include both positive and negative statements and better reflect real-world reporting style. We use NLG (Natural Language Generation) metrics and CE (Clinical Efficacy) metrics as evaluation metrics. For NLG, we report BLEU and ROUGE-L (F1) scores computed on the Chinese report text. For CE metrics, we employ an LLM-based evaluator (ChatGPT-5.2) that matches each ground-truth (GT) statement to the most similar generated statement, since clinically correct findings may be expressed with different wording (e.g., "left lung" vs. "left upper lobe"). The evaluator outputs four accuracy scores, A-all, A-pos, F-all, and F-pos, which comprehensively quantify clinical efficacy by separately assessing anatomical and finding correctness on both positive and negative descriptions, highlighting the advantage of our method.

In Table 1, we can see that our method consistently outperforms LLaVA-Med [6] and PET2Rep [15] in both NLG and CE metrics, demonstrating clear advantages in generating clinically faithful whole-body PET/CT reports. Here,

Table 1. Comparison study results on structured and free-text reports. All values are reported in $[0, 1]$. † denotes the language model that takes structured reports as input. B-3/B-4: BLEU-3/BLEU-4; R-L: ROUGE-L (F1). A-all/F-all: overall Anatomy/Finding accuracy (positive and negative descriptions); A-pos/F-pos: accuracy on positive lesions only. In PET2Rep [15], "Sep" and "Fus" denote the two PET/CT input strategies, "Separate" and "Fused," respectively. Best results are in bold. * indicates a statistically significant improvement of our method over all competing baselines (excluding our variants) on the CE metrics under the same setting ($p < 0.05$).

Method	LLMs/VLMs	NLG Metrics			**CE Metrics**			
		B-3	B-4	R-L	A-all	**A-pos**	F-all	**F-pos**
Structured Report								
LLaVA-Med	Llama-7B (ViT)	0.507	0.488	0.765	-	0.286	-	0.765
Ours	Llama-7B (CNN)	0.706	0.681	0.957	-	0.499	-	0.852
	Llama-13B (CNN)	0.719	0.694	0.962	-	0.502	-	0.855
	Llama-7B (ViT)	0.785	0.758	0.981	-	0.519	-	0.868
	Llama-13B (ViT)	**0.796**	**0.769**	**0.985**	-	**0.522**	-	**0.877**
Free-text Report								
S-to-F	QWen3-8B †	0.314	0.269	**0.316**	0.499	0.520	0.810	0.654
LLaVA-Med	Llama-13B (ViT)	0.143	0.105	0.273	0.296	0.300	0.757	0.608
PET2Rep (Sep)	Qwen3-VL-235B	0.201	0.179	0.255	0.491	0.413	0.695	0.635
PET2Rep (Fus)	Qwen3-VL-235B	0.216	0.196	0.258	0.504	0.420	0.695	0.521
Ours	Llama-13B (CNN)	0.274	0.232	0.299	0.646	0.571	0.720	0.611
	Llama-13B (ViT)	**0.326**	**0.282**	0.309	**0.649***	**0.594***	**0.817***	**0.670***

CNN/ViT denote 3D-CNN/3D-ViT vision encoders respectively, and LLaMA-7B/13B denote LLaMA-Chinese-7B/13B. It is shown that increasing the language model size from 7B to 13B further improves performance, while the choice of visual encoder has an even larger impact: 3D-ViT consistently achieves higher CE scores than 3D-CNN, suggesting that ViT-style visual representations are more compatible with LLM-based parsing and generation for whole-body PET/CT reporting. Finally, S-to-F denotes a structured-to-free strategy that uses structured-report text to predict free-text reports. The results indicate that structured reports provide helpful cues, but their information content remains limited.

3.4 Ablation Study

We conduct ablation study to validate the effectiveness of (i) local entity alignment with global relation alignment (*Alignment*), and (ii) the two-stage report generation training pipeline (*Structured Report*). The baseline is trained end-to-end on free-text reports using the same architecture as ours. As shown in Table 2, adding *Alignment* consistently improves performance, indicating that local entity alignment and global relation alignment benefits the report generation task. Meanwhile, the *Structured Report* yields a larger gain than using Alignment alone, and combining them achieves the best results. This suggests

Table 2. Ablation results for free-text report generation under CE metrics (all values in $[0, 1]$). All experiments use LLaMA-13B (ViT). CE metrics are defined the same as in the comparison study.

Method	CE Metrics			
	A-all	A-pos	F-all	F-pos
baseline	0.602	0.548	0.794	0.556
+*Alignment*	0.614	0.560	0.773	0.603
+*Structured Report*	0.630	0.581	0.809	0.651
+*Alignment* +*Structured Report*	**0.649**	**0.594**	**0.817**	**0.670**

that pretraining on structured, positive-only reports is a key factor for improving clinical report generation, as it helps the model focus on lesion-relevant semantics before learning free-text report templates.

4 Conclusion

In this paper, we present a local lesion-level guidance framework for whole-body PET/CT report generation. Inspired by the sparsity of clinically meaningful findings and the common many-to-one lesion-description pattern in clinical practice, we propose (1) a SetTransformer-based aggregation module to compress variable lesion candidates into a fixed entity set for many-to-one alignment, (2) local entity alignment and global relation alignment to enable image-text pretraining, and (3) a two-stage report generation training pipeline (structured reports to free-text reports) to mitigate template-driven bias. We also build a multi-center dataset containing **3,015** whole-body PET/CT image-report pairs from **three** institutions, with **120,549** annotated lesions. Experimental results demonstrate that the proposed framework improves performance under clinically reliable evaluations of generated reports.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China (grant numbers 82441023, U23A20295, 62131015, 82394432), the China Ministry of Science and Technology (S20240085, STI2030-Major Projects-2022ZD0209000, STI2030-Major Projects-2022ZD0213100), Shanghai Municipal Central Guided Local Science and Technology Development Fund (No. YDZX20233100001001), The Key R&D Program of Guangdong Province, China (grant number 2023B0303040001), HPC Platform of ShanghaiTech University, the special fund of Beijing Clinical Key Specialty Construction Program, P. R. China (2022), China Medical Health Development Foundation, Peking University Third Hospital and United-Imaging Research Institution Intelligential Imaging Joint Research & Development Center Foundation, and Key Clinical Project of Peking University Third Hospital (BYSYZD2019038 and BYSYZD2023016). We also acknowledge the Department of Nuclear Medicine, The First Affiliated Hospital of Guangzhou Medical University, for providing validation data.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chen, K., Xu, W., Li, X.: The Potential of Gemini and GPTs for Structured Report Generation based on Free-Text 18F-FDG PET/CT Breast Cancer Reports. Academic Radiology **32**(2), 624–633 (2025)
2. Huh, J., Park, H.J., Ye, J.C.: Breast Ultrasound Report Generation using LangChain. arXiv preprint arXiv:2312.03013 (2023)
3. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
4. Jiao, W., Shang, K., Li, H., Yan, K., Zhang, J., Yang, G., Guo, L., Wan, Y., Yang, X., Jin, D., et al.: Vision-Language Models for Automated 3D PET/CT Report Generation. arXiv preprint arXiv:2511.20145 (2025)
5. Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., Teh, Y.W.: Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks. In: International conference on machine learning. pp. 3744–3753. PMLR (2019)
6. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
7. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. Advances in Neural Information Processing Systems **36**, 34892–34916 (2023)
8. Maqbool, D., Lee, C., Huemann, Z., Church, S.D., Larson, M.E., Perlman, S.B., Romero, T.A., Warner, J.D., Lubner, M., Tie, X., et al.: PETAR: Localized Findings Generation with Mask-Aware Vision-Language Modeling for PET Automated Reporting. arXiv preprint arXiv:2510.27680 (2025)
9. Nguyen, H.T., Nguyen, D.T., Nguyen, T.M.D., Nguyen, T.T., Truong, T.N., Pham, H.H., Barthelemy, J., Tran, M.Q., Nguyen, T.T., Nguyen, Q.V.H., et al.: Toward a Vision-Language Foundation Model for Medical Data: Multimodal Dataset and Benchmarks for Vietnamese PET/CT Report Generation. arXiv preprint arXiv:2509.24739 (2025)
10. Ouis, M.Y., Akhloufi, M.A.: Deep learning for report generation on chest X-ray images. Computerized Medical Imaging and Graphics **111**, 102320 (2024)
11. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning Transferable Visual Models From Natural Language Supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
12. Wang, S., Zhao, Z., Ouyang, X., Liu, T., Wang, Q., Shen, D.: Interactive computer-aided diagnosis on medical image using large language models. Communications Engineering **3**(1), 133 (2024)
13. Wray, R., Yeh, R.: Multimodal Large Language Model Based PET/CT Report Generation: Capabilities and Comparison of ChatGPT-4 Versus ChatGPT-3.5. Journal of Nuclear Medicine **65**(supplement 2), 242447 (2024)
14. Yu, M., Gao, Y., Shu, Y., Chen, Y., Liu, J., Jiang, C., Sun, K., Cui, Z., Zhang, W., Zhan, Y., et al.: Location-Guided Automated Lesion Captioning in Whole-Body PET/CT Images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 348–357. Springer (2025)
15. Zhang, Y., Zhang, W., Ling, Z., Feng, G., Peng, S., Chen, D., Liu, Y., Zhang, H., Wang, S., Li, L., et al.: PET2Rep: Towards Vision-Language Model-Drived Automated Radiology Report Generation for Positron Emission Tomography. arXiv preprint arXiv:2508.04062 (2025)

16. Zhong, T., Zhao, W., Zhang, Y., Pan, Y., Dong, P., Jiang, Z., Kui, X., Shang, Y., Yang, L., Wei, Y., et al.: ChatRadio-Valuer: A Chat Large Language Model for Generalizable Radiology Report Generation Based on Multi-institution and Multi-system Data. arXiv preprint arXiv:2310.05242 (2023)