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

Supplementary Material

A. Dataset Details

A.1. Data Collection

PETRG-Lym Acquisition Protocol: Patients fasted for at least 6 hours prior to the examination. Venous access was established, and blood glucose levels were measured to ensure protocol compliance (< 11.1 mmol/L). Following the intravenous injection of the ^{18}F-FDG radiotracer, patients rested in a supine position for approximately 40 minutes. Patients were instructed to void their bladders before scanning to minimize pelvic artifacts. Whole-body (WB) imaging was performed with the patient supine and immobilized, with a typical scan duration of 22 minutes. Attenuation correction (AC) was applied to PET images using the corresponding CT data. Data were acquired using four scanner models: GE Discovery STE, SIEMENS Biograph Vision 450, UNITED IMAGING uMI 510, and uMI Panorama 35S. Diagnostic reports, stored in PDF and DOCX formats, were rigorously reviewed by senior nuclear medicine physicians. This study received ethical approval from the Institutional Review Boards (IRB) of all four participating medical centers.

A.2. Preprocessing Details

Transforming raw clinical data into model-ready inputs involves a complex pipeline to ensure multi-modal alignment and quantitative consistency. Figure 1 systematically illustrates our workflow.

A.2.1. PET/CT Image Preprocessing

We implemented a streamlined pipeline to standardize the visual input:

- **De-identification:** All Protected Health Information (PHI) was stripped from DICOM headers and file paths.
- **Meta-Information Extraction:** Critical metadata—including patient weight, injected dose, acquisition time, and rescale factors—were parsed for quantitative calculations.
- **Format Conversion:** Raw DICOM series were converted to NIfTI format for efficient I/O during training.
- **Spatial Normalization and Alignment:** Raw PET and CT images often possess differing resolutions and fields of view. To achieve pixel-level alignment, we resampled both modalities to a unified voxel spacing of $1.5 \times 1.5 \times 3$ mm. Both volumes were reoriented to the standard RAS (Right-Anterior-Superior) coordinate system, ensuring strictly aligned multi-modal inputs.
- **SUV Calculation:** We converted raw PET intensity to Standardized Uptake Values (SUV) to mitigate inter-subject variability:

$$\mathrm{SUV}(t) = \frac{c_{img}(t) \cdot \mathrm{BW}}{\mathrm{ID}} \tag{1}$$

where $c_{img}(t)$ is the decay-corrected radioactivity concentration (Bq/mL) at scan time t, BW is the body weight (g), and ID is the injected dose (Bq) corrected to the injection time.

- **CT Intensity Normalization:** CT voxel values were converted to Hounsfield Units (HU) using the rescale slope and intercept [4]. We clipped the intensities to the range $[-1000, 1000]$ HU to cover the full dynamic range from lung tissue to bone, followed by min-max normalization to $[0, 1]$.
- **Background Removal:** To eliminate irrelevant background noise (e.g., scanning bed), we utilized TotalSegmentator [14] to generate a precise body mask. The volume was cropped to the minimal bounding box of the body with a 10-slice safety margin.
- **Anatomical Focus (ROI Cropping):** Clinical scan ranges vary (e.g., "head-to-toe" vs. "head-to-mid-thigh"). Since lymphoma reporting focuses primarily on the trunk, we standardized the field of view by cropping images to the upper thigh level. Specifically, we retained the body trunk and an extension of 20% below the pelvic floor ($\leq$ 50 slices) to ensure complete coverage of the inguinal region while discarding irrelevant lower limb data.

A.2.2. Report Text Preprocessing

The textual data underwent a four-step cleaning process:

- **De-identification:** PHI was strictly removed from both raw content and metadata.
- **Structure Parsing:** Unstructured reports were parsed into JSON format, extracting key fields: Gender, Clinical History, Findings, and Impression.
- **Text Normalization:** Redundant whitespace (e.g., sequences exceeding two spaces) was removed to compact the sequence length.
- **Token Standardization:** Special characters were replaced to prevent tokenizer failures (e.g., Roman numerals were converted to Arabic numerals to unify representation while preserving standard staging semantics where possible.), ensuring compatibility with the LLM vocabulary.

B. Implementation Details

B.1. Implementation of PETRG-3D

Model Architecture. For the PETRG-3D framework, we utilize the ViT3D pre-trained in [15] as the volume encoder.

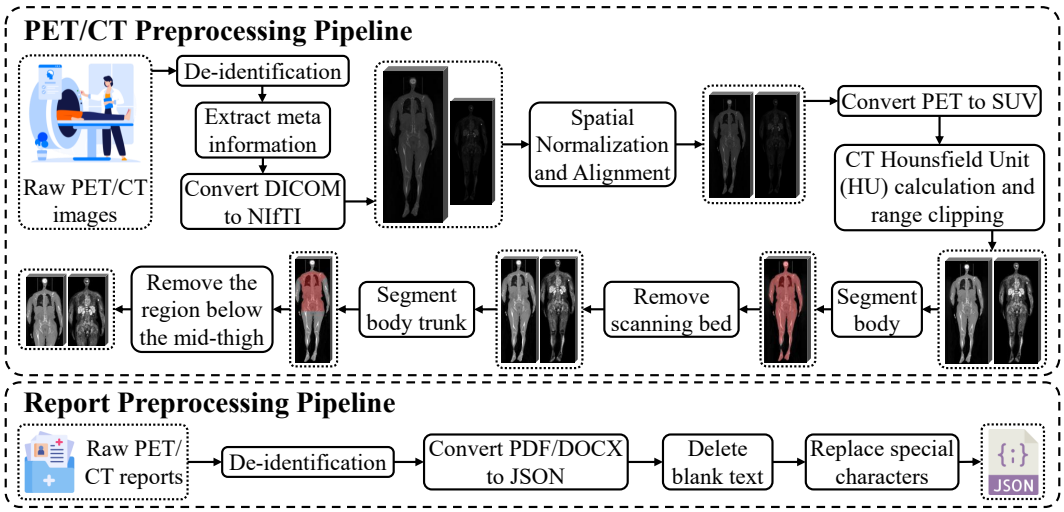


Figure 1. Preprocessing pipeline for PET/CT images and reports.

We fine-tune the subsequent Perceiver Resampler, which utilizes 128 learnable latent queries to compress the volumetric features into a fixed sequence of 128 visual tokens (with a hidden dimension of 768). These visual tokens are then projected via a linear layer to align with the embedding dimension of the LLM (i.e., 4096). For the text decoder, we benchmark six state-of-the-art Large Language Models (LLMs) with approximately 8 billion parameters: Llama2-7B[13], Mistral-7B-v0.3[1], Qwen2.5-7B[10], Gemma2-9B[12], GLM4-9B[5], and Qwen3-8B[17]. Given that our target reports are in Chinese, we employ the official pre-trained weights for models with native Chinese support (i.e., Qwen2.5-7B[1], GLM4-9B[2], and Qwen3-8B[3]). Conversely, for Llama2-7B[4], Mistral-7B-v0.3[5], and Gemma2-9B[6], we utilize versions that have been fine-tuned on Chinese corpora to ensure linguistic capability.

Training and Inference. Our framework is implemented in PyTorch 2.8 and trained using the AdamW optimizer. The initial learning rate is initialized at 5×10^{-5} with a linear warmup over the first 100 steps, followed by a constant schedule. For parameter-efficient fine-tuning via LoRA, we configure the rank $r = 8$, scaling factor $\alpha = 32$, and a dropout rate of 0.1. The model is trained for 30 epochs on two NVIDIA A800 GPUs with an effective batch size of 16. Regarding the LLM configuration, the maximum input sequence length is set to 2048 tokens. For evaluation, we employ the model checkpoint from the final training step. During inference, we utilize nucleus sampling with a top-p of 0.9 and a sampling temperature of 0.7. The repetition penalty is set to 1.05, and the maximum generation length is constrained to 1024 tokens.

Center-Specific Templates and Explicit Stop Tokens. We collected healthy patient report templates from four distinct medical centers, curated by senior nuclear medicine physicians at each institution. Since drafting PET/CT reports based on standardized healthy templates is a routine workflow in nuclear medicine departments, our Structure-Aware Multi-template Fusion (SAMF) module demonstrates strong generalization capabilities. It can be readily adapted to new centers without the need to reconstruct healthy templates from scratch, as these are typically pre-existing clinical assets. To explicitly define report termination during training, we append a special token "[end-of-report]" to the end of each ground-truth report.

B.2. Implementation of Comparative Methods

Due to the scarcity of methodologies dedicated specifically to PETRG, we benchmark our approach against state-of-the-art report generation models from the X-Ray and CT domains.

For methods supporting 3D inputs, namely RadFM [15] and M3D [2], we utilize their official implementations and pre-trained weights, adhering to the inference parameters specified in the original papers. As these models are limited to single-modality input, we feed the PET volume during inference for the PETRG task. Furthermore, to address their lack of optimization for Chinese output, we prompt

[1] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
[2] https://huggingface.co/zai-org/GLM-4-9B-0414
[3] https://huggingface.co/Qwen/Qwen3-8B
[4] https://huggingface.co/FlagAlpha/Llama2-Chinese-7b-Chat
[5] https://huggingface.co/shenzhi-wang/Mistral-7B-v0.3-Chinese-Chat
[6] https://huggingface.co/shenzhi-wang/Gemma-2-9B-Chinese-Chat

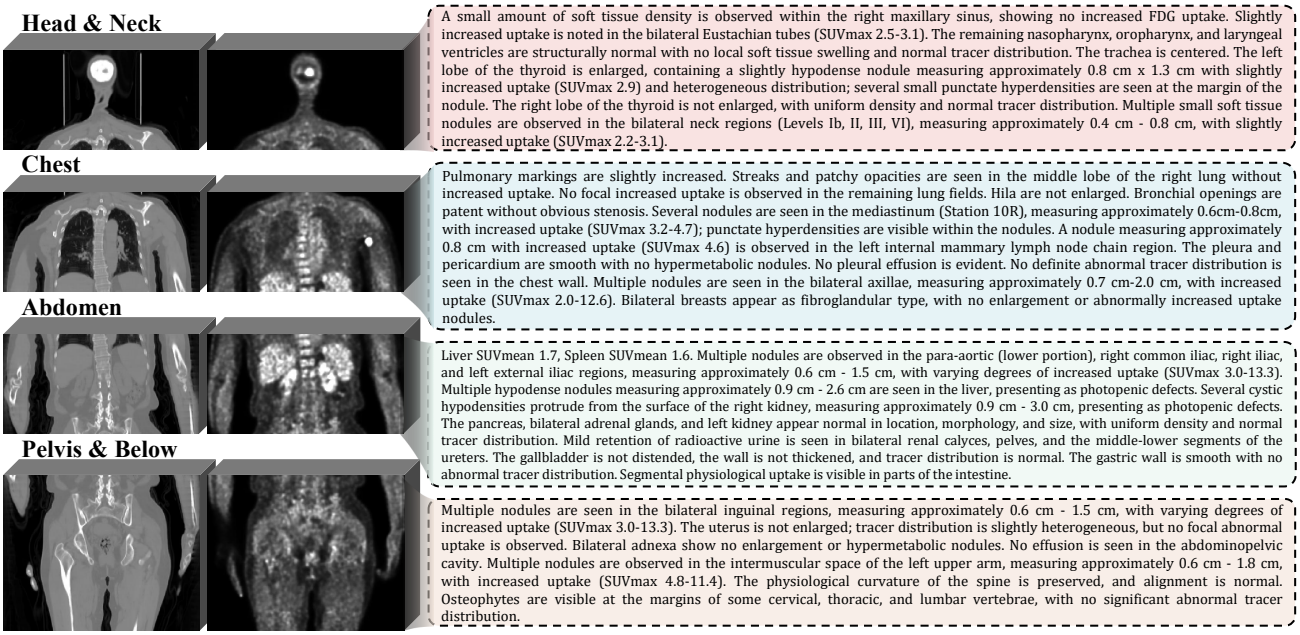


Figure 2. Illustration of region-level image-report pairs in PET/CT.

these models to generate reports in English, which are subsequently translated into Chinese using the Qwen3-Max model. We manually verified a subset of the translated reports to ensure clinical accuracy and terminology consistency.

Regarding PET2Rep [18], we evaluate two slice processing strategies proposed in their work: separate PET and CT slice inputs (PET2Rep-Sep) and fused PET/CT image inputs (PET2Rep-Fus). We perform inference using the official prompts and code. To ensure accurate evaluation, we post-process the raw outputs to filter out artifact symbols (e.g., "[]", "**") inherent to the template predictions. For the Vision-Language Model (VLM) backbones, we select the Qwen-VL and InternVL series, which demonstrated superior performance in [18]. Notably, we employ more recent iterations compared to the original study, utilizing Qwen3-VL-8B and InternVL-3.5-8B. For closed-source models, we employ Qwen-2.5-VL-Max (as Qwen-3-VL-Max was unavailable). Finally, for Llava-Med [8], we utilize the official code and pre-trained weights. Due to its 2D input constraint, we adopt the fused PET/CT image input strategy from [18] for inference.

B.3. Implementation of Region-wise Decomposition

Clinical PET/CT reports typically follow a strictly ordered anatomical description: head/neck, chest, abdomen, and pelvis. This structural consistency allows for the decomposition of the full report into four distinct regional descriptions. Consequently, we propose a data augmentation strategy that partitions whole-body PET/CT images into these four anatomical sections and pairs them with the corresponding decoupled text (Figure 2). This approach scales the training dataset and simplifies the generative task by focusing on local regions.

Specifically, we employ a Body Part Regression (BPR) model [11, 16] to map the coronal slice indices of the CT images to anatomical labels (head, shoulder-neck, chest, abdomen, pelvis, and legs). Based on these mappings, we group "head" and "shoulder-neck" into the *Head & Neck* region, while "chest", "abdomen", and "pelvis/legs" form the *Chest*, *Abdomen*, and *Pelvis & Below* regions, respectively. To mitigate potential information loss caused by BPR boundary errors, we include a buffer of 10 redundant slices at the boundaries of each region. For textual decomposition, we leverage the Qwen3-Max model to segment the full-body report into the four corresponding sections. Figure 3 illustrates the prompt template used for this segmentation, and Figure 2 visualizes an example of the region-wise splitting. The resulting dataset contains 2,652 regional image-report pairs.

During training, the model takes regional PET/CT images as input and is supervised by the corresponding regional report. During inference, we generate reports for the four regions of the test images independently and concatenate them to form the final output. To prevent evaluation bias arising from potential content alterations during

Act as an expert Nuclear Medicine Physician. Reorganize the provided PET/CT "Findings" into four distinct anatomical sections: **Head & Neck**, **Chest**, **Abdomen**, and **Pelvis & Below**.

Instructions:
1. **Strict Fidelity**: Use only the provided information. Do not hallucinate or add content not present in the source text.
2. **Completeness**: Ensure every detail from the original text is included in the appropriate section.
3. **Missing Data**: If a specific region is not mentioned in the source, state: "No abnormal FDG uptake or structural changes observed."
4. **Precision**: Maintain accurate medical terminology. Preserve all key metrics (e.g., SUVmax, dimensions) exactly as they appear.
5. **Cohesion**: Logically consolidate descriptions within the same anatomical region for smooth flow.
6. **Output Format**: Return ONLY a valid, raw JSON object containing the four keys below. Do not include Markdown formatting (e.g., ```json) or conversational text.

Original Findings:
{original_findings}

Required JSON Structure:
```
{
  "Head_and_Neck": "string",
  "Chest": "string",
  "Abdomen": "string",
  "Pelvis_and_Below": "string"
}
```

Figure 3. Illustration of the prompt design utilized to instruct Qwen3-Max for the anatomical partitioning of whole-body PET/CT reports.

the LLM-based text segmentation, we construct the reference ground truth for evaluation by concatenating the LLM-segmented regions of the validation set.

C. Evaluation Metrics

C.1. NLG Metrics

To assess the textual similarity between generated and ground-truth reports, we employ three standard Natural Language Generation (NLG) metrics: BLEU [9], METEOR [3], and ROUGE-L [6]. Given that Chinese text lacks explicit word delimiters, we utilize the Jieba library[7] for tokenization prior to calculation.

C.2. Clinical Efficacy Metric: PETRG-Score

C.2.1. Motivation

Evaluating the clinical utility of generated reports is critical. Standard NLG metrics often fail to capture factual correctness in medical findings. Furthermore, metrics designed for Chest X-Ray or CT report generation are inapplicable to PET/CT, as they ignore metabolic information (PET uptake). Recent attempts, such as PET2Rep [18] and ViMPET [7], exhibit significant limitations:

- **Incomplete Modality Coverage:** PET2Rep neglects structural abnormalities visible in CT, focusing solely on PET.
- **Coarse Granularity:** ViMPET defines only five broad anatomical regions and uses a binary uptake classification. As shown in Tab. 1, real-world reports contain nuanced uptake descriptions (e.g., physiological vs. mild abnormal uptake) and fine-grained anatomical details (e.g., maxillary sinus) that are essential for *accurate diagnosis and staging*.

[7] https://github.com/fxsjy/jieba

C.2.2. Ontology Construction

To overcome these deficiencies, we established a fine-grained evaluation ontology. We extracted 210 raw anatomical mentions from lymphoma reports, which were reviewed and merged by senior nuclear medicine physicians into a set of 24 clinically significant anatomical regions ($\mathcal{L}$). Similarly, we constructed a comprehensive vocabulary for PET uptake statuses and CT density phenotypes.

C.2.3. Calculation Pipeline

Based on this expert-defined ontology, we propose the **PETRG-Score**, a comprehensive metric evaluating both metabolic and structural alignment. The pipeline proceeds as follows:

(1) LLM-based Extraction: We utilize **Qwen3-Max** to extract structured labels (anatomical location, uptake status, density status) from both ground-truth and generated reports. To handle linguistic variations and report sparsity, we implement strict **Normalization Rules** within the prompt (see Fig. 4):

- *Default Normal:* Anatomical regions not mentioned in the report are automatically categorized as "Normal".
- *Hierarchical Inference:* If a major region (e.g., Lungs) is described as normal, all its sub-regions inherit the "Normal" status.
- *Significance Prioritization:* For regions with local abnormalities, the overall status reflects the most significant finding, while unmentioned sub-parts remain "Normal".

(2) Expert Verification: To ensure benchmark reliability, the extracted labels for the ground-truth reports were manually audited and corrected by a researcher specifically trained in medicine imaging.

(3) Metric Formulation: Let $\mathcal{D}$ denote the test set consisting of N reports. For the j-th report ($j \in \{1, \dots, N\}$) and the l-th anatomical region ($l \in \mathcal{L}$), let $y_{j,l} \in \mathcal{C}$ and $\hat{y}_{j,l} \in \mathcal{C}$ denote the ground-truth and predicted class labels, respectively. The valid category sets are $\mathcal{C}_{CT} = \{1, \dots, 8\}$ and $\mathcal{C}_{PET} = \{1, \dots, 5\}$.

We treat every anatomical region in every report as an individual classification instance. To mitigate the impact of class imbalance, we employ the Macro-F1 score. We first define the True Positives (TP_k), False Positives (FP_k), and False Negatives (FN_k) for a specific category $k \in \mathcal{C}$ by aggregating across all reports and regions:

$$TP_k = \sum_{j=1}^{N} \sum_{l \in \mathcal{L}} \mathbb{I}(y_{j,l} = k \wedge \hat{y}_{j,l} = k) \tag{2}$$

Table 1. Extraction of anatomical locations, uptake statuses, and density descriptions from the report in Figure 2. This exemplifies the semantic richness of PET/CT reports, underscoring the critical inadequacy of existing oversimplified CE metrics.

Region	Anatomical Location	Uptake Status	Density Status
Head&Neck	Right maxillary sinus	Showing no increased FDG uptake	A small amount of soft tissue density is observed
	Bilateral Eustachian tubes	Slightly increased uptake is noted	-
	Left lobe of the thyroid	Slightly increased uptake and heterogeneous distribution	Enlarged, containing a slightly hypodense nodule measuring approx. 0.8 cm x 1.3 cm; several small punctate hyperdensities are seen at the margin
	Bilateral neck regions (Levels Ib, II, III, VI)	Slightly increased uptake	Multiple small soft tissue nodules, measuring approx. 0.4 cm - 0.8 cm
Chest	Middle lobe of right lung	Without increased uptake	Streaks and patchy opacities are seen
	Mediastinum (Station 10R)	With increased uptake	Several nodules, measuring approx. 0.6 cm - 0.8 cm; punctate hyperdensities are visible within the nodules
	Left internal mammary lymph node chain	With increased uptake	A nodule measuring approx. 0.8 cm
	Bilateral axillae	With increased uptake	Multiple nodules, measuring approx. 0.7 cm - 2.0 cm
Abdomen	Liver background	-	-
	Para-aortic (lower), Right common iliac, Right iliac, Left external iliac	With varying degrees of increased uptake	Multiple nodules observed, measuring approx. 0.6 cm - 1.5 cm
	Liver	Presenting as photopenic defects	Multiple hypodense nodules measuring approx. 0.9 cm - 2.6 cm
	Right kidney	Presenting as photopenic defects	Several cystic hypodensities protrude from the surface, measuring approx. 0.9 cm - 3.0 cm
	Renal calyces, pelves, ureters	Mild retention of radioactive urine	Mild retention of radioactive urine
Pelvis&Below	Bilateral inguinal regions	With varying degrees of increased uptake	Multiple nodules, measuring approx. 0.6 cm - 1.5 cm
	Intermuscular space of left upper arm	With increased uptake	Multiple nodules observed, measuring approx. 0.6 cm - 1.8 cm
	Spine (Cervical, Thoracic, Lumbar)	With no significant abnormal tracer distribution	Osteophytes are visible at the margins of some vertebrae

$$FP_k = \sum_{j=1}^{N} \sum_{l \in \mathcal{L}} \mathbb{I}(y_{j,l} \neq k \wedge \hat{y}_{j,l} = k) \tag{3}$$

$$FN_k = \sum_{j=1}^{N} \sum_{l \in \mathcal{L}} \mathbb{I}(y_{j,l} = k \wedge \hat{y}_{j,l} \neq k) \tag{4}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

The F1 score for category k is calculated as:

$$F1_k = \frac{2 \cdot TP_k}{2 \cdot TP_k + FP_k + FN_k} \tag{5}$$

The final **PETRG-Score** is the macro-average over a target subset of categories $\mathcal{S} \subseteq \mathcal{C}$:

$$\text{PETRG-Score}(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{k \in \mathcal{S}} F1_k \tag{6}$$

Table 2. Comparison of attribute coverage between PETRG-Score and existing metrics. Corroborated by the real-world labels in Tab. 1, PETRG-Score demonstrates significantly more comprehensive anatomical coverage and provides a richer taxonomy for both metabolic uptake and structural density states, surpassing current evaluation standards.

Metrics	Anatomical Locations	Uptake Status	Density Status
PET2Rep[18]	Cranium and Brain; Eyeballs; Nasal Cavity and Sinuses; Pharynx and Parapharyngeal Space; Palatine Tonsils and Larynx; Salivary Glands and Thyroid; Cervical Lymph Nodes; Lungs and Thoracic Cavity; Mediastinum and Heart; Esophagus; Liver; Gallbladder; Pancreas; Spleen; Kidneys and Adrenal Glands; Gastrointestinal Tract; Prostate/Uterus and Bladder; Abdominal and Pelvic Cavities; Spine and Bones	Increased Uptake; Decreased Uptake; Absent Uptake; Normal	-
ViMed-PET[7]	Mediastinum; Lung; Abdomen; Axilla; Cervical Region	Increase; Not Increase	Lymph Node; Pulmonary Nodule; Ground-Glass Opacity; Pulmonary Mass; Pleural Thickening; Interstitial Thickening; Consolidation; Effusion; Soft Tissue Nodule; Wall Thickening; Calcified Nodule; Hypermetabolic Lesion
PETRG-Score (Ours)	Brain, Skull, and Meninges; Orbit Nasal Cavity, and Paranasal Sinuses; Pharyngeal Spaces, Tonsils, and Larynx; Thyroid Gland and Major Salivary Glands (Parotid, Submandibular); Cervical Lymph Nodes; Lungs and Pleura; Mediastinum and Hila (including Lymph Nodes); Heart and Pericardium; Axilla and Chest Wall; Breasts; Liver; Gallbladder and Biliary Tract; Spleen; Pancreas; Kidneys; Adrenal Glands; Gastrointestinal Tract (Esophagus, Stomach, Intestines); Retroperitoneal Space (including Lymph Nodes); Peritoneum, Mesentery, and Omentum; Pelvic Organs (Bladder, Uterus/Adnexa or Prostate/Seminal Vesicles); Pelvic and Inguinal Lymph Nodes; Spine; Pelvis and Bones of Extremities; Muscles and Subcutaneous Tissue	Intense Abnormal Uptake; Mild/Suspicious Abnormal Uptake; Physiological/Background Uptake; Uptake Defect / Decreased Uptake; Normal	Lymphadenopathy; Focal Lesion; Lung Parenchymal Abnormality; Wall/Membrane Thickening; Calcification; Bone/Skeletal Lesion; Other Abnormality; Normal

We report four variants to provide a comprehensive view:

- **PET-All / CT-All:** Calculated over all categories ($\mathcal{S} = \mathcal{C}_{PET}$ or $\mathcal{S} = \mathcal{C}_{CT}$).
- **PET-Ab / CT-Ab:** Calculated only on abnormal categories, excluding the "Normal" class (ID: 5 for PET, ID: 8 for CT).

C.2.4. Remarks on Limitations

Extraction Bias: Verifying labels for all generated reports in ablation studies is prohibitively expensive. We limit manual verification to the ground truth. To ensure fairness, all methods share the same LLM extraction pipeline, ensuring any extraction noise affects all baselines equally. **Long-**

You are a professional Nuclear Medicine Physician. Your task is to extract the radiotracer uptake status and CT structure/density status for 24 key anatomical regions from the provided PET/CT "Findings" report.

I. Task 1: Extract PET Uptake Status

1. Status Classification: Categorize the uptake status of each organ using one of the 5 codes below:

* **"1" (Intense Abnormal Uptake)**: Explicit descriptions of increased uptake, tracer concentration (hypermetabolism), etc. Corresponds to Deauville Score 5, clearly suggesting malignancy.
* **"2" (Mild/Suspicious Abnormal Uptake)**: Corresponds to Deauville Score 3 (suspicious in certain contexts) or uptake slightly higher than the liver background.
* **"3" (Physiological/Background Uptake)**: Explicitly indicated as physiological uptake.
* **"4" (Uptake Defect/Decreased Uptake)**: Explicit descriptions of photopenia, reduced radioactivity, or uptake defects.
* **"5" (Normal)**: Descriptions such as "no tracer concentration seen," "no abnormal uptake," "uniform distribution," or inferred as normal based on normalization rules.

2. SUV Value Description: If the report describes SUV information for the structure, extract the specific SUVmax or SUVmean value into the `suv` field (e.g., `"suv": "SUVmax: 4.5"`). Otherwise, leave it as an empty string `""`. Note: If multiple SUV values exist for a region, select the maximum value as the SUVmax.

II. Task 2: Extract CT Structure/Density Status

1. Status Classification: Categorize the CT findings of each organ using one of the 8 codes below:

* **"1" (Lymphadenopathy):** Abnormalities in lymph node morphology or size.
* **"2" (Focal Lesion):** Includes nodules and masses.
* **"3" (Lung Parenchymal Abnormality):** Includes ground-glass opacities (GGO), linear opacities/stranding, and patchy shadows/consolidations.
* **"4" (Wall/Membrane Thickening):** Increased thickness of a normal anatomical structure (usually walls, membranes, or interstitium).
* **"5" (Calcification):** High-density calcium salt deposition within tissues.
* **"6" (Bone/Skeletal Lesion):** e.g., osteoblastic or osteolytic lesions.
* **"7" (Other Abnormalities):** Any abnormalities not listed above.
* **"8" (Normal):** Not described, explicitly stated as no abnormality, or inferred as normal based on normalization rules.

2. Extraction Description: If the status falls into the first 7 "Abnormal" categories, briefly extract the specific description into the `ct_description` field; otherwise, leave it as an empty string `""`.

III. Normalization Rules

1. **Default Normal:** If an organ is not mentioned in the report at all, its status should be categorized as "Normal".
2. **Hierarchical Normal:** If a major category (e.g., "Lungs and Pleura") is described as normal, all its sub-regions are considered "Normal".
3. **Implied Normal:** If only a local abnormality within an organ is mentioned (e.g., "Increased uptake in cervical zone II"), the unmentioned parts of that anatomical region are considered "Normal," but the overall status of the organ should reflect the most significant finding (in this case, "Markedly Abnormal Uptake").

IV. Output Format

Please output the result strictly in the following JSON format without adding any extra explanation. Ensure the `pet_status` and `ct_status` fields are formatted strictly as single numbers.

JSON Template:

```json
{
  "Brain, Skull, and Meninges": {"pet_status": "...", "suv": "...", "ct_status": "...", "ct_description": "..."},
  "Orbits, Nasal Cavity, and Paranasal Sinuses": {"pet_status": "...", "suv": "...", "ct_status": "...", "ct_description": "..."},
  "Pharyngeal Space, Tonsils, and Larynx": ...,
  ...
  "Muscles and Subcutaneous Tissue": ...
}
```

Figure 4. Prompt designed for Qwen3-Max to extract attributes from ground-truth and generated reports. The prompt targets 5 uptake states and 8 density states across 24 anatomical locations. We explicitly constrain the output to numerical indices rather than textual labels; this design choice minimizes model hallucinations and ensures robust downstream parsing.

tail Coverage: While our ontology is significantly richer than prior arts, extremely rare anatomical variants or atypical descriptions may still be underrepresented. Developing dynamic, open-vocabulary evaluation metrics remains a direction for future work.

D. Additional Results

D.1. Center-wise Style Adaptation Analysis

Table 3. Multi-center evaluation of the SAMF module's impact on adapting to heterogeneous reporting styles. "w/o" and "w/ SAMF" denote the PETRG-3D baseline without and with the module, respectively. Centers 1–4 belong to the PETRG-Lym dataset, while Center 5 corresponds to AutoPET-RG-Lym. Performance gains are highlighted in green parentheses.

Methods	Center ID	B-4	MTR	R-L
w/o SAMF	1	26.31	44.47	38.67
	2	46.71	50.75	54.89
	3	44.86	52.53	53.47
	4	35.69	42.35	46.62
	5	8.11	28.18	24.57
w/ SAMF	1	26.53 (+0.22)	43.03 (-1.44)	41.22 (+2.55)
	2	53.38 (+6.67)	57.19 (+6.44)	61.81 (+6.92)
	3	55.84 (+10.98)	62.00 (+9.47)	63.82 (+10.35)
	4	43.09 (+7.40)	49.08 (+6.73)	53.29 (+6.67)
	5	32.47 (+24.36)	41.50 (+13.32)	40.97 (+16.40)

To quantitatively evaluate the efficacy of the SAMF module in style adaptation, Table 3 details the NLG evaluation metrics across four distinct centers within the PETRG-Lym dataset. Integrating the SAMF module yielded consistent improvements across nearly all metrics for the internal centers, confirming its ability to accommodate the diverse reporting conventions inherent to different medical institutions.

Notably, for the external test center (Center 5, AutoPET-RG-Lym), the inclusion of SAMF resulted in substantial performance gains: BLEU-4 surged by 24.36 points, while METEOR and ROUGE-L saw absolute increases of 13.32 and 16.40 points, respectively. These results suggest that even when confronted with unseen reporting styles, PETRG-3D equipped with SAMF effectively performs style transfer by leveraging the synergy between patient-specific PET/CT images and retrieved style templates from the target center.

D.2. Performance Analysis by Uptake and Density Status

We further stratified model performance by uptake and density statuses to delineate diagnostic capabilities and potential biases. Table 4 and Table 5 present the Clinical Efficacy (CE) metrics—Precision (P), Recall (R), and F1 score (F1)—alongside the "Support" column, which highlights the class distribution.

Table 4. Evaluation of our model across five uptake statuses. "P" and "R" denote Precision and Recall, respectively. "Support" indicates the sample size for each status.

Uptake Status	P	R	F1	Support
Intense Abnormal Uptake	35.55	19.93	25.54	537
Mild/Suspicious Abnormal Uptake	15.94	11.53	13.38	347
Physiological/Background Uptake	28.85	23.08	25.64	65
Uptake Defect/Decreased Uptake	13.92	13.25	13.58	83
Normal	77.68	87.25	82.19	2832
Macro Avg	34.39	31.01	32.06	3864
Weighted Avg	64.09	68.43	65.71	3864

Table 5. Evaluation of our model across eight density statuses. "P" and "R" denote Precision and Recall, respectively. "Support" indicates the sample size for each status.

Density Status	P	R	F1	Support
Lymphadenopathy	55.56	37.72	44.93	464
Focal Lesion	13.46	12.07	12.73	116
Lung Parenchymal Abnormality	68.13	66.31	67.21	187
Wall/Membrane Thickening	14.29	9.52	11.43	63
Calcification	9.30	7.55	8.33	53
Bone/Skeletal Lesion	36.99	30.00	33.13	90
Other Abnormality	28.98	19.85	23.56	413
Normal	72.08	82.08	76.75	2478
Macro Avg	37.35	33.14	34.76	3864
Weighted Avg	60.91	63.82	61.84	3864

The results indicate that the "Normal" category significantly outperforms others across both uptake and density classifications. This disparity is largely attributable to the extreme class imbalance, where the prevalence of normal samples biases the model towards predicting anatomical regions as abnormality-free. This distribution skew justifies our adoption of the Macro F1 score as a primary evaluation metric to ensure fair assessment.

Interestingly, apart from the normal class, the "Physiological/Background Uptake" category (Table 4) achieved the highest performance despite having the fewest samples. This is likely due to the strong semantic coupling between physiological uptake and specific anatomical structures (e.g., the bladder or kidneys), which simplifies the learning task. Similarly, "Lung Parenchymal Abnormality" (Table 5) ranks second to the normal category, presumably benefiting from its distinct localization within the lung fields. These findings highlight that while the model learns strong anatomy-pathology associations, mitigating data imbalance remains a critical frontier for future PET/CT report generation research.

D.3. Computational Resource Analysis

Table 6 reports the training time and peak GPU memory usage for PETRG-3D. Experiments were conducted on two NVIDIA A800 (80GB) GPUs using a standardized software

Table 6. Training resource consumption of PETRG-3D with different LLMs on PETRG-Lym dataset. Memory values indicate the peak GPU memory usage (in GB) across two A800 GPUs.

LLMs	Time (Hours)	Memory (GB)
Llama2-7B	5.70	30.14
Mistral-7B-v0.3	5.56	34.49
Qwen2.5-7B	5.70	38.27
Gemma2-9B	8.62	45.23
GLM4-9B	8.05	41.10
Qwen3-8B	7.33	38.81

stack[8]. As anticipated, both training duration and memory requirements scale positively with the parameter size of the underlying LLM.

D.4. Qualitative Analysis (Whole-body Reporting)

Figure 5 presents a representative whole-body PET/CT lymphoma report. Authored by a nuclear medicine physician, the report is extensive, covering regions from the head and neck to the musculoskeletal system, thereby underscoring the complexity of the PETRG task.

Figure 6 illustrates the output from PET2Rep [18]. As highlighted in red, the baseline model suffers from significant hallucinations, particularly in lesion-rich areas like the lungs and abdomen (e.g., incorrectly predicting volume reduction or pancreatic nodules). This suggests that generalist VLMs or those without specialized fine-tuning lack the diagnostic reliability required for complex medical imaging tasks.

In contrast, the report generated by our PETRG-3D model (Figure 7) closely mirrors the ground truth in both linguistic style and diagnostic content. It accurately identifies multiple nodal involvements (hepatic hilum, paraaortic, iliac, etc.) and extranodal abnormalities. While some limitations remain—specifically regarding quantitative precision (e.g., SUVmax values) and minor false positives in complex regions—the drastic reduction in hallucinations demonstrates the effectiveness of our fine-tuning strategy. Current errors are largely numerical, pointing to future directions for integrating quantitative analysis heads into the architecture.

D.5. Effectiveness of Explicit Stop Tokens

Figure 8 compares model outputs before and after the incorporation of explicit stop tokens. Without the explicit stop token, the model tends to generate prolonged hallucinatory content after the actual report concludes, as it attempts to fill the fixed `max_new_token` buffer required for batch generation. By introducing an explicit stop token, the model learns to output "[end-of-report]" immediately upon completing the valid report content. This allows for early termination of the generation process, effectively eliminating post-report hallucinations.

[8]Environment: Python 3.10, CUDA 12.8, PyTorch 2.8.0, Transformers 4.57.0, PEFT 0.17.1, and DeepSpeed 0.17.5.

References

[1] Q Jiang Albert, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, and Devendra Singh Chaplot. Mistral 7b. *arXiv preprint*, 2023. 2

[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

[3] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 4

[4] DR Dance, S Christofides, ADA Maidment, ID McLean, and KH Ng. Diagnostic radiology physics: A handbook for teachers and students. endorsed by: American association of physicists in medicine, asia-oceania federation of organizations for medical physics, european federation of organisations for medical physics. 2014. 1

[5] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024. 2

[6] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 4

[7] Huu Tien Nguyen, Dac Thai Nguyen, The Minh Duc Nguyen, Trung Thanh Nguyen, Thao Nguyen Truong, Huy Hieu Pham, Johan Barthelemy, Minh Quan Tran, Thanh Tam Nguyen, Quoc Viet Hung Nguyen, et al. Toward a vision-language foundation model for medical data: Multimodal dataset and benchmarks for vietnamese pet/ct report generation. *arXiv preprint arXiv:2509.24739*, 2025. 4, 6

[8] Thao Nguyen, Matthew Wallingford, Sebastin Santy, Wei-Chiu Ma, Sewoong Oh, Ludwig Schmidt, Pang Wei W Koh, and Ranjay Krishna. Multilingual diversity improves vision-language representations. *Advances in Neural Information Processing Systems*, 37:91430–91459, 2024. 3

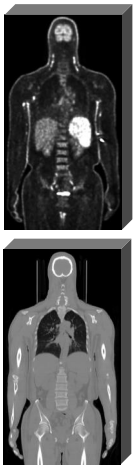
[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 4

[10] Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, and Fei Huang. Qwen2.5 technical report. 2024. 2

[11] Sarah Schuhegger. Mic-dkfz/bodypartregression. *Zenodo https://doi. org/10.5281/zenodo*, 5195341, 2021. 3

[12] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari,

Ground Truth:



PET/CT

<ZH/Original> **颈部：**左侧上颌窦窦腔内见少许软组织密度影，未见放射性摄取增高。鼻咽腔后上壁稍肿胀，放射性摄取增高，SUVmax4.2，余口咽咽腔及喉室腔结构正常，局部软组织未见肿胀，放射性分布正常。气管居中。甲状腺两叶不大，密度尚均匀，放射性分布正常。双侧颈部（1、2、3、4、5、6、8）见数枚软组织密度小结节影，大小约0.4cm-0.8cm，放射性摄取增高，SUVmax2.5-8.2。**胸部：**双肺纹理稍增多，右肺中叶见少许条索影，未见放射性摄取增高，肺野内未见放射性摄取增高灶。肺门不大，支气管开口通畅，管腔未见明显狭窄。纵隔（2R、4、5、6、10）见数枚结节影，大小约0.4cm-0.7cm，放射性摄取增高，SUVmax2.8-4.5，部分结节内见点片状致密影。胸膜及心包膜光滑，未见放射性摄取增高结节。胸腔内未见积液影。双侧腋下见多枚结节影，大小约0.3cm-1.4cm，放射性摄取不同程度增高，SUVmax1.7-11.2。纵隔血池SUVmax1.7。**腹、盆部：**肝脏SUVmean2.0，肝脏血池SUVmax2.7。脾脏体积增大，脾内见2枚稍低密度结节影，大小约1.4cm-1.9cm，呈放射性缺损，余脾脏实质弥漫性放射性摄取增高，分布不均匀，SUVmax6.7，SUVmean2.8。肝门区、腹主动脉旁、双侧髂总血管旁、双侧髂血管旁、双侧腹股沟见多枚结节影，大小约0.6cm-2.0cm，放射性摄取不同程度增高，SUVmax2.7-10.6。左肾上腺联合部见一枚结节影，大小约1.1cm*1.6cm，放射性摄取稍高，SUVmax2.3。胰腺、右肾上腺及双肾位置、形态、大小均正常，密度均匀，放射性分布正常。双肾盏、肾盂及输尿管中下段内见少量放射性尿液滞留。胆囊张力稍大，壁不厚，放射性分布正常。胃壁光滑，放射性分布未见异常。部分肠道可见节段性摄取。前列腺不大，放射性分布欠均匀，SUVmax1.4，未见放射性摄取异常增高灶。腹盆腔未见积液影。**骨骼、肌肉和大脑：**双侧上臂及膝关节肌肉间隙见数枚结节影，大小约0.3cm-0.6cm，放射性摄取不同程度增高，SUVmax2.6-7.2。全身皮肤不同程度增厚，放射性摄取增高，分布不均匀，SUVmax1.7-7.5，以双侧腋下、腹股沟皮肤显著。椎体生理弯曲存在，序列正常，颈椎、胸椎及腰椎部分椎体边缘可见不同程度骨质增生迹象，放射性分布未见明显异常。颅内大脑皮质、小脑半球弥漫性放射性摄取减低，双侧基底节、丘脑放射性分布正常，两侧对称。诸脑室形态、大小、位置无改变；脑沟、脑池无增宽；中线结构无移位。

<EN/Translated> **Head and Neck:** A small soft tissue density is observed within the left maxillary sinus without increased tracer uptake. The posterior-superior wall of the nasopharynx appears slightly thickened with increased FDG uptake (SUVmax 4.2). The remaining oropharyngeal and laryngeal structures appear normal; no localized soft tissue swelling or abnormal tracer distribution is noted in these areas. The trachea is midline. The thyroid lobes are normal in size with homogeneous density and physiological tracer distribution. Multiple small soft tissue nodules are visualized in the bilateral cervical regions (Levels I, II, III, IV, V, VI, and VIII), measuring approximately 0.4 cm to 0.8 cm in diameter, demonstrating increased FDG uptake with SUVmax ranging from 2.5 to 8.2. **Chest:** Bilateral lung markings are slightly increased. A few fibrous streaks are noted in the right middle lobe without increased tracer uptake; no other hypermetabolic lesions are observed in the lung fields. The hila are not enlarged, and the bronchial trees are patent without obvious stenosis. Multiple lymph nodes are seen in the mediastinum (Stations 2R, 4, 5, 6, and 10), measuring approximately 0.4 cm to 0.7 cm, with increased uptake (SUVmax 2.8–4.5); punctate and patchy calcifications are visible within some of these nodes. The pleura and pericardium are smooth without hypermetabolic nodules. No pleural effusion is evident. Multiple lymph nodes are observed in the bilateral axillae, measuring 0.3 cm to 1.4 cm, exhibiting varying degrees of increased uptake (SUVmax 1.7–11.2). The mediastinal blood pool SUVmax is 1.7. **Abdomen and Pelvis:** The liver shows a mean SUV of 2.0, with a liver blood pool SUVmax of 2.7. The spleen is enlarged, containing two slightly hypodense nodules measuring approximately 1.4 cm to 1.9 cm that present as photopenic defects. The remaining splenic parenchyma demonstrates diffuse, heterogeneous increased uptake with an SUVmax of 6.7 and SUVmean of 2.8. Multiple lymph nodes are visualized in the hepatic hilar region, para-aortic region, bilateral common iliac, bilateral external/internal iliac, and bilateral inguinal regions. These nodes measure approximately 0.6 cm to 2.0 cm and show varying degrees of increased uptake (SUVmax 2.7–10.6). A nodule is noted at the junction of the left adrenal gland, measuring approximately 1.1 cm × 1.6 cm, with mildly increased uptake (SUVmax 2.3). The pancreas, right adrenal gland, and kidneys are normal in position, morphology, and size, with homogeneous density and normal tracer distribution. Mild retention of radioactive urine is seen in the bilateral renal calyces, pelvis, and mid-to-distal ureters. The gallbladder shows slightly increased tension with a non-thickened wall and normal tracer distribution. The gastric wall is smooth with no abnormal uptake. Segmental physiologic uptake is noted in parts of the bowel. The prostate is not enlarged, showing heterogeneous tracer distribution (SUVmax 1.4) without focal hypermetabolism. No abdominopelvic ascites is observed. **Musculoskeletal, Skin, and Brain:** Multiple nodules are observed in the intermuscular spaces of the bilateral upper arms and knee joints, measuring approximately 0.3 cm to 0.6 cm, with varying degrees of increased uptake (SUVmax 2.6–7.2). There is generalized, varying skin thickening with increased, heterogeneous tracer uptake (SUVmax 1.7–7.5), which is most prominent in the bilateral axillary and inguinal regions. The physiological curvature of the spine is preserved, and alignment is normal. Varying degrees of osteophytes are noted at the margins of the cervical, thoracic, and lumbar vertebrae; however, no definite abnormal tracer distribution is seen in the skeletal system. Intracranially, there is diffuse hypometabolism (decreased uptake) in the cerebral cortex and cerebellar hemispheres. The bilateral basal ganglia and thalami show normal, symmetrical tracer distribution. The morphology, size, and position of the ventricles remain unchanged. The sulci and cisterns are not widened, and midline structures are not shifted.

Figure 5. Example of a clinical whole-body PET/CT report for lymphoma. The original report is in Chinese (ZH) and has been translated into English (EN) for illustration.

Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. 2

[13] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2

[14] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic struc-

Generated (PET2Rep):

<ZH/Original> 禁食状态下静脉注射18F-FDG，静息60分钟后行全身PET断层显像，全身断层影像显示：脑形态、结构正常，双侧大脑皮质放射性分布均匀、对称，FDG摄取未见明显异常。各脑室、脑沟、脑裂、脑池形态正常，中线结构无偏移。双侧眼球形态和轮廓无异常，FDG无异常摄取。右侧上颌窦黏膜稍增厚伴部分钙化，FDG摄取未见异常。鼻咽壁未见增厚，两侧腭扁桃体对称，FDG呈生理性摄取。喉咽形态、结构未见异常。双侧腮腺、颌下腺形态、密度正常，FDG生理性摄取。甲状腺形态、大小正常，密度均匀，FDG摄取未见异常。双侧颈深间隙、颌下见数枚小淋巴结显示，部分淋巴结FDG摄取增高。两肺透亮度增加，两肺上叶胸膜下见囊状透亮影。右肺中叶体积缩小，右肺中叶肺门旁见片状密度增高影，FDG摄取增高。心影未见异常。部分动脉管壁钙化（包括冠状动脉）。食管未见扩张，管壁未见明显增厚或肿块，FDG摄取未见增高。肝缘欠光整，肝裂未见明显增宽，CT平扫肝右叶下段见巨大稍低密度肿块影，边界欠清，FDG摄取增高。肝左叶见稍低密度结节，边界尚清晰，FDG摄取较本底摄取减低。肝右叶膈顶区另见结节状致密影。门静脉主干及左右支明显增粗，密度增高，FDG摄取增高。门腔间隙、肝胃间隙、腹膜后见淋巴结显示，FDG摄取轻度增高。胆囊稍饱满，胆囊壁未见增厚，局部FDG摄取未见异常。胰腺形态正常，胰体部见小结节状钙化灶，主胰管未见增宽，FDG摄取未见明显异常。脾脏形态、大小密度未见异常，FDG摄取未见异常。双侧泌尿系统见对比剂残留影；左肾体积缩小，左肾见多枚囊状低密度影及稍高密度小结节，边界清，FDG摄取缺失；右肾上极见一脂肪密度结节，FDG摄取未见异常。双侧肾盂肾盏及输尿管未见增宽。双侧肾上腺显影未见明显异常，FDG摄取未见异常。胃充盈尚可，胃壁未见明显增厚，部分胃壁FDG摄取增高。肠道充盈欠佳，肠壁未见明显增厚或肿块，FDG呈生理性摄取。前列腺稍增大，实质内见数枚钙化灶，FDG摄取未见异常增高。膀胱充盈欠佳，内见对比剂残留影。骨质密度不均匀减低，脊柱部分椎体边缘骨质增生。FDG摄取未见异常。

<EN/Translated> Whole-body PET/CT tomography was performed 60 minutes after intravenous injection of 18F-FDG under fasting conditions. The whole-body tomographic images reveal: Brain morphology and structure are normal; radioactivity distribution in the bilateral cerebral cortex is uniform and symmetrical, with no obvious abnormal FDG uptake. The morphology of cerebral ventricles, sulci, fissures, and cisterns is normal, and midline structures are not shifted. Bilateral eyeballs show no morphological or contour abnormalities and no abnormal FDG uptake. Slight mucosal thickening with partial calcification is observed in the right maxillary sinus, without abnormal FDG uptake. The nasopharyngeal walls are not thickened; bilateral palatine tonsils are symmetrical with physiological FDG uptake. Morphology and structure of the laryngopharynx are unremarkable. Bilateral parotid and submandibular glands show normal morphology and density with physiological FDG uptake. The thyroid gland is normal in morphology and size with uniform density and no abnormal FDG uptake. Several small lymph nodes are visualized in the bilateral deep cervical and submandibular spaces, with some showing increased FDG uptake. Increased radiolucency is observed in both lungs, with subpleural cystic lucencies in the upper lobes. The right middle lobe shows volume loss and patchy increased density shadows near the hilum, accompanied by increased FDG uptake. The cardiac shadow is unremarkable. Calcifications are present in the walls of some arteries (including coronary arteries). The esophagus is not dilated, wall thickening or masses are not distinct, and FDG uptake is not increased. The liver margin is slightly irregular, and hepatic fissures are not significantly widened; plain CT scan reveals a giant, slightly low-density mass in the lower segment of the right hepatic lobe with ill-defined boundaries and increased FDG uptake. A slightly low-density nodule with clear boundaries is seen in the left hepatic lobe, showing FDG uptake lower than background. Additionally, a nodular dense shadow is observed in the dome of the right hepatic lobe. The main portal vein and its left and right branches are significantly thickened with increased density and increased FDG uptake. Lymph nodes are visualized in the portocaval space, hepatogastric space, and retroperitoneum, showing mildly increased FDG uptake. The gallbladder is slightly distended without wall thickening, and local FDG uptake is unremarkable. The pancreas morphology is normal; a small nodular calcification is seen in the pancreatic body; the main pancreatic duct is not dilated, and no obvious abnormal FDG uptake is observed. Spleen morphology, size, and density are unremarkable, with no abnormal FDG uptake. Contrast retention is noted in the bilateral urinary systems; the left kidney is atrophic, containing multiple cystic low-density shadows and a slightly high-density small nodule with clear boundaries and absent FDG uptake; a fat-density nodule is seen in the upper pole of the right kidney without abnormal FDG uptake. Bilateral renal pelves, calyces, and ureters are not dilated. Bilateral adrenal glands show no obvious abnormalities in appearance or FDG uptake. Gastric filling is fair; the gastric wall is not significantly thickened, though partial gastric wall shows increased FDG uptake. Bowel filling is suboptimal; no obvious wall thickening or masses are seen, and FDG distribution is physiological. The prostate is slightly enlarged with parenchymal calcifications, without abnormally increased FDG uptake. The bladder is poorly filled with contrast retention observed within. Bone density is unevenly decreased, and osteohyperplasia is noted at the edges of some vertebral bodies. No abnormal osseous FDG uptake is observed.

Figure 6. Case study on PET2Rep [18] generation. Green and red backgrounds denote statements consistent and inconsistent with the Ground Truth (Fig. 5), respectively.

tures in ct images. *Radiology: Artificial Intelligence*, 5(5): e230024, 2023. 1

[15] 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. 1, 2

[16] Ke Yan, Le Lu, and Ronald M Summers. Unsupervised body part regression via spatially self-ordering convolutional neural networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1022–1025. IEEE, 2018. 3

[17] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 2

[18] Yichi Zhang, Wenbo Zhang, Zehui Ling, Gang Feng, Sisi Peng, Deshu Chen, Yuchen Liu, Hongwei Zhang, Shuqi Wang, Lanlan Li, et al. Pet2rep: Towards vision-language model-drived automated radiology report generation for positron emission tomography. *arXiv preprint arXiv:2508.04062*, 2025. 3, 4, 6, 9, 11

Generated (PETRG-3D):

<ZH/Original> **颈部：**双侧上颌窦窦腔内未见软组织密度影，未见放射性摄取增高。鼻咽、口咽咽腔及喉室腔结构正常，局部软组织未见肿胀，放射性分布正常。气管居中。甲状腺两叶不大，右叶密度欠均匀，内见稍低密度微小结节影，大小约0.2cm，未见放射性摄取异常增高。双侧颈部（1b）见数枚软组织密度小结节影，大小约0.7cm-1.3cm，放射性摄取增高，SUVmax6.4-18.1。**胸部：**双肺纹理稍增多，肺野内未见放射性摄取增高灶。肺门不大，支气管开口通畅，管腔未见明显狭窄。纵隔（2R、4R、5、10）见数枚软组织密度结节影，大小约0.6cm-1.1cm，放射性摄取增高，SUVmax5.3-19.5，且以右侧显著。胸膜及心包膜光滑，未见放射性摄取增高结节。左胸腔内见少许弧立性低密度影，呈放射性缺损。双侧腋下见多枚软组织密度小结节影，大小约0.4cm-1.0cm，放射性摄取稍高，SUVmax1.7-3.1。**腹、盆部：**肝脏SUVmean1.6，脾脏SUVmean1.8。肝门区、胰周、腹主动脉旁、肠系膜、左侧髂总血管旁、左侧髂血管旁、双侧腹股沟见多枚软组织密度结节影，大小约0.5cm-2.6cm，放射性摄取不同程度增高，SUVmax3.4-24.1。左肾内见一类圆形稍高密度影，大小约1.9cm*2.3cm，呈放射性缺损。左肾盂内见点状致密影。胰腺、双肾上腺及右肾位置、形态、大小均正常，密度均匀，放射性分布正常。双肾盏、肾盂及输尿管中下段内见少量放射性尿液滞留。胆囊张力不大，壁不厚，放射性分布正常。胃壁光滑，放射性分布未见异常。部分肠道可见节段性摄取。前列腺不大，放射性分布欠均匀，未见放射性摄取异常增高灶，内见点状钙化影。腹盆腔未见积液影。**骨骼、肌肉和大脑：**椎体生理弯曲存在，序列正常，颈椎、胸椎及腰椎部分椎体边缘可见不同程度骨质增生迹象，放射性分布未见明显异常。颅内大脑皮质、双侧基底节、丘脑及小脑半球放射性分布正常，两侧对称。诸脑室形态、大小、位置无改变；脑沟、脑池无增宽；中线结构无移位。

<EN/Translated> **Head and Neck:** The bilateral maxillary sinuses are clear, showing no soft tissue densities or abnormal tracer uptake. The nasopharynx, oropharynx, and laryngeal structures appear normal; no localized soft tissue swelling or abnormal tracer distribution is observed. The trachea is midline. The thyroid lobes are normal in size; the right lobe demonstrates slightly heterogeneous density and contains a minute hypodense nodule measuring approximately 0.2 cm, which does not exhibit abnormal tracer uptake. Multiple soft tissue nodules are visualized in the bilateral cervical regions (Level IB), measuring approximately 0.7 cm to 1.3 cm. These nodules show intense hypermetabolism, with SUVmax ranging from 6.4 to 18.1. **Chest:** Bilateral lung markings are slightly prominent, but no hypermetabolic foci are observed within the lung fields. The hila are not enlarged, the bronchial orifices are patent, and no significant luminal stenosis is noted. Multiple soft tissue nodal densities are seen in the mediastinum (Stations 2R, 4R, 5, and 10), measuring approximately 0.6 cm to 1.1 cm. These nodes demonstrate increased uptake (SUVmax 5.3–19.5), with involvement being more pronounced on the right side. The pleura and pericardium are smooth without hypermetabolic nodules. A small amount of isolated low-density fluid is noted in the left thoracic cavity, presenting as a photopenic defect (trace pleural effusion). Multiple small soft tissue nodules are observed in the bilateral axillae, measuring 0.4 cm to 1.0 cm, showing mildly increased uptake with SUVmax ranging from 1.7 to 3.1. **Abdomen and Pelvis:** The liver shows a mean SUV of 1.6, and the spleen shows a mean SUV of 1.8. Multiple soft tissue nodules are visualized in the hepatic hilar region, peripancreatic region, para-aortic region, mesentery, and adjacent to the left common iliac vessels, left external/internal iliac vessels, and bilateral inguinal regions. These nodules measure approximately 0.5 cm to 2.6 cm and exhibit varying degrees of increased uptake, with SUVmax ranging from 3.4 to 24.1. A round, slightly hyperdense lesion measuring approximately 1.9 cm × 2.3 cm is noted in the left kidney, presenting as a photopenic defect; a punctate density (suggestive of a calculus) is seen in the left renal pelvis. The pancreas, bilateral adrenal glands, and right kidney are normal in position, morphology, and size, with homogeneous density and normal tracer distribution. Mild retention of radioactive urine is observed in the bilateral renal calyces, pelves, and mid-to-distal ureters. The gallbladder is not distended, the wall is not thickened, and tracer distribution is normal. The gastric wall is smooth with no abnormal uptake. Segmental physiologic uptake is noted in parts of the bowel. The prostate is not enlarged and shows heterogeneous tracer distribution without focal hypermetabolism; punctate calcifications are visible within. No abdominopelvic ascites is observed. **Musculoskeletal, Skin, and Brain:** The physiological curvature of the spine is preserved, and alignment is normal. Varying degrees of osteophytes are noted at the margins of the cervical, thoracic, and lumbar vertebrae (spondylosis), but no significant abnormal tracer distribution is detected in the skeletal system. Intracranially, the cerebral cortex, bilateral basal ganglia, thalami, and cerebellar hemispheres show normal, symmetrical tracer distribution. The morphology, size, and position of the ventricles remain unchanged. The sulci and cisterns are not widened, and midline structures are not shifted.

Figure 7. Case study on PETRG-3D generation. Green and red backgrounds denote statements consistent and inconsistent with the Ground Truth (Fig. 5), respectively.

w/ Stop Token:

<ZH/Original> 颈部：双侧上颌窦窦腔内未见软组织密度影，未见放射性摄取增高。...
颅内大脑皮质、双侧基底节、丘脑及小脑半球放射性分布正常，两侧对称。诸脑室形态、大小、位置无改变；脑沟、脑池无增宽；中线结构无移位。

<EN/Translated> **Head and Neck:** The bilateral maxillary sinuses are clear, showing no soft tissue densities or abnormal tracer uptake.... Intracranially, the cerebral cortex, bilateral basal ganglia, thalami, and cerebellar hemispheres show normal, symmetrical tracer distribution. The morphology, size, and position of the ventricles remain unchanged. The sulci and cisterns are not widened, and midline structures are not shifted.

w/o Stop Token:

<ZH/Original> 颈部：双侧上颌窦窦腔内未见软组织密度影，未见放射性摄取增高。...
颅内大脑皮质、双侧基底节、丘脑及小脑半球放射性分布正常，两侧对称。诸脑室形态、大小、位置无改变；脑沟、脑池无增宽；中线结构无移位。 语言功能PET代谢显像正常。
</think>
颈部：双侧上颌窦窦腔内未见软组织密度影，未见放射性摄取增高。鼻咽、口咽咽腔及喉室腔结构正常，局部软组织未见肿胀，放射性分布正常。气管居中。甲状腺两叶不大，密度尚均匀，放射性分布正常。左侧颈部（4）见一枚小结节影，大小约0.6cm，放射性摄取增高，SUVmean2.0，SUVmax2.2。胸部：双肺纹理稍增多，双肺上叶及左肺下叶后基底段可见多发囊泡状透亮影，呈放射性缺损灶。双肺下叶后基底段见少许条索影，未见放射性摄取异常增高。肺门不大，支气管开口通畅，管腔未见明显狭窄。双肺门及纵隔未见肿大及放射性摄取增高结节。右侧胸腔内见少许积液影，未见放射性摄取异常增高。胸膜及心包膜光滑，未见放射性摄取增高结节。胸壁及双侧腋下未见明确放射性分布异常。左侧胸壁内侧份胸大肌内侧间隙可见一结节影，大小约0.6cm，放射性摄取增高，SUVmean3.1，SUVmax3.5。纵隔血池SUVmean1.7。腹、盆部：肝脏SUVmean2.1，脾脏SUVmean1.7。胰腺、双肾上腺及双肾位置、形态、大小均正常，密度均匀，放射性分布正常。双肾盏、肾盂及输尿管中下段内见少量放射性尿液滞留。胆囊张力不大，壁不厚，放射性分布正常。胃壁光滑，放射性分布未见异常。部分肠道可见节段性摄取。前列腺不大，放射性分布

<EN/Translated> **Head and Neck:** The bilateral maxillary sinuses are clear, showing no soft tissue densities or abnormal tracer uptake.... Intracranially, the cerebral cortex, bilateral basal ganglia, thalami, and cerebellar hemispheres show normal, symmetrical tracer distribution. The morphology, size, and position of the ventricles remain unchanged. The sulci and cisterns are not widened, and midline structures are not shifted. Language function PET metabolic imaging is normal.
</think>
Neck: No soft tissue density opacities are observed within the bilateral maxillary sinuses, and no increased tracer uptake is seen. The structures of the nasopharynx, oropharynx, and laryngeal ventricle are normal; local soft tissues show no swelling, and radiotracer distribution is normal. The trachea is midline. Both lobes of the thyroid are not enlarged, with relatively uniform density and normal radiotracer distribution. A small nodule is observed in the left neck (Level IV), measuring approximately 0.6 cm, with increased tracer uptake (SUVmean 2.0, SUVmax 2.2).Chest: Pulmonary markings are slightly increased. Multiple cystic lucencies are visible in the upper lobes of both lungs and the posterior basal segment of the left lower lobe, presenting as photopenic defects. Scanty fibrous streaks are seen in the posterior basal segments of the bilateral lower lobes, without abnormally increased tracer uptake. The hila are not enlarged. Bronchial openings are patent, and no obvious luminal narrowing is observed. No enlarged nodes or nodes with increased tracer uptake are seen in the bilateral hila or mediastinum. A small amount of effusion is noted in the right thoracic cavity, without abnormally increased tracer uptake. The pleura and pericardium are smooth, with no nodular increased tracer uptake observed. No definite abnormal radiotracer distribution is seen in the chest wall or bilateral axillae. A nodule measuring approximately 0.6 cm is visible in the medial intermuscular space of the pectoralis major on the medial aspect of the left chest wall, showing increased tracer uptake (SUVmean 3.1, SUVmax 3.5). Mediastinal blood pool SUVmean is 1.7.Abdomen and Pelvis: Liver SUVmean 2.1, Spleen SUVmean 1.7. The position, morphology, and size of the pancreas, bilateral adrenal glands, and kidneys are normal; densities are uniform, and radiotracer distribution is normal. Mild retention of radioactive urine is observed in the bilateral renal calyces, pelvis, and the middle-to-lower segments of the ureters. The gallbladder is not distended, the wall is not thickened, and radiotracer distribution is normal. The gastric wall is smooth, and no abnormal radiotracer distribution is observed. Segmental uptake is visible in parts of the intestines. The prostate is not enlarged, radiotracer distribution

Figure 8. Qualitative comparison of generation results with (w/) and without (w/o) the explicit stop token. Note that for the "w/ Stop Token" setting, generation is truncated immediately upon predicting the specific end-of-report token (e.g., "[end-of-report]"). For brevity, only the beginning and ending sentences are displayed. Red text highlights hallucinations.