# ReMAP-PET: Beyond Visual Understanding - Learning Region-Guided Metabolic Alignment Semantics from Brain PET

**Anonymous ACL submission**


## Abstract

Positron Emission Tomography (PET) reveals brain metabolism and is clinically central to neurodegenerative disease assessment, yet existing 3D brain foundation models treat PET as generic volumetric data, missing the structured regional metabolic information that distinguishes it from structural neuroimaging. To address these limitations, we propose ReMAP-PET, a framework that moves beyond visual encoding by supervising a partially-tuned MedicalNet 3D ResNet-50 with brain regional standardized uptake value ratio (SUVR) profiles through joint regression and contrastive objectives, enabling the encoder to learn the metabolic semantics underlying PET modality. On 1015 paired PET–SUVR samples, ReMAP-PET achieves 0.070 SUVR MAE and 77.8% PET→SUVR Recall@1, substantially outperforming five frozen pretrained baselines. We further connect the metabolic embedding to clinical language via contrastive alignment with frozen BioClinicalBERT and demonstrate end-to-end PET-to-report generation through SUVR-constrained verbalization. Linear probing on diagnostic classification and cognitive regression tasks confirms that the embeddings retain clinically relevant information without task-specific fine-tuning. Our results show that grounding PET encoders in regional metabolic semantics—rather than treating PET as generic volumetric data—yields representations that are structured, interpretable, and language-compatible, pointing to a new direction for metabolic-aware PET understanding.


## 1 Introduction

Accurately modeling brain metabolic patterns is a core challenge in functional neuroimage analysis and neurodegenerative disease research, with important implications for understanding disease progression and supporting clinical diagnosis (Perovnik et al., 2023). As one of the clinical gold-standard functional modalities for dementia diagnosis, fluorodeoxyglucose (FDG)-PET reflects neuronal activity through glucose metabolism. Unlike structural MRI, which captures anatomical changes, FDG-PET reveals early functional abnormalities and cross-regional metabolic degeneration patterns, and has been widely shown to be closely associated with early screening, disease progression, and cognitive decline in Alzheimer's disease (AD) (Xie et al., 2024). More importantly, PET possesses explicit quantitative medical properties: standardized uptake value ratios (SUVR) measured across brain regions provide physiologically meaningful descriptions of regional metabolism (Teune et al., 2010). For example, AD related hypometabolism typically first appears in the posterior cingulate cortex, precuneus, and temporoparietal regions. Compared with raw voxel intensities, these regional metabolic measurements exhibit substantially stronger structural organization and clinical interpretability (Bailly et al., 2015; Gunn et al., 2015). Learning clinically interpretable metabolic representations from PET therefore matters directly for downstream cognitive disease analysis.

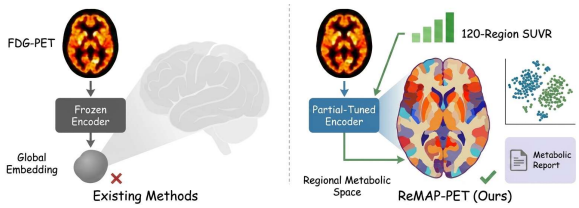


Figure 1: Existing methods treat PET as generic 3D volumes, extracting visual features that lack metabolic semantics (left). ReMAP-PET introduces region-level metabolic supervision to teach the encoder the physiological meaning behind PET imaging (right), producing representations that encode both visual and metabolic structure and enable cross-modal retrieval, clinical interpretation, and report generation.

In recent years, the development of pretrained 3D medical image encoders has made brain representation learning increasingly feasible, enabling

effective transfer to downstream neuroimaging tasks under limited annotation settings (Yang et al., 2024; Fang et al., 2025). However, nearly all existing frameworks are still built around structural MRI, with objectives primarily focused on visual texture modeling, global image representation learning, or anatomical structure reconstruction, while lacking explicit modeling of brain metabolic patterns themselves (Zhou et al., 2025; Falconnier, 2025). As a result, when directly applied to PET scans, these methods still tend to treat PET as an ordinary 3D visual image without leveraging the metabolic information that PET inherently provides. This limits the model's ability to capture the spatial organization of brain metabolism and to align representations with the structured clinical signals available in PET. A natural question is whether pretrained encoders can retain their general visual capability while also learning to encode the regional metabolic structure specific to PET. Rather than relying solely on image-level supervision, we aim to map PET into an interpretable metabolic space constrained by region-level measurements, so that the resulting representations capture neurodegenerative functional patterns and can support PET retrieval, automated report generation, and vision-language model (VLM) integration.

To address these challenges, we propose **ReMAP-PET** (**Re**gion-guided **M**etabolic **A**lignment with **P**artial-tuned **PET** Encoders), a method that grounds PET representation learning in regional metabolic activity. As illustrated in Figure 1, ReMAP-PET uses 120-region SUVR profiles as structured supervision to reshape the embedding space of a pretrained MedicalNet 3D ResNet-50 (Chen et al., 2019), updating only the final residual stage (`layer4`) while preserving generic anatomical knowledge in the earlier layers. A joint objective combining SUVR regression and bidirectional PET–SUVR contrastive alignment produces embeddings that are both numerically predictive and semantically structured. We further connect this metabolic representation to clinical language: lightweight projection heads align the frozen PET embedding with region-text summaries encoded by frozen BioClinicalBERT (Alsentzer et al., 2019), and an end-to-end report generation pipeline verbalizes the predicted SUVR profile into clinical-style metabolic summaries whose factual content is fully constrained by the underlying measurements.

Our experiments show three main results. First, regression alone produces accurate SUVR predictions but a degenerate embedding space; the contrastive term is what creates the structure needed for retrieval and downstream transfer. Second, partial tuning is architecture-dependent: the same last-block recipe that produces large gains on MedicalNet's ResNet yields negligible or negative effects on ViT-based and U-Net-based encoders, suggesting that ResNet's feature hierarchy provides a uniquely suitable bottleneck for metabolic adaptation. Third, linear probing on seven clinical tasks confirms that the metabolic embeddings carry clinically relevant information without task-specific fine-tuning, and the gains in metabolic alignment do not come at the expense of clinical signal.

Our contributions are as follows:

- We propose a metabolic-aware representation learning paradigm for FDG-PET: instead of treating PET as generic volumetric data, we use 120-region SUVR profiles as structured metabolic supervision with joint regression and contrastive objectives, enabling the encoder to learn regional metabolic semantics beyond visual features.
- We show that partial tuning of the final ResNet stage is uniquely effective for PET metabolic adaptation—the same recipe does not transfer to ViT or U-Net architectures—providing an empirical finding about when and why partial tuning works.
- We connect the learned metabolic embedding to clinical language through PET–text contrastive alignment and SUVR-constrained report generation, completing an image-to-representation-to-language pipeline for FDG-PET.

## 2 Related Work

**3D brain foundation models.** Pretrained 3D medical and brain encoders have become common in neuroimaging representation learning. MedicalNet provides a general 3D transfer-learning backbone, while BrainIAC (Tak et al., 2026), BrainFM (Liu et al., 2025), AnatCL (Barbano et al., 2026), and BrainSegFounder (Cox et al., 2024) explore different forms of brain-specific pretraining. Other volumetric backbones such as SAM-Med3D (Wang et al., 2025) and Swin-UNETR (Hatamizadeh et al., 2021) are also widely used, with recent surveys summarizing this line of work. However, these models are largely designed for structural MRI, segmentation, or generic 3D representation learning, and do not explicitly

model the regional metabolic patterns that are central to FDG-PET.

**Medical vision-language alignment.** Image–text contrastive learning has been widely used to align visual features with language (Radford et al., 2021). Medical variants such as ConVIRT (Zhang et al., 2022), MedCLIP (Wang et al., 2022), PubMedCLIP (Eslami et al., 2023), and BiomedCLIP (Zhang et al., 2024) extend this idea to reports, captions, or biomedical text. Most of them focus on 2D modalities and use report- or label-level supervision. In the neuroimaging domain, recent efforts explore ROI-level multimodal alignment and cross-modality translation using structured regional features (Xie et al., 2023), though primarily for structural MRI rather than PET. For PET/CT reporting, region-aware models such as PETAR (Maqbool et al., 2025) and Reg2RG (Chen et al., 2025) generate clinical text with explicit spatial grounding; however, these rely on free-form supervision and do not constrain report content to measured biomarkers. FDG-PET offers a different setting: language can be grounded in quantitatively measured regional metabolism, and the report content can be fully constrained by the SUVR measurement to avoid hallucination—a safety property that free-form supervision does not provide.

**Parameter-efficient tuning.** Full fine-tuning is often difficult in medical imaging because datasets are small and heterogeneous. Parameter-efficient methods, including adapters (Houlsby et al., 2019), LoRA (Hu et al., 2021), prompt tuning (Lester et al., 2021), and visual prompt tuning (Jia et al., 2022), adapt pretrained models while keeping most parameters fixed. Similar ideas have been used in medical image analysis, especially for adapting large segmentation models with limited annotations (Zhang and Liu, 2023). However, how partial tuning should be applied to 3D PET encoders remains less well studied.

**Clinical semantics and structured biomarkers.** Clinical text and structured biomarkers have received increasing attention in neuroimaging analysis. BERT (Devlin et al., 2019) and BioClinicalBERT have shown the usefulness of pretrained language models for medical semantics. In brain PET, region-level measurements such as SUVR provide physiologically meaningful descriptions of metabolic status. Existing methods still mainly focus on image–text or image–label alignment, while joint modeling of 3D PET and structured metabolic profiles remains underexplored.

# 3 Methods

## 3.1 Problem Setup

As shown in Figure 2, the pipeline has two training stages followed by downstream probing. Each subject is a triplet $(x_i, s_i, t_i)$: a 3D FDG-PET volume $x_i$, a 120-region SUVR vector $s_i \in \mathbb{R}^{120}$ derived from the brain atlas (Rolls et al., 2015), and a short text summary $t_i$ deterministically generated from $s_i$. In Stage 1, we train a PET encoder $f_\theta$ jointly with a SUVR encoder $g_\phi$ so that both modalities map into a shared embedding space; supervision comes from SUVR regression together with a cross-modal contrastive term. In Stage 2, $f_\theta$ and a pretrained biomedical text encoder are both frozen, and only two small projection heads are trained to align the PET embedding with $t_i$. We then evaluate the learned features by linear probing on clinical endpoints with no task-specific fine-tuning. Dataset statistics, splits, and preprocessing are given in Appendix A.

## 3.2 ReMAP-PET: Region-Grounded Metabolic Alignment

The PET encoder in ReMAP-PET is a MedicalNet 3D ResNet-50. Rather than fine-tuning the whole network, we keep all early stages frozen and update only the last residual stage:

$$\theta = \{\theta^{\text{frozen}}_{\text{stem:layer3}},\ \theta^{\text{trainable}}_{\text{layer4}}\}. \tag{1}$$

The intuition is straightforward: the early layers already encode generic anatomical structure from large-scale pretraining, and what is missing is a way to re-weight the highest-level features toward metabolic patterns specific to PET. We find in Section 4 that this recipe is not universal. Applying the same protocol to ViT-based backbones (BrainIAC, SAM-Med3D) or to BrainFM's U-Net yields much smaller, and sometimes negative, gains. ResNet's last stage seems to act as a natural semantic bottleneck, which the transformer and U-Net backbones do not provide in the same way.

The SUVR vector is encoded by a small MLP $g_\phi$ trained from scratch, giving $z_i^{\text{suvr}} = g_\phi(s_i)$. On the PET side, a linear head $h_\psi$ predicts the SUVR vector from the embedding, $\hat{s}_i = h_\psi(z_i^{\text{pet}})$, and is supervised by mean squared error,

$$\mathcal{L}_{\text{reg}} = \frac{1}{B}\sum_{i=1}^{B} \|\hat{s}_i - s_i\|_2^2. \tag{2}$$

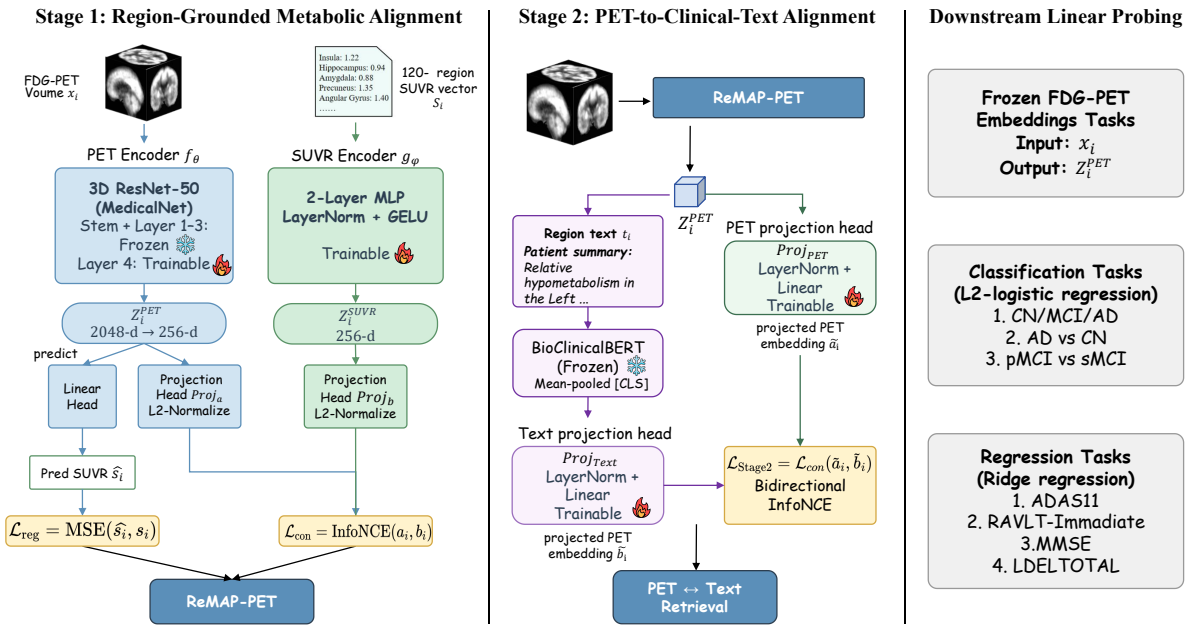


Figure 2: Overview of ReMAP-PET. **Stage 1** (left) aligns a partially-tuned 3D PET encoder with structured 120-region SUVR profiles via joint regression and contrastive objectives. **Stage 2** (center) connects the frozen metabolic embedding to clinical language through lightweight projection heads paired with frozen BioClinicalBERT. **Downstream probing** (right) evaluates the learned representations on diagnostic classification and cognitive regression tasks with no encoder fine-tuning.

The regression loss alone is enough to make the PET embedding numerically predictive of the SUVR profile, but it does not by itself produce an embedding space with a useful global structure. We therefore add a bidirectional contrastive term between normalized projections $a_i, b_i$ of the PET and SUVR embeddings, with similarity $\ell_{ij} = a_i^\top b_j / \tau$, where $\tau$ is a learnable scalar initialized to 0.07 that controls the sharpness of the softmax distribution over negative pairs and is updated jointly with the model parameters via gradient descent:

$$\mathcal{L}_{\text{con}} = \tfrac{1}{2}\big[\text{CE}(\ell, y) + \text{CE}(\ell^\top, y)\big], y_i = i. \quad (3)$$

The full first-stage objective is $\mathcal{L}_{\text{stage1}} = \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}$, with the two weights tuned on validation; we report the chosen values together with the ablation in Section 4.

To put ReMAP-PET in context, we compare it against five pretrained 3D encoders spanning ResNet, ViT, U-Net, and Swin architectures (Section 4.1). Each baseline is evaluated both fully frozen and with the same last-block partial-tuning recipe; the architecture-specific definitions of "last block" and the full comparison are reported in Section 4.2.

### 3.3 Linking PET to Clinical Text

The second stage connects the metabolic embedding to natural language. Writing free-form radiology-style text for each scan would risk hallucinated findings, so we instead generate $t_i$ from $s_i$ by a fixed rule: we rank the 120 regions by SUVR, take the five lowest and five highest, map their identifiers to readable English names, and insert them into a short template that reports relative hypo- and hyper-metabolism. An optional LLM rewriting step is allowed only to improve fluency; it is explicitly forbidden to introduce regions, numbers, or diagnostic claims that are not already in the template. As a result, every factual statement in $t_i$ is determined by the SUVR measurement. The template and rewriting prompt are reproduced in Appendix A.

For alignment, we use BioClinicalBERT as the text encoder and take the mean-pooled last hidden state as the sentence representation. Both the text encoder and the first-stage PET encoder $f_\theta$ are kept frozen, and we train only two small projection heads (LayerNorm followed by a linear layer),

$$\tilde{a}_i = \text{Proj}_{\text{pet}}(z_i^{\text{pet}}), \qquad \tilde{b}_i = \text{Proj}_{\text{text}}(\bar{h}_i^{\text{text}}), \quad (4)$$

using the same bidirectional contrastive loss as in the first stage. Because nothing inside the two

encoders changes, any cross-modal alignment that emerges here can only come from the structure that the first stage has already put into the PET embedding space; this makes Stage 2 a fairly direct test of how well that structure transfers to language.

### 3.4 Clinical Downstream Probing

To assess whether the metabolic embeddings retain clinically relevant information, we freeze the Stage 1 encoder and fit linear probes on seven clinical endpoints: three diagnostic classification tasks (CN/MCI/AD, AD vs. CN, and pMCI vs. sMCI conversion) evaluated with logistic regression, and four cognitive score regressions (ADAS11, MMSE, RAVLT-Immediate, LDELTOTAL) evaluated with ridge regression. The regularization strength is selected on the validation split and each probe is evaluated once on the held-out test set; no encoder weights are updated. As an empirical ceiling, we apply the same probes to the ground-truth SUVR vectors. Hyperparameter grids and additional evaluation details are given in Appendix B.

## 4 Experiments

### 4.1 Setup

We work with an internal cohort of 1015 subjects, each providing a paired FDG-PET scan and 120-region SUVR profile, split by subject into 710 training, 152 validation, and 153 test cases. ADNI-derived clinical labels are matched to all subjects but are never seen during Stage 1 or Stage 2 training; they are used only for the linear probes in Section 4.5. We compare ReMAP-PET against MedicalNet, BrainIAC, BrainFM, SAM-Med3D, and SwinUNETR, each evaluated fully frozen with only an MLP probe trained on top. For Stage 2, every PET encoder is paired with frozen BioClinicalBERT. Optimization settings, input resolutions, and the full list of metrics are given in Appendix A; unless stated otherwise, all numbers are reported on the held-out test split.

### 4.2 Stage 1: Aligning PET with Regional Metabolism

Table 1 compares ReMAP-PET against the five frozen baselines on the eight Stage 1 metrics. ReMAP-PET is the strongest method across the board. Against the best frozen baseline (BrainFM, a recent brain foundation model), it cuts SUVR MAE by roughly 28% (from 0.097 to 0.070) and more than doubles PET→SUVR Recall@1 (from 0.37 to 0.78). Frozen MedicalNet, which shares the same backbone as ReMAP-PET, sits near the bottom of the table; its features alone clearly do not encode PET-specific metabolic semantics, and the gap to ReMAP-PET is entirely attributable to the partial-tuning recipe described in Section 3. Figure 3 visualizes this contrast: the frozen MedicalNet similarity matrix is near-uniform, whereas ReMAP-PET produces a strong diagonal indicating accurate cross-modal pairing. Figure 4 further shows that the predicted SUVR profile closely matches the ground truth across all 120 regions for a representative test subject.

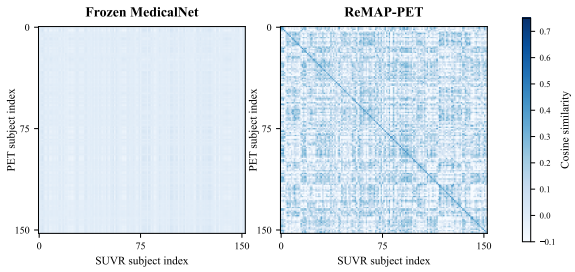


Figure 3: PET–SUVR cosine similarity matrices on the 153-subject test set. Left: Frozen MedicalNet produces near-uniform similarity (R@1 = 2.6%). Right: ReMAP-PET yields a strong diagonal, indicating that each PET embedding is most similar to its paired SUVR embedding (R@1 = 77.8%).

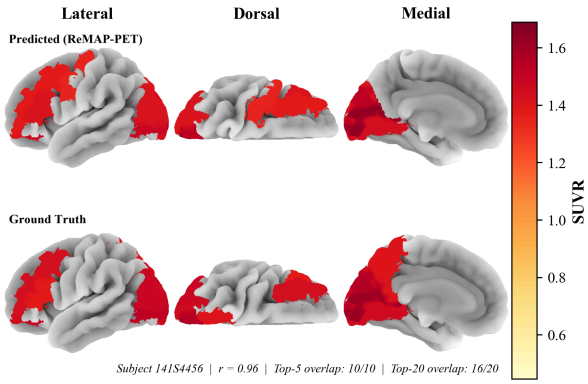


Figure 4: Predicted versus ground-truth 120-region SUVR profiles for a representative test subject. The close alignment between predicted (blue) and true (orange) SUVR values across all brain regions illustrates the quality of ReMAP-PET's regional metabolic reconstruction.

A key question is whether these gains come from the partial-tuning recipe itself or from the specific backbone. Table 2 applies the same last-block partial-tuning protocol to all five encoder architectures under identical training conditions.

| Encoder | *SUVR Prediction* | | *Correlation* | | *Retrieval R@1* | | *Top-5 Region* | |
|---|---|---|---|---|---|---|---|---|
| | MAE↓ | RMSE↓ | Pearson↑ | Spearman↑ | P→S↑ | S→P↑ | High↑ | Low↑ |
| MedicalNet | 0.117±.009 | 0.153 | 0.739 | 0.845 | 0.026 | 0.046 | 0.429 | 0.729 |
| BrainIAC | 0.125±.009 | 0.161 | 0.707 | 0.815 | 0.013 | 0.020 | 0.343 | 0.680 |
| BrainFM | 0.097±.007 | 0.126 | 0.838 | 0.885 | 0.373 | 0.458 | 0.471 | 0.736 |
| SAM-Med3D | 0.115±.009 | 0.147 | 0.829 | 0.858 | 0.163 | 0.196 | 0.463 | 0.724 |
| SwinUNETR | 0.100±.008 | 0.129 | 0.824 | 0.881 | 0.118 | 0.144 | 0.507 | 0.744 |
| **ReMAP-PET** | **0.070**±.005 | **0.089** | **0.920** | **0.935** | **0.778** | **0.922** | **0.637** | **0.766** |

Table 1: Stage 1 results on the held-out test set (153 subjects). Baselines use the named encoder fully frozen with only an MLP probe trained on top; ReMAP-PET uses MedicalNet 3D ResNet-50 with `layer4` unfrozen. P→S / S→P are PET-to-SUVR / SUVR-to-PET retrieval. Best in , second best underlined. MAE column shows bootstrap 95% CI half-widths ($B$=1000).

The recipe broadly helps MedicalNet's ResNet across all eight metrics, partially improves SAM-Med3D and SwinUNETR on retrieval, but degrades BrainFM on most metrics and barely affects BrainIAC. A plausible explanation is that ResNet's final bottleneck stage concentrates high-level semantic features in a single, clearly delineated block, making it a natural site for task-specific adaptation. By contrast, ViT self-attention blocks distribute information more uniformly across layers, and U-Net skip connections couple the deepest stage to the decoder, so tuning one block in isolation is less effective. This finding suggests that partial tuning is not a universal parameter-efficient alternative to full fine-tuning; its success depends on whether the architecture provides a clean semantic bottleneck for adaptation.

### 4.3 Stage 2: Aligning PET with Clinical Text

Stage 2 is designed as a diagnostic of the Stage 1 embedding: both the PET encoder (frozen after Stage 1) and BioClinicalBERT are fixed, so any retrieval signal must come from the metabolic structure that Stage 1 has learned. This means Stage 2 does not introduce new NLP capability; rather, it tests whether the regional patterns captured in Stage 1 are organized well enough to transfer to a language-grounded space through small projection heads alone. Table 3 shows that ReMAP-PET achieves the best retrieval accuracy in both directions and the highest factual overlap between the retrieved and reference region summaries.

The absolute Recall@1 numbers look small, but the candidate pool contains 153 test subjects, so random chance is below $0.7\%$. Across all encoders the Stage 2 ranking mirrors Stage 1, with BrainFM again the strongest baseline and BrainIAC the weakest. The high region-overlap scores (Low $0.72$, High $0.58$) further suggest that even when ReMAP-PET does not retrieve the exact paired summary, the retrieved text describes a metabolically similar subject—a clinician receiving such a summary is getting actionable region-level information regardless of whether it comes from the exact paired subject, which is arguably the more clinically meaningful notion of correctness. Extended retrieval numbers (R@5) are reported in Appendix C.

### 4.4 From Retrieval to Report Generation

A natural follow-up to retrieval is whether the metabolic representation can also drive end-to-end report generation: given a PET scan alone, can we produce a metabolic summary that says the right thing? We test this by predicting the 120-region SUVR vector with ReMAP-PET, taking the five lowest- and five highest-metabolism regions, and passing them through Qwen3 (Yang et al., 2025) as a surface realizer; applying the same procedure to the ground-truth SUVR yields a per-subject reference paragraph.

Because both the predicted and reference reports are produced by the same verbalizer over a discrete region list, the only thing that varies between them is which regions the PET encoder identifies. These overlap with the reference at exactly the Top-5 high-/low rates already reported for Stage 1 (Table 1), confirming that the verbalization step introduces no additional factual loss. Against ground-truth reference reports, the generated summaries achieve BLEU-1 of $0.631$ and ROUGE-L of $0.544$; residual surface differences reflect phrasing variation rather than factual disagreement. We view this as a read-out of the learned metabolic representation

| Backbone | *SUVR Prediction* | | *Correlation* | | *Retrieval R@1* | | *Top-5 Region* | |
|---|---|---|---|---|---|---|---|---|
| | MAE↓ | RMSE↓ | Pearson↑ | Spearman↑ | P→S↑ | S→P↑ | High↑ | Low↑ |
| **ReMAP-PET (ResNet)** | **0.070** | **0.089** | **0.920** | **0.935** | **0.778** | **0.922** | **0.637** | **0.766** |
| MedicalNet (ResNet) – | 0.117 | 0.153 | 0.739 | 0.845 | 0.026 | 0.046 | 0.429 | 0.729 |
| BrainIAC (ViT) – | 0.125 | 0.161 | 0.707 | 0.815 | 0.013 | 0.020 | 0.343 | 0.680 |
| BrainIAC (ViT) • | 0.119 | 0.155 | 0.731 | 0.847 | 0.007 | 0.026 | 0.401 | 0.729 |
| BrainFM (U-Net) – | 0.097 | 0.126 | 0.838 | 0.885 | 0.373 | 0.458 | 0.471 | 0.736 |
| BrainFM (U-Net) • | 0.104 | 0.134 | 0.811 | 0.843 | 0.261 | 0.320 | 0.475 | 0.682 |
| SAM-Med3D (ViT) – | 0.115 | 0.147 | 0.829 | 0.858 | 0.163 | 0.196 | 0.463 | 0.724 |
| SAM-Med3D (ViT) • | 0.092 | 0.117 | 0.862 | 0.881 | 0.255 | 0.346 | 0.515 | 0.731 |
| SwinUNETR (Swin) – | 0.100 | 0.129 | 0.824 | 0.881 | 0.118 | 0.144 | 0.507 | 0.744 |
| SwinUNETR (Swin) • | 0.100 | 0.130 | 0.824 | 0.864 | 0.235 | 0.307 | 0.418 | 0.726 |

Table 2: Partial-tuning recipe applied across five encoder architectures. Gray : frozen encoder + MLP probe; blue : last-block partial tuning. Green marks metrics that improve under partial tuning, red marks degradation. The first row (MedicalNet frozen) is the starting point of ReMAP-PET.

| Encoder | *R@1* | | *Region Overlap* | |
|---|---|---|---|---|
| | P→T↑ | T→P↑ | Low↑ | High↑ |
| MedicalNet | 0.020 | 0.020 | 0.658 | 0.435 |
| BrainIAC | 0.007 | 0.007 | 0.659 | 0.264 |
| BrainFM | 0.026 | 0.052 | 0.698 | 0.460 |
| SAM-Med3D | 0.052 | 0.026 | 0.663 | 0.465 |
| SwinUNETR | 0.046 | 0.052 | 0.694 | 0.459 |
| **ReMAP-PET** | **0.098** | **0.144** | **0.719** | **0.578** |

Table 3: Stage 2 PET-text retrieval. All rows pair the named PET encoder (frozen) with frozen BioClinicalBERT. P→T / T→P are PET-to-text / text-to-PET retrieval; Region Overlap measures factual agreement between the query subject's reference summary and the summary retrieved for it. R@5 values are in Appendix C.

rather than a separate generation model. A per-subject prediction example and qualitative report comparison are provided in Appendix F.

### 4.5 Clinical Probing

Table 4 reports linear probes on the three diagnostic classification tasks and on ADAS11—the cognitive score most directly tied to disease severity—with probes trained on the ground-truth 120-region SUVR vector as a non-trivial upper bound. The remaining three cognitive regression tasks (MMSE, RAVLT-Immediate, LDELTOTAL) are deferred to Appendix D.

ReMAP-PET achieves the highest AUROC on all three classification tasks: on the binary diagnostic extremes (AD vs. CN) it reaches 0.946, slightly above the SUVR ceiling (0.943). The gap is small (0.003) and within bootstrap uncertainty, but it is not contradictory: the SUVR ceiling reflects what a linear model can extract from 120 scalar values, whereas the 2048-dimensional PET embedding retains additional spatial and textural cues that the SUVR summary discards. ReMAP-PET is also the top encoder on the more challenging 3-way and pMCI vs. sMCI settings. We use linear probing rather than task-specific fine-tuning because our goal is to evaluate the intrinsic quality of the learned representation, not to maximize clinical performance; this is the standard protocol for representation evaluation in self-supervised learning. The regression numbers tell a more mixed story: no single encoder dominates and the SUVR ceiling is close to or slightly above every PET encoder. We read this as evidence that, once a backbone has reasonable 3D anatomical features, much of the clinically predictive signal is already accessible to a linear probe; the value added by ReMAP-PET is not primarily in raw clinical accuracy but in producing an embedding that is simultaneously good at SUVR prediction, retrieval, and downstream classification. The gains ReMAP-PET shows on Stage 1 metrics do not come at the cost of clinical signal—a failure mode we were initially concerned about.

### 4.6 The Role of Contrastive Alignment

Figure 5 sweeps the contrastive weight $\lambda_{\text{con}}$ with $\lambda_{\text{reg}} = 1.0$ fixed, using the MedicalNet `layer4` setup. The most informative comparison is at the two endpoints. With $\lambda_{\text{con}} = 0$ (pure regression), the model achieves its lowest SUVR error (MAE 0.055) but its retrieval collapses to chance (R@1

| Encoder | *AUROC*↑ | | | *ADAS11* |
|---|---|---|---|---|
| | 3-way | AD/CN | pMCI/sMCI | MAE↓ |
| MedicalNet | 0.747±.060 | 0.929±.069 | 0.719 | 3.990 |
| BrainIAC | 0.652±.064 | 0.809±.114 | 0.660 | 4.598 |
| BrainFM | 0.693±.067 | 0.856±.099 | 0.675 | 4.282 |
| SAM-Med3D | 0.700±.066 | 0.851±.108 | 0.759 | 3.953 |
| SwinUNETR | 0.684±.072 | 0.912±.078 | 0.737 | 4.072 |
| **ReMAP-PET** | **0.752**±.062 | **0.946**±.059 | **0.787** | **3.873** |
| *SUVR (ceiling)* | *0.774* | *0.943* | *0.835* | *3.820* |

Table 4: Linear probing on three diagnostic tasks and ADAS11. The SUVR row uses the ground-truth 120-region SUVR as an upper bound. 3-way and AD/CN columns show bootstrap 95% CI half-widths ($B$=1000). Full regression results are in Appendix D.

$\approx 0.007$, essentially $1/153$). The PET encoder has learned to predict SUVR values numerically, but its embedding space carries no useful global structure. Switching on the contrastive term at $\lambda_{\text{con}} = 0.1$ immediately raises R@1 to 0.68, and increasing it further trades a small amount of regression accuracy for additional retrieval quality.

We pick $\lambda_{\text{con}} = 0.2$ because it sits in the elbow of this tradeoff: regression MAE rises by less than a third compared to the pure-regression model, while R@1 reaches 0.78. More broadly, the ablation makes a point worth stating directly: regression alone is enough to fit SUVR values, but it is not enough to produce a representation that is useful for anything else. The contrastive term is what turns SUVR prediction into a structured metabolic embedding.

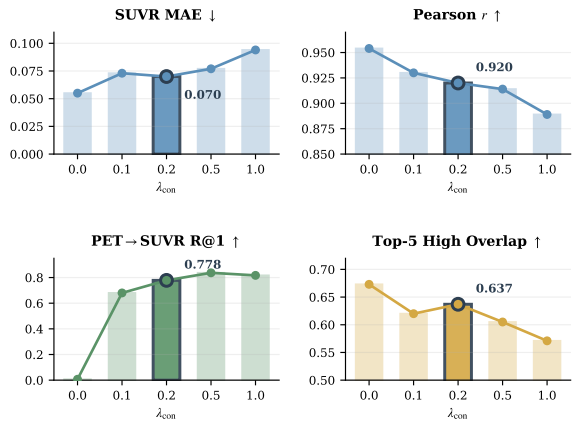


Figure 5: Contrastive weight ablation ($\lambda_{\text{reg}} = 1.0$ fixed, MedicalNet `layer4`). The highlighted bar marks the selected configuration ($\lambda_{\text{con}} = 0.2$). Pure regression ($\lambda_{\text{con}} = 0$) yields the lowest MAE but near-zero retrieval; adding contrastive alignment dramatically improves R@1 with modest regression cost.

Beyond Recall@1, Table 5 reports MRR and median rank for both retrieval directions. ReMAP-PET places the correct partner at rank 1 for the majority of subjects (MRR 0.86 PET→SUVR, 0.95 SUVR→PET; median rank 1 in both directions), while BrainFM frozen is the closest baseline (median rank 2) and BrainIAC barely aligns at all (median rank 75).

| Model | *P→S* | | *S→P* | |
|---|---|---|---|---|
| | MRR↑ | MedR↓ | MRR↑ | MedR↓ |
| **ReMAP-PET** | **0.857** | **1** | **0.949** | **1** |
| BrainFM | 0.533 | 2 | 0.597 | 2 |
| BrainFM | 0.446 | 3 | 0.480 | 3 |
| SAM-Med3D | 0.299 | 7 | 0.363 | 4 |
| SAM-Med3D | 0.446 | 3 | 0.499 | 3 |
| SwinUNETR | 0.244 | 9 | 0.280 | 7 |
| SwinUNETR | 0.382 | 5 | 0.462 | 3 |
| BrainIAC | 0.037 | 75 | 0.067 | 50 |

Table 5: Extended PET–SUVR retrieval metrics. Gray : frozen encoder; blue : last-block partial tuning. MedR is the median rank of the correct cross-modal partner. **Bold**: best; underline: second best.

## 5 Conclusion

We presented ReMAP-PET, an approach that grounds FDG-PET representation learning in regional metabolic semantics. By supervising only the final ResNet stage with 120-region SUVR profiles, the method learns embeddings that are predictive of regional metabolism, retrievable across modalities, and clinically useful under linear probing without task-specific fine-tuning. A comparison across five encoder families shows that the effectiveness of partial tuning depends on whether the backbone provides a clean semantic bottleneck, an observation that may carry over to other 3D medical adaptation settings. The learned metabolic structure transfers to a text-grounded space through frozen projection heads, and predicted SUVR profiles can be verbalized into reports whose content is fully determined by the measurement, sidestepping hallucination risks common in free-form medical text generation. We hope this work encourages further exploration of structured physiological supervision for building interpretable and language-compatible representations in functional neuroimaging. Future work should validate the approach on external cohorts with different tracers and scanners.

**Limitations.** Several limitations should be noted. First, all experiments use a single ADNI-derived cohort. Since PET acquisition requires the injection of radioactive tracers into participants, collecting PET imaging data is considerably more complex and costly than acquiring conventional clinical imaging such as MRI. As a result, publicly available brain imaging datasets containing PET scans are substantially smaller and less abundant worldwide. Under this practical constraint, this study first validated the proposed framework using the largest available collection of FDG-PET scans from the largest neuroimaging databases, ADNI. Nevertheless, the generalizability of the model to other PET tracers, imaging protocols, or populations still requires further evaluation on external datasets. Second, the Stage 2 text descriptions are generated from a deterministic template rather than from free-form clinical reports, which limits the linguistic diversity of the training signal. Third, the partial-tuning finding is empirical—we provide a structural explanation for why ResNet benefits while ViT and U-Net do not, but a formal theoretical account is lacking. Finally, the current evaluation is restricted to linear probing; task-specific fine-tuning or integration into a full clinical decision pipeline may reveal different performance patterns.

## References


Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 72–78.

Matthieu Bailly, Christophe Destrieux, Caroline Hommet, Karl Mondon, Jean-Philippe Cottier, Emilie Beaufils, Emilie Vierron, Johnny Vercouillie, Méziane Ibazizene, Thierry Voisin, and 1 others. 2015. Precuneus and cingulate cortex atrophy and hypometabolism in patients with alzheimer's disease and mild cognitive impairment: Mri and 18f-fdg pet quantitative analysis using freesurfer. *BioMed research international*, 2015(1):583931.

Carlo Alberto Barbano, Matteo Brunello, Benoit Dufumier, and Marco Grangetto. 2026. Anatomical foundation models for brain MRIs. *Pattern Recognition Letters*, 199:178–184.

Sihong Chen, Kai Ma, and Yefeng Zheng. 2019. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*.

Zhixuan Chen, Yequan Bie, Haibo Jin, and Hao Chen. 2025. Large language model with region-guided referring and grounding for ct report generation. *Preprint*, arXiv:2411.15539.

Joseph Cox, Peng Liu, Skylar E Stolte, Yunchao Yang, Kang Liu, Kyle B See, Huiwen Ju, and Ruogu Fang. 2024. Brainsegfounder: Towards 3d foundation models for neuroimage segmentation. *Medical Image Analysis*, 97:103301.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Sedigheh Eslami, Christoph Meinel, and Gerard de Melo. 2023. PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1181–1193, Dubrovnik, Croatia. Association for Computational Linguistics.

Pierre Falconnier. 2025. Representation learning for 3d brain imaging: A comparative study of pre-trained encoders, foundation models and self-supervised learning methods. Master's thesis, KTH Royal Institute of Technology.

Mengjie Fang, Zipei Wang, Sitian Pan, Xin Feng, Yunpeng Zhao, Dongzhi Hou, Ling Wu, Xuebin Xie, Xu-Yao Zhang, Jie Tian, and 1 others. 2025. Large models in medical imaging: Advances and prospects. *Chinese Medical Journal*, 138(14):1647–1664.

Roger N Gunn, Mark Slifstein, Graham E Searle, and Julie C Price. 2015. Quantitative imaging of protein targets in the human brain with pet. *Physics in Medicine & Biology*, 60(22):R363–R411.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. 2021. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. *Preprint*, arXiv:1902.00751.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. *Preprint*, arXiv:2203.12119.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *Preprint*, arXiv:2104.08691.

Peirong Liu, Oula Puonti, Xiaoling Hu, Karthik Gopinath, Annabel Sorby-Adams, and Juan Eugenio Iglesias. 2025. A modality-agnostic multi-task foundation model for human brain imaging. *Preprint*, arXiv:2509.00549.

Danyal Maqbool, Changhee Lee, Zachary Huemann, Samuel D. Church, Matthew E. Larson, Scott B. Perlman, Tomas A. Romero, Joshua D. Warner, Meghan Lubner, Xin Tie, Jameson Merkow, Junjie Hu, Steve Y. Cho, and Tyler J. Bradshaw. 2025. Petar: Localized findings generation with mask-aware vision-language modeling for pet automated reporting. *Preprint*, arXiv:2510.27680.

Matej Perovnik, Tomaž Rus, Katharina A Schindlbeck, and David Eidelberg. 2023. Functional brain networks in the evaluation of patients with neurodegenerative disorders. *Nature Reviews Neurology*, 19(2):73–90.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763.

Edmund T Rolls, Marc Joliot, and Nathalie Tzourio-Mazoyer. 2015. Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. *Neuroimage*, 122:1–5.

Divyanshu Tak, Biniam Garomsa, Anna Zapaishchykova, Tafadzwa Chaunzwa, Juan Carlos Climent Pardo, Zezhong Ye, John Zielke, Yashwanth Ravipati, Suraj Pai, Sri Vajapeyam, Maryam Mahootiha, Mitchell Parker, Luke Pike, Ceilidh Smith, Ariana Familiar, Kevin Liu, Sanjay Prabhu, Omar Arnaout, Pratiti Bandopadhayay, and 6 others. 2026. A generalizable foundation model for analysis of human brain MRI. *Nature Neuroscience*.

Laura K Teune, Anna L Bartels, Bauke M de Jong, Antoon TM Willemsen, Silvia A Eshuis, Jeroen J de Vries, Joost CH van Oostrom, and Klaus L Leenders. 2010. Typical cerebral metabolic patterns in neurodegenerative brain diseases. *Movement Disorders*, 25(14):2395–2404.

Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianbin Li, Jianpin Chen, Yanzhou Su, Ziyan Huang, and Yiqing Shen. 2025. Sam-med3d: a vision foundation model for general-purpose segmentation on volumetric medical images. *IEEE Transactions on Neural Networks and Learning Systems*, 36:17599–17612.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. *Preprint*, arXiv:2210.10163.

Guoyang Xie, Yawen Huang, Jinbao Wang, Jiayi Lyu, Feng Zheng, Yefeng Zheng, and Yaochu Jin. 2023. Cross-modality neuroimage synthesis: A survey. *Preprint*, arXiv:2202.06997.

Lijun Xie, Jihua Zhao, Ye Li, and Jie Bai. 2024. Pet brain imaging in neurological disorders. *Physics of Life Reviews*, 49:100–111.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Hongxi Yang, Chang Yuwen, Xuelian Cheng, Hengwei Fan, Xin Wang, and Zongyuan Ge. 2024. Deep learning: A primer for neurosurgeons. *Computational Neurosurgery*, pages 39–70.

Kaidong Zhang and Dong Liu. 2023. Customized segment anything model for medical image segmentation. *Preprint*, arXiv:2304.13785.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, and 5 others. 2024. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1).

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. 2022. Contrastive learning of medical visual representations from paired images and text. *Preprint*, arXiv:2010.00747.

Xinliang Zhou, Chenyu Liu, Zhisheng Chen, Kun Wang, Yi Ding, Ziyu Jia, and Qingsong Wen. 2025. Brain foundation models: A survey on advancements in neural signal processing and brain discovery. *IEEE Signal Processing Magazine*, 42(5):22–35.

# A Implementation Details

## A.1 Dataset and Preprocessing

The dataset contains 1015 paired PET–SUVR samples derived from an internal ADNI-based cohort, including subjects with Alzheimer's disease (AD), stable mild cognitive impairment (sMCI), progressive mild cognitive impairment (pMCI), and cognitively normal (CN) controls, split by subject into 710 training, 152 validation, and 153 held-out test cases. ADNI-derived clinical metadata are matched for all subjects and are used only for the linear probing experiments.

PET volumes are resampled to isotropic $1.5 \times 1.5 \times 1.5$ mm spacing, center-cropped or padded to $96^3$ voxels, and intensity-normalized to zero mean and unit variance per volume. SAM-Med3D uses $128^3$ inputs because of its learned positional embeddings. SUVR profiles are min–max normalized to $[0, 1]$ per subject. The background region is excluded, leaving $s_i \in \mathbb{R}^{120}$.

Per-subject min–max normalization removes absolute uptake differences caused by scanner variability, injection dose, and individual physiology, so that the model learns relative regional metabolic patterns rather than absolute tracer concentrations. All regression targets and prediction-error metrics (MAE, RMSE) are computed in this normalized [0,1] space. Figures that show SUVR profiles (e.g., Figure 4, Figure 6) display values mapped back to the original clinical scale for interpretability.

SUVR profiles are derived from the ADNI FDG-PET processing pipeline. Each PET volume is co-registered to the subject's MRI, spatially normalized to MNI space, and smoothed with an 8 mm FWHM Gaussian kernel. Regional uptake is extracted using the AAL3 atlas (Rolls et al., 2015) (120 cortical and subcortical regions after excluding the background label). SUVR values are computed as the ratio of regional uptake to a cerebellar reference region. No partial-volume correction is applied; harmonization across scanners relies on the per-subject normalization described above.

## A.2 Backbones and Baselines

The ReMAP-PET encoder uses MedicalNet 3D ResNet-50. The backbone consists of a stem block, four bottleneck stages, and global average pooling, yielding a 2048-dimensional feature vector before projection. ReMAP-PET keeps the stem through `layer3` frozen and updates only `layer4`. The SUVR encoder $g_\phi$ is a two-layer MLP with LayerNorm and GELU.

We compare against four additional pretrained 3D encoders: BrainIAC, BrainFM, SAM-Med3D, and SwinUNETR. Each backbone is evaluated in two regimes: fully frozen with an MLP head, and last-block partial tuning using the same training recipe as ReMAP-PET.

## A.3 Optimization

Stage 1 uses AdamW with learning rate $1 \times 10^{-5}$, weight decay $1 \times 10^{-4}$, batch size 4, and 50 epochs. The contrastive temperature $\tau$ is initialized to 0.07. Unless otherwise stated, the loss weights are $\lambda_{\text{reg}} = 1.0$ and $\lambda_{\text{con}} = 0.2$, selected from the sweep in Table 8.

Stage 2 uses AdamW with learning rate $1 \times 10^{-4}$, batch size 16, and 20 epochs. All experiments use mixed-precision training on NVIDIA A40 GPUs.

## A.4 Region-Text Construction

The controlled text summary used in Stage 2 is generated deterministically from the SUVR profile. For each subject, we rank the 120 regions by SUVR, select the five lowest- and five highest-metabolism regions, map atlas identifiers to readable English names, and insert them into the template:

> FDG-PET regional metabolic summary. Relatively low metabolism is observed in [low-5 regions]. Relatively high uptake is observed in [high-5 regions].

An optional Qwen3 rewriting step is used only as a surface realizer. The prompt instructs the model to preserve the given region names and polarity labels, and not to introduce new regions, measurements, or diagnostic claims.

# B Evaluation Protocols

## B.1 Stage 1 and Stage 2 Metrics

Stage 1 reports SUVR prediction error (MAE, RMSE), rank correlation between predicted and true SUVR profiles (Pearson $r$, Spearman $\rho$), bidirectional PET–SUVR retrieval, and Top-5 high/low-metabolism region overlap.

Stage 2 reports bidirectional PET–text retrieval and retrieved-text region overlap. The main text reports Recall@1 and factual region overlap; full Recall@5 results are given in Appendix C.

### B.2 Clinical Probing

For each Stage 1 encoder, PET embeddings are extracted with the encoder frozen. Classification probes use $L_2$-regularized logistic regression, with

$$C \in \{0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0\}.$$

Regression probes use Ridge regression, with

$$\alpha \in \{0.1, 0.3, 1.0, 3.0, 10.0, 30.0, 100.0\}.$$

The regularization strength is selected on the validation split, and the selected probe is evaluated once on the test split. No encoder parameters are updated during probing.

Classification tasks are evaluated with AUROC in the main text; balanced accuracy and macro-F1 are also computed. Regression tasks are evaluated with MAE and Pearson $r$, with Spearman $\rho$ included in the appendix tables where useful.

## C Additional PET–Text Retrieval Results

Table 6 gives the full Stage 2 retrieval results, including Recall@5. These numbers complement the compact single-column table in the main text.

| Encoder | *P→T* | | *T→P* | | *Overlap* | |
|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@1 | R@5 | Low | High |
| MedicalNet | 0.020 | 0.144 | 0.020 | 0.092 | 0.658 | 0.435 |
| BrainIAC | 0.007 | 0.039 | 0.007 | 0.046 | 0.659 | 0.264 |
| BrainFM | 0.026 | 0.183 | 0.052 | 0.163 | 0.698 | 0.460 |
| SAM-Med3D | 0.052 | 0.170 | 0.026 | 0.163 | 0.663 | 0.465 |
| SwinUNETR | 0.046 | 0.196 | 0.052 | 0.144 | 0.694 | 0.459 |
| ReMAP-PET | 0.098 | 0.392 | 0.144 | 0.340 | 0.719 | 0.578 |

Table 6: Full Stage 2 PET–text retrieval results. All rows use frozen BioClinicalBERT on the text side; only the PET encoder changes. P→T denotes PET-to-text retrieval and T→P denotes text-to-PET retrieval.

## D Additional Clinical Probing Results

The main text reports the three diagnostic classification tasks and ADAS11. Table 7 reports the remaining cognitive regression tasks. The pattern is similar to ADAS11: ReMAP-PET is competitive, but no single encoder dominates every score.

## E Additional Ablation and Retrieval Details

### E.1 Full Loss-Weight Sweep

The main text uses a compact loss-weight table. Table 8 includes the full set of Stage 1 metrics for the $\lambda_{\text{con}}$ sweep.

### E.2 Extended PET–SUVR Retrieval Metrics

Table 9 reports the full retrieval statistics behind the compact retrieval discussion in the main text.

## F Per-Subject SUVR Prediction Example

Figure 6 shows a representative test subject (selected as the subject whose per-region Pearson $r$ is closest to the test-set median). The scatter plot confirms tight agreement between predicted and true SUVR across all 120 regions ($r = 0.948$). The Top-5 overlap bars show that ReMAP-PET correctly identifies 3 of 5 high-metabolism and 4 of 5 low-metabolism regions for this subject.

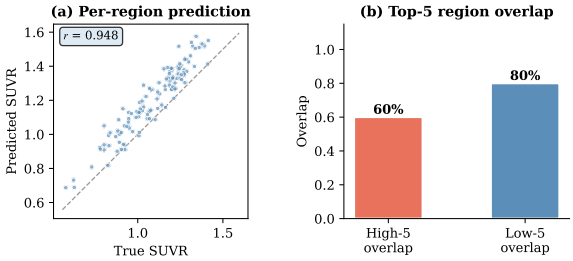


Figure 6: Per-subject SUVR prediction for a representative test subject (median Pearson $r$). (a) Predicted vs. true SUVR across 120 regions. (b) Top-5 high/low metabolism region overlap.

## G Report Verbalization Examples

The two boxes below show verbalized reports for the representative test subject in Figure 6. Both are produced by the same Qwen3 prompt; they differ only in whether the input regions come from predicted or ground-truth SUVR. The two reports agree on the main metabolic pattern (bilateral cerebellar and right temporal hypometabolism with frontoparietal preservation), with differences concentrated in nearby regions around the Top-5 boundary.

| Encoder | *MMSE* | | | *RAVLT Immediate* | | | *LDELTOTAL* | | |
|---|---|---|---|---|---|---|---|---|---|
| | MAE↓ | $r$ ↑ | $\rho$ ↑ | MAE↓ | $r$ ↑ | $\rho$ ↑ | MAE↓ | $r$ ↑ | $\rho$ ↑ |
| MedicalNet | **1.698** | 0.631 | 0.525 | 9.152 | 0.519 | 0.427 | **3.389** | 0.485 | 0.451 |
| BrainIAC | 2.044 | 0.380 | 0.322 | 9.762 | 0.386 | 0.283 | 3.581 | 0.342 | 0.323 |
| BrainFM | 1.969 | 0.506 | 0.410 | 9.170 | 0.472 | 0.379 | 3.562 | 0.384 | 0.370 |
| SAM-Med3D | 1.850 | 0.531 | 0.414 | **9.013** | 0.520 | 0.381 | 3.613 | 0.421 | 0.358 |
| SwinUNETR | 1.884 | 0.513 | 0.387 | 9.416 | 0.462 | 0.379 | 3.454 | 0.390 | 0.401 |
| ReMAP-PET | 1.715 | 0.620 | 0.520 | 9.047 | 0.547 | 0.430 | 3.409 | 0.477 | 0.439 |
| *SUVR* | *1.721* | *0.617* | *0.554* | *9.183* | *0.534* | *0.487* | *3.338* | *0.494* | *0.484* |

Table 7: Additional cognitive regression probing results. Bold marks the best PET encoder per MAE column; the SUVR row uses the ground-truth 120-region SUVR vector as an empirical ceiling.

| $\lambda_{\text{con}}$ | *SUVR Prediction* | | *Correlation* | | *Retrieval R@1* | | *Top-5 Region* | |
|---|---|---|---|---|---|---|---|---|
| | MAE↓ | RMSE↓ | Pearson↑ | Spearman↑ | P→S↑ | S→P↑ | High↑ | Low↑ |
| 0.0 (reg. only) | **0.055** | **0.071** | **0.954** | **0.954** | 0.007 | 0.007 | **0.673** | **0.800** |
| 0.1 | 0.073 | 0.096 | 0.930 | 0.941 | 0.680 | 0.915 | 0.620 | 0.773 |
| 0.2 | 0.070 | 0.089 | 0.920 | 0.935 | 0.778 | 0.922 | 0.637 | 0.766 |
| 0.5 | 0.077 | 0.098 | 0.914 | 0.934 | **0.837** | 0.915 | 0.605 | 0.753 |
| 1.0 | 0.094 | 0.119 | 0.889 | 0.910 | 0.817 | **0.928** | 0.571 | 0.760 |

Table 8: Full loss-weight ablation with $\lambda_{\text{reg}} = 1.0$ fixed. Pure regression yields the lowest SUVR error but gives near-random retrieval. The selected setting, $\lambda_{\text{con}} = 0.2$, balances regression accuracy and cross-modal structure.

| Model | *P→S* | | | *S→P* | | |
|---|---|---|---|---|---|---|
| | R@5↑ | MRR↑ | MedR↓ | R@5↑ | MRR↑ | MedR↓ |
| ReMAP-PET | **0.980** | **0.857** | **1** | **0.987** | **0.949** | **1** |
| BrainFM – | 0.732 | 0.533 | 2 | 0.765 | 0.597 | 2 |
| BrainFM • | 0.654 | 0.446 | 3 | 0.706 | 0.480 | 3 |
| SAM-Med3D – | 0.418 | 0.299 | 7 | 0.562 | 0.363 | 4 |
| SAM-Med3D • | 0.706 | 0.446 | 3 | 0.680 | 0.499 | 3 |
| SwinUNETR – | 0.353 | 0.244 | 9 | 0.412 | 0.280 | 7 |
| SwinUNETR • | 0.536 | 0.382 | 5 | 0.634 | 0.462 | 3 |
| BrainIAC • | 0.039 | 0.037 | 75 | 0.059 | 0.067 | 50 |

Table 9: Extended PET–SUVR retrieval metrics. – denotes a frozen encoder with an MLP head; • denotes last-block partial tuning. MedR is the median 1-based rank of the correct cross-modal partner.

**Generated Report (from ReMAP-PET Predicted SUVR)**

FDG-PET imaging demonstrates relatively reduced glucose metabolism involving the bilateral inferior cerebellar hemispheres and inferior vermis, as well as the right superior and middle aspects of the anterior temporal pole. In contrast, relatively preserved to increased radiotracer uptake is noted within the left lateral orbitofrontal cortex, left middle frontal gyrus, left angular gyrus, left posterior cingulate cortex, and right anterior orbitofrontal region. Overall, the examination demonstrates a heterogeneous metabolic pattern characterized by right temporopolar and bilateral inferior cerebellar hypometabolism with relative frontoparietal and cingulate preservation.

**Reference Report (from Ground-Truth SUVR)**

FDG-PET imaging demonstrates relatively reduced glucose metabolism involving the bilateral cerebellar hemispheres, anterior cerebellar vermis, and superior right temporal pole. Conversely, relatively preserved to increased radiotracer uptake is observed within the left anterior and lateral orbitofrontal cortex, left middle frontal gyrus, left inferior parietal cortex, and left posterior cingulate cortex. Overall impression: The study demonstrates a heterogeneous metabolic pattern characterized by bilateral cerebellar and right temporal hypometabolism with relative left frontoparietal and posterior cingulate preservation.