brat: Aligned Multi-View Embeddings for Brain MRI Analysis

Under Review

Abstract

*We present brat (**b**rain **r**eport **a**lignment **t**ransformer), a multi-view representation learning framework for brain magnetic resonance imaging (MRI) trained on T1-contrast enhanced MRIs paired with clinical reports. Brain MRIs present unique challenges due to the presence of numerous, highly varied, and often subtle abnormalities that are localized to a few slices within a 3D volume. To address these challenges, we introduce the largest brain MRI dataset to date, containing approximately 80,000 3D scans with corresponding radiology reports, and propose a multi-view pre-training approach inspired by advances in document retrieval. We develop an implicit query-feature matching mechanism and introduce concepts from quality-diversity to obtain multi-view embeddings of MRIs that are aligned with the clinical features given by report sentences. We evaluate our approach across multiple vision-language and vision tasks, demonstrating substantial performance improvements. By publicly releasing our suite of model weights, we aim to facilitate further research in brain MRI analysis.*

1. Introduction

Magnetic resonance imaging (MRI) is standard-of-care imaging performed for diagnosis and management of neurodegenerative diseases as well as cancers occurring in the brain. Modern vision-language models (VLM) have shown the capability to generate descriptive free-text for natural, and even medical images [49]. With the rapid increase in the number of diagnostic imaging performed compared to the number of radiologists, the availability of models that can reliably map medical images to radiology reports could assist radiologists in accelerating report generation and reducing time and effort needed to perform diagnosis [23]. Furthermore, vision-language pre-training (VLP) has been shown to produce strong vision backbones for robust tumor segmentation [34]. However, the scarcity of large-scale image-text datasets and the limited generalization of conventional vision-language methods designed for 2D images have hindered the development of effective vision-language approaches for complex 3D imaging modalities.

Clinical reports for medical images, especially detailed cross-sectional (3D) scans such as CT and MRI, often contain rich, diverse information, with multiple sections addressing different aspects of the scan. Existing methods that aim to learn joint representations of medical images and their reports often overlook this complexity. Instead, it's common to adopt architectures similar to those used in natural image captioning, where descriptions typically consist of a single sentence, and are represented via a single embedding. This approach fails to fully leverage the richness of clinical narratives as a learning signal for better image representations [11, 25].

We address the limitation of prior works by introducing the largest existing multimodal brain MRI dataset, containing approximately 80,000 3D MRIs and paired reports,

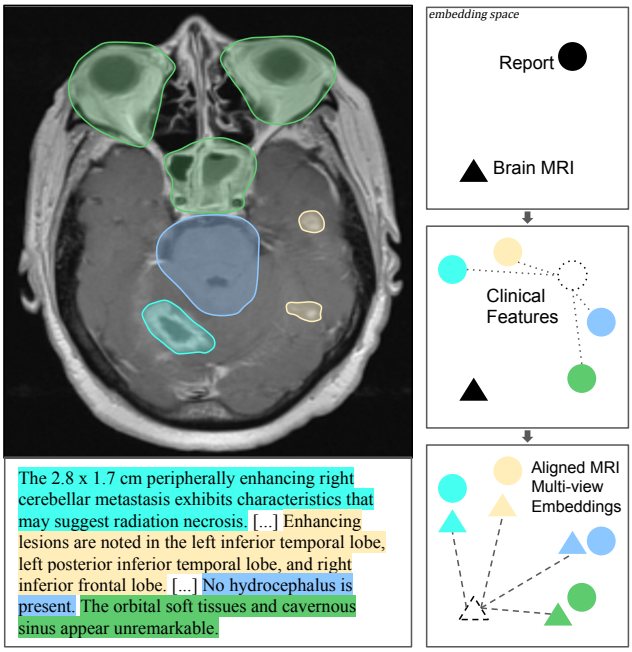


Figure 1. *(Left)* Brain MRI reports contain rich and diverse information that relate to different aspects of the image. Colours are used to associate report sentences with corresponding regions on the scan. The report was cut off ([...]) to only contain findings that can be seen on this 2D slice. *(Right)* Drawing parallels to multi-vector retrieval Zhang et al. [52], we align multi-view embeddings of the MRI with clinical features (sentences) given in the reports.

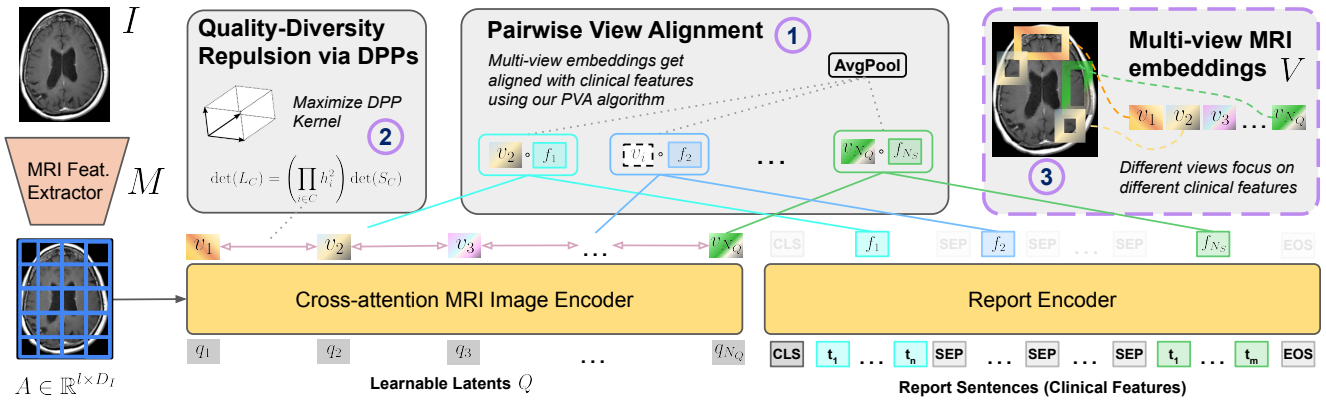


Figure 2. Our brat framework. Our Pairwise View Alignment (PVA) algorithm (described in Section 3.2) and quality-diversity via Determinental Point Processes (DPPs) (described in Section 3.2) lead to clinically aligned multi-view embeddings of the MRI.

and propose a new VLP framework called brat (**b**rain **r**eport **a**lignment **t**ransformer). brat is based on the premise that, similar to documents, clinical reports can consist of extensive text describing a wide range of findings. Multi-view embedding approaches have been used in document retrieval to represent the diversity of information in a document and then mapped to sub-units of the user query [22]. In our work, the learnable multi-view embeddings are used to represent the 3D brain MRI, whereas the corresponding reports are sub-divided into sentence embeddings or sub-units. By aligning these sub-units with brain MRIs via contrastive learning, we implicitly encourage the multi-view image embeddings to represent the clinical features described by the sentences (see Figure 1). We ensure that distinct image views are aligned with sub-unit sentence embeddings through a Pairwise View Alignment (PVA) matching algorithm. We further enhance diversity of the multi-view embeddings via quality diversity (QD) loss based on Determinental Point Processes [26] (DPPs).

brat sets a new benchmark for image-to-text and text-to-image retrieval in both brain MRIs and lung CTs. We also pre-train and release a range of vision backbones using brat that lead to significant performance improvements.

Our contributions are as follows:

(i) We propose a new multi-view vision-language representation learning framework for complex 3D medical images that draws parallels to document representation learning.

(ii) We introduce concepts from Quality-Diversity by applying DPPs that encourage diverse and aligned multi-view embeddings.

(iii) We pre-train and release a suite of brain MRI foundation models trained on the largest existing dataset of paired brain MRIs and reports.

(iv) The model is evaluated across a wide range of datasets and tasks and shows promising results.

2. Related Work

Multi-vector document retrieval. Document retrieval, which involves retrieving documents based on user queries, has seen significant improvements through the use of multi-vector retrieval methods, instead of traditional single-vector approaches. In single-vector retrieval, documents are represented as a single embedding, limiting the ability to capture the diversity of information in documents [21, 32]. In contrast, multi-vector embeddings offer more versatile query-document interactions, which better represents the variety of information in documents. One of the important works is ColBERT [22, 38], which computes query-document similarity by selecting the most similar document token for each query token and aggregating the similarities across a document. However, these ideas have not yet been applied to image-text datasets, even though text like radiology reports is often of document length. Zhang et al. [52] emphasized that documents typically contain multiple semantic units, each potentially relevant to different queries, and proposed using multi-view embeddings to represent these diverse aspects. Drawing inspiration from their method, we treat brain MRI scans as "documents" and report sentences as corresponding "queries", to enhance representation and retrieval.

Vision-Language Pre-training (VLP). VLP on large-scale datasets of paired images and captions is an effective way to learn image-representations for both vision [10, 37, 41] and vision-language tasks [1, 8, 45]. Vision-language datasets occur naturally in the medical imaging domain, as radiologists routinely write reports to describe findings in medical scans. Large-scale public datasets are predominantly available for chest X-rays [5, 16, 20], and as such most approaches for medical imaging focus on this domain [46, 47, 53, 54]. Recently there have been efforts to publish datasets in more advanced imaging modalities, such

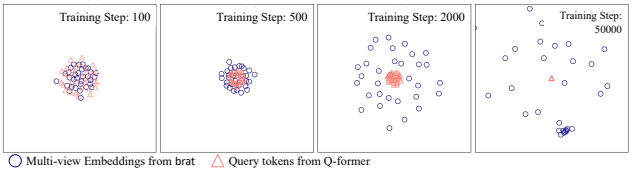


Figure 3. Conventional query tokens collapse into a single representation as training progresses. The multi-view embeddings of brat , on the other hand, are diverse and spread out. The plot was obtained by using multi-dimensional scaling to plot the 32 query tokens on a 2D plane based on their pairwise euclidean distances averaged across a random sample of 32 images.

as lung CTs [9]. Some existing models attempt to capture the fine-grained features of medical images by aligning local image feature patches with text tokens [15, 44]. However, this is limited by the fact that individual text tokens are not necessarily representative of clinical features and image feature patches are spatially restricted to a specific location in the image. In addition, chest X-rays are 2D images and their reports usually only contain 2-3 image-descriptive sentences. 3D scans such as brain MRIs, typically have reports that are several times larger. Recently the first models have been trained on large 3D lung CTs and report datasets, relying mainly on scale to achieve good results [49].

Some VLMs focus on using learnable latent variables that learn to efficiently compress visual representations [18, 27]. BLIP-2 [27] employs a Q-Former model that uses "querying tokens" as learnable latents that align cross-modal representations. We adopt a similar architecture, with the querying tokens representing the learnable multi-view embeddings that are aligned to clinical features but can freely attend to the image, enabling to capture features that are not localized spatially in the image.

Quality-diversity of learned features. Increasing diversity to avoid informational collapse is common in self-supervised learning (SSL), where information maximization methods are used to achieve this [4, 50]. In terms of diversity for representing different aspects of an input, Zhang et al. [52] in document retrieval has a similar goal and uses a loss that penalizes high pairwise similarity between embedding vectors. Determinental point processes (DPPs), on the other hand, have been used in particular for recommender systems, summarisation, or dataset/batch sampling, where trade-offs between quality and diversity are desired [26, 28, 39]. Although these trade-offs are also important considerations in (multi-view) representation learning, DPPs have not yet been applied in this domain. We show that they are well-suited for this and lead to better performance than simple pairwise similarity reduction.

Brain MRI analysis. The scarcity of large-scale brain MRI datasets (and lack of image-text datasets) has led researchers to pool smaller public datasets [31] and focus on SSL. SSL tasks involve reversing various image augmentations [33, 43, 51], masked image modeling [3, 13, 19, 42], and contrastive losses [7]. Other efforts aim to learn models that can be generalized across MRI modalities [24, 48]. Downstream applications are often focused on segmentation, with notable datasets being the BraTS series [2, 30], on which we evaluate our models. Our work uses the largest pre-training dataset combining brain MRI with paired radiology reports to learn an effective representation of pathologies visible in brain MRIs.

3. Methods

3.1. The AnonBrain Dataset

We collected a comprehensive dataset of brain MRI scans and their corresponding clinical reports from a cancer center, covering the period from 2012 to 2017. These scans were primarily obtained to monitor brain metastases and tumors in cancer patients, resulting in a dataset rich in positive findings (89.7% of scans show abnormalities, average of 134 words or 8 sentences per report) and representative of a diverse patient population. We also collected the clinical reports corresponding to these images, as well as extensive demographic data, primary diagnosis, ongoing chemotherapy and radiotherapy treatments, and survival information, which will be utilized in future studies.

Our dataset includes 77,228 brain MRI image-report pairs from 24,262 unique patients. To develop our model, we performed a patient-wise split of the data into 75,142 examples for training, 945 for development, and 1,141 for the test set.

As the focus of this work is on learning image representations from brain MRIs, we ensured that all report content was visually grounded in the corresponding images. For example, keyword filtering revealed that 94% of reports make references to prior scans. To efficiently remove these references, information from excluded MRI modalities, and protected health information (PHI), we developed a PHI-enabled GPT-4 based pipeline. This pipeline re-wrote reports and extracted structured data simultaneously. Liu et al. [29] demonstrated that GPT-4 performs well on radiology report processing; and indeed we found that our pipeline achieved an annotation accuracy of 96% on a gold standard set of 50 manually annotated reports. Annotating all the reports cost approximately $1,600, which is significantly lower than the cost of expert annotation. More information on the dataset is provided in Appendix 8.

We are exploring options to make this dataset available to the wider research community.

Characteristic	Value
Word Count (Q1, Median, Q3)	115, 134, 156
Sentence Count (Q1, Median, Q3)	7, 9, 11
Age (Q1, Median, Q3)	45, 58, 68
Any Abnormality (%)	87.9
Prior Surgery (%)	38.1
Enhancing Lesions (%)	47.6
Midline Shift (%)	5.4
White Matter Changes (%)	43.6
Pituitary Gland Abnormality (%)	2.3
Hydrocephalus (%)	2.6
Biggest Mass Length (%)	<1cm (17.7), 1-2cm (14.5) >2cm (18.8)
Enhancing Lesion Count (%)	1 (27.0), 2-6 (23.0) 7-15 (1.6), >15 (2.3)
# of Unique Surgeries	32,428
# of Enhancing Lesion Locations	95,815

Table 1. Brain MRI dataset characteristics. For reference, the Conceptual Captions [40] dataset has 10 tokens (less than 10 words) per image.

3.2. The brat Framework

brat is a vision-language contrastive pre-training framework that represents images via aligned multi-view embeddings. To obtain the multi-view embeddings, we adopt a similar base architecture as Q-Former [27], with brat using learnable latents that extract multi-view embeddings by cross-attenting to the MRI features (see Fig. 2). A 3D vision model M is used to extract features from an MRI image I, resulting in a set of feature maps $M(I) = A \in \mathbb{R}^{l \times D_I}$ with l feature maps of dimension D_I. We mainly use Densenet-121 [14] as M, which outperformed ViT and Resnet-50 in our preliminary experiments. The set of learnable latent tokens $Q = [q_1, \ldots, q_{N_Q}]$ where $q_l \in \mathbb{R}^{D_Q}$, interact with the image encoder features A to extract a set of image-informed multi-view embeddings $E_I(Q, I) = V$, where $V = [v_1, \ldots, v_{N_Q}]$ with $v_i \in \mathbb{R}^{D_V}$. The text encoder E_R takes a brain MRI radiology report and returns sentence embeddings that capture the clinical features described in them: $E_R(R) = F$, where $F = [f_1, \ldots, f_{N_S}]$ with $f_i \in \mathbb{R}^{D_F}$ representing the i-th sentence. We obtain sentence embeddings f_i by averaging all token embeddings of the sentence. As $D_F = D_V$, D_F is used in the rest of the paper for clarity. Training and implementation details are provided in the Appendix.

Pairwise View Alignment Existing approaches that use latent variables to learn to extract image features, such as Q-Former, often exhibit embedding collapse, where the learned latents converge into a single representation [6] (an illustration is given in Figure 3). Volumetric brain MRI scans contain diverse visual elements that are very complex due to highly variable appearance of pathology such as primary brain tumors and metastastic tumors, as well as impact on normal tissue from such tumors such as mass effect pushing normal tissue structures or excessive accumulation of cerebrospinal fluid. Capturing such a complex set of findings requires larger flexibility. However, given the complexity of brain MRIs, which frequently contain a diverse range of findings—analogous to how documents comprise distinct semantic units—we hypothesize that multi-view embeddings provide a more suitable representation for brain MRI images. Multi-view embeddings have been shown to improve document representations by capturing different semantic elements within a text [52]. Inspired by this, we assume that individual sentences in radiology reports correspond to distinct clinical features, and we aim for our multi-view embeddings to encapsulate these same clinical characteristics (see Figure 1). To achieve this, we introduce a two-step approach: (1) Pairwise View Alignment (PVA) to align embeddings with clinically meaningful features, and (2) quality-diversity repulsion using determinantal point processes (DPPs) to encourage diversity in the learned representations.

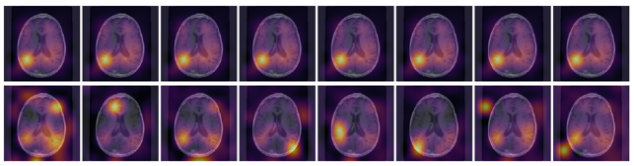


Figure 4. Juxtaposition of 8 query tokens from Q-Former (upper row) and the same 8 query tokens from brat (lower row). The collapsed Q-Former queries all attend to the same image regions, whereas the multi-view embeddings of brat focus on distinct features.

PVA ensures that each image view embedding can only be matched to a single sentence feature, preserving the granularity of the report in the image representation. Embeddings are matched based on their cosine similarities. A step-by-step description of the approach is given in Algorithm 1.

Quality-Diversity via DPPs Empirically, we find that pairwise view alignment alone does not sufficiently encourage diverse features in multi-view embeddings. Thus, we adopt ideas from quality-diversity (QD) to address this issue. The idea behind QD is to have many diverse solutions to tackle a problem from different angles. This fits our problem well, as we want the different multi-view embeddings to focus on different features of the image. We consider as feature diversity the diversity of the attention maps over the image feature maps A of the multi-view embeddings V. A naive option to address this would e.g. be pairwise repulsion between the embeddings v_i by maximizing their cosine dissimilarities. However, to avoid the embeddings collaps-

Algorithm 1 Pairwise View Alignment

Input: Normalized multi-view embeddings of an image $V = [v_1, \ldots, v_{N_Q}] \in \mathbb{R}^{N_Q \times D_F}$ and normalized clinical features matrix of a report $F = [f_1, \ldots, f_{N_S}] \in \mathbb{R}^{N_S \times D_F}$

Output: Match pairs of multi-view embeddings and clinical features P_M

Algorithm:

Compute similarities between all multi-view embeddings V and clinical features F: $S_{v,f} \leftarrow VF^T \in \mathbb{R}^{N_Q \times N_S}$

Initialize list P_{all} with all similarity pairs $\langle v_i, f_j \rangle = S_{v,f}[i,j]$ sorted in descending order of similarity

Initialize an empty list of matched pairs P_M

while P_{all} is not empty **do**

  Select a pair of indices (s_v, s_f) with the highest similarity (top of stack): $\langle v_{s_v}, f_{s_f} \rangle = \max(P_{\text{all}})$

  Add (s_v, s_f) to the list of matched pairs P_M

  Remove all pairs from P_{all} with indices s_v for image features or s_f for report features (each query token and sentence feature can only be matched once)

end while

ing into unimodal representations (where all the attention focuses on a single feature map) and to capture overall diversity across all multi-view embeddings, we model quality and diversity using DPPs [26]. We show in Table 2 that this is a crucial step.

DPPs are distributions over subsets of a fixed ground set that attribute higher probability to sets that are diverse. In our case, we want to maximize the probability of our set of multi-view embeddings V under the DPP. We consider quality-diversity with respect to the cross-attention maps $C = [c_1, \ldots, c_{N_Q}]$ where $c_j \in \mathbb{R}^{D(=n_{fx} \cdot n_{fy} \cdot n_{fz})}$ contains the attention values from multi-view embedding v_j to the 3D feature maps representing the image. As *quality* of each embedding token we use Shannon entropy of its cross-attention map, denoted as $h_i \in \mathbb{R}^+$:

$$h_i = \mathcal{H}(c_i) = -\sum_k c_i(k) \log c_i(k), \tag{1}$$

where k indexes over spatial positions. A higher entropy implies a more uniform attention distribution, which we interpret as higher quality. This prevents the attention maps from collapsing to trivial, high-diversity solutions where each query attends to a single point. The attention maps c_i themselves are considered as the *diversity features*. The DPP kernel matrix L_{ij} can be written as:

$$L_{ij} = h_i c_i^T c_j h_j. \tag{2}$$

The DPP for a selecting a subset C' is given by:

$$P_L(C') \propto \det(L_{C'}), \tag{3}$$

In our case $C = C'$, as we consider repulsion between all image tokens. $\det(L_C)$ can be decomposed as follows:

$$\det(L_C) = \left(\prod_{i \in C} h_i^2 \right) \det(S_C), \tag{4}$$

where S_C is the similarity matrix between all attention maps c_i.

The determinant of the kernel matrix L_C corresponds to the squared volume of the parallelepiped spanned by the vectors $h_i c_i$ for each i in C. By maximizing the product $\prod_{i \in C} h_i^2$, we encourage each embedding token to have high entropy, corresponding to a large magnitude in the feature space. By maximizing the determinant $\det(S_C)$, where S_C captures the pairwise similarities between attention maps, we ensure that the directions c_i are as different as possible, promoting diversity among the tokens. This approach naturally prevents the embeddings from collapsing into a single representation by encouraging both high quality (non-collapsed attention) and diverse (distinct attention patterns) embeddings.

In practice, we define the DPP loss by taking the negative log-determinant of the kernel matrix L

$$\mathcal{L}_{\text{DPP}} = -\log \det(L_C + \epsilon I), \tag{5}$$

where ϵI is a small diagonal matrix added for numerical stability.

brat loss To obtain the overall image-report similarity, we aggregate the multi-view embedding sentence similarities by mean-averaging:

$$S_{R,I} = S_{I,R} = \frac{1}{|P_M|} \sum_{(i,j) \in P_M} S_{v,f}[i,j] \tag{6}$$

As such, we get our contrastive losses as follows:

$$\mathcal{L}^{(I|R)} = -\log \left(\frac{\exp(S_{I,R}/\tau)}{\sum_k \exp(S_{I,R_k}/\tau)} \right) \tag{7}$$

$$\mathcal{L}^{(R|I)} = -\log \left(\frac{\exp(S_{R,I}/\tau)}{\sum_m \exp(S_{R,I_m}/\tau)} \right) \tag{8}$$

We also use the same "Image-grounded Text Generation" (ITG) loss as in BLIP-2 [27], as we found it to help performance. Our final loss is thus given as:

$$\mathcal{L} = \frac{\mathcal{L}^{(I|R)} + \mathcal{L}^{(R|I)}}{2} + \mathcal{L}_{\text{DPP}} + \mathcal{L}_{\text{ITG}} \tag{9}$$

3.3. Downstream Models

Our brat framework provides both a pre-trained vision backbone M and a model $E_I(Q, I)$ for extracting multi-view embeddings. We refer to brat-viz for the vision backbone only, and brat for multi-view framework.

Figure 5 illustrates how the brat weights can be modularized for different downstream tasks. Different task-specific heads, such as an MLP for classification, a language model for report generation, or a segmentation decoder, can appended to either brat or brat-viz. Experiments on different such configurations are provided in the next section.

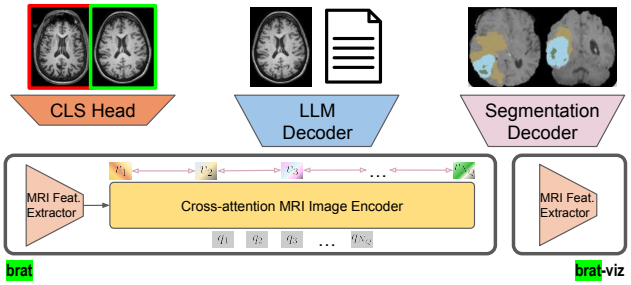


Figure 5. We connect two configurations of brat with various decoders to evaluate the benefits of our pre-training on downstream tasks.

4. Experiments and Results

This paper presents a new vision-language pre-training method for 3D medical scans and document-length reports. In this section, we demonstrate the benefits of our approach over other pre-training methods, both in terms of pre-training metrics, as well as on downstream tasks including tumor and metastases segmentation, Alzheimer's classification, and brain MRI report generation.

4.1. VLP Performance: Image-Text Retrieval

We evaluate brat on image-text retrieval tasks on both AnonBrain and BIMCV-R, an external public benchmark of lung CTs and corresponding reports. We computed key retrieval metrics such as recall@k and mean and median rank. For AnonBrain we are also providing finding-based metrics, where "P@5 (F)" corresponds to how frequently each of the 5 retrieved samples contain at least one common positive finding (F) with the ground-truth match. "R@5 (F)"corresponds to the frequency of finding at least one sample containing exactly the same labels as the ground-truth in the top-5 samples (recall).

AnonBrain We evaluated brat against multiple baselines: CLIP, Q-Former (the base of brat), brat with the traditional Colbert matching algorithm [22] instead of PVA, brat without QD, brat with simple pairwise repulsion as used in [52] instead of DPPs, and a Q-Former with QD. Results with ViT and ResNet-50 backbone models are provided for completeness. For simplicity, we only included MRIs that have at least one positive finding (around 90% of our original dataset) in evaluation, as negative reports usually apply to all negative images. As shown in Table 2, except mean rank, brat sets the benchmark on all metrics. The lower mean rank can generally be explained by the model making higher confidence predictions, and this can be adjusted by selecting the weights at earlier training steps. The QD component improved performance, suggesting reliance of PVA on QD for effective learning. Q-Former did not show benefits with QD repulsion, potentially because the diversity is of less use when the query tokens are not encouraged to be aligned with diverse clinical features. Simple pairwise repulsion also does not match the performance improvements obtained by using DPPs. We also find that the document-typical Colbert algorithm for matching multi-view emeddings does not perform as well as PVA. Qualitative examples of images and corresponding reports retrieved by brat are shown in Figure 6.

BIMCV-R To demonstrate the generalizability of our framework, we also pre-trained it from scratch on the BIMCV-R dataset, a publicly available dataset of lung CT scans paired with radiology reports. Similar to the original paper, we find that conventional contrastive loss approaches such as a basic Q-Former or CLIP perform very poorly (see Table 3). Notably, without the need for self-supervised techniques employed by MedFinder, brat achieves comparable performance purely by leveraging textual supervision. We also identify certain quality issues within the BIMCV-R dataset, detailed in Appendix 6.3, which may contribute to the generally lower performance observed on this benchmark. Despite these limitations, our results show that brat can be effectively applied off-the-shelf to other medical imaging modalities with complex visuals and lengthy reports. We note a greater discrepancy in mean rank on BIMCV-R, but limited methodological details in prior work and unavailable model weights make direct comparison difficult, leaving some aspects open to further examination.

4.2. Downstream Tasks

In this section, we showcase how our pre-training is beneficial for a wide range of downstream tasks.

Brain MRI Report Generation. VLP naturally suits radiology report generation, as the visual embeddings already align with text features. To evaluate our pre-trained backbones, we freeze the vision backbone and assess how well a language model can extract image-grounded information from it. We use Llama-3.2-1B [12], providing either multi-view embeddings or image feature maps to the LLM via a bridging MLP. Training and evaluation occur on the AnonBrain dataset, and we provide both LLM-

Methods	Text to Image							Image to Text						
	R@1 ↑	R@5 ↑	R@10 ↑	R@5 (F) ↑	P@5 (F) ↑	MdR ↓	MnR ↓	R@1 ↑	R@5 ↑	R@10 ↑	R@5 (F) ↑	P@5 (F) ↑	MdR ↓	MnR ↓
CLIP	0.146	0.407	0.564	0.894	0.718	8.0	35.8	0.159	0.431	0.569	0.853	0.748	8.0	37.2
QFormer	0.154	0.377	0.529	0.867	0.672	10.0	32.5	0.146	0.368	0.532	0.837	0.703	9.0	34.9
Colbert	0.125	0.370	0.509	0.889	0.680	10.0	**31.9**	0.113	0.326	0.487	0.810	0.732	11.0	36.0
brat w/o QD	0.173	0.458	0.615	0.894	0.711	**6.0**	37.3	0.171	0.449	0.606	0.875	0.745	7.0	**34.5**
brat w/ PR	0.099	0.349	0.497	0.875	**0.723**	11.0	36.6	0.109	0.328	0.478	0.818	0.696	11.0	39.2
QFormer w/ QD	0.155	0.370	0.542	0.851	0.701	10.0	34.8	0.152	0.381	0.529	0.817	0.699	10.0	33.6
brat	**0.205**	**0.493**	**0.666**	**0.911**	0.718	**6.0**	124.1	**0.201**	**0.481**	**0.645**	**0.882**	**0.752**	**6.0**	96.3
brat vit	0.015	0.066	0.117	0.661	0.410	385.0	404.6	0.016	0.066	0.129	0.604	0.473	357.0	401.0
brat resnet	0.095	0.292	0.436	0.843	0.654	13.0	109.4	0.131	0.343	0.462	0.809	0.640	12.0	62.0

Table 2. Evaluation results for text-to-image and image-to-text retrieval on AnonBrain. For the "↑" metrics higher is better and for the "↓" metrics lower is better. "R@5 (finding)" and "P@5 (finding)" indicate the recall and precision at 5 for the finding task.

Methods	Text to Image					Image to Text				
	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
CLIP4clip [21]	0.003	0.015	0.022	717.0	735.9	0.003	0.008	0.015	722.0	738.7
3D-MIR [1]	0.011	0.047	0.103	121.1	152.3	0.012	0.040	0.088	134.9	162.4
MedFinder (Resnet-50)	0.028	0.087	0.203	**68.9**	81.3	0.029	0.088	0.197	71.2	**80.7**
MedFinder (ViT-base)	0.027	0.089	**0.214**	75.4	**80.1**	0.027	0.090	**0.203**	72.3	81.9
Q-Former	0.007	0.025	0.048	223.0	371.7	0.000	0.015	0.034	225.0	365.8
brat	**0.030**	**0.109**	0.165	71.0	283.0	**0.036**	**0.103**	0.182	**67.0**	282.0

Table 3. Evaluation results for text-to-image and image-to-text retrieval on BIMCV-R, a lung CT dataset. The compared results, except Q-Former, are taken from Chen et al. [9]. It's unclear how the median ranks happened to be reported as non discrete values.

based metrics (GREEN metric [35]) and natural language metrics. We compare brat to training from scratch, Q-Former pre-training, and classification-based pre-training ("CLS"), using either the vision backbone or multi-view embeddings as LLM input. Results are in Table 4. We can see that the vision-language pre-training leads to significant improvements over no pre-training or classification pre-training. brat also leads to additional improvements over simple QFormer pre-training. This is the first work to provide report generation capabilities for brain MRIs that were trained on a large-scale dataset. We show that, overall, the ability of VLMs to generate reports for brain MRIs is in line with other radiographic modalities, such as chest X-rays. Example reports are shown in the Appendix.

Alzheimer Classification To investigate whether our pre-training generalizes to non-cancer-focused brain MRI datasets, we evaluated brat on ADNI [36], a brain MRI dataset investigating the progression of Alzheimer's disease. We split the cohort into training (n=1,932), validation (n=384), and hold-out test (n=291) sets. Brain MRIs are either "cognitive normal", "mild cognitive impairment" (MCI), or "Alzheimer's disease". In Figure 7, we show performance on binary classification (Normal or Alz.) for 1, 10, and 100% training data. As results on ADNI vary significantly based on random seeds and selected subsets of the training data, we launch 10 runs for each setting and bootstrap from these results to obtain 95% confidence intervals. Table 7 in the Appendix contains more extensive results. We show that our pre-training leads to consistent and significant improvements across all settings. Vision-language pre-training, in general, leads to significant performance improvements over classification pre-training. These results underline that the brat pre-training framework is a great starting point for brain MRI problems across a wide range of domains.

Segmentation Tasks We also evaluate brat on one of the most common downstream applications in brain MRI analysis: tumor segmentation. We use BraTS2021 [2], containing gliomas, and BraTS2023-METS [30], containing brain metastases. To isolate the benefit of the brat pre-training framework, we only used T1 modalities and did not include many of post-processing steps typically included for these datasets, such as patch-based processing. We follow the conventional approach for this dataset in doing 4-fold cross-validation [17], however, we also use three random seeds for each run so we can again obtain confidence intervals. We use the benchmark appropriate evaluation metrics Dice (Brats2021) and lesion-wise Dice (Brats2023), given for three overlapping regions: whole tumor, tumor core, and enhancing tumor. Figure 8 shows that our pre-training improves performance for the metastases, but not for the gliomas. This may be explained by the fact the gliomas are generally large and visible to even non-expert, meaning that the core task is more separating pixels precisely rather than

Backbone	Pre-training	LLM	GREEN (LLM Eval)					NLG Metrics						
			All	FP	FN	Location	Severity	METEOR	CIDEr	Rouge-L	Bleu-1	Bleu-2	Bleu-3	Bleu-4
Densenet-121	None	Llama 3.2-1B	0.300	0.110	0.190	0.750	0.850	0.117	0.039	0.180	0.177	0.103	0.065	0.042
Densenet-121	Classification	Llama 3.2-1B	0.310	0.115	0.195	0.760	0.860	0.102	0.049	0.187	0.124	0.072	0.048	0.033
Densenet-121	QFormer	Llama 3.2-1B	0.375	0.138	0.287	0.840	0.911	0.131	0.079	0.216	0.201	0.123	0.081	0.056
Densenet-121	brat	Llama 3.2-1B	0.390	0.150	0.300	**0.860**	**0.920**	**0.134**	0.098	0.214	**0.241**	**0.142**	**0.091**	**0.061**
QFormer	QFormer	Llama 3.2-1B	0.360	0.130	0.280	0.820	0.900	0.125	0.105	0.210	0.190	0.115	0.078	0.053
QFormer	brat	Llama 3.2-1B	**0.402**	**0.172**	**0.318**	0.852	0.917	0.128	**0.114**	**0.219**	0.197	0.121	0.081	0.056

Table 4. The backbone is always frozen, except for "None" pre-training. The GREEN metric is obtained using a 7B parameter LLM. Four GREEN scores, relating to false findings (FP), missing findings (FN), false findings (FP), and accuracy of severity and location specification of findings are provided. Two additional metrics used in GREEN, missing or hallucinated references to prior scans are omitted as we removed these references from our dataset and therefore our models all score a 100% on these metrics.

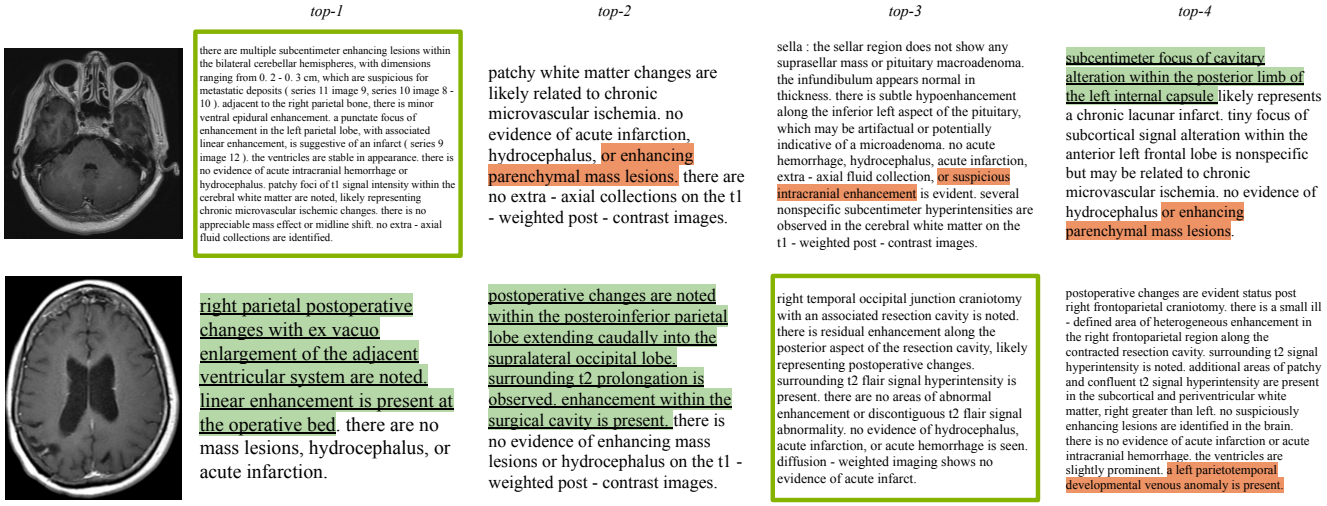


Figure 6. Qualitative examples showing the top-4 output of our brat model for image-to-text retrieval on a reduced dev set of 315 examples. On this subset the median rank achieved was 2. Enboxed examples are correct. Green (and underlined) sections are passages that are clinically correct, even though they are from a different MRI. In red are passages that don't correspond to the MRI.

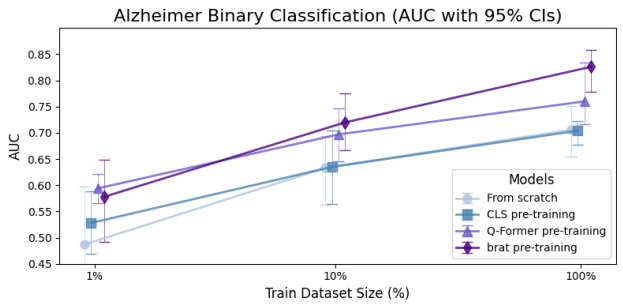


Figure 7. Comparison of brat pre-training to alternative pre-training methods for Alzheimer classification on ADNI.

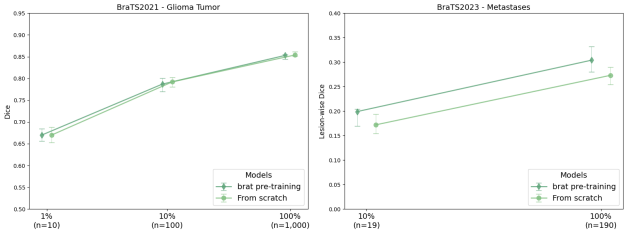


Figure 8. Comparison of brat pre-training to random initialisation for tumor and metastases segmentation on BraTS2021 and BraTS2023. Scores are averaged across the three tumor regions.

requiring anatomical understanding of brain MRIs. More detailed results are provided in Appendix Table 8 and 9.

5. Conclusion

We have introduced two ideas novel to vision-language representation learning: multi-view embeddings adopted from document retrieval and DPPs to maximize the quality-diversity of these embeddings. Our approach demonstrates promising results when applied to images paired with long reports, including both brain MRI and lung CT datasets. The proposed brat framework is architecture-agnostic and compatible with a variety of image and text encoders. The flexibility of the learnable multi-view embeddings also naturally allows to extend the input beyond imaging data. This

is promising for medical image analysis, where patient context and lab results can provide crucial cues for diagnosis.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022. 2

[2] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021. 3, 7

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. 3

[4] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *10th International Conference on Learning Representations, ICLR 2022*, 2022. 3

[5] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 2020. 2

[6] Vinod Kumar Chauhan, Anshul Thakur, Odhran O'Donoghue, Omid Rohanian, Soheila Molaei, and David A Clifton. Continuous patient state attention model for addressing irregularity in electronic health records. *BMC Medical Informatics and Decision Making*, 2024. 4

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020. 3

[8] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. In *International Conference on Learning Representations*, 2023. 2

[9] Yinda Chen, Che Liu, Xiaoyu Liu, Rossella Arcucci, and Zhiwei Xiong. Bimcv-r: A landmark dataset for 3d ct text-image retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024. 3, 7

[10] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[11] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2019. 1

[12] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 6, 1

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3

[14] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 4

[15] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 3

[16] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, 2019. 2

[17] Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024. 7

[18] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, 2021. 3

[19] Jue Jiang, Neelam Tyagi, Kathryn Tringale, Christopher Crane, and Harini Veeraraghavan. Self-supervised 3d anatomy segmentation using self-distilled masked image transformer (smit). In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 556–566. Springer, 2022. 3

[20] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 2019. 2

[21] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 2

[22] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 2020. 2, 6

[23] Kleanthis Konstantinidis. The shortage of radiographers: A global crisis in healthcare. *Journal of medical imaging and radiation sciences*, 2024. 1

[24] Aishik Konwer, Xiaoling Hu, Joseph Bae, Xuan Xu, Chao Chen, and Prateek Prasanna. Enhancing modality-agnostic representations via meta-learning for brain tumor segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

[25] Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of bert. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. 1

[26] Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 2012. 2, 3, 5

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023. 3, 4, 5

[28] Ping Li, Qinghao Ye, Luming Zhang, Li Yuan, Xianghua Xu, and Ling Shao. Exploring global diverse attention via pairwise temporal relation for video summarization. *Pattern Recognition*, 2021. 3

[29] Qianchu Liu, Stephanie Hyland, Shruthi Bannur, Kenza Bouzid, Daniel Castro, Maria Wetscherek, Robert Tinn, Harshita Sharma, Fernando Pérez-García, Anton Schwaighofer, et al. Exploring the boundaries of gpt-4 in radiology. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 3

[30] Ahmed W Moawad, Anastasia Janas, Ujjwal Baid, Divya Ramakrishnan, Rachit Saluja, Nader Ashraf, Leon Jekel, Raisa Amiruddin, Maruf Adewole, Jake Albrecht, et al. The brain tumor segmentation (brats-mets) challenge 2023: Brain metastasis segmentation on pre-treatment mri. *arXiv preprint arXiv:2306.00838*, 2023. 3, 7

[31] Asbjørn Munk, Jakob Ambsdorf, Sebastian Llambias, and Mads Nielsen. Amaes: Augmented masked autoencoder pretraining on public brain mri data for 3d-native segmentation. *arXiv preprint arXiv:2408.00640*, 2024. 3

[32] Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, et al. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. 2

[33] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*. Springer, 2016. 3

[34] Yujin Oh, Sangjoon Park, Hwa Kyung Byun, Yeona Cho, Ik Jae Lee, Jin Sung Kim, and Jong Chul Ye. LLM-driven multimodal target volume contouring in radiation oncology. *Nature Communications*, (9816), 2024. 1

[35] Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Md, Michael Moseley, Curtis Langlotz, Akshay Chaudhari, et al. Green: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 374–390, 2024. 7

[36] Ronald Carl Petersen, Paul S Aisen, Laurel A Beckett, Michael C Donohue, Anthony Collins Gamst, Danielle J Harvey, CR Jack Jr, William J Jagust, Leslie M Shaw, Arthur W Toga, et al. Alzheimer's disease neuroimaging initiative (adni) clinical characterization. *Neurology*, 74(3): 201–209, 2010. 7

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 2021. 2

[38] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022. 2

[39] Akanksha Saran, Safoora Yousefi, Akshay Krishnamurthy, John Langford, and Jordan T Ash. Streaming active learning with deep neural networks. In *International Conference on Machine Learning*, 2023. 3

[40] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018. 4

[41] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 2

[42] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20730–20740, 2022. 3

[43] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*, 2008. 3

[44] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. In *Advances in Neural Information Processing Systems*, 2022. 3

[45] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research*, 2022. 2

[46] Rhydian Windsor, Amir Jamaludin, Timor Kadir, and Andrew Zisserman. Vision-language modelling for radiological imaging and reports in the low data regime. In *Medical Imaging with Deep Learning*, 2023. 2

[47] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2

[48] Wentian Xu, Matthew Moffat, Thalia Seale, Ziyun Liang, Felix Wagner, Daniel Whitehouse, David Menon, Virginia Newcombe, Natalie Voets, Abhirup Banerjee, et al. Feasibility and benefits of joint learning from mri databases with different brain diseases and modalities for segmentation. In *Medical Imaging with Deep Learning*, 2024. 3

[49] Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, et al. Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162*, 2024. 1, 3

[50] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, 2021. 3

[51] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision*. Springer, 2016. 3

[52] Shunyu Zhang, Yaobo Liang, Ming Gong, Daxin Jiang, and Nan Duan. Multi-view document representation learning for open-domain dense retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022. 1, 2, 3, 4, 6

[53] Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 2023. 2

[54] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, 2022. 2

brat: Aligned Multi-View Embeddings for Brain MRI Analysis

Supplementary Material

6. Implementation and Training Details

In this section, we discuss the implementation details of our pre-training and downstream evaluation. All code and models will be made public.

6.1. Pre-Training

Training parameters were determined empirically, with the final set provided in Table 6. Contrary to the general assumption that big batches lead to improved performance for image-text contrastive learning, our results consistently improved for comparitvely small batch sizes, in the range of 25-32. This allowed us to train each model on a single A/H100 GPU. Our experiments also found that the Q-Former's language modeling loss consistently improved performance across nearly all configurations, while image-text matching did not yield benefits, leading us to omit the image-text matching loss. We also found that using a biomedically pre-trained BERT outperformed the standard BERT pre-training version in all evaluated scenarios. All model weights were selected based on the best average metrics on the development set. For AnonBrain, all models were trained with the same image processing: $1mm \times 1mm \times 1mm$ voxel spacing, intensity normalization, and resizing to $32 \times 256 \times 256$. Preliminary analysis on AnonBrainalso showed lowered performance with standard data augmentation such as Gaussian noise, image rotation and translation, as well as random view cropping, and was removed from subsequent analyses.

6.2. Downstream Tasks

We evaluated our pre-training methods by fine-tuning on several downstream tasks. When feasible, our hyperparameters were selected via grid search. The ADNI hyperparameter are given in Table 5. For ADNI, image preprocessing was performed using Clinica's t1-volume-tissue-segmentation pipeline. For report generation, we follow the parameters chooses in the Llama paper [12]. We used a batch size of 1024 and learning rate of 0.0002. We use an AdamW optimizer with a cosine decay and a warm-up ratio of 0.3. For segmentation, we used nnUNet as the baseline model, fine-tuning it with an initial learning rate of 1e-2 and a weight decay of 3e-5. The training pipeline included standard nnUNet pre-processing, data augmentation was not used. Result model weights were selected based on the highest mean Dice score for BraTS-2021 and the best Lesion-Wise metrics for BraTS-2023-METS on the validation set. All model weights were selected based on the best average metrics on the development set.

Parameter	1%	10%	100%
Batch Size	16	32	32
Learning Rate	1.00E-06	1.00E-05*	1.00E-05*
Training Precision		Bfloat16	
Augmentation	Yes	No	No
Trained Layers	MLP Only	All	All
MLP size		2 layers	

Table 5. Implementation details of our Alzheimer classification downstream task. *For ViT we used 1.00E-06 across all data amounts. Augmentation consisted of: random flipping, random intensity scaling, random intensity shifting, adding gaussian noise, gaussian smoothing, random contrast adjustment, and random low resolution simulation. More details can be found in our code.

Parameter	all models on AnonBrain	brat on BIMCV-R
Batch Size	32	25
3D Vision Model M	Densenet-121/ViT/ResNet-50	Densenet-121
Weights Init. of M	None	
Architecture of $E_{I/R}$	BERT-base[1]	
Weights Init. of $E_{I/R}$	BiomedBERT[2]	
Learning Rate M	5.00E-04*	
Learning Rate $E_{I/R}$	5.00E-05	
Max. Text Length E_R	256	
N_Q (# of Query Tokens)	32	
Cross-Attention Frequency	2	
Max. Number of Sentences	20	
Training Precision	Bfloat16	
Augmentation	None	

Table 6. Implementation details of our pre-training. Except for BIMCV-R , the batch size was chosen to be maximal given compute resources. *For ViT, we used a lower learning rate of 1.00E-07.

6.3. BIMCV-R Dataset

We found quality issues with the BIMCV-R dataset that may explain the overall lower performance obtained on this dataset compared to AnonBrain. Figure 9 shows how for some images the middle slice (depicted) is already no longer in the lung, suggesting that the scan mainly depicts other body parts. Several images also seem to depict localizer scans, which makes it difficult to connect them to radiology reports. Appropriate processing of these images would likely lead to significant performance improvements.

[h]

7. Additional Results

In this section, we provide more detailed results and examples.

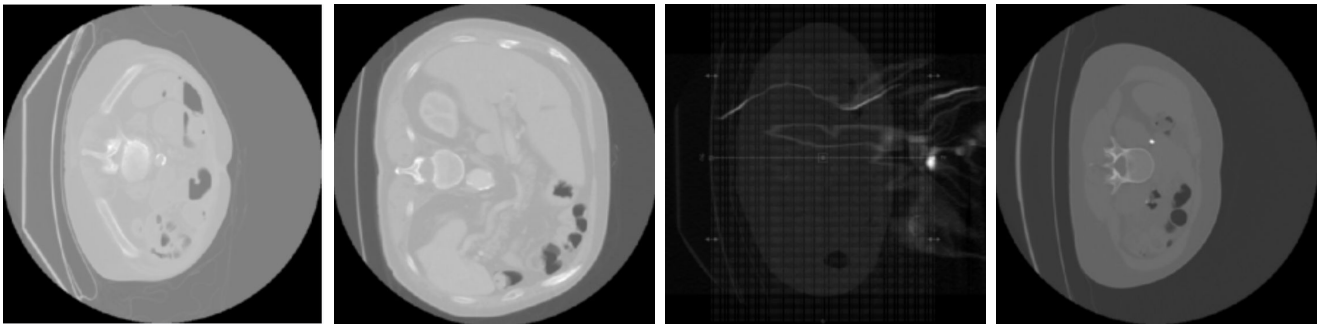


Figure 9. BIMCV-R example images of localizer scans or where the middle slice is already in the abdomen or pelvis.

7.1. Alzheimer's Classification

More detailed results for Alzheimer's classification are provided in Table 7.

7.2. Report Generation

Figure 10 shows examples of generated reports.

7.3. Segmentation

8. Dataset Details

In this section, we discuss AnonBrain, the largest ever dataset of brain MRIs used to train an AI model.

8.1. Raw Dataset

We collected a comprehensive dataset of brain MRI scans and their corresponding clinical reports from a cancer center, covering the period from 2012 to 2017. MRI sessions typically consists of multiple MRI modality scans (e.g. FLAIR), however, in this first iteration we focus on T1-post contrast MRIs, the most informative one. In order to extract T1 post-contrast scans, we generated a long, clinician-validated list of keywords that are typically used to refer to these scans. The list contained over 50 expressions such as "Axial T1 post SENSE" or "Ax T1 POST". Around 3,000 sessions that didn't include T1 post contrast imaging were removed. The DICOMS were converted to NIFTI and we standardized the intensity values by thresholding the images at the 99th percentile and rescaling them to a range of 0-800, converting the final values to 16-bit integers. All MRI sessions were connected to a patient data storage, which enabled us to obtain patient's demographic information, treatments and diagnoses. We provide an overview of key patient data in Figure 11.

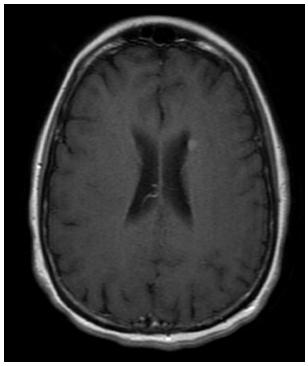
8.2. GPT-4 Processing

We processed the thousands of medical reports in parallel using GPT-4 with Python's asyncio framework. Each report underwent two GPT-4 calls: one for rewriting and another for answering specific questions. A ThreadPoolExecutor handled asynchronous API calls, with a logit bias reducing certain temporal medical terms (e.g., "increase," "new"). The temperature was set to 0.0 and top_p to 1.0 for deterministic outputs. Reports were sorted by length before processing, and asyncio.gather() improved throughput over sequential execution. Processing 80,000 reports took 48 hours. Comparisons with GPT-3.5 showed GPT-4's clear advantages, though this was before GPT-4o's release.

Figure 12 shows an example raw report including the references to prior scans and image modalities other than T1 post contrast. Figure 13 shows the prompt that was used to make GPT-4 remove these references and any PHI data. To annotate findings from the report, we used the prompt showing in Figure 14. An example of this execution is shown in Figure 15.

Pre-training Approaches		1% Training Data (n=19)				10% Training Data (n=193)				100% Training Data (n=1,932)			
Vision Model M	Weight Init.	Alz.	Normal	MCI	μ	Alz.	Normal	MCI	μ	Alz.	Normal	MCI	μ
Densenet-121	Random	0.523	0.513	0.527	0.521 [0.495, 0.547]	0.640	0.560	0.498	0.567 [0.535, 0.596]	0.724	0.629	0.535	0.629 [0.608, 0.649]
	CLS	0.514	0.517	0.511	0.514 [0.487, 0.538]	0.614	0.598	0.523	0.578 [0.555, 0.602]	0.720	0.628	0.556	0.635 [0.612, 0.650]
	Q-Former	0.565	0.525	0.486	0.526 [0.506, 0.547]	0.688	0.627	0.550	0.623 [0.604, 0.640]	0.747	0.662	0.581	0.663 [0.651, 0.681]
	brat	0.560	0.559	0.505	0.543 [0.497, 0.579]	0.720	0.644	0.518	0.628 [0.606, 0.653]	0.793	0.687	0.505	0.661 [0.650, 0.672]
ResNet-50	Random	0.497	0.566	0.541	0.535 [0.497, 0.569]	0.516	0.529	0.541	0.530 [0.498, 0.561]	0.590	0.525	0.528	0.548 [0.514, 0.586]
	brat	0.527	0.531	0.532	0.530 [0.500, 0.556]	0.621	0.456	0.452	0.510 [0.490, 0.532]	0.636	0.542	0.533	0.569 [0.519, 0.612]
ViT	Random	0.517	0.485	0.473	0.492 [0.471, 0.512]	0.554	0.491	0.502	0.515 [0.495, 0.532]	0.528	0.473	0.515	0.505 [0.476, 0.533]
	brat	0.518	0.498	0.460	0.491 [0.458, 0.523]	0.607	0.555	0.467	0.543 [0.522, 0.561]	0.622	0.521	0.450	0.531 [0.513, 0.551]

Table 7. Evaluation results (AUC scores) for different initialisations using 1%, 10%, and 100% of training data. "Alz." stands for Alzheimer's disease and "MCI" for mild cognitive impairment. The column μ is the average of the per-class AUC scores computed on the balanced test set; only this column displays the confidence interval.

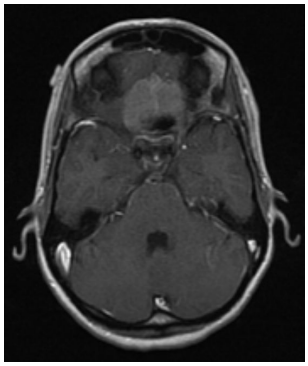


An enhancing lesion approximately 0.5 cm in size is noted within the left caudate nucleus, exhibiting minimal precontrast T1 hyperintensity without surrounding edema. This finding is suggestive of a late subacute lacunar infarct or may be related to vascular enhancement. There is no evidence of hydrocephalus, acute infarction, or extra-axial collections.

Ground-truth

The enhancing lesion within the left caudate head measures 0.6 cm. There are no enhancing parenchymal, dural, or leptomeningeal lesions. No evidence of hydrocephalus or acute infarction is present on the T1-weighted post-contrast images.

Generated



Imaged with frameless stereotactic fiducial markers, the anterior cranial fossa floor meningioma measures 4 x 3.9 cm; it fills the olfactory grooves and extends along the planum sphenoidale and orbital roofs; there is mild hyperostosis; mass effect is observed upon the paramedian inferior frontal lobes with associated T1 hyperintense changes; no suspicious enhancing mass lesions are identified; ventricles are unremarkable; no diffusion restriction to indicate acute ischemia is present.

Ground-truth

The MRI brain demonstrates an enhancing, ill-defined, heterogeneously enhancing, and heterogeneous lesion centered in the right olfactory groove with extension into the right medial aspect of the anterior cranial fossa. The lesion is closely applied to the right planum sphenoidale and cribriform plate. There is associated edema involving the anterior and inferior aspects of the right frontal lobe. The differential diagnosis includes a small meningioma as well as tumor such as esthesioneuroblastoma. No additional suspiciously enhancing lesions are identified in the brain parenchyma. There is no evidence of hydrocephalus, midline shift, or herniation.

Generated

Figure 10. Two example MRI reports generated by our VLM with a brat backbone. The reports largely capture the correct image findings.

Pre-training Approaches		1% Training Data (n=12)			10% Training Data (n=120)			100% Training Data (n=1200)		
Vision Model M	Weight Init.	Whole Tumor	Tumor Core	Enhanced Tumor	Whole Tumor	Tumor Core	Enhanced Tumor	Whole Tumor	Tumor Core	Enhanced Tumor
Densenet-121	Random	0.780	0.646	0.585	0.875	0.791	0.710	0.903	0.865	0.779
Densenet-121	brat	0.796	0.633	0.580	0.870	0.785	0.707	0.903	0.864	0.776

Table 8. Segmentation performance (Dice scores) for different pre-training initialisations using 1%, 10%, and 100% of the training data. The values correspond to the Dice scores for the Whole Tumor, Tumor Core, and Enhanced Tumor regions.

Pre-training Approaches		1% Training Data (n=12)			10% Training Data (n=120)			100% Training Data (n=1200)		
Vision Model M	Weight Init.	Whole Tumor	Tumor Core	Enhanced Tumor	Whole Tumor	Tumor Core	Enhanced Tumor	Whole Tumor	Tumor Core	Enhanced Tumor
Densenet-121	Random	0.780	0.646	0.585	0.875	0.791	0.710	0.922	0.854	0.761
brat	brat	0.796	0.633	0.580	0.870	0.785	0.707	0.925	0.867	0.762

Table 9. Segmentation performance (Lesion-wise Dice scores) for different pre-training initialisations using 1%, 10%, and 100% of the training data. The values correspond to the lesion-wise Dice scores for the Whole Tumor, Tumor Core, and Enhanced Tumor regions.

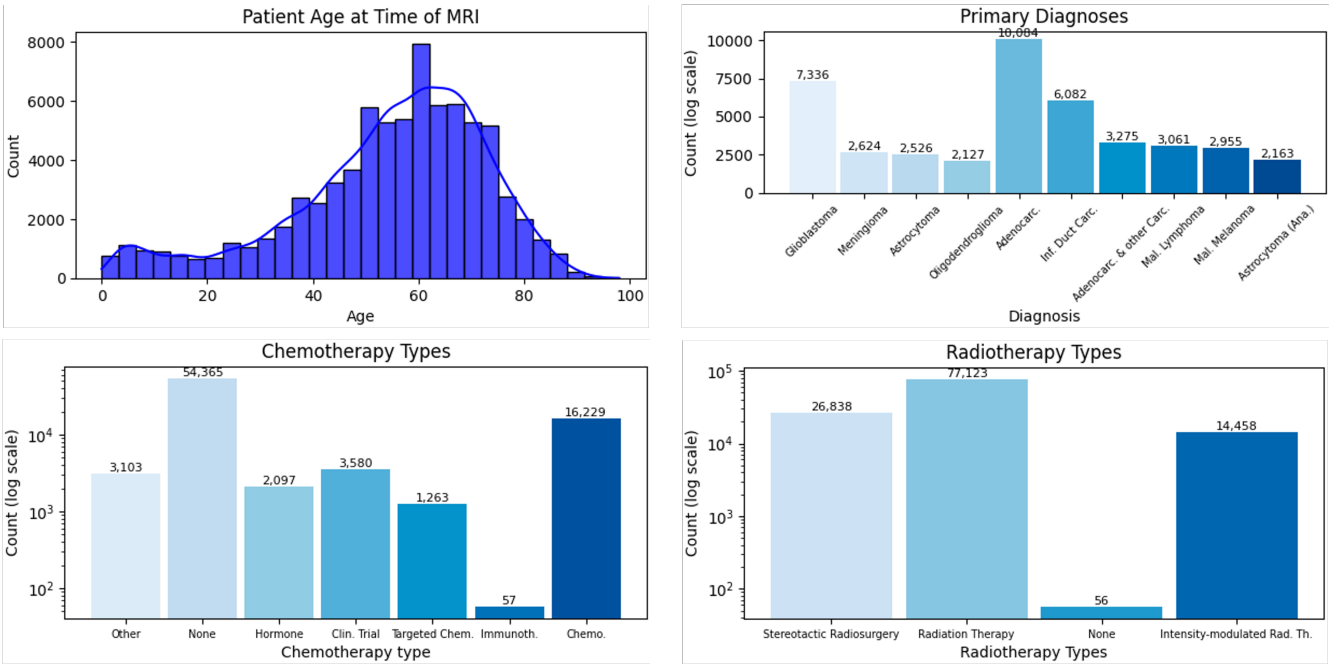


Figure 11. Participant demographics. Primary diagnoses refers to the primary cancer diagnosis for the patients for whomst the scan was ordered. Chemotherapy and radiotherapy types show a count of all the types of chemo/radio sessions assigned to the patients in this dataset.

Again status post left-sided craniotomy with stable postoperative changes and with slight increase in the heterogeneously enhancing mass lesion centered in the left temporal lobe which now measures 7.5 x 4.8 cm on image 13 series 14 from 6.7 x 4.7 cm, though the enhancement within it is more irregular and less intense than before. The mass is not completely imaged on the perfusion sequence but there is hyperperfusion inferiorly within the nodular enhancing component which is incompletely demonstrated. The surrounding hyperintense T2/FLAIR infiltrating nonenhancing signal abnormality is stable consistent with nonenhancing tumor/edema. No new discontinuous suspiciously enhancing brain lesions. There is slightly increased dilatation of the ventricles with slightly increased hyperintense T2/FLAIR signal in the periventricular white matter particularly about the frontal horns and atrium, suggesting transependymal flow of CSF from a communicating hydrocephalus. Stable mild midline shift to the right without significant downward herniation. No acute intracranial hemorrhage, infarct, or new extra-axial collections.

Figure 12. An example report showing references to prior scans in blue and descriptions of findings not visible on T1 post-contrast scans in yellow.

You are a highly experienced radiologist. Re-write the given brain MRI report and only modify the following:

(a) Leave out any details not visible on T1-weighted post-contrast images. Note that T2/FLAIR hyperintensities can often be seen on T1 Images. Observations related to e.g. perfusion, plasma volume or K trans cannot be seen and should be excluded.
(b) Leave out any terms that suggest temporal change or progression (e.g. dates, “new”, “increased”, “previous”, “now”, “compared to”, “since last”, “more”, “less”, etc.)
(c) Remove any PHI.

Figure 13. The final prompt that was used to re-write the reports and remove PHI and information not visible from the T1 post contrast images.

You are a highly experienced radiologist. Accurately answer the questions below based on the given brain MRI report. Your output must be in json format.

(a) For each question, choose the appropriate answer (wording must match exactly). If answers are mutually exclusive, choose one. If multiple answers can apply, list all that are true, separated by semicolons (";").
(b) If the MRI report does not contain information to answer a specific question, use the default answer indicating a normal status.
(c) Note the following assumptions: meningiomas are considered enhancing lesions; burr holes and ventriculostomy and Ommaya catheters are considered prior surgery; punctate lesions are less than 1cm.

Questions (Answer options):
Is there evidence of prior surgery? (Yes / No)
What kind of surgery was performed? (NA / left frontal craniotomy; right frontal craniotomy; left parietal craniotomy; right parietal craniotomy; left temporal or pterional craniotomy; right temporal or pterional craniotomy ; left occipital craniotomy; right occipital craniotomy)
Are there any enhancing lesions? (Yes / No)
What is the length of the biggest mass lesion? (NA / Less than 1cm / 1 to 2cm / More than 2cm)
Which side of the brain has more enhancing lesions? (NA / Left / Right)
List all the locations that contain enhancing lesions. (NA / Left frontal lobe; Right frontal lobe; Left parietal lobe; Right parietal lobe; Left temporal lobe; Right temporal lobe; Left occipital lobe; Right occipital lobe; Left thalamus or basal ganglia; Right thalamus or basal ganglia; Cerebellum; brainstem; cervical spinal cord)
How many enhancing lesions are there? (NA / One / Between 2 and 6 / Between 7 and 15 / More than 15)
Is there a herniation or midline shift? (Yes / No)
Are there any signs of white matter disease (e.g., leukoaraiosis or leukoencephalopathy)? (Yes / No)
Is the pituitary gland normal in appearance? (Yes / No)
Are there abnormalities in the sella or parasellar regions? (Yes / No)
Where is there evidence of invasion into or compression of adjacent structures? (Nowhere / ventricles; brainstem; cranial nerves)

Brain MRI report: ${Insert processed report}$

Figure 14. The final prompt that was used to extract structured information from the reports.

Original Report

FINDINGS: BRAIN: The peripherally enhancing centrally cystic/necrotic metastasis within inferior right frontal lobe bulging into frontal horn of the right lateral ventricle has not significantly changed in size when allowing for differences in image acquisition. Extensive right anterior hemispheric vasogenic edema which extends into the basal ganglia and anterior thalamus appears mildly improved. Mass effect resulting in right-to-left midline shift has mildly improved as well. The Ommaya catheter remains in place with its tip within necrotic core of the right inferior frontal lobe mass. The nodular dural based thickening along the anteroinferior margin of the right frontal burr hole has increased now measuring 1.8 x 1.3 cm on the axial images, having previously measured 1.1 x 0.8 cm. On the coronal images, it measures 1.5 x 0.8 cm and likely represent focal dural based tumor. The previously seen superficial focus of enhancement along the right inferior frontal gyrus appears less conspicuous. Small enhancing lesion adjacent to upper aspect of the Ommaya catheter has increased in size now measuring 4 to 5 mm. Right high parafalcine parietal hemorrhagic lesion has not significantly changed. Background white matter changes are stable.
IMPRESSION: Compared to December 21, 2011; 1. Mild progression of dural based disease along the anteroinferior margin of the right frontal burr hole. 2. Mild progression of the metastasis adjacent to upper aspect of the right Ommaya catheter. 3. Stable hemorrhagic lesion within high right parafalcine parietal lobe. 4. The heterogeneous right inferior frontal lobe mass its stable in size however, right hemispheric edema and mass effect have mildly improved.

↓

Rewritten Report

The peripherally enhancing centrally cystic/necotic metastasis is located within the inferior right frontal lobe, bulging into the frontal horn of the right lateral ventricle. There is extensive right anterior hemispheric vasogenic edema extending into the basal ganglia and anterior thalamus. Mass effect results in right-to-left midline shift. The Ommaya catheter is in place with its tip within the necrotic core of the right inferior frontal lobe mass. There is nodular dural based thickening along the anteroinferior margin of the right frontal burr hole, measuring 1.8 x 1.3 cm on the axial images. On the coronal images, it measures 1.5 x 0.8 cm and likely represents a focal dural based tumor. There is a superficial focus of enhancement along the right inferior frontal gyrus. A small enhancing lesion is adjacent to the upper aspect of the Ommaya catheter, measuring 4 to 5 mm. There is a right high parafalcine parietal hemorrhagic lesion. Background white matter changes are present.

↓

Structured Report

- Is there evidence of prior surgery? Yes
- What kind of surgery was performed? Right frontal craniotomy
- Are there any enhancing lesions? Yes
- Are there any non-enhancing lesions? Yes
- What is the length of the biggest lesion? 1 to 2 cm
- Which side of the brain has more enhancing lesions? Right
- List all the locations that contain enhancing lesions. Right frontal lobe; Right thalamus or basal ganglia
- How many enhancing lesions are there? Between 2 and 6
- Is there a herniation or midline shift? Yes
- Is there hydrocephalus? No
- What kind of hydrocephalus is there? NA
- Are there white matter changes suggestive of other pathologies? Yes
- Are there abnormalities in the cortical or subcortical grey matter suggestive of other pathologies? No
- Is the pituitary gland normal in appearance? Yes
- Are there abnormalities in the sella or parasellar regions? No
- Where is there evidence of invasion into or compression of adjacent structures? Ventricles

Figure 15. An example report and how it was processed with out pipeline.