MUSEVLP: CLINICALLY GROUNDED AND ANATOMICALLY CONSISTENT VISION-LANGUAGE PRETRAINING FOR MULTI-SEQUENCE BRAIN MRI

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain MRI sequences depict the same anatomy through complementary contrasts, posing a central challenge for vision–language pretraining: *grounding sequence-specific clinical evidence while learning sequence-invariant anatomical relationships*. Existing coarse examination-level image-report alignment leaves both aspects unresolved, obscuring the sequence and anatomical sources of individual findings while providing no explicit supervision for shared anatomical relations. We introduce **MuseVLP**, a multi-sequence vision-language pretraining framework that is clinically grounded and anatomically consistent, addressing these challenges through multi-granular clinical grounding and relational consistency. MuseVLP decomposes examination-level reports into sequence-compatible descriptions and anatomically localized findings, replacing ambiguous global alignment with *explicit* alignment between clinical evidence and its supporting sequences and regions. Building on these grounded representations, we propose sequence-invariant anatomical relational consistency (ARC), which matches each region's distribution of inter-region similarities to an exponential moving average (EMA) reference aggregated from other sequences of the same examination, encouraging invariance in anatomical relations without requiring identical regional features. We pretrain MuseVLP on 87,675 MR-RATE examinations spanning T1-weighted, T2-weighted, FLAIR, and susceptibility-weighted MRI with paired radiology reports, and evaluate it across five brain MRI datasets for linear-probe and zero-shot classification and report generation. MuseVLP consistently outperforms existing vision–language pretraining methods and transfers effectively to external datasets, suggesting that anatomical relations complement sequence-specific clinical grounding to learn transferable brain MRI representations.

1 INTRODUCTION

Brain magnetic resonance imaging (MRI) is inherently multi-sequence, with complementary contrasts such as T1-weighted, T2-weighted, fluid-attenuated inversion recovery (FLAIR), and susceptibility-weighted imaging (SWI) depicting different manifestations of the same underlying anatomy and pathology (Tae et al., 2025; Haller et al., 2021). Clinical interpretation depends on identifying both the imaging contrast that supports a finding and its anatomical location, then integrating evidence across sequences. This creates a central challenge for representation learning: *capturing sequence-specific clinical evidence while learning the anatomical relationships shared across sequences* (Chalcroft et al., 2026).

Learning transferable brain MRI representations from expert annotations is costly and limits supervision to predefined clinical labels. Visual self-supervised pretraining reduces this dependence by learning directly from unlabeled images (He et al., 2022; Tak et al., 2026), but does not explicitly connect visual patterns to the clinical concepts used for radiological interpretation. Vision-language pretraining (VLP) provides a natural source of supervision by leveraging routinely acquired radiology reports (Zhang et al., 2022; Wang et al., 2022b; Wood et al., 2026; Fu et al., 2026). Its effectiveness, however, depends on establishing meaningful correspondences between report content and the visual evidence actually available in each MRI sequence and anatomical region (Huang et al., 2021; Müller et al., 2022; Qiu et al., 2026b).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

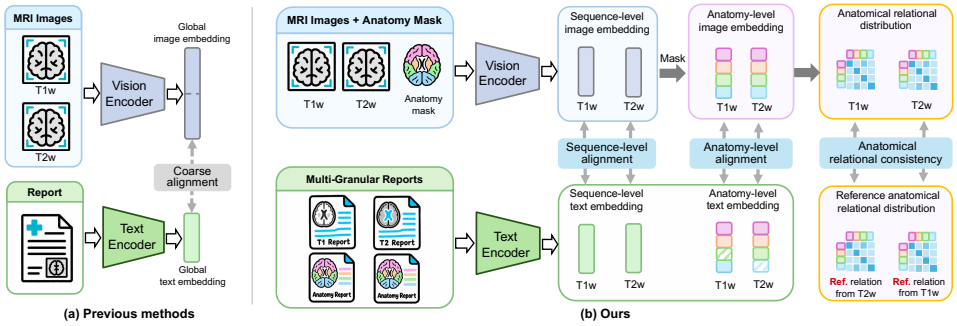

Figure 1: **Conceptual comparison with prior VLP methods**. Prior VLP methods rely on coarse examination-level alignment, overlooking sequence-specific and anatomical grounding. MuseVLP instead aligns at both sequence and anatomy levels while enforcing cross-sequence anatomical relational consistency.

Coarse examination-level image-report alignment (Fig. 1a) leaves two coupled challenges unresolved. The first is *clinical evidence attribution*. A comprehensive report integrates findings from multiple sequences and anatomical locations, whereas individual sequences provide different supporting evidence (Lei et al., 2025; Lteif et al., 2026). Pairing the complete report with every sequence can therefore assign appearance descriptions to contrasts that do not support them. Likewise, global alignment provides limited guidance about which anatomical region corresponds to each finding (Müller et al., 2022; Qiu et al., 2026b). Effective supervision must resolve both dimensions: which clinical content is compatible with a sequence, and where that content is anatomically grounded. The second challenge is *sequence-invariant anatomical relational learning*. Even when individual sequences and regions are aligned with appropriate text, these correspondences do not explicitly constrain how regional representations relate across sequences. Complementary contrasts depict the same underlying anatomy, yet can produce substantially different regional appearances (Rui et al., 2025; Qiu et al., 2026a). We hypothesize that their shared anatomical organization can be captured through the relative relations among regions, without requiring the regional features being identical. This motivates consistency at the relational level, allowing anatomical structure to provide a shared learning signal alongside contrast-specific clinical supervision.

To address these challenges, we introduce **MuseVLP**, a clinically grounded and anatomically consistent vision-language pretraining framework for multi-sequence brain MRI (Fig. 1b). MuseVLP decomposes examination-level reports into sequence-compatible descriptions and anatomically localized findings, which enables multi-granular vision-language alignment: sequence-level alignment associates each imaging contrast with compatible clinical semantics, while anatomy-level alignment connects localized findings to the corresponding regional visual representations. Building on these grounded representations, we introduce sequence-invariant Anatomical Relational Consistency (ARC), which aligns each sequence's inter-regional similarity distribution with an EMA-derived reference aggregated from the other available sequences. By matching anatomical relationships rather than features directly, ARC promotes sequence-invariant anatomical structure while preserving sequence-specific representations.

We pretrain MuseVLP on 87,675 MR-RATE examinations containing four MRI sequences paired with radiology reports (Forithmus, 2026), and evaluate it on five brain MRI datasets, where it consistently outperforms existing vision-language pretraining methods in linear-probe classification, zero-shot classification, and report generation, while showing strong transfer to external datasets.

Contributions. The contributions of this work are summarized as follows: (**i**) MuseVLP enables clinically grounded multi-sequence MRI pretraining through sequence- and anatomy-level image-text alignment; (**ii**) ARC promotes cross-sequence anatomical relational consistency using regional similarity distributions and complementary-sequence EMA references, without enforcing identical regional features; and (**iii**) evaluations on five brain MRI datasets demonstrate improved classification, report generation, and external transfer, validating the complementary benefits of clinical grounding and relational learning.

2 Related Work

Multi-sequence MRI pretraining. Multi-sequence MRI provides complementary views of the same underlying anatomy and pathology, which has motivated a range of pretraining strategies that exploit cross-sequence information. Most existing methods remain primarily visual. BrainMVP combines cross-modal reconstruction, modality-aware contrastive pretraining, and representation distillation (Rui et al., 2025); sequence-invariant contrastive pretraining promotes information shared across MRI contrasts (Chalcroft et al., 2026); and MARS further models anatomy-invariant information together with sequence-dependent variation (Qiu et al., 2026a). These approaches show that complementary sequences provide useful supervision beyond learning from each scan independently, but they do not explicitly incorporate the clinical semantics available in radiology reports.

Recent work has therefore begun to connect multi-sequence MRI with language supervision. UniBrain introduces sequence- and examination-level vision–language correspondence (Lei et al., 2025), while ReMIND explores vision–language pretraining for multi-sequence brain MRI (Lteif et al., 2026). Yet such supervision is still largely defined at the sequence or examination level, leaving the connection between sequence-specific clinical evidence and localized anatomical findings less explored. In contrast, MuseVLP introduces clinical supervision at both sequence and anatomy levels while further enforcing consistency of anatomical relationships across sequences.

Medical vision-language pretraining (VLP). Medical VLP initially focused on global image-text correspondence, with methods such as ConVIRT, MedCLIP, and BioViL learning clinically meaningful visual representations from radiology reports (Zhang et al., 2022; Wang et al., 2022b; Boecking et al., 2022). As global alignment provides limited localization of clinical evidence, later work has progressively moved toward finer-grained grounding. GLoRIA and LoVT associate localized visual representations with report semantics (Huang et al., 2021; Müller et al., 2022), MGCA introduces alignment across multiple semantic granularities (Wang et al., 2022a), and fVLM and LIGHT further incorporate explicit anatomical grounding in volumetric imaging (Shui et al., 2025; Qiu et al., 2026b). Despite this progress, fine-grained VLP primarily focuses on grounding individual anatomical representations with their corresponding clinical semantics. Although different anatomical structures are inherently related, such relationships are rarely modeled explicitly during pretraining. In contrast, our MuseVLP exploits the shared anatomical organization across complementary MRI sequences and introduces cross-sequence anatomical relational consistency, enabling fine-grained clinical grounding together with sequence-invariant anatomical relationship learning.

3 Methodology

3.1 Problem Formulation and Framework Overview

Problem formulation. Multi-sequence brain MRI provides complementary contrasts of the same underlying anatomy, while an examination-level radiology report integrates rich clinical information across the examination. Given paired MRI examinations and radiology reports, our goal is to learn transferable visual representations that capture clinically relevant semantics from the associated reports. Formally, let $\mathcal{M}$ denote the predefined set of M MRI sequences and $\mathcal{D} = \{(\mathcal{X}_i, R_i)\}_{i=1}^{N}$ the training set, where $\mathcal{X}_i = \{\boldsymbol{x}_{i,m}\}_{m\in\mathcal{M}}$ denotes the multi-sequence MRI inputs of examination i and its radiology report R_i. We denote the visual tokens of sequence m as $\boldsymbol{h}_{i,m} = f_\theta(\boldsymbol{x}_{i,m})$. The learning objective is to establish clinically meaningful correspondence between $\boldsymbol{v}_{i,m}$ and the clinical information described in R_i.

Overview of MuseVLP. Fig. 2 presents the overview of the proposed MuseVLP, which consists of two complementary components: multi-granular vision-language alignment and sequence-invariant anatomical relational consistency (ARC). MuseVLP decomposes radiology reports into clinical facts and constructs sequence- and anatomy-level supervision to ground visual representations at the corresponding granularities. Building on the anatomy-level representations, ARC then models anatomical relationships and constructs a reference relational distribution from complementary sequences to provide relational supervision. MuseVLP enables visual representations to capture more precise clinical semantics and meaningful anatomical relationships from radiology reports.

Next, we detail these two components and the overall training objective.

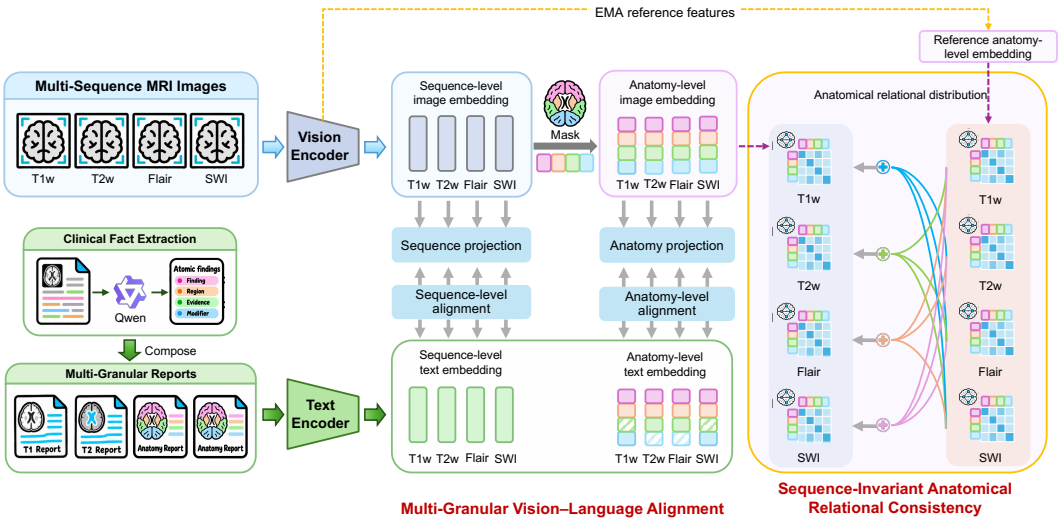

Figure 2: **Overview of MuseVLP**, which consists of two complementary components. Multi-granular vision-language alignment constructs sequence- and anatomy-level clinical supervision from radiology reports and aligns it with the corresponding MRI representations. Sequence-invariant anatomical relational consistency further exploits complementary sequences from the same examination to capture anatomical relationships shared across MRI contrasts.

3.2 Multi-Granular Vision-Language Alignment

An examination-level radiology report integrates clinical evidence across complementary MRI sequences and anatomical locations. Using the complete report to supervise every sequence therefore yields a coarse image-text correspondence, because not all reported findings are supported by every sequence and some are localized to specific anatomies. To provide more precise clinical grounding, we decompose each report into clinical facts and recompose them into sequence- and anatomy-level targets matched to the corresponding visual granularities.

Multi-granular report composition. For examination i, a semantic parser decomposes radiology report R_i into clinical facts. The extracted facts are then recomposed according to sequence compatibility and anatomical localization. For target sequence m, the resulting sequence-level description is denoted by T_i^m and retains clinical information supported by that sequence. The anatomy-level description for anatomy k is denoted by $T_i^{m,k}$ and further localizes the relevant clinical findings according to the predefined anatomy ontology.

Visual and textual representations. For each MRI sequence $\boldsymbol{x}_{i,m}$, a shared 3D visual encoder f_θ produces visual tokens $\boldsymbol{h}_{i,m} = f_\theta(\boldsymbol{x}_{i,m}), \boldsymbol{h}_{i,m} \in \mathbb{R}^{P \times c_v}$, where P denotes the number of visual tokens and c_v the feature dimension. Global pooling gives the sequence-level representation $\boldsymbol{z}_{i,m} = \mathrm{GAP}(\boldsymbol{h}_{i,m})$. For anatomy k, its anatomical mask is projected onto the visual token grid to obtain binary support $\boldsymbol{b}_{i,m,k} \in \{0,1\}^P$, and a learnable anatomy query $\boldsymbol{q}_k$ aggregates the supported tokens as $\boldsymbol{a}_{i,m,k} = \mathrm{Attn}(\boldsymbol{q}_k, \boldsymbol{h}_{i,m}[\boldsymbol{b}_{i,m,k} = 1])$. Separate visual projection heads map $\boldsymbol{z}_{i,m}$ and $\boldsymbol{a}_{i,m,k}$ to $\hat{\boldsymbol{z}}_{i,m}$ and $\hat{\boldsymbol{a}}_{i,m,k}$, while the text encoder maps T_i^m and $T_i^{m,k}$ to $\boldsymbol{t}_{i,m}$ and $\boldsymbol{t}_{i,m,k}$, respectively. All projected visual and textual representations are ℓ_2-normalized.

Sequence-level alignment. At the sequence level, each MRI representation is aligned with its sequence-compatible clinical description. For sequence m, let n_m denote the number of valid sequence-text pairs in the mini-batch. The cross-modal similarity between the visual representation of examination i and the textual representation of examination j is $s_{ij}^m = \langle \hat{\boldsymbol{z}}_{i,m}, \boldsymbol{t}_{j,m} \rangle / \tau_{\text{seq}}$. With matched image-text pairs on the diagonal, we use a symmetric contrastive objective averaged over the available MRI sequences:

$$\mathcal{L}_{\text{seq}} = -\frac{1}{M} \sum_{m=1}^{M} \frac{1}{2n_m} \sum_{i=1}^{n_m} \left[\log \frac{\exp(s_{ii}^m)}{\sum_{j=1}^{n_m} \exp(s_{ij}^m)} + \log \frac{\exp(s_{ii}^m)}{\sum_{j=1}^{n_m} \exp(s_{ji}^m)} \right]. \tag{1}$$

Thus, contrastive alignment is performed separately within each MRI sequence, preserving sequence-specific clinical grounding.

Anatomy-level alignment. At anatomy level, anatomically localized clinical descriptions are aligned with the corresponding anatomy-level visual representations. For sequence m and anatomy k, the cross-modal similarity between examinations i and j is $s_{ij}^{m,k} = \alpha_k \langle \hat{\boldsymbol{a}}_{i,m,k}, \boldsymbol{t}_{j,m,k} \rangle + \beta_k$, where α_k and β_k are learnable anatomy-specific calibration parameters. Given $n_{m,k}$ valid image-text pairs for sequence m and anatomy k, the anatomy-level objective is

$$\mathcal{L}_{\text{ana}} = -\frac{1}{M}\sum_{m=1}^{M}\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_{m,k}^2}\sum_{i,j=1}^{n_{m,k}}\log\sigma\Big[(2\mathbf{1}[i=j]-1)\, s_{ij}^{m,k}\Big]. \tag{2}$$

The sigmoid formulation σ evaluates image-text correspondences independently within each sequence-anatomy pair, so different sequences and different anatomical groups are not directly contrasted with one another.

3.3 Sequence-Invariant Anatomical Relational Consistency

Multi-granular vision-language alignment grounds clinical semantics at sequence and anatomy levels, yet anatomy-level representations are learned largely in isolation, overlooking the associations among anatomies within an examination. These inter-anatomy associations provide complementary information beyond independently grounded anatomy-level representations. To exploit this information, we introduce sequence-invariant Anatomical Relational Consistency (ARC), which explicitly models anatomical relations and uses complementary MRI sequences to provide relational guidance for visual representation learning.

Anatomical relational distribution. To characterize inter-anatomy relations, each anatomy-level representation is first projected into a relational embedding space as $\boldsymbol{d}_{i,m,k} = g_{\text{rel}}(\hat{\boldsymbol{a}}_{i,m,k})$. For anchor anatomy k, let $\ell \in \{1,\dots,K\} \setminus \{k\}$ index the remaining anatomies. We compute

$$\boldsymbol{r}_{i,m,k} = [\text{sim}(\boldsymbol{d}_{i,m,k}, \boldsymbol{d}_{i,m,\ell})]_{\ell \neq k}, \qquad \boldsymbol{p}_{i,m,k} = \text{softmax}\left(\frac{\boldsymbol{r}_{i,m,k}}{\tau_{\text{rel}}}\right), \tag{3}$$

where $\boldsymbol{p}_{i,m,k} \in \mathbb{R}^{K-1}$ denotes the normalized relational distribution of anchor anatomy k over the remaining anatomies, and τ_{rel} is the temperature parameter.

Reference anatomical relational distribution. To obtain a stable relational target, an exponential moving average (EMA) branch mirrors the visual-relational pathway and produces *reference* embeddings $\boldsymbol{d}_{i,m,k}^{\text{ref}}$. Its parameters are updated as $\theta^{\text{ref}} \leftarrow \mu\theta^{\text{ref}} + (1-\mu)\theta$, where μ denotes the EMA momentum, and gradients are not propagated through the EMA branch. The corresponding reference relational distribution $\boldsymbol{p}_{i,m,k}^{\text{ref}}$ is computed following Eq. (3). For the current sequence m, its reference relational distribution distributions of the same anchor anatomy from the *remaining* sequences are aggregated as

$$\bar{\boldsymbol{p}}_{i,m,k}^{\text{ref}} = \frac{1}{M-1}\sum_{s \neq m}\boldsymbol{p}_{i,s,k}^{\text{ref}}, \tag{4}$$

yielding a reference distribution that integrates relational estimates from complementary MRI sequences of the same examination.

Anatomical relational consistency. ARC aligns the relational distribution of the current sequence with the reference distribution using a soft-label cross-entropy objective:

$$\mathcal{L}_{\text{arc}} = -\mathbb{E}_{i,m,k}[\bar{\boldsymbol{p}}_{i,m,k}^{\text{ref}} \log \boldsymbol{p}_{i,m,k}], \tag{5}$$

which encourages more reliable anatomical relational patterns for visual representation learning.

3.4 Overall Training Objective

MuseVLP jointly optimizes sequence-level and anatomy-level visual-language alignment and and sequence-invariant anatomical relational consistency. The overall pretraining objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seq}} + \lambda_{\text{ana}}\mathcal{L}_{\text{ana}} + \lambda_{\text{arc}}\mathcal{L}_{\text{arc}}, \tag{6}$$

where λ_{ana} and λ_{arc} weight anatomy-level alignment and relational consistency, respectively.

Table 1: Linear-probe classification on four brain MRI benchmarks.

Method	ADNI			OASIS			UCSF-PDGM			MR-RATE-Heldout		
	AUC	Acc.	F1	AUC	Acc.	F1	AUC	Acc.	F1	AUC	ACC*	F1
Image-only pretraining												
MAE	0.740	0.667	0.556	0.806	0.723	0.683	0.827	0.758	0.556	0.658	0.606	0.069
BrainIAC	0.739	0.708	0.517	0.763	0.702	0.682	0.798	0.717	0.562	0.621	0.606	0.061
MARS	0.688	0.656	0.548	0.713	0.638	0.605	0.872	0.727	0.509	0.667	0.611	0.062
NeuroVFM	0.820	0.688	0.605	0.733	0.617	0.571	0.918	0.808	0.667	0.718	0.664	0.090
3DINO-ViT	0.827	0.719	0.649	0.796	0.702	0.650	0.909	0.828	0.653	0.712	0.667	0.091
Vision-language pretraining												
CT-CLIP	0.705	0.646	0.528	0.746	0.681	0.615	0.734	0.677	0.385	0.610	0.577	0.066
CT-VocabFine	0.751	0.688	0.583	0.696	0.681	0.595	0.720	0.727	0.491	0.660	0.611	0.063
CT-LiPro	0.771	0.719	0.620	0.730	0.723	0.667	0.771	0.697	0.500	0.638	0.600	0.064
COLIPRI	0.780	0.688	0.571	0.798	0.681	0.595	0.860	0.828	0.679	0.653	0.609	0.073
UniBrain	0.787	0.740	0.667	0.757	0.660	0.600	0.841	0.808	0.612	0.684	0.622	0.069
MuseVLP (**ours**)	**0.854**	**0.760**	**0.709**	**0.828**	**0.745**	**0.714**	**0.927**	**0.831**	**0.694**	**0.777**	**0.706**	**0.113**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Pretraining dataset. We pretrain MuseVLP on the official training partition of MR-RATE (Forithmus, 2026), using atlas-space brain MRI volumes paired with radiology reports. After excluding examinations without valid anatomical masks or multi-granular clinical supervision, the resulting cohort contains 87,675 examinations and is denoted as **MR-RATE-Pretrain**. Each examination is treated as one training instance, with up to 12 volumes retained while preserving sequence diversity. All downstream evaluations on MR-RATE use a disjoint held-out cohort, denoted as **MR-RATE-Heldout**, to ensure no overlap with pretraining.

Evaluation tasks and datasets. We evaluate the pretrained representations on brain MRI classification and MRI report generation. Classification is assessed under both linear-probe (LP) and zero-shot (ZS) settings. ADNI (Weiner et al., 2010) and OASIS (Marcus et al., 2007) are used for AD versus cognitively normal (CN) classification from T1-weighted MRI, UCSF-PDGM (Calabrese et al., 2022) for IDH-mutant versus IDH-wildtype classification from T1w, T1CE, T2w, and FLAIR, and MR-RATE-Heldout for multi-label recognition of 25 brain abnormalities from multi-sequence MRI. MRI report generation is evaluated on MR-RATE-Heldout and the external RadGenome-Brain benchmark (Lei et al., 2026). Dataset splits and case counts are provided in the appendix A.

Implementation details. We use a 3D ViT-Base (Dosovitskiy et al., 2021) as the shared visual encoder, initialized through masked image pretraining (He et al., 2022) on MR-RATE-Pretrain. MRI volumes are reoriented to RAS space, intensity normalized, resampled to 1-mm isotropic spacing, and spatially processed to $96 \times 108 \times 96$. Each volume is partitioned into $12 \times 12 \times 12$ patches. When multiple acquisitions of the same sequence are available, their encoded visual features are averaged. We use Qwen3.5-397B-A17B as the offline semantic parser for report processing and BiomedVLP-CXR-BERT (Boecking et al., 2022) as the text encoder. The anatomical segmentations released with MR-RATE are mapped to the atlas space used for pretraining and consolidated into 21 anatomical groups shared by anatomy-level textual grounding and mask-guided visual aggregation. We pretrain MuseVLP for 20 epochs using AdamW (Loshchilov & Hutter, 2019) with a learning rate of 1×10^{-4} and an effective batch size of 256. The loss weights are set to $\lambda_{\text{ana}} = 1$ and $\lambda_{\text{arc}} = 0.5$, with $\tau_{\text{seq}} = 0.07$ and $\tau_{\text{rel}} = 0.1$. The EMA momentum is set to $\mu = 0.996$.

Evaluation metrics. For LP classification, we report area under the ROC curve (AUC), accuracy, and F1 score. ZS classification is evaluated using AUC and average precision (AP), with mean AP (mAP) used for the multi-label MR-RATE-Heldout benchmark. For MR-RATE-Heldout, classification metrics are macro-averaged across the 25 abnormalities, with ACC* denoting macro balanced accuracy. For MRI report generation, we report BLEU-4 (Papineni et al., 2002), ROUGE-1 and ROUGE-L (Lin, 2004), BERTScore (Zhang et al., 2020), and RaTEScore (Zhao et al., 2024).

Comparison methods. We compare MuseVLP with representative image-only and vision-language pretraining methods. Image-only baselines include MAE (He et al., 2022), BrainIAC (Tak et al., 2026), MARS (Qiu et al., 2026a), NeuroVFM (Kondepudi et al., 2026), and 3DINO-ViT (Xu et al., 2025). Vision-language baselines include CT-CLIP and its VocabFine and LiPro variants (Hamamci et al., 2026), COLIPRI (Wald et al., 2025), and the brain-MRI-specific UniBrain (Lei

Table 2: Zero-shot classification across four brain MRI benchmarks.

Method	ADNI		OASIS		UCSF		MR-RATE-Heldout	
	AUC	AP	AUC	AP	AUC	AP	AUC	mAP
CT-CLIP	0.648	0.444	0.582	0.491	0.436	0.177	0.486	0.034
CT-VocabFine	0.598	0.404	0.679	0.592	0.537	0.237	0.484	0.037
CT-LiPro	0.516	0.375	0.678	0.531	0.439	0.191	0.508	0.041
COLIPRI	0.448	0.288	0.524	0.476	0.482	0.193	0.546	0.044
UniBrain	0.676	0.485	0.502	0.441	0.366	0.161	0.596	0.072
MuseVLP (**ours**)	**0.752**	**0.592**	**0.750**	**0.604**	**0.681**	**0.368**	**0.665**	**0.101**

et al., 2025). For LP classification, the visual encoder is frozen and only a linear classifier is trained. For ZS classification, all vision-language methods use the same fixed clinical prompts without task-specific training. For report generation, all pretrained visual encoders share the same downstream architecture and training protocol.

4.2 Main Results

Linear-probe classification. Table 1 evaluates the transferability of frozen visual representations across four brain MRI benchmarks. MuseVLP achieves the best performance across all reported metrics on all benchmarks. Importantly, the gains extend from multi-sequence UCSF-PDGM and MR-RATE-Heldout to the T1-only ADNI and OASIS benchmarks, indicating that representations learned from multi-sequence pretraining remain transferable when only a single MRI contrast is available downstream. The largest AUC gain over the strongest baseline occurs on MR-RATE-Heldout (0.777 vs. 0.718), where multiple abnormalities must be recognized from complementary MRI contrasts. This larger improvement on the multi-sequence task is consistent with the motivation of explicitly grounding sequence-dependent and anatomically localized clinical information.

Zero-shot classification. We next examine whether the learned image-text correspondence can be transferred directly without task-specific training. As shown in Table 2, MuseVLP achieves the highest AUC and AP/mAP across all four benchmarks. Compared with the strongest baseline, the AUC gains are 0.076 on ADNI, 0.071 on OASIS, 0.144 on UCSF-PDGM, and 0.069 on MR-RATE-Heldout. The particularly large improvement on UCSF-PDGM is accompanied by an AP gain of 0.131, suggesting that the advantage is not limited to more linearly separable visual representations but also extends to direct clinical image-text correspondence.

Fig. 3 further shows the spatial and sequence-specific evidence underlying this correspondence; this examination contains gliotic signal changes in the right periventricular white matter. Compared with UniBrain, MuseVLP produces attribution that is more consistently aligned with the report-linked anatomical region across T1w, T2w, FLAIR, and SWI. The difference is most pronounced on T2w, where MuseVLP yields a stronger and more spatially concentrated response within the target region, whereas UniBrain shows weaker and more diffuse attribution. These results suggest that MuseVLP achieves stronger image-text correspondence with more anatomically relevant and sequence-sensitive visual evidence.

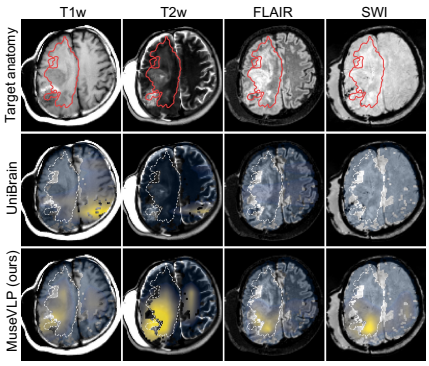

Figure 3: Sequence-specific anatomical attribution

MRI report generation. Table 3 evaluates whether the learned visual representations preserve sufficiently rich clinical information for free-text generation. On MR-RATE-Heldout, MuseVLP achieves the best performance across all metrics, improving BLEU-4 from 23.07 to 25.71 and ROUGE-L from 40.76 to 43.34 over the strongest corresponding baselines, while increasing RaTEScore from 66.90 to 68.23. The simultaneous improvements in lexical and clinically oriented metrics suggest that the learned representations preserve both descriptive detail and clinically relevant information. The advantage remains on the external RadGenome-Brain benchmark, where MuseVLP achieves 21.99 BLEU-4 and 70.38 RaTEScore and obtains the best result on every reported metric. Although the margins are smaller than on MR-RATE-Heldout, maintaining the improvement across lexical and semantic metrics indicates that the learned clinical information transfers beyond the pretraining distribution rather than reflecting only in-domain report style.

Table 3: MRI report generation results [%] on MR-RATE-Heldout and external RadGenome-Brain.

Visual Encoder	MR-RATE-Heldout					RadGenome-Brain				
	B-4 ↑	R-1 ↑	R-L ↑	BERT ↑	RaTE ↑	B-4 ↑	R-1 ↑	R-L ↑	BERT ↑	RaTE ↑
Scratch	9.34	30.38	23.14	85.04	58.39	12.48	38.29	32.73	88.64	58.10
MAE	19.89	45.72	36.24	88.40	65.99	19.11	51.05	45.36	92.12	66.99
BrainIAC	17.38	43.14	32.78	87.63	65.16	17.31	48.99	43.68	91.40	65.76
MARS	19.66	45.42	35.98	88.26	65.42	17.60	49.61	44.07	91.05	65.84
NeuroVFM	23.07	48.63	40.76	88.98	66.90	20.18	52.45	46.58	92.36	69.12
3DINO-ViT	19.03	44.39	35.40	88.28	64.54	20.94	54.00	47.40	92.41	70.05
CT-CLIP	13.63	38.12	29.93	86.66	62.54	9.10	32.00	28.12	86.36	55.85
COLIPRI	17.85	43.13	34.37	87.82	64.02	16.94	47.47	43.00	91.42	63.74
UniBrain	18.96	44.63	34.80	88.18	65.35	5.52	26.09	21.72	84.84	55.46
MuseVLP (**ours**)	**25.71**	**50.99**	**43.34**	**89.23**	**68.23**	**21.99**	**54.84**	**47.45**	**92.52**	**70.38**

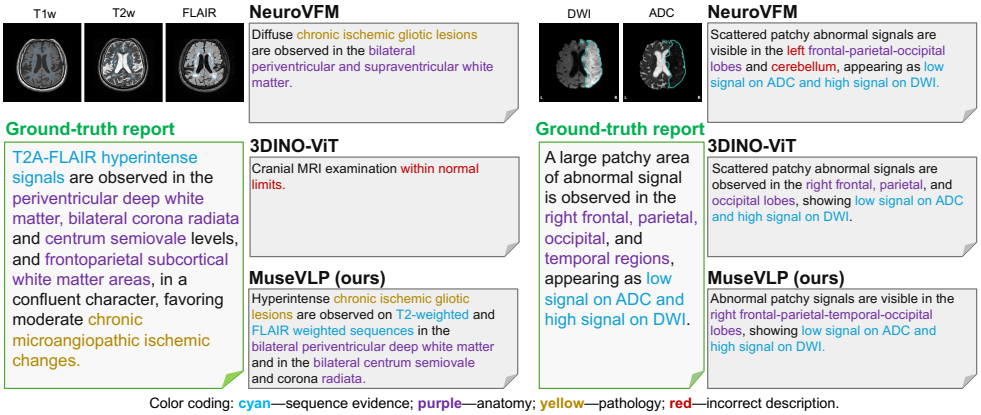

Figure 4: **Qualitative comparison for MRI report generation.** The examples are from MR-RATE-Heldout and external RadGenome-Brain.

Fig. 4 presents a representative example to highlight specific factual differences. On MR-RATE-Heldout, NeuroVFM gives an overly limited white-matter distribution, whereas MuseVLP recovers the broader corona-radiata, centrum-semiovale, and frontoparietal involvement described in the reference together with the T2/FLAIR abnormality. On RadGenome-Brain, 3DINO-ViT captures the ADC/DWI abnormality but *miss* part of the temporal lobar extent, while MuseVLP correctly preserves the right frontal-parietal-temporal-occipital involvement.

4.3 Model Analysis

Ablation studies. Table 4 presents the ablation study results on the proposed components. The baseline corresponds to conventional global image-report alignment through CLIP loss. Introducing sequence-level grounding improves all four benchmarks, increasing AUC by 0.013, 0.021, and 0.031 on ADNI, OASIS, and UCSF-PDGM, respectively, while improving RadGenome-Brain BLEU-4 by 1.69 points. This consistent change shows that resolving sequence-dependent clinical evidence provides a more precise training signal than assigning the same examination-level report to every MRI sequence.

Table 4: Component ablation study.

Variants	ADNI AUC	OASIS AUC	UCSF AUC	RadGenome B-4
Baseline	0.801	0.742	0.872	18.11
$+\mathcal{L}_{\text{seq}}$	0.814	0.763	0.903	19.80
$+\mathcal{L}_{\text{ana}}$	0.826	0.789	0.897	20.30
$+\mathcal{L}_{\text{mse}}$	0.788	0.791	0.904	20.10
$+\mathcal{L}_{\text{arc}}$ (our MuseVLP)	**0.854**	**0.828**	**0.927**	**21.99**

Adding anatomy-level grounding further improves ADNI, OASIS, and RadGenome-Brain, with gains of 0.012, 0.026, and 0.50 BLEU-4 over sequence-level grounding alone, although UCSF-PDGM shows a small decrease of 0.006 AUC. The mixed change indicates that localized clinical grounding contributes complementary information but is not uniformly sufficient across downstream tasks. Direct MSE-based relational matching also produces inconsistent effects, decreasing ADNI AUC from 0.826 to 0.788 and RadGenome-Brain BLEU-4 from 20.30 to 20.10. In contrast, ARC improves every benchmark over the same sequence-anatomy baseline, with gains of 0.028, 0.039, and 0.030 AUC on ADNI, OASIS, and UCSF-PDGM and 1.69 BLEU-4 points on RadGenome-

Brain. These results indicate that the benefit does not arise merely from adding a cross-sequence constraint; using ARC provides a more effective learning signal.

Representation geometry. We next analyze the learned feature space on MR-RATE disease classification. Fig. 5 visualizes the same downstream samples using the coarse global CLIP baseline and the complete MuseVLP across representative abnormalities. Under global alignment, positive cases are more dispersed and frequently intermixed with different parts of the negative population. In contrast, MuseVLP produces more compact disease-related groups with fewer isolated positive samples. The pattern is observed across abnormalities with different positive-sample frequencies, suggesting that the change is not specific to a single category. Although t-SNE provides only a qualitative two-dimensional projection, the more coherent local organization is consistent with the improvements observed in linear probing.

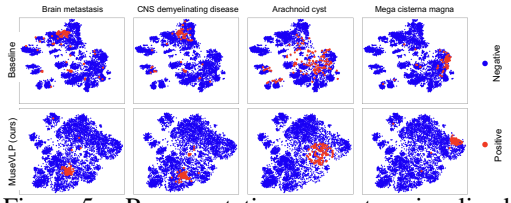

Figure 5: Representation geometry visualized by t-SNE.

Label efficiency. We further examine the transferability of the learned representations under limited downstream annotations. Fig. 6 reports linear-probe AUC using 1%, 5%, 10%, 20%, 40%, and 100% of the labeled training data on ADNI, OASIS, UCSF-PDGM, and MR-RATE-Heldout. Identical labeled subsets are used across methods at each annotation fraction. MuseVLP maintains consistent advantages across the four benchmarks, including under limited annotation budgets. The gains in the low-label regime indicate that the learned representations transfer effectively with substantially reduced task-specific supervision.

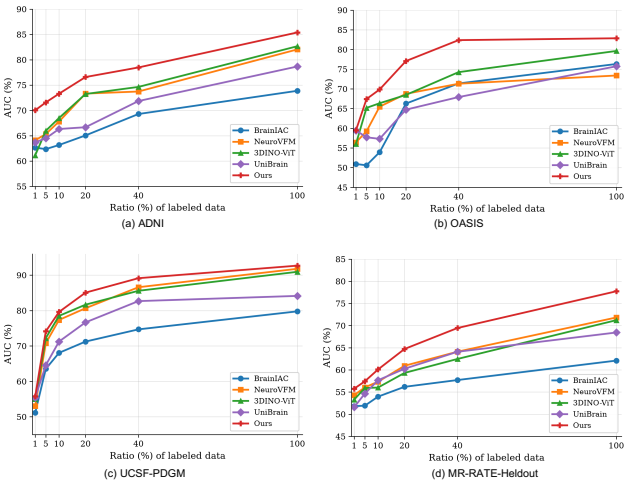

Figure 6: Label efficiency under low annotation budgets.

Please refer to Appendix F for anatomy-aware visual attribution and CLIP/SigLIP alignment ablations.

5 Conclusion

Our results highlight an important distinction in multi-sequence brain MRI representation learning: clinical evidence is often sequence-dependent, whereas the underlying anatomical relation is shared across sequences. MuseVLP explicitly separates these two forms of supervision, grounding clinical semantics at the sequence and anatomy levels while regularizing the relationships among anatomical representations across complementary contrasts. The consistent gains across classification, zero-shot recognition, report generation, and low-label transfer suggest that these two sources of information are complementary rather than interchangeable. More broadly, our findings indicate that modeling what should remain invariant across sequences can be as important as modeling the sequence-specific information they uniquely reveal.

Limitations and future work. MuseVLP currently relies on predefined anatomical masks and a fixed anatomical ontology, which may limit its flexibility for abnormalities that do not conform well to atlas-defined regions. Future work could replace this fixed partition with learned or pathology-adaptive anatomical representations. In addition, the quality of multi-granular supervision depends on the accuracy of offline report decomposition and anatomical mapping; jointly learning or refining this grounding could reduce dependence on preprocessing heuristics.

AI USE STATEMENT

In this work, we used generative AI tools to assist with radiology report processing and manuscript refinement. Specifically, a large language model was used as an offline semantic parser to extract structured clinical facts from radiology reports and support the construction of sequence- and anatomy-level supervision. Generative AI tools were also used for limited language polishing and proofreading. We did not use generative AI to formulate the core research hypothesis, design the proposed method, generate experimental results, or make final scientific interpretations and conclusions. All AI-assisted outputs used in the study were reviewed by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses de-identified brain MRI scans and radiology reports from existing research datasets. No new human participants were recruited and no additional clinical data were collected. We follow the data-use requirements and access conditions associated with each dataset, and the corresponding source studies document their applicable ethical review and informed-consent procedures. All experiments were conducted for research purposes only and do not provide clinical decisions or diagnostic recommendations. We have adhered to the ICLR Code of Ethics throughout the study and have taken care to ensure responsible use of clinical imaging data and transparent reporting of experimental procedures.

REPRODUCIBILITY STATEMENT

We have made extensive efforts to ensure the reproducibility of our work. The main paper and appendix describe the preprocessing pipeline, report construction procedure, anatomical grouping, model architecture, training objectives, hyperparameter settings, dataset splits, downstream evaluation protocols, and evaluation metrics. The proposed objectives and training procedure are specified in detail in the Methodology section, with additional implementation details provided in the appendix. We plan to release the source code and relevant pretrained weights after the double-blind review to facilitate reproducibility and future research.

REFERENCES

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie L. Hyland, Maria Wetscherek, Tristan Naumann, Aditya V. Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision-language processing. In *Computer Vision – ECCV 2022*, pp. 1–21. Springer, 2022. doi: 10.1007/978-3-031-20059-5_1.

Evan Calabrese, Javier E. Villanueva-Meyer, Jeffrey D. Rudie, Andreas M. Rauschecker, Ujjwal Baid, Spyridon Bakas, Soonmee Cha, John T. Mongan, and Christopher P. Hess. The university of california san francisco preoperative diffuse glioma MRI dataset. *Radiology: Artificial Intelligence*, 4(6), 2022. doi: 10.1148/ryai.220058.

Liam Chalcroft, Jenny Crinion, Cathy J. Price, and John Ashburner. Unified 3d MRI representations via sequence-invariant contrastive learning. In *Simulation and Synthesis in Medical Imaging*, pp. 63–74. Springer, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Forithmus. MR-RATE: A vision-language foundation model and dataset for magnetic resonance imaging. Hugging Face Dataset and Official Code Release, 2026. Dataset release; accompanying paper forthcoming.

Yibing Fu, Yunpeng Zhao, Zhitao Zeng, Cheng Chen, and Yueming Jin. Unleashing the power of image-tabular self-supervised learning via breaking cross-tabular barriers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4049–4057, 2026.

Sven Haller, E. Mark Haacke, Majda M. Thurnher, and Frederik Barkhof. Susceptibility-weighted imaging: Technical essentials and clinical neurologic applications. *Radiology*, 299(1):3–26, 2021. doi: 10.1148/radiol.2021203071.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, Weicheng Dai, Murong Xu, Hadrien Reynaud, Muhammed Furkan Dasdelen, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah Alanbay, Anjany Sekuboyina, Berkan Lafci, Ahmet Kaplan, Zhiyong Lu, Malgorzata Polacin, Bernhard Kainz, Christian Bluethgen, Kayhan Batmanghelich, Mehmet Kemal Ozdemir, and Bjoern Menze. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering*, 10:1610–1628, 2026. doi: 10.1038/s41551-025-01599-y.

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*, pp. 16000–16009, 2022. doi: 10.1109/CVPR52688.2022.01553.

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*, pp. 3942–3951, 2021. doi: 10.1109/ICCV48922.2021.00391.

Akhil Kondepudi, Akshay Rao, Chenhui Zhao, Yiwei Lyu, Samir Harake, Soumyanil Banerjee, Jacob Ogle, Rushikesh Joshi, Anna-Katharina Meissner, Xinhai Hou, Cheng Jiang, Asadur Chowdury, Ashok Srinivasan, Brian Athey, Vikas Gulani, Aditya Pandey, Honglak Lee, and Todd Hollon. Health system learning enables generalist neuroimaging models. *Nature Medicine*, 32:2831–2837, 2026. doi: 10.1038/s41591-026-04497-1.

Jiayu Lei, Lisong Dai, Haoyun Jiang, Chaoyi Wu, Xiaoman Zhang, Yao Zhang, Jiangchao Yao, Weidi Xie, Yanyong Zhang, Yuehua Li, Ya Zhang, and Yanfeng Wang. UniBrain: Universal brain MRI diagnosis with hierarchical knowledge-enhanced pre-training. *Computerized Medical Imaging and Graphics*, 122:102516, 2025. doi: 10.1016/j.compmedimag.2025.102516.

Jiayu Lei, Xiaoman Zhang, Chaoyi Wu, Lisong Dai, Ya Zhang, Yanyong Zhang, Yanfeng Wang, Weidi Xie, and Yuehua Li. Interpretable brain MRI report generation anchored by lesion topography. *IEEE Journal of Biomedical and Health Informatics*, 30(7):5935–5950, 2026. doi: 10.1109/JBHI.2025.3646647.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, 2004.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Diala Lteif, Shuyue Jia, Subhrangshu Bit, Artem Kaliaev, Asim Z. Mian, Juan E. Small, Balamurugan Mangaleswaran, Bryan A. Plummer, Sarah A. Bargal, Rhoda Au, and Vijaya B. Kolachalama. Vision-language framework for multi-sequence brain magnetic resonance imaging. *medRxiv*, 2026. doi: 10.64898/2026.03.30.26349106. Preprint.

Daniel S. Marcus, Tracy H. Wang, Jamie Parker, John G. Csernansky, John C. Morris, and Randy L. Buckner. Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 19 (9):1498–1507, 2007. doi: 10.1162/jocn.2007.19.9.1498.

Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert. Joint learning of localized representations from medical images and reports. In *Computer Vision – ECCV 2022*, pp. 685–701. Springer, 2022. doi: 10.1007/978-3-031-19809-0_39.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Zelin Qiu, Xi Wang, Zhuoyao Xie, Juan Zhou, Yu Wang, Lingjie Yang, Xinrui Jiang, Juyoung Bae, Moo Hyun Son, Qiang Ye, Dexuan Chen, Rui Zhang, Tao Li, Neeraj Ramesh Mahboobani, Varut Vardhanabhuti, Xiaohui Duan, Yinghua Zhao, and Hao Chen. Large-scale multi-sequence pretraining for generalizable MRI analysis in versatile clinical applications. *Nature Biomedical Engineering*, 2026a. doi: 10.1038/s41551-026-01740-5.

Zhaoyu Qiu, Yuxiao Liu, Jianrui Li, Lei Jin, Zhongxiang Ding, Feng Shi, and Dinggang Shen. LIGHT: Learning image-text grounding for hierarchical tumor localization and subtype classification in multimodal brain MRI. *IEEE Transactions on Medical Imaging*, 2026b. doi: 10.1109/TMI.2026.3718445.

Shaohao Rui, Lingzhi Chen, Zhenyu Tang, Lilong Wang, Mianxin Liu, Shaoting Zhang, and Xiaosong Wang. Multi-modal vision pre-training for medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5164–5174, 2025.

Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua Ye, Tingbo Liang, Qi Zhang, and Ling Zhang. Large-scale and fine-grained vision-language pretraining for enhanced CT image understanding. In *The Thirteenth International Conference on Learning Representations*, 2025.

Woo-Suk Tae, Byung-Joo Ham, Sung-Bom Pyun, and Byung-Jo Kim. Current clinical applications of structural MRI in neurological disorders. *Journal of Clinical Neurology*, 21(4):277–293, 2025. doi: 10.3988/jcn.2025.0185.

Divyanshu Tak, Biniam A. Garomsa, Anna Zapaishchykova, Tafadzwa L. Chaunzwa, Juan Carlos Climent Pardo, Zezhong Ye, John Zielke, Yashwanth Ravipati, Suraj Pai, Sri Vajapeyam, Maryam Mahootiha, Mitchell Parker, Luke R. G. Pike, Ceilidh Smith, Ariana M. Familiar, Kevin X. Liu, Sanjay Prabhu, Omar Arnaout, Pratiti Bandopadhayay, Ali Nabavizadeh, Sabine Mueller, Hugo J. W. L. Aerts, Raymond Y. Huang, Tina Y. Poussaint, and Benjamin H. Kann. A generalizable foundation model for analysis of human brain MRI. *Nature Neuroscience*, 29(4):945–956, 2026. doi: 10.1038/s41593-026-02202-6.

Tassilo Wald, Ibrahim Ethem Hamamci, Yuan Gao, Sam Bond-Taylor, Harshita Sharma, Maximilian Ilse, Cynthia Lo, Olesya Melnichenko, Noel C. F. Codella, Maria Teodora Wetscherek, Klaus H. Maier-Hein, Panagiotis Korfiatis, Valentina Salvatelli, Javier Alvarez-Valle, and Fernando Pérez-García. Comprehensive language-image pre-training for 3d medical image understanding. *arXiv preprint arXiv:2510.15042*, 2025.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems*, 35:33536–33549, 2022a.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3876–3887. Association for Computational Linguistics, 2022b. doi: 10.18653/v1/2022.emnlp-main.256.

Michael W. Weiner, Paul S. Aisen, Clifford R. Jack, William J. Jagust, John Q. Trojanowski, Leslie Shaw, Andrew J. Saykin, John C. Morris, Nigel Cairns, Laurel A. Beckett, Arthur Toga, Robert Green, Sarah Walter, Holly Soares, Peter Snyder, Eric Siemers, William Potter, Patricia E. Cole, Mark Schmidt, and Alzheimer's Disease Neuroimaging Initiative. The alzheimer's disease neuroimaging initiative: Progress report and future plans. *Alzheimer's & Dementia*, 6(3):202–211.e7, 2010. doi: 10.1016/j.jalz.2010.03.007.

David A. Wood, Emily Guilhem, Sina Kafiabadi, Ayisha Al Busaidi, Kishan Dissanayake, Ahmed Hammam, Nina Mansoor, Matthew Townend, Siddharth Agarwal, Yiran Wei, Asif Mazumder, Gareth J. Barker, Peter Sasieni, Sébastien Ourselin, James H. Cole, Nikhil Nair, Anil Geetha, Chike Onyekwuluje, Rob Dineen, Permesh Dhillon, Carolyn Costigan, Kavi Fatania, Mark Igra,

Rebecca Nichols, Janak Saada, Arne Juette, Ramona-Rita Barbara, Hilmar Spohr, Thomas C. Booth, and MIDI Consortium Group. Self-supervised text-vision alignment for automated brain MRI abnormality detection: A multicenter study (ALIGN study). *Radiology: Artificial Intelligence*, 8(2):e240619, 2026. doi: 10.1148/ryai.240619.

Tony Xu, Sepehr Hosseini, Chris Anderson, Anthony Rinaldi, Rahul G. Krishnan, Anne L. Martel, and Maged Goubran. A generalizable 3d framework and model for self-supervised learning in medical imaging. *npj Digital Medicine*, 2025. doi: 10.1038/s41746-025-02035-w.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020.

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 *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pp. 2–25. PMLR, 2022.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. RaTEScore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15004–15019. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.836.

Appendix of Paper: "MuseVLP: Clinically Grounded and Anatomically Consistent Vision-Language Pretraining for Multi-sequence Brain MRI"

This appendix complements the main paper with additional details and analyses of MuseVLP. Specifically, it includes:

- **Dataset details:** composition of MR-RATE-Pretrain and exact downstream dataset splits;
- **Multi-granular clinical supervision:** anatomical ontology, report decomposition and composition, and the distributions of sequence- and anatomy-level supervision;
- **Pretraining implementation:** model architecture, projection modules, optimization settings, and ARC configuration;
- **Downstream evaluation protocols:** linear probing, zero-shot classification, and MRI report generation;
- **Detailed MR-RATE results:** the 25 brain abnormality categories and their class-wise classification performance;
- **Additional representation analyses:** anatomical visual grounding, comparisons of CLIP- and SigLIP-style objectives for multi-granular vision-language alignment.

A Dataset Details

MR-RATE pretraining cohort. The final MR-RATE-Pretrain cohort contains 87,675 examinations from the official training partition. Because one examination may include multiple MRI sequences and repeated acquisitions, pretraining statistics are reported at the MRI volume level. After sequence filtering, the cohort contains 627,267 volumes. We retain at most 12 volumes per examination, resulting in 617,152 volumes used in each pretraining epoch. For examinations exceeding this limit, volumes are grouped by MRI sequence and sampled in a round-robin manner to preserve sequence diversity, with the selected subset fixed across epochs. As shown in Fig. S1, T1w, T2w, FLAIR, and SWI together account for 99.98% of the resulting training volumes.

Downstream splits. Table S1 reports the exact dataset splits used for downstream evaluation. All MR-RATE downstream examinations are excluded from MR-RATE-Pretrain.

B Multi-Granular Clinical Supervision

B.1 Anatomical Ontology

As the Table S2 show,We consolidate the anatomical segmentations released with MR-RATE into $K = 21$ mask-supported anatomical groups. Left and right structures are retained separately. The same ontology is used for anatomy-level textual grounding and mask-guided visual aggregation.

B.2 Report Decomposition and Composition

Radiology reports are processed offline using Qwen3.5-397B-A17B. Each examination-level report is decomposed into source-traceable clinical facts containing the clinical finding, anatomical scope, explicitly stated sequence evidence, and associated modifiers. The extracted facts are then converted into sequence- and anatomy-conditioned supervision using deterministic rules.

Sequence assignment. Sequence aliases are canonicalized to T1w, T2w, FLAIR, and SWI. For a target sequence, we retain facts explicitly supported by that sequence. Facts without explicit sequence evidence are shared across sequence views only when they describe sequence-independent cranial findings and do not encode contrast-specific imaging appearances. Evidence explicitly attributed to another sequence is excluded. Findings supported only by unavailable sequences, such as DWI or ADC, are retained in the structured fact records but excluded from the four sequence-conditioned supervision views.

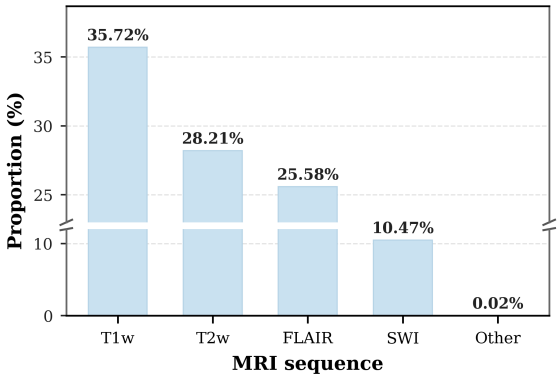

Figure S1: **MRI sequence composition of MR-RATE-Pretrain.** Proportions are reported at the MRI volume level after applying the 12-volume-per-examination limit.

Table S1: Dataset splits used for downstream evaluation.

Dataset	MRI sequences	Split
ADNI	T1w	500 / 125
OASIS	T1w	188 / 47
UCSF-PDGM	T1w, T1CE, T2w, FLAIR	396 / 99
MR-RATE-Heldout	T1w, T2w, FLAIR, SWI	7,448 / 1,862
RadGenome-Brain	Multi-sequence MRI	536 / 76 / 152

Anatomical assignment. Anatomical mentions are normalized using their reported structure, location, and laterality and resolved to fine-grained atlas labels using a fixed rule-based mapping. Specific structures are matched before broader lobe- or hemisphere-level descriptions, and laterality is preserved throughout the mapping. The resolved labels are then consolidated into the 21 anatomical groups defined in Table S2, which are shared by anatomy-level textual supervision and mask-guided visual aggregation. Mentions that cannot be reliably mapped to the supported atlas are omitted from anatomy-level supervision.

For target sequence m, the resulting description T_i^m contains sequence-compatible clinical information, while $T_i^{m,k}$ further localizes the relevant content to anatomy k. Table S3 illustrates the resulting multi-granular supervision.

B.3 Multi-Granular Supervision Distribution

We further summarize the distribution of the constructed supervision at the sequence and anatomy levels in Fig. S2. For sequence-level supervision, we count sequence-conditioned report views containing at least one explicitly sequence-linked finding and exclude shared fallback text. Among 87,675 pretraining examinations, 82,675 such views are identified, comprising 8,376 T1w (10.13%), 26,952 T2w (32.60%), 17,867 FLAIR (21.61%), and 29,480 SWI (35.66%) views. For anatomy-level supervision, the final pretraining set contains 772,199 valid textual targets across the 21 anatomical groups. For visualization, left and right groups are merged into 12 anatomical families, and their proportions are shown in Fig. S2(b).

C Pretraining Implementation Details

Table S4 summarizes the implementation details complementary to those reported in the main paper.

For each MRI sequence, duplicate acquisitions belonging to the same sequence are encoded independently and their visual features are averaged before sequence-level alignment. Anatomy-level representations are obtained by mask-guided aggregation using the corresponding anatomical sup-

Table S2: The 21 anatomical groups used by MuseVLP. Paired structures are represented as separate left and right groups.

Anatomical group family	Groups
Brain Stem	1
Basal Ganglia	Left / Right
Cerebellum	Left / Right
Cerebral White Matter	Left / Right
Frontal Cortex	Left / Right
Medial Temporal Limbic Region	Left / Right
Occipital Cortex	Left / Right
Parietal Cortex	Left / Right
Temporal Cortex	Left / Right
Thalamus and Ventral Diencephalon	Left / Right
Ventricular System	1
Cerebellar Vermis	1
Total	**21**

Table S3: Example of multi-granular clinical supervision constructed from an examination-level radiology report.

Level	Target	Example clinical supervision
Clinical fact	Multi-sequence lesions	Numerous millimetric lesions involving the basal ganglia, thalamus, midbrain, cerebellar tonsil, and vermis, with T1 hypointensity, T2/FLAIR hyperintensity, and susceptibility-related signal on SWI.
Sequence level	T1w	Numerous millimetric lesions are present in the reported anatomical locations and are hypointense on T1-weighted MRI.
Sequence level	FLAIR	Numerous millimetric lesions are hyperintense on FLAIR, with additional gliotic signal around the reported cerebellar hemorrhagic foci.
Anatomy level	Right Basal Ganglia, T2w	Numerous millimetric lesions are present in the right basal ganglia and are hyperintense on T2-weighted MRI.
Anatomy level	Brain Stem, SWI	Numerous millimetric lesions are present in the midbrain with susceptibility signal potentially related to calcification or hemorrhage.

port. The EMA branch is updated without gradient propagation and provides the reference relational distributions used by ARC.

D Downstream Evaluation Protocols

Unless otherwise specified, MAE and MuseVLP use visual encoders pretrained on MR-RATE-Pretrain. For the external comparison methods, we use their officially released or project-provided pretrained weights with the native architectures and source pretraining described in the corresponding publications.

D.1 Linear Probing

For all linear-probe experiments, the pretrained visual encoder is frozen. The mean of the final visual tokens is used as the volume-level representation. No text encoder or anatomical mask is used during downstream probing.

For multi-sequence datasets, representations from the available MRI sequences are aggregated at the examination level before classification. Table S5 summarizes the dataset-specific protocols.

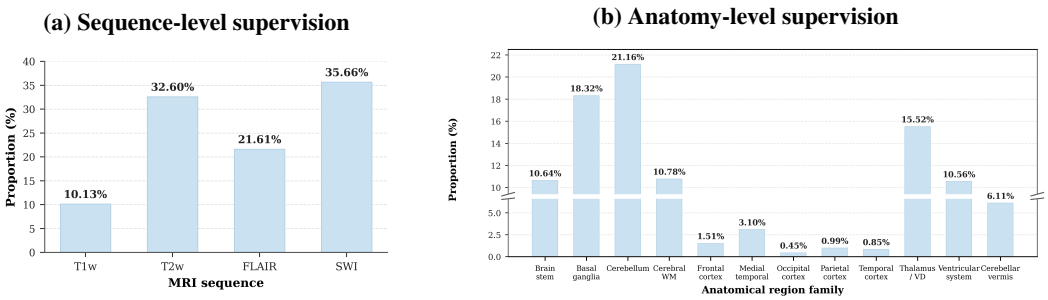

Figure S2: **Distribution of multi-granular clinical supervision. (a)** Distribution of sequence-level textual supervision across MRI sequences. **(b)** Distribution of anatomy-level textual supervision across anatomical families. The 21 anatomical groups used during training are merged into 12 families for visualization.

Table S4: Pretraining configuration of MuseVLP.

Setting	**Configuration**
Visual encoder	3D ViT-Base, initialized by masked image pretraining on MR-RATE-Pretrain
Input size	$96 \times 108 \times 96$
Patch size	$12 \times 12 \times 12$, producing an $8 \times 9 \times 8$ token grid
Visual hidden dimension	768
Text encoder	BiomedVLP-CXR-BERT
Sequence projection	$768 \to 256$
Anatomy projection	$768 \to 256$
Anatomical groups	21
Relation projector	$256 \to 512 \to 512 \to 512$
Relation similarity	Cosine similarity
Optimizer	AdamW
Learning rate	1×10^{-4}
AdamW β_1, β_2	$(0.9, 0.98)$
Weight decay	0.1
Training epochs	20
Effective batch size	256
Learning-rate schedule	Cosine decay with 200 warm-up updates
Precision	BF16
Training hardware	$8\times$ NVIDIA H200 GPUs
Sequence temperature	$\tau_{\text{seq}} = 0.07$
Relation temperature	$\tau_{\text{rel}} = 0.1$
Loss weights	$\lambda_{\text{ana}} = 1$, $\lambda_{\text{arc}} = 0.5$
EMA momentum	$\mu = 0.996$

D.2 Zero-Shot Classification

Zero-shot evaluation uses the frozen image and text encoders without task-specific optimization. Both image and text representations are ℓ_2-normalized.

Given positive and negative clinical descriptions, the classification score is

$$s(\boldsymbol{x}) = \cos(\boldsymbol{v}, \boldsymbol{t}^+) - \cos(\boldsymbol{v}, \boldsymbol{t}^-). \tag{S1}$$

The same fixed prompts are used for all compared vision-language methods. For each target disease c, we use the positive template "There is c." and the negative template "There is no c.". Accordingly, ADNI and OASIS use Alzheimer's disease as the target concept, UCSF-PDGM uses IDH mutation status, and MR-RATE-Heldout uses the corresponding disease name for each of the 25 abnormal-

Table S5: Linear-probing protocols for downstream classification.

Dataset	Input	Classifier	Feature aggregation
ADNI	T1w	Logistic regression	Single-sequence feature.
OASIS	T1w	Logistic regression	Single-sequence feature.
UCSF-PDGM	T1w, T1CE, T2w, FLAIR	Logistic regression	Mean pooling over available sequences.
MR-RATE-Heldout	T1w, T2w, FLAIR, SWI	One-vs-rest logistic regression	Mean pooling over available sequences.

Note. All datasets use an 80/20 train-test split. Features are standardized using statistics computed from the training set only. Logistic regression uses balanced class weights, $C = 1.0$, and a decision threshold of 0.5. These hyperparameters are fixed across methods, and no validation-based model selection is performed. For MR-RATE-Heldout, one binary classifier is fitted independently for each abnormality, and the reported metrics are macro-averaged over the 25 selected brain abnormalities. ACC* denotes macro balanced accuracy.

ities. For multi-sequence examinations, visual representations are aggregated at the examination level before similarity computation.

D.3 MRI Report Generation

For report generation, all pretrained visual encoders are evaluated under a shared downstream generation framework while retaining their native visual architectures and tokenizations. The visual encoder is frozen throughout downstream training. MAE and MuseVLP are pretrained on MR-RATE-Pretrain, whereas the external baselines use their released or project-provided pretrained checkpoints with the native architectures and source pretraining described in the corresponding publications.

For each available MRI sequence, we extract the final spatial feature map or patch-token sequence before global pooling and remove special tokens when present. For visual encoder b, the resulting representation is denoted as $\mathbf{X}_b \in \mathbb{R}^{N_b \times d_b}$, where N_b and d_b denote its native number of spatial tokens and feature dimension, respectively. To bound the computational cost of the shared report decoder, representations with more than 576 spatial tokens per sequence are adaptively average-pooled to at most 576 tokens while preserving their spatial organization whenever the native feature-grid geometry is available. Encoders with fewer native tokens retain their original token counts rather than being artificially upsampled. Accordingly, BrainIAC retains 216 tokens per sequence and 3DINO-ViT retains 343 tokens per sequence, whereas the remaining dense encoders use at most 576 tokens per sequence. The native feature dimension is 512 for CT-CLIP, 1024 for 3DINO-ViT, and 768 for the remaining evaluated encoders.

We use `healx/gpt-2-pubmed-medium` as the common report decoder and keep its parameters frozen. For each visual encoder, a method-specific trainable linear projection followed by LayerNorm maps the native feature dimension d_b to the 1,024-dimensional decoder space:

$$\widetilde{\mathbf{X}}_b = \mathrm{LN}\left(\mathbf{X}_b \mathbf{W}_b + \mathbf{b}_b\right), \qquad \mathbf{W}_b \in \mathbb{R}^{d_b \times 1024}.$$

The projected visual representations are injected into the decoder through trainable pseudo-attention key and value projections at each decoder layer. Thus, downstream optimization updates only the visual projection and visual-conditioning modules, while both the pretrained visual encoder and language decoder remain fixed.

For multi-sequence examinations, representations from the available FLAIR, T1w, T2w, and SWI sequences are concatenated in a fixed order. Missing sequences are handled by zero padding together with an attention mask that excludes padded positions. The maximum visual-token budget is 2,304 tokens per examination, corresponding to four sequences with at most 576 tokens per sequence. Models with lower native spatial resolution naturally use fewer tokens, e.g., at most 864 tokens for BrainIAC and 1,372 tokens for 3DINO-ViT.

Table S6: Training and inference configuration for MRI report generation.

Configuration	Setting
Visual encoder	Frozen
Report decoder	`healx/gpt-2-pubmed-medium`
Decoder parameters	Frozen
Trainable modules	Visual projection and conditioning modules
Maximum visual tokens	2,304 per examination
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	1×10^{-4}
Adam β_1, β_2	0.9, 0.999
Training epochs	20
Precision	BF16
Gradient clipping	1.0
LR schedule	200-step warm-up, then constant
Decoding	Greedy decoding without sampling
Maximum generation length	300 tokens

All methods share the same frozen report decoder, adapter design, downstream training split, optimization settings, and inference protocol within each benchmark, while retaining the native architecture, spatial resolution, and channel dimensionality of each pretrained visual encoder. The input dimension of the visual projection is determined by the native feature dimension of each encoder. This evaluation therefore compares the transferability of pretrained visual representations under a unified downstream generation framework, rather than performing a backbone-controlled comparison with identical visual architectures.

MR-RATE-Heldout uses the 7,448 / 1,862 downstream train/test split, whereas RadGenome-Brain follows its official 536 / 76 / 152 train, validation, and test split. The training and inference configurations are summarized in Table S6.

All models are trained for 20 epochs, and the final checkpoint is used for evaluation without validation-based model selection. Generated reports are evaluated using BLEU-4, ROUGE-1, ROUGE-L, BERTScore, and RaTEScore, following the metrics reported in the main paper.

E MR-RATE Brain Abnormality Label Set

MR-RATE-Heldout label set and per-class results. MR-RATE-Heldout evaluates multi-label classification over 25 brain abnormalities. The same label set is used for linear-probe and zero-shot evaluation. Table S7 reports the per-class linear-probe performance.

F Additional Representation Analysis

F.1 Anatomy-aware Visual Attribution

We further examine whether the disease-related evidence used for zero-shot classification is spatially associated with the corresponding anatomy described in the report. Using the zero-shot classification score defined in Sec. D.2 as the attribution target, we backpropagate its gradient to the L10 visual Transformer features and compute Grad-CAM on the $8 \times 9 \times 8$ patch grid. The resulting attribution map is upsampled to the input resolution for visualization. Contours indicate the report-associated atlas regions and are shown only as anatomical references.

As shown in Fig. S3, sequence-level grounding produces relatively diffuse responses, whereas adding anatomy-level supervision yields stronger and more spatially concentrated responses within the report-associated regions. This suggests that anatomy-level clinical grounding improves the spatial specificity of disease-related visual evidence.

Table S7: Per-class linear-probe classification results on MR-RATE-Heldout.

Brain abnormality	**Pos.**	**AUC**	**ACC***	**F1**
Brain metastasis	43	.953	.911	.146
Intracranial meningioma	13	.755	.671	.018
CNS demyelinating disease	37	.881	.804	.086
Lacunar infarct	10	.724	.620	.013
Silent cerebral microhemorrhage	3	.831	.630	.004
Cavernous hemangioma	9	.630	.577	.007
Subdural intracranial hemorrhage	1	.989	.945	.005
Gliosis	652	.657	.613	.481
Cerebral atrophy	134	.900	.820	.333
Encephalomalacia	12	.909	.779	.038
Arachnoid cyst	69	.866	.798	.178
Empty sella syndrome	71	.604	.568	.057
Intracranial aneurysm	3	.310	.394	.000
Chiari malformation	8	.749	.729	.012
Schwannoma	4	.809	.661	.006
Pineal cyst	23	.747	.685	.027
Rathke's pouch cyst	2	.347	.385	.000
Cerebral edema	88	.964	.907	.323
Cerebellar degeneration	4	.828	.557	.004
Mega cisterna magna	19	.984	.926	.094
Cavum septum pellucidum	16	.766	.681	.024
Watershed infarct	2	.870	.694	.005
Choroid plexus cyst	3	.803	.760	.007
Glioma	7	.956	.887	.038
Pituitary adenoma	6	.602	.649	.008
Macro average	–	**.777**	**0.706**	**0.113**

Note. Pos. denotes the number of positive test cases for each abnormality. ACC* denotes balanced accuracy. Macro averages are computed across all 25 abnormalities.

Table S8: Comparison of CLIP- and SigLIP-style objectives for sequence- and anatomy-level vision–language alignment. Classification results are linear-probe AUC, and report generation is evaluated by BLEU-4 on RadGenome-Brain.

Setting	$\mathcal{L}_{\text{seq}}$	$\mathcal{L}_{\text{ana}}$	**UCSF-PDGM**	**ADNI**	**OASIS**	**RadGenome B4**
–	SigLIP	CLIP	.816	.781	.774	19.52
–	SigLIP	SigLIP	.824	.804	.787	18.36
–	CLIP	CLIP	.903	.828	.814	20.07
MuseVLP (ours)	CLIP	SigLIP	**.927**	**.854**	**.828**	**21.99**

F.2 Additional ablation

We compare CLIP- and SigLIP-style objectives for sequence- and anatomy-level vision-language alignment while keeping the remaining training configuration unchanged. The results are summarized in Table S8.

The results show that the two supervision granularities favor different alignment objectives. CLIP-style alignment performs better at the sequence level, where batch-wise contrastive competition helps preserve discriminative global clinical semantics. In contrast, SigLIP-style alignment is more effective at the anatomy level, where anatomical descriptions can be semantically similar across examinations and independent pairwise matching reduces unnecessary competition between such pairs. The combination of sequence-level CLIP and anatomy-level SigLIP achieves the best overall performance and is adopted in MuseVLP.

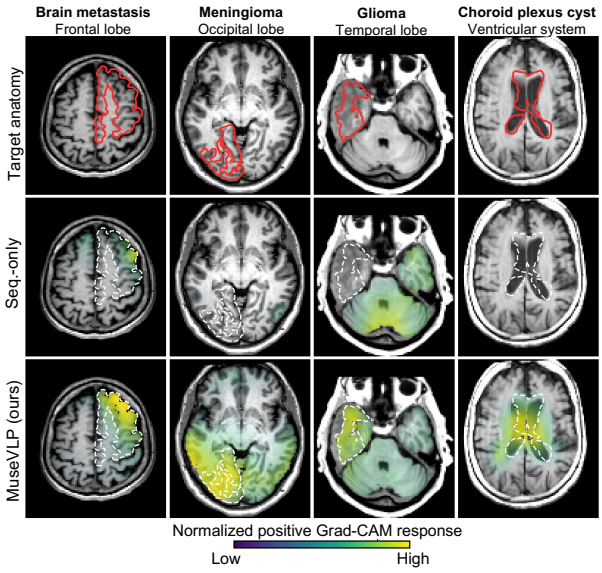

Figure S3: Attribution maps comparing sequence-level grounding alone with MuseVLP. Contours indicate report-linked atlas regions.