RETRIEVAL-AUGMENTED ANATOMY-AWARE REASONING FOR INTERPRETABLE ALZHEIMER'S DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate Alzheimer's disease (AD) diagnosis from structural MRI (sMRI) increasingly relies on deep learning, yet existing models often provide limited insight into the anatomical evidence underlying their predictions. Recent vision–language approaches improve interpretability by generating diagnostic rationales, but global image–text alignment does not guarantee that anatomy-related statements are grounded in the corresponding brain regions, leading to statement–region attention drift. We propose **RAR**, a retrieval-augmented anatomy-aware reasoning framework that explicitly connects anatomical evidence, diagnostic reasoning, and prediction. RAR comprises three components: an *Anatomy-Aware Visual Encoder* that guides visual attention toward AD-critical regions using subject-specific anatomical priors; a *Retrieval-Augmented Reasoning Module* that enriches the current sMRI representation with compact rationale context from relevant training cases; and an *Anatomy-Grounded Rationale Generator* that aligns anatomy-related rationale tokens with their supporting brain regions through token-specific anatomical priors. Experiments on ADNI, NACC, OASIS-3, and the external AIBL cohort demonstrate competitive diagnostic performance, improved rationale quality, and a 20.6-point gain in token–region grounding Dice. These results support explicitly linking regional imaging evidence to diagnostic reasoning for more interpretable AD diagnosis. Code will be made publicly available.

1 INTRODUCTION

Alzheimer's disease (AD) is a prevalent neurodegenerative disorder characterized by progressive cognitive decline and irreversible neuronal degeneration 2023 Alzheimer's Association Report (2023). In clinical practice, sMRI is widely used to assess neurodegeneration because it provides non-invasive and accessible measurements of structural alterations associated with AD, including hippocampal and entorhinal cortical atrophy Frisoni et al. (2010); Garg et al. (2022), motivating substantial progress in deep learning-based AD classification from sMRI. Existing models Qiu et al. (2022); Zhou et al. (2025); Kang et al. (2025) can distinguish cognitively normal controls (NC), mild cognitive impairment (MCI), and AD, but typically provide limited insight into the anatomical evidence underlying their predictions.

Medical vision-language models (VLMs) have increasingly been explored for generating diagnostic text from medical images Li et al. (2023); Moor et al. (2024), evolving from radiology report generation Chen et al. (2020); Park et al. (2025) toward explicit diagnostic rationales that organize observed evidence into a stepwise path toward a diagnostic conclusion Zhang et al. (2025). Related efforts in AD include clinician-oriented report generation Wang et al. (2026) and guideline-guided diagnostic rationale generation Chen et al. (2026).

However, generated anatomical statements are not grounded in their corresponding image regions. Existing approaches are mainly optimized for diagnostic accuracy, textual quality, or high-level image–text correspondence, with limited supervision linking individual anatomy-related statements to region-level sMRI evidence. Consequently, a model may describe "hippocampal atrophy" while attending primarily to the ventricles (Vent), reflecting statement–region attention drift (Figure 1).

To address this challenge, we propose **RAR** (**R**etrieval-Augmented **A**natomy-Aware **R**easoning), a reasoning-driven framework for interpretable AD diagnosis from sMRI. RAR establishes an

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.

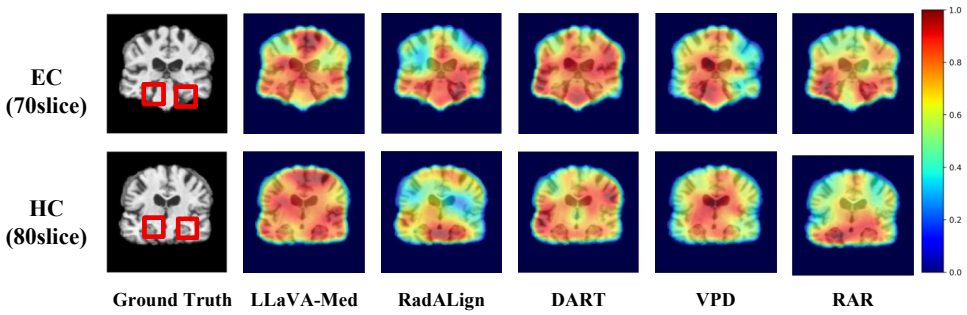

Figure 1: Cross-attention maps during the generation of entorhinal cortex (EC)- and hippocampus (HC)-related tokens in a representative AD case.

evidence-to-reasoning-to-diagnosis pathway through three components: an *Anatomy-Aware Visual Encoder* that learns anatomically focused visual representations, a *Retrieval-Augmented Reasoning Module* that retrieves case-relevant reasoning context, and an *Anatomy-Grounded Rationale Generator* that produces a stepwise diagnostic rationale aligned with the corresponding anatomical evidence. Our contributions are summarized as follows:

- We propose RAR to explicitly connect anatomical evidence, diagnostic reasoning, and prediction within a unified evidence-to-reasoning-to-diagnosis framework.
- We develop anatomical grounding in both visual representation learning and rationale generation, reducing statement–region attention drift and improving token–region grounding.
- We introduce retrieval-augmented reasoning to augment current MRI evidence with case-relevant rationale representations. Evaluation across four cohorts demonstrates competitive diagnostic performance with improved rationale quality and anatomical grounding.

2 Related Work

2.1 Visual Representation Learning for AD Diagnosis

Representation learning for AD diagnosis has advanced from task-driven supervision toward incorporating cross-modal semantic guidance and explicit anatomical priors. Task-specific models such as 3DNesT learn discriminative structural representations directly from 3D sMRI Kang & Liu (2023), but rely mainly on diagnostic supervision. Biomedical vision–language methods, including BiomedCLIP Zhang et al. (2024), BiomedCoOp Koleilat et al. (2025), and ClinicCLIP Ko & Park (2025), enrich visual representations with clinical semantics but offer limited explicit guidance on localizing disease-relevant evidence. Anatomy-aware methods such as SGCN Zhou et al. (2025), BAR Zhang et al. (2023), AAGN Jiang & Miao (2024), PPAL Diao et al. (2025), and E^{2}AD Wang et al. (2026) incorporate ROI structure, segmentation priors, anatomical concepts, or cross-modal anatomical dependencies to learn spatially meaningful representations. However, these methods provide limited modeling of how regional evidence is translated into explicit diagnostic reasoning.

2.2 Medical Vision–Language Generation and Diagnostic Reasoning

Beyond representation learning, medical vision–language models aim to translate visual evidence into clinically meaningful text. Early efforts focused on radiology report generation, such as R2Gen with memory-driven report generation Chen et al. (2020). Recent medical VLMs, including LLaVA-Med Li et al. (2023), extend this capability toward open-ended medical image understanding, while methods such as RadAlign and DART improve generation through concept-level alignment and disease-aware contextual enhancement, respectively Gu et al. (2025); Park et al. (2025). Beyond global image–text correspondence, AdaMatch-Cyclic explores fine-grained patch–word alignment for explainable medical image-report generation by explicitly associating visual regions with textual concepts Chen et al. (2024).

More recent approaches explore diagnostic reasoning by organizing visual evidence into explicit reasoning steps. VPD distills programmatic reasoning traces into VLMs Hu et al. (2024), and AD-oriented methods such as E^2AD and AD-Reasoning generate anatomically informed reports or guideline-guided diagnostic rationales Wang et al. (2026); Chen et al. (2026). However, existing methods mainly optimize diagnostic correctness or textual quality, without explicitly supervising whether anatomy-related statements are grounded in their corresponding brain regions.

Additionally, post-hoc explainability methods, including gradient-based attribution Selvaraju et al. (2017); Sundararajan et al. (2017) and perturbation-based evaluation Samek et al. (2017), identify image regions associated with model predictions. While useful for visualizing diagnostic evidence, such methods primarily explain classifier-level decisions and do not explicitly link individual generated statements to their supporting anatomical regions. RAR instead focuses on token-to-region grounding during rationale generation, explicitly linking generated reasoning steps with their supporting anatomical evidence.

3 METHOD

We propose RAR, a retrieval-augmented anatomy-aware reasoning framework for interpretable AD diagnosis from sMRI, as illustrated in Figure 2. RAR follows an evidence-to-reasoning-to-diagnosis pathway through three components: an *Anatomy-Aware Visual Encoder*, a *Retrieval-Augmented Reasoning Module*, and an *Anatomy-Grounded Rationale Generator*. To address the mismatch between generated anatomical statements and their supporting image regions, RAR introduces anatomical grounding at two levels: subject-specific anatomical priors guide visual representation learning, while token-level supervision aligns anatomy-related rationale tokens with their corresponding brain regions. provides relevant case-level reasoning context for the current sMRI.

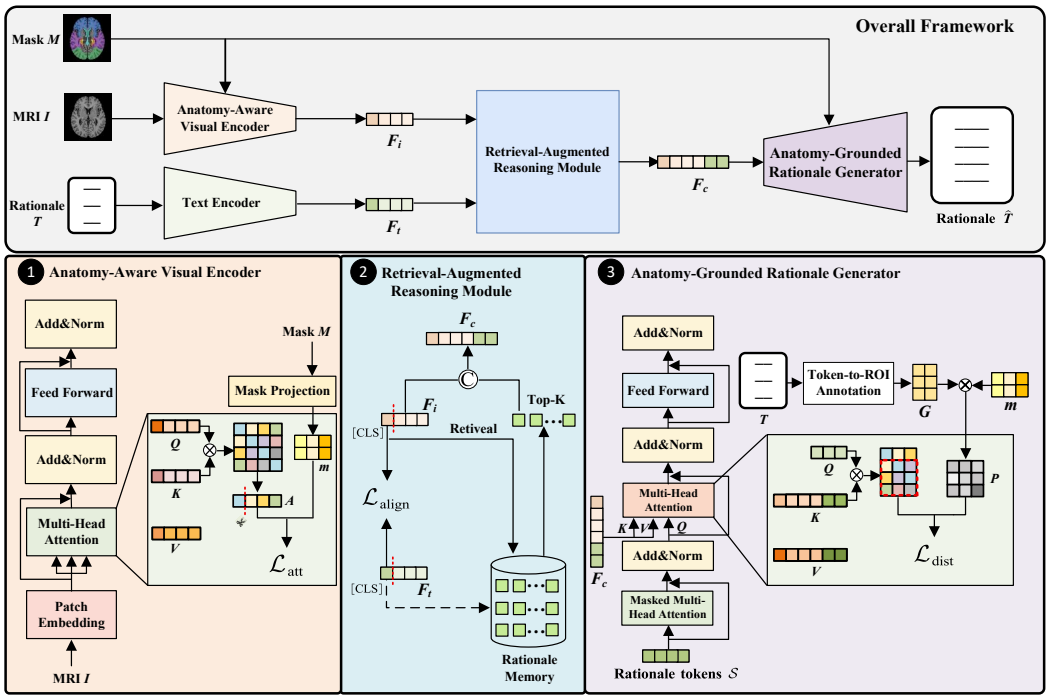

Figure 2: Overview of RAR. During training, an sMRI scan I, anatomical mask M, and diagnostic rationale T are processed by three components: (1) an Anatomy-Aware Visual Encoder for anatomy-guided representation learning, (2) a Retrieval-Augmented Reasoning Module for case-level rationale retrieval, and (3) an Anatomy-Grounded Rationale Generator for token-to-region grounding. At inference, only the sMRI scan and the fixed rationale memory are required.

3.1 Anatomical Mask and Diagnostic Rationale Construction

Anatomical Mask. MRI scans are preprocessed using the standardized T1-linear workflow in Clinica Routier et al. (2021); Wen et al. (2020), followed by FreeSurfer v7.4.1 Fischl (2012) with aseg and Desikan–Killiany–Tourville (DKT) parcellations Klein & Tourville (2012) to derive subject-specific anatomical masks. Preprocessing details are provided in the Supplementary Material. Candidate ROIs are selected based on established patterns of AD neuropathology Braak et al. (2006); Jack et al. (2010); Serrano-Pozo et al. (2026) and a targeted review of 20 neuroimaging studies. Six regions identified as AD-critical in more than half of the studies ($> 10/20$) are retained (Supplementary Fig. 1), with the frontal lobe additionally included given its involvement in later, more widespread neocortical AD pathology Serrano-Pozo et al. (2026). This yields seven AD-critical regions: entorhinal cortex (EC), hippocampus (HC), amygdala, temporal, parietal, and frontal lobes, and ventricles (Vent). HC and amygdala are derived from aseg, and EC and cortical lobes from DKT. The resulting mask M assigns each ROI a unique label, with all remaining voxels treated as non-target regions, providing subject-specific anatomical priors for representation learning and rationale grounding.

Diagnostic Rationale. For each of the seven AD-critical ROIs and hemispheres, we compute the atrophy rate $a = (\mu_{\mathrm{NC}} - v)/\mu_{\mathrm{NC}}$ and standardized volume deviation (Z-score) Torok et al. (2018) $z = (v - \mu_{\mathrm{NC}})/\sigma_{\mathrm{NC}}$, where v is the regional volume and $\mu_{\mathrm{NC}}, \sigma_{\mathrm{NC}}$ are estimated exclusively from training-set NC subjects. Positive a (negative z) indicates volume reduction, and the converse indicates relative enlargement, as may occur in the ventricles. These measurements form a structured anatomical evidence trace, which Qwen2.5-14B-Instruct Qwen Team (2024) verbalizes into a diagnostic rationale T. Diagnostic labels are used only to standardize target verbalization during rationale construction and are never provided to RAR at inference; thus, T serves as an anatomically structured diagnostic rationale. Reference thresholds, prompt templates, and an example are provided in Supplementary Sections B–C and Fig. 2.

3.2 Anatomy-Aware Visual Encoder

To obtain anatomically focused visual representations, we regularize CLS-to-patch attention using subject-specific anatomical priors during training. Each sMRI scan $I \in \mathbb{R}^{H\times W\times D}$ is partitioned into $N = 512$ non-overlapping 3D patches. After linear embedding, a learnable classification token ([CLS]) is prepended and the sequence is processed by a 3D ViT encoder, yielding $F_i \in \mathbb{R}^{(N+1)\times d}$ with global representation $f_i^{\mathrm{cls}} = F_i[0] \in \mathbb{R}^d$.

Anatomical Mask Projection. The subject-specific anatomical mask M is projected onto the same patch grid. For each ROI $r \in \{1, \ldots, 7\}$, we define $m_r \in \{0, 1\}^{1\times N}$, where $m_{r,j} = 1$ if patch j overlaps ROI r. Collectively, $m = [m_1; \ldots; m_7] \in \{0, 1\}^{7\times N}$. We further define the non-target indicator $\tilde{m} \in \{0, 1\}^N$ for patches outside all selected ROIs, where $\tilde{m}_j = 1 - \max_r m_{r,j}$.

Anatomy-Guided Attention Regularization. Diagnostic supervision specifies *what* to predict but not *where* supporting structural evidence should be localized. We therefore regularize the normalized CLS-to-patch attention $A \in \mathbb{R}^{1\times N}$ as

$$\mathcal{L}_{\text{att}} = \left(1 - \sum\nolimits_r w_r \frac{A m_r^\top}{\|m_r\|_1 + \epsilon}\right) + \tilde{\lambda} \frac{A \tilde{m}^\top}{\|\tilde{m}\|_1 + \epsilon}, \tag{1}$$

where $\|m_r\|_1 = \sum_j m_{r,j}$ and $w_r \geq 0$, $\sum_r w_r = 1$, are learnable normalized ROI weights. The first term promotes high attention density within AD-critical ROIs, with ROI-size normalization preventing large regions from dominating the objective and preserving the contribution of small but important structures such as EC and HC. The second term suppresses attention to non-target regions; we set $\tilde{\lambda} = 0.5$ to apply weaker suppression because these regions may still contain diagnostic information. This supervision is used only during training and requires no masks at inference.

3.3 RETRIEVAL-AUGMENTED REASONING MODULE

Inspired by case-based clinical reasoning, we augment the current sMRI representation with relevant case-level reasoning context. Specifically, we first align global sMRI and rationale representations in a shared metric space and then retrieve compact rationale representations from training cases.

Image–Rationale Alignment. Let $f_i^{\text{cls}} \in \mathbb{R}^d$ denote the global sMRI representation. The paired rationale T is encoded by PubMedBERT Gu et al. (2021), whose [CLS] embedding $f_t^{\text{cls}} \in \mathbb{R}^d$ serves as its global representation. For a mini-batch $\mathcal{B}$, we define $\mathbf{Sim}_{kj} = \text{sim}\big(f_{i_k}^{\text{cls}}, f_{t_j}^{\text{cls}}\big)/\tau$ and optimize

$$\mathcal{L}_{\text{align}} = \frac{1}{2}\left[\text{CE}(\mathbf{Sim}, \mathbf{y}) + \text{CE}(\mathbf{Sim}^\top, \mathbf{y})\right], \tag{2}$$

where k and j index samples in $\mathcal{B}$, $\mathbf{y}_k = k$ denotes the matched pair, $\text{sim}(\cdot,\cdot)$ is cosine similarity, and τ is a temperature parameter. This alignment makes cross-modal similarity meaningful for retrieving rationale representations relevant to the current sMRI.

Rationale Memory Retrieval. We maintain a memory of rationale [CLS] embeddings from training cases. Given f_i^{cls}, we retrieve the top-K most similar entries, excluding its paired rationale during training, with indices denoted by $\mathcal{N}_K$, and append them to the visual tokens:

$$F_c = \text{Concat}\left(F_i, \{f_{t_j}^{\text{cls}}\}_{j\in\mathcal{N}_K}\right). \tag{3}$$

Retrieving compact rationale [CLS] embeddings rather than full rationale token sequences provides case-level reasoning context while retaining the complete sMRI representation. At inference, the rationale memory is fixed and contains training cases only. We use $K = 4$ throughout.

3.4 ANATOMY-GROUNDED RATIONALE GENERATOR

Global image–text alignment does not guarantee that anatomy-related rationale tokens attend to their corresponding brain regions, leading to statement–region attention drift. To address this challenge, we first associate rationale tokens with AD-critical ROIs, then project these token–ROI associations onto the subject-specific sMRI patch grid to construct token-specific anatomical priors. These priors regularize the generator's visual cross-attention toward the corresponding brain regions.

Token-to-ROI Annotation. Each training rationale T is divided into J predefined segments $\{u_j\}_{j=1}^J$. Each segment is tokenized separately and concatenated into a fixed-length sequence $\mathcal{S} = \{s_t\}_{t=1}^L$ ($L = 256$), while preserving the segment membership. For each segment u_j, Qwen2.5 assigns ROI relevance scores to each segment based on its clinical semantics, yielding a normalized relevance vector $\mathbf{r}_j \in [0,1]^7$ with $\sum_{k=1}^7 r_{j,k} = 1$. Each s_t inherits its segment-level ROI relevance:

$$g_t = \mathbf{r}_j, \qquad \text{if } s_t \in u_j. \tag{4}$$

This yields the token–ROI correspondence $G = [g_1; \ldots; g_L] \in [0,1]^{L\times 7}$, with tokens in the same segment sharing the same ROI relevance distribution.

Anatomical Prior Projection. We project $G \in [0,1]^{L\times 7}$ onto the visual patch grid using the ROI–patch mask $m \in \{0,1\}^{7\times N}$. The resulting token–patch relevance matrix $\widetilde{P}$ is row-normalized to obtain token-specific anatomical priors:

$$\widetilde{P} = Gm, \qquad P_{t,n} = \frac{\widetilde{P}_{t,n} + \epsilon}{\sum_{n'=1}^N \widetilde{P}_{t,n'} + N\epsilon}. \tag{5}$$

Thus, $P_t \in [0,1]^N$ with $\sum_n P_{t,n} = 1$ defines the expected spatial distribution of visual evidence for token s_t.

Anatomy-Grounded Attention Regularization. Let $A_t^{\text{ca}} \in [0,1]^N$ denote the head-averaged cross-attention over visual patch tokens when predicting s_t, after excluding non-visual tokens in F_c and renormalizing over the remaining N patches. We minimize

$$\mathcal{L}_{\text{dist}} = \frac{1}{L}\sum_{t=1}^L \text{KL}\left(P_t \,\|\, A_t^{\text{ca}}\right). \tag{6}$$

Thus, $\mathcal{L}_{\text{dist}}$ explicitly constrains *where* each rationale token attends, rather than enforcing only global semantic consistency.

3.5 Training and Inference

Training. An auxiliary linear classifier on f_i^{cls} predicts NC, MCI, and AD with cross-entropy $\mathcal{L}_{\text{cls}}$. The rationale generator is trained with the autoregressive next-token prediction loss:

$$\mathcal{L}_{\text{gen}} = -\sum_{t=1}^{L} \log p(s_t \mid s_{<t}, F_c). \tag{7}$$

The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{gen}}\mathcal{L}_{\text{gen}} + \lambda_{\text{att}}\mathcal{L}_{\text{att}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}}, \tag{8}$$

Inference. Given an sMRI scan I, the Anatomy-Aware Visual Encoder produces visual features F_i with its global representation f_i^{cls}. The latter global representation retrieves the top-K rationale [CLS] embeddings from the fixed rationale memory, which is combined with F_i to form F_c. The Anatomy-Grounded Rationale Generator then produces $\hat{T}$, from which the final diagnostic label is derived. No anatomical masks or reference rationales are required.

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets. We evaluate RAR on four multi-center public sMRI datasets: ADNI (1232), AIBL (664), NACC (1490), and OASIS-3 (349) for NC/MCI/AD classification. Models are trained on pooled ADNI+NACC+OASIS-3 using subject-stratified 7:2:1 splits and externally evaluated on AIBL.

Metrics and Implementation. We report ACC, SEN, SPE, and AUC for diagnosis, with SEN, SPE, and AUC computed in a one-vs-rest manner and macro-averaged across the three classes, and BLEU-4, ROUGE-L, and METEOR for rationale generation. For generation-based models, ACC, SEN, and SPE are computed from the generated diagnostic label, while AUC is computed from the NC/MCI/AD probability distribution at the diagnostic decoding step. RAR is optimized with Adam (lr 1×10^{-4}) for 100 epochs on an L20 GPU.

Table 1: Classification performance on the internal and external AIBL test sets.

Method	Internal test				External test (AIBL)			
	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
3DNesT	<u>0.7654</u>	0.7641	0.8213	0.7745	0.7028	0.7159	0.7953	0.7401
BiomedCLIP	0.7029	0.7323	0.7964	0.7583	0.6519	0.6817	0.7431	0.7052
BiomedCoOp	0.7238	**0.7989**	0.8125	0.7538	0.6681	0.7371	0.7527	0.7146
ClinicCLIP	0.7136	<u>0.7891</u>	0.8212	0.7623	0.6597	0.7315	0.7588	0.7108
SGCN	0.7489	0.7765	0.8421	0.7644	0.6967	0.7097	0.7918	0.7339
BAR	0.6774	0.6665	0.8289	<u>0.8258</u>	**0.7147**	0.6279	<u>0.8219</u>	<u>0.8149</u>
AAGN	0.7523	0.7544	<u>0.8552</u>	0.7651	0.6841	0.7015	0.7862	0.7213
PPAL	0.6968	0.6856	0.8374	**0.8654**	0.6935	<u>0.7373</u>	**0.8576**	**0.8891**
AD-Reasoning	0.7430	0.6488	0.8509	0.8082	0.2671	0.6558	0.7102	0.7512
RAR	**0.7743**	0.7615	**0.8726**	0.7841	<u>0.7143</u>	**0.7480**	0.8029	0.7615

4.2 Classification Performance Comparison

We compare RAR with discriminative, vision–language, and anatomy-aware baselines under identical data splits, preprocessing, and NC/MCI/AD protocol; originally multimodal methods are evaluated using sMRI only. As shown in Table 1, RAR achieves the best internal ACC and macro SPE, improving ACC over 3DNesT, AAGN, and BiomedCoOp by 0.89, 2.20, and 5.05 percentage points, respectively. On external AIBL, RAR maintains competitive performance, improving ACC over

3DNesT and BiomedCoOp by 1.15 and 4.62 points and achieving the highest SEN. Notably, PPAL achieves higher AUC on both internal and external tests, indicating stronger threshold-independent ranking performance, whereas RAR yields higher accuracy under the reported decision rule. Overall, RAR preserves competitive diagnostic performance while additionally providing anatomically grounded rationale generation to improve reasoning transparency and evidence alignment.

Table 2: Rationale generation quality on the same AD rationale corpus.

Method	ACC	SEN	SPE	AUC	BLEU-4	ROUGE-L	METEOR
LLaVA-Med	0.2345	0.3333	0.6667	0.5000	0.3239	0.5501	0.3223
RadAlign	0.5442	0.5261	0.7617	0.6439	0.3128	0.5487	0.3189
DART	0.5232	0.5418	0.6808	0.5908	0.3222	0.5581	0.3283
AdaMatch-Cyclic	0.5290	0.4993	0.7473	0.4698	<u>0.5265</u>	<u>0.6799</u>	**0.5913**
VPD	0.5531	0.5138	0.7521	0.6329	0.3124	0.5417	0.3152
AD-Reasoning	<u>0.7430</u>	<u>0.6488</u>	<u>0.8509</u>	**0.8082**	0.3686	0.4848	0.5474
RAR	**0.7743**	**0.7615**	**0.8726**	<u>0.7841</u>	**0.5814**	**0.7049**	<u>0.4018</u>

4.3 Rationale Generation Evaluation

We compare RAR with representative rationale-generation baselines on the same AD rationale corpus (Supplementary Section E). As shown in Table 2, RAR achieves the best ACC, SEN, and SPE, together with the highest BLEU-4 and ROUGE-L. AdaMatch achieves the highest METEOR, whereas AD-Reasoning attains the highest AUC. These results indicate that RAR better balances diagnostic correctness with phrase- and sequence-level reference consistency. Since textual similarity alone does not establish anatomical grounding, token-to-region correspondence is evaluated separately below.

Table 3: Ablation study of the key components in RAR.

AA	RS	RA	AG	ACC	SEN	SPE	AUC	BLEU-4	ROUGE-L	METEOR
×	×	×	×	0.6672	0.6653	0.7416	0.7046	N/A	N/A	N/A
✓	×	×	×	0.7349	0.7418	0.8584	0.7645	N/A	N/A	N/A
✓	✓	×	×	0.7263	0.7235	0.8326	0.7561	0.4213	0.4722	0.3031
✓	✓	✓	×	0.7458	0.7393	0.8573	0.7753	0.4712	0.5821	0.3314
✓	✓	×	✓	0.7244	0.7211	0.8282	0.7574	0.4261	0.4635	0.3112
✓	✓	✓	✓	**0.7743**	**0.7615**	**0.8726**	**0.7841**	**0.5814**	**0.7049**	**0.4018**

Note: AA: anatomy-aware visual encoding; RS: rationale supervision; RA: retrieval augmentation; AG: anatomical grounding via attention regularization; N/A: no generation branch.

4.4 Ablation Study

Table 3 evaluates anatomy-aware visual encoding (AA), rationale supervision (RS), retrieval augmentation (RA), and anatomical grounding (AG). AA consistently improves all diagnostic metrics over the backbone (ACC/SEN/SPE/AUC: +6.77/+7.65/+11.68/+5.99 percentage points), confirming the benefit of anatomy-aware representation learning. Adding RS enables rationale generation but slightly decreases diagnostic performance, suggesting a modest trade-off between discriminative prediction and rationale supervision, as the model shifts from optimizing only diagnostic discrimination to jointly learning diagnosis and rationale generation. RA subsequently improves both diagnosis and generation, increasing ACC/SEN/SPE/AUC by 1.95/1.58/2.47/1.92 percentage points and BLEU-4/ROUGE-L/METEOR by 4.99/10.99/2.83 percentage points over AA+RS. AG alone yields only minor and inconsistent changes in standard diagnostic and generation metrics, consistent with its primary role in anatomical grounding rather than direct optimization of these metrics. When combined with RA, however, AG further improves ACC/SEN/SPE/AUC by 2.85/2.22/1.53/0.88 percentage points and BLEU-4/ROUGE-L/METEOR by 11.02/12.28/7.04 percentage points, achieving

the best overall diagnosis and rationale generation performance. These results suggest complementary roles: RA provides case-level reasoning context, while AG grounds the generated rationale in anatomical evidence from the current sMRI.

4.5 Anatomy Grounding Evaluation

Visual–Retrieval Attention Allocation. Following Sinha et al. Sinha et al. (2026), we first quantify how decoder cross-attention is distributed between current-image visual tokens and retrieved rationale embeddings. Let R_{vis} and R_{CoT} denote the fractions of attention assigned to the 512 visual tokens and four retrieved rationale [CLS] embeddings, respectively, with $R_{\text{vis}} + R_{\text{CoT}} = 1$. Because the two groups differ substantially in size, uniform attention would yield $R_{\text{vis}} = 0.9922$ and $R_{\text{CoT}} = 0.0078$. RAR w/o AG yields $R_{\text{vis}} = 0.9916$ and $R_{\text{CoT}} = 0.0085$, whereas full RAR yields 0.9946 and 0.0054, respectively. Thus, both variants rely predominantly on MRI-derived visual tokens, while AG further shifts attention away from retrieved rationale embeddings and toward current-image representations. This is consistent with the intended role of retrieval as compact case-level reasoning context rather than a substitute for the current sMRI.

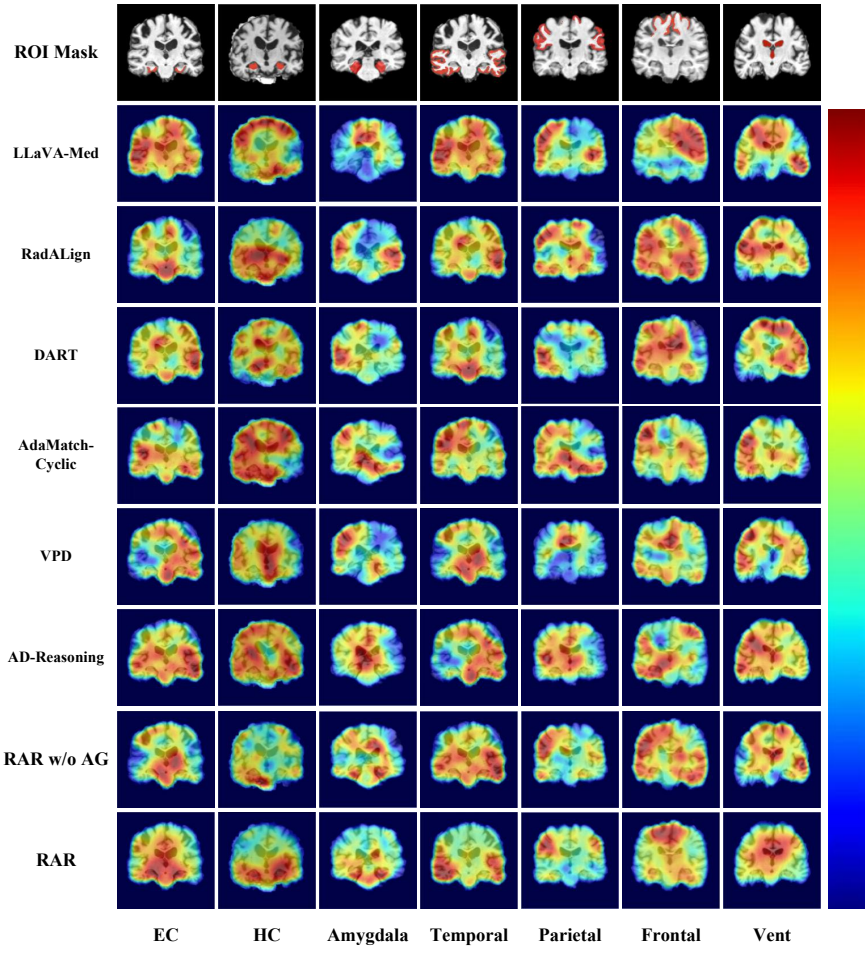

Figure 3: Token-level visual cross-attention for seven AD-critical ROIs in a representative AD case.

Token-level ROI Alignment. We next examine whether anatomy-related rationale tokens attend to their corresponding brain regions, directly addressing statement–region attention drift. Figure 3 visualizes generator cross-attention for seven ROI-related rationale steps in a representative AD case. Existing generation baselines generally exhibit diffuse or spatially inconsistent attention. RAR w/o AG already captures substantial responses within the corresponding ROIs; however, its

attention remains less spatially concentrated and complete. In contrast, full RAR shows more consistent and comprehensive correspondence between anatomy-related tokens and their target ROIs across the seven regions. Notably, no anatomical masks are provided at inference, indicating that the region-specific attention patterns learned during training transfer to mask-free rationale generation.

Table 4: Dice between ROI-token attention maps and FreeSurfer anatomical masks on the test set. Results are reported as mean$_{\pm\mathrm{std}}$. p values are computed using two-sided paired t-tests on per-subject Dice scores between RAR and RAR w/o AG.

ROI	LLaVA-Med	RadAlign	DART	AdaMatch	VPD	AD-Reasoning	RAR w/o AG	RAR	p
EC	$0.324_{\pm 0.091}$	$0.386_{\pm 0.083}$	$0.357_{\pm 0.087}$	$0.341_{\pm 0.089}$	$0.372_{\pm 0.085}$	$0.423_{\pm 0.080}$	$0.523_{\pm 0.124}$	$\mathbf{0.718_{\pm 0.092}}$	6.57×10^{-4}
HC	$0.347_{\pm 0.086}$	$0.412_{\pm 0.081}$	$0.379_{\pm 0.084}$	$0.363_{\pm 0.085}$	$0.395_{\pm 0.082}$	$0.441_{\pm 0.078}$	$0.487_{\pm 0.151}$	$\mathbf{0.734_{\pm 0.087}}$	7.85×10^{-4}
Amygdala	$0.302_{\pm 0.103}$	$0.354_{\pm 0.092}$	$0.328_{\pm 0.098}$	$0.314_{\pm 0.100}$	$0.341_{\pm 0.095}$	$0.385_{\pm 0.089}$	$0.442_{\pm 0.148}$	$\mathbf{0.681_{\pm 0.103}}$	8.12×10^{-4}
Temporal	$0.315_{\pm 0.097}$	$0.367_{\pm 0.088}$	$0.342_{\pm 0.093}$	$0.329_{\pm 0.095}$	$0.356_{\pm 0.090}$	$0.398_{\pm 0.084}$	$0.516_{\pm 0.132}$	$\mathbf{0.695_{\pm 0.098}}$	7.54×10^{-4}
Parietal	$0.301_{\pm 0.112}$	$0.335_{\pm 0.101}$	$0.316_{\pm 0.106}$	$0.308_{\pm 0.109}$	$0.426_{\pm 0.113}$	$0.362_{\pm 0.097}$	$0.498_{\pm 0.145}$	$\mathbf{0.667_{\pm 0.112}}$	9.75×10^{-4}
Frontal	$0.336_{\pm 0.098}$	$0.403_{\pm 0.103}$	$0.353_{\pm 0.112}$	$0.358_{\pm 0.079}$	$0.324_{\pm 0.104}$	$0.387_{\pm 0.112}$	$0.531_{\pm 0.129}$	$\mathbf{0.702_{\pm 0.101}}$	6.25×10^{-4}
Vent	$0.357_{\pm 0.088}$	$0.416_{\pm 0.089}$	$0.371_{\pm 0.089}$	$0.345_{\pm 0.089}$	$0.394_{\pm 0.088}$	$0.401_{\pm 0.088}$	$0.413_{\pm 0.162}$	$\mathbf{0.654_{\pm 0.126}}$	8.43×10^{-4}
Avg.	0.326	0.382	0.349	0.337	0.373	0.400	0.487	**0.693**	$< \mathbf{0.001}$

To quantify token–region grounding over the full test set, we compute Dice between token-specific cross-attention maps and the corresponding FreeSurfer ROI masks. As shown in Table 4, existing rationale-generation methods achieve relatively limited anatomical alignment, with average Dice ranging from 0.326 to 0.400. RAR w/o AG already improves the average Dice to 0.487, suggesting that the proposed visual–reasoning architecture provides meaningful anatomical cues even without explicit grounding supervision. Adding AG further raises the average Dice to 0.693, an improvement of 20.6 percentage points over RAR w/o AG and 29.3 points over the strongest competing baseline (AD-Reasoning, 0.400). The improvement is consistent across all seven ROIs, with gains of 16.9–24.7 percentage points over RAR w/o AG and significant improvements for every ROI ($p < 0.001$). These results directly support the motivation of AG: while retrieval and rationale supervision can provide anatomically relevant context, they do not by themselves ensure precise statement–region correspondence. AG explicitly addresses this gap by aligning anatomy-related rationale tokens with their supporting regions in the current sMRI, thereby substantially reducing statement–region attention drift.

5 Conclusion

We presented RAR, a retrieval-augmented anatomy-aware reasoning framework for interpretable AD diagnosis from sMRI. RAR connects anatomical evidence, diagnostic reasoning, and prediction through an evidence-to-reasoning-to-diagnosis pathway. Its Anatomy-Aware Visual Encoder uses subject-specific anatomical priors to guide visual attention toward AD-critical regions; the Retrieval-Augmented Reasoning Module supplements the current sMRI representation with compact rationale context from relevant training cases; and the Anatomy-Grounded Rationale Generator introduces token-level anatomical priors to align anatomy-related rationale tokens with their supporting brain regions during generation. Experiments across ADNI, NACC, OASIS-3, and the external AIBL cohort show that RAR maintains strong diagnostic performance while improving rationale generation quality. Ablation and grounding analyses further demonstrate the complementary benefits of retrieval and anatomical grounding, with substantially improved token–region alignment. Overall, our results suggest that explicitly linking regional imaging evidence to diagnostic reasoning provides a promising direction for more interpretable and clinically meaningful AD diagnosis.

References

2023 Alzheimer's Association Report. 2023 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 19(4):1598–1695, 2023.

Heiko Braak, Irina Alafuzoff, Thomas Arzberger, Hans Kretzschmar, and Kelly Del Tredici. Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. *Acta Neuropathologica*, 112(4):389–404, 2006.

Qiuhui Chen, Yushan Deng, Xuancheng Yao, and Yi Hong. AD-Reasoning: Multimodal guideline-guided reasoning for Alzheimer's disease diagnosis, 2026. Preprint; under review.

Wenting Chen, Linlin Shen, Jingyang Lin, Jiebo Luo, Xiang Li, and Yixuan Yuan. Fine-grained image-text alignment in medical imaging enables explainable cyclic image-report generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9494–9509. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.514.

Z. Chen, Y. Song, T. Chang, and X. Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1439–1449, 2020.

Yueqin Diao, Huihui Fang, Hanyi Yu, Yuning Wang, Yaling Tao, Ziyan Huang, Si Yong Yeo, and Yanwu Xu. Prior-guided prototype aggregation learning for Alzheimer's disease diagnosis. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 15974, pp. 483–492, 2025.

Bruce Fischl. FreeSurfer. *NeuroImage*, 62(2):774–781, 2012.

G. B. Frisoni, N. C. Fox, C. R. Jack, P. Scheltens, and P. M. Thompson. The clinical use of structural MRI in alzheimer disease. *Nature Reviews Neurology*, 6(2):67–77, 2010.

Neha Garg, Mahipal Singh Choudhry, and Rajesh M. Bodade. A review on Alzheimer's disease classification from normal controls and mild cognitive impairment using structural MR images. *Journal of Neuroscience Methods*, 384:109745, 2022.

D. Gu, Y. Gao, Y. Zhou, M. Zhou, and D. Metaxas. RadAlign: Advancing radiology report generation with vision-language concept alignment. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 15966, pp. 484–494, 2025.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2021.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9590–9601, 2024.

Clifford R. Jack, David S. Knopman, William J. Jagust, Leslie M. Shaw, Paul S. Aisen, Michael W. Weiner, Ronald C. Petersen, and John Q. Trojanowski. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119–128, 2010.

Hongchao Jiang and Chunyan Miao. Anatomy-aware gating network for explainable Alzheimer's disease diagnosis. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 15005, pp. 90–100, 2024.

X. Kang, J. Lin, K. Zhao, et al. Structural MRI-based computer-aided diagnosis models for alzheimer disease: Insights into misclassifications and diagnostic limitations. *Radiology: Artificial Intelligence*, 7(3):e240123, 2025.

Xiaopeng Kang and Yong Liu. 3dnest: A hierarchical local self-attention model for alzheimer's disease diagnosis. In *2023 International Conference on Neuromorphic Computing (ICNC)*, pp. 1–6. IEEE, 2023. doi: 10.1109/ICNC59488.2023.10462762.

Arno Klein and Jason Tourville. 101 labeled brain images and a consistent human cortical labeling protocol. *Frontiers in Neuroscience*, 6:171, 2012.

Hanbin Ko and Chang-Min Park. Bringing CLIP to the clinic: Dynamic soft labels and negation-aware learning for medical analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25897–25906, 2025.

Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. BiomedCoOp: Learning to prompt for biomedical vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14766–14776, 2025.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

M. Moor, Q. Huang, S. Wu, et al. Med-flamingo: a multimodal medical few-shot learner. In *Proceedings of Machine Learning Research (PMLR)*, volume 225, pp. 353–367, 2024.

Sang-Jun Park, Keun-Soo Heo, Dong-Hee Shin, Young-Han Son, Ji-Hye Oh, and Tae-Eui Kam. DART: Disease-aware image-text alignment and self-correcting re-alignment for trustworthy radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15580–15589, 2025.

S. Qiu, P. S. Joshi, M. I. Miller, et al. Multimodal deep learning for Alzheimer's disease dementia assessment. *Nature Communications*, 13:3404, 2022.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Alexandre Routier, Ninon Burgos, Mauricio Díaz, Michael Bacci, Simona Bottani, Omar El-Rifai, Sabrina Fontanella, Pietro Gori, Jérémy Guillon, Alexis Guyot, Ravi Hassanaly, Thomas Jacquemont, Pascal Lu, Arnaud Marcoux, Tristan Moreau, Jorge Samper-González, Marc Teichmann, Elina Thibeau-Sutre, Ghislain Vaillant, Junhao Wen, Adam Wild, Marie-Odile Habert, Stanley Durrleman, and Olivier Colliot. Clinica: An open-source software platform for reproducible clinical neuroscience studies. *Frontiers in Neuroinformatics*, 15:689675, 2021.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 2017. doi: 10.1109/TNNLS.2016.2599820.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Alberto Serrano-Pozo, Valentina Escott-Price, Lea T. Grinberg, Tharick Pascoal, Marc Suarez-Calvet, Bruno Dubois, and Reisa A. Sperling. Alzheimer's disease. *The Lancet*, 407:2241–2262, 2026.

Sanchit Sinha, Guangzhi Xiong, Bohan Liu, Zhenghao He, and Aidong Zhang. Attention-guided fine-tuning of multimodal large language models improves chain-of-thought reasoning. *arXiv preprint arXiv:2606.01558*, 2026.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pp. 3319–3328. PMLR, 2017.

Justin Torok, Pedro D. Maia, Fon Powell, Sneha Pandya, and Ashish Raj. A method for inferring regional origins of neurodegeneration. *Brain*, 141(3):863–876, 2018. doi: 10.1093/brain/awx371.

Chenhui Wang, Sirong Piao, Zhihao Chen, Tao Chen, Zhaoyang Li, Tongrui Zhang, Yuxin Li, Xing-Ming Zhao, Hongming Shan, Alzheimer's Disease Neuroimaging Initiative, and Australian Imaging Biomarkers and Lifestyle flagship study of aging. E^2AD: Enhanced and explainable Alzheimer's disease detection framework via anatomy- and relation-aware cross-modal knowledge distillation. *Medical Image Analysis*, 112:104099, 2026.

Junhao Wen, Elina Thibeau-Sutre, Mauricio Diaz-Melo, Jorge Samper-González, Alexandre Routier, Simona Bottani, Didier Dormont, Stanley Durrleman, Ninon Burgos, and Olivier Colliot. Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Medical Image Analysis*, 63:101694, 2020.

Lintao Zhang, Jinjian Wu, Lihong Wang, Li Wang, David C. Steffens, Shijun Qiu, Guy G. Potter, and Mingxia Liu. Brain anatomy-guided MRI analysis for assessing clinical progression of cognitive impairment with structural MRI. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 14227, pp. 109–119, 2023.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1631–1662, 2025.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1):AIoa2400640, 2024.

Houliang Zhou, Lifang He, Brian Y. Chen, Li Shen, and Yu Zhang. Multi-modal diagnosis of Alzheimer's disease using interpretable graph convolutional networks. *IEEE Transactions on Medical Imaging*, 44(1):142–153, 2025.