CausalRAG-AD: Multimodal MRI Classification and Guideline-Compliant MRI Captioning for Alzheimer's Diagnosis

Ramisa Farha
Department of Computer Science
Morgan State University
Baltimore, Maryland, USA
rafar2@morgan.edu

Md Mahmudur Rahman
Department of Computer Science
Morgan State University
Baltimore, Maryland, USA
md.rahman@morgan.edu

Fahmi Khalifa
Dept. Electrical & Computer Engineering
Morgan State University
Baltimore, Maryland, USA
fahmi.khalifa@morgan.edu

***Abstract*—Alzheimer's disease (AD) is a degenerative disease, and current AI-based diagnostic procedures often lack calibrated probabilities, clear causal explanations, and comprehensive reporting. We propose CausalRAG-AD, an end-to-end framework that (i) trains calibrated transformer classifiers for AD staging from structural MRI and (ii) generates *NIA-AA* guideline–compliant, evidence-grounded MRI captions. The proposed architecture augments Video Vision Transformer (ViViT) with ROI-gated attention (*CausalRAG-ViViT*) from FreeSurfer hippocampus/ventricle/temporal z-scores and a light clinical gate; per-fold temperature scaling yields calibrated probabilities. It is evaluated against *ViT+BiLSTM* and *ViViT-lite* baselines using Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. For binary AD vs. Cognitively Normal (CN), *CausalRAG-ViViT* achieves the strongest calibration (ECE 0.078). In triclass CN / mild cognitive impairment (MCI) / AD, *CausalRAG-ViViT* reaches the best Acc (0.519) with competitive F1 (0.452). The proposed system produced 4,427 structured per-visit MRI reports. Every report contained all required sections, and most explicitly documented MRI provenance (scanner/vendor and field strength; 93.6%). All pipeline-generated numbers—MRI z-scores, MMSE, and CDR-SB—were replicated exactly from source tables (100% fidelity), and Aβ42 values matched the source (93.9%). When the necessary biomarkers were available, the *AT(N)* labels in the report consistently matched with our rule-based determination (100%).**

***Index Terms*—Alzheimer's disease, ADNI, clinical captioning, medical image classification, explainable artificial intelligence, clinical report generation, retrieval-augmented generation**

I. Introduction

Alzheimer's disease (AD) affects $\approx$ 55 million individuals globally, with ventricular enlargement and hippocampal atrophy serving as key neurodegenerative biomarkers [1]. Research on AD relies heavily on magnetic resonance imaging (MRI)–based staging; yet, many machine learning–based algorithms provide limited clinical traceability and generate *uncalibrated* probability projections. The reliability and clinical significance of textual outputs are undermined by the frequent absence of explicit connections to quantitative biomarkers.

Various studies have been proposed in literature for AD classification using Alzheimer's Disease Neuroimaging Initiative (ADNI) data [2] and MRI-based transformers. Robust ADNI baselines include slice-sequence hybrids and video-style transformers. Leveraging a Vision-Transformer (ViT) [3] to extract per-slice features and a bidirectional long short-term memory (BiLSTM) to model inter-slice correlations, ViT+BiLSTM achieves significant AD vs. CN classification on ADNI [4]. Recent work (*ViTranZheimer*) finds excellent multiclass performance on AD/MCI/CN [5]. Video-ViT/ViViT versions treat 3D MRI as a spatiotemporal "video" [6], acquiring tubelet embeddings and sharing attention across slices. Moreover, solid transformer baselines have been established for AD classification in previous research [4], [5], and clinical reporting is effectively grounded [7]–[10]. To facilitate retrieval, captioning, and VQA of brain MRI, vision-language models match 3D encoders with clinical text [11].

In this study, we propose **CausalRAG-AD**, an MRI-based end-to-end framework, that (1) trains calibrated transformer classifiers for AD staging and (2) generates evidence-grounded MRI captions. Unlike free-text summaries, our captioning focuses on National Institute on Aging–Alzheimer's Association (NIA-AA) [12] Amyloid/Tau/Neurodegeneration (AT(N)) with distinct evidence lines and numerical validation. The NIA-AA AT(N) contain quantitative evidence, causal reasoning, and attainable counterfactuals. A key advantage of the proposed approach is the ability to transform longitudinal knowledge and assess grounded reporting and discrimination and reliability. The retrieval-augmented generation (RAG) indicates that the developed system first retrieves structured; per-visit evidence, ROI z-scores (hippocampus, ventricles, temporal cortex); clinical scores (MMSE, CDR-SB); available Cerebrospinal fluid (CSF) values with assay/date provenance, and cohort cut-points before generating the NIA-AA AT(N) report. The "causal" denotes explicit, checkable rationales (which domain crossed which threshold and by how much) and minimal counterfactuals for N, rather than causal discovery.

For evaluation, we used a unified ADNI dataset that includes 5,557 T1-weighted Magnetization-Prepared Rapid Gradient Echo (MP-RAGE) series from $\sim$1,230 participants spanning ADNI 1/GO/2/3. We retrain our model CausalRAG-ViViT and other baselines (*ViT + BiLSTM*, *Video Vision Transformer*

(ViViT-lite)) on the same cohort using frozen, subject-wise 5-fold splits (by subject, not by visit) to guarantee equitable, realistic comparisons. The core model, CausalRAG-ViViT, explicitly emphasises faithfulness, calibration, and interpretability while trading off headline accuracy. It enhances ViViT with ROI–gated attention provided by FreeSurfer-derived hippocampal, ventricular, and temporal-lobe z-scores and a lightweight clinical gate.

II. Dataset and Preprocessing

We retrieve 5,557 T1-weighted MP–RAGE series from ADNI 1/GO/2/3 spanning ∼1,230 subjects (mixed 1.5T/3T; GE / Siemens / Philips). Before conversion, each visit's raw DICOMs are collected from the archive, de-duplicated, validated, and converted to NIfTI format using `dcm2niix` [13]. This step preserves vendor intensity scaling (Philips rescale/scale-slope fields) and BIDS-style sidecars. We then reorient volumes to Right–Anterior–Superior (RAS).

In brief: N4 bias correction → skull-strip → (optional) MNI affine → 1 mm resample → masked clip → masked z-score → crop/pad to 128^3.

We apply this standardized pipeline uniformly across visits and log QC flags per visit. For each visit, we store the preprocessed volume, brain mask, and QC flags such as skull-strip method and whether MNI affine was applied. QC *(quality control)* requires a nonempty mask (>500 voxels), finite intensities, and a shape 128^3.

The remaining visits are matched by Roster ID (RID) within ± 90 days after merging preprocessed visits with ADNI clinical tables using `EXAMDATE`. For **866** subjects, this yields **3,754** labeled visits (CN=**1,064**, MCI=**1,932**, AD=**758**); **146** preprocessed visits remain unlabeled after fuzzy matching (total preprocessed ≈**3,900**).

To quantify ROIs for gating and reporting, we use FreeSurfer-derived outputs when available (UCSF FreeSurfer v7) [16] to compute intracranial-volume (ICV)-adjusted z-scores for bilateral hippocampal and ventricular volumes and temporal cortical thickness. If ROI z-scores are missing for a visit, the model falls back to imaging-only features. These z-scores drive ROI-gated attention and populate the evidence lines in the *AT(N)* (Amyloid/Tau/Neurodegeneration) captions.

III. Methods

The proposed end-to-end CausalRAG-AD pipeline is schematized in Fig. 1. We begin by outlining the two coupled components of the system and how they are evaluated. The two components of the CausalRAG-AD framework are (1) guideline-aligned reporting that converts model/biomarker evidence into NIA-AA AT(N) captions with numerical validation, and (2) calibrated MRI classifiers for AD staging. Our reporting module follows RAG pattern: for each visit we first retrieve structured evidence (ROI z-scores, MMSE/CDR-SB, available CSF values with assay/date provenance, and cohort cut-points) and then render an NIA-AA AT(N) aligned narrative directly from these fields. The same unified ADNI 1/GO/2/3 cohort and frozen subject-wise 5-fold splits are used for training and evaluation of all models.

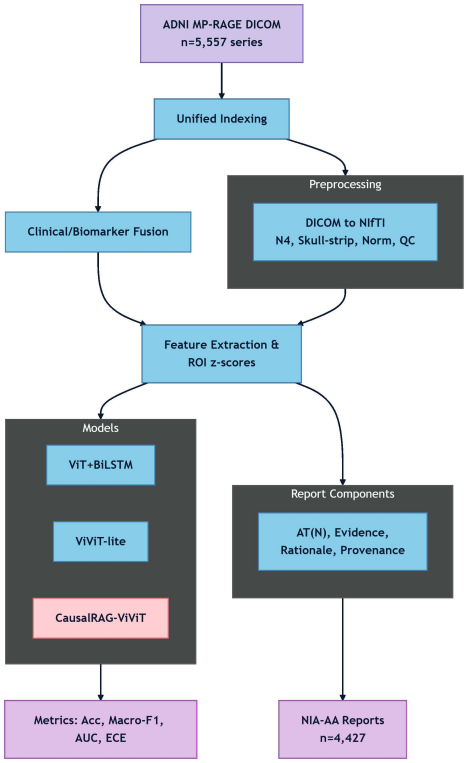


Fig. 1. CausalRAG-AD workflow: data → preprocess → join → ROI z-scores → calibrated 5-fold training → evaluation → *AT(N)* reports *(retrieve structured evidence → render report; RAG)*.

Our core classifier is CausalRAG-ViViT, a ViViT-style backbone augmented with a light clinical gate (age/sex) and FreeSurfer ROI z-scores (hippocampus L/R, ventricles, temporal cortex) that guide attention by adding small, learnable gates to the attention logits. The gates act additively on the logits, preserving end-to-end learning while nudging focus toward regions linked to neurodegeneration. Qualitatively, ROI gating shifts attention toward hippocampal/ventricular/temporal regions compared to an ungated ViViT baseline.

We evaluate two classification tasks: (i) binary AD vs. CN and (ii) three-way CN/MCI/AD using the T1w MP-RAGE series data outlined in Section II. We use the standardized preprocessing, ROI z-scores, and clinical/CSF joins described in Section II.

We consider two baselines for comparison. **ViT+BiLSTM** uses a 2D ViT to encode slices and a bidirectional LSTM to model slice order, producing a volume-level prediction. **ViViT-lite** treats a volume as a short "video" with tubelet embeddings and multi-head self-attention (MHSA) across space–slice. Our proposed **CausalRAG-ViViT** retains the ViViT backbone but augments attention logits with two small gates:

$$\underbrace{\mathrm{AttnLogits}}_{\text{ViViT}} + \underbrace{g_{\mathrm{ROI}}(z_{\mathrm{hip}}, z_{\mathrm{vent}}, z_{\mathrm{temp}})}_{\text{ROI-gated}} + \underbrace{g_{\mathrm{clin}}(\mathrm{age},\ \mathrm{sex})}_{\text{clinical gate}}.$$

No ROI is hard-masked; gates are linear projections with trained scaling. Prioritizing faithfulness, calibration, and in-

terpretability may trade a small amount of headline accuracy.

All models are trained from scratch on the same cohort using frozen subject-wise 5-fold cross-validation. We report pooled out-of-fold Accuracy, Macro-F1, AUC, and Expected Calibration Error (ECE). We use AdamW with linear warmup and cosine decay, light 3D augmentation, class-balanced batches, gradient clipping, exponential moving average (EMA), mixed precision, and early stopping on Macro-F1. For calibration, we fit temperature scaling per fold on that fold's validation split and evaluate reliability curves and ECE before and after calibration [17]. This yields probability estimates suitable for thresholding and abstention policies.

For each eligible visit, the system retrieves ROI z-scores ($z_{\text{hip}}^{L/R}$, $\bar{z}_{\text{hip}}$, z_{temp}, z_{vent}), diagnosis, MMSE, CDR-SB, and (when available) CSF values (Aβ42 or Aβ42/40, p-tau) together with assay/date provenance and the cohort cut-points used for thresholding. The renderer then generates a structured AT(N) report directly from these retrieved fields adding an explicit Evidence line (value vs. cut-point), a brief rationale (which domain crossed and by how much), and a minimal counterfactual to flip N and exports a compact PNG + JSON artifact.

AT(N) rules (cut-points): Neurodegeneration:

$$N^{+} \iff \bar{z}_{\text{hip}} \le -1 \ \vee \ z_{\text{temp}} \le -1 \ \vee \ z_{\text{vent}} \ge +1. \quad (1)$$

CSF thresholds (cohort μ, σ):

$$\begin{aligned} A^{+} &\iff \frac{\text{A}\beta42}{\text{A}\beta40} \le \mu - \sigma \ \vee \ \text{A}\beta42 \le \mu - \sigma, \\ T^{+} &\iff \text{p-tau} \ge \mu + \sigma. \end{aligned} \quad (2)$$

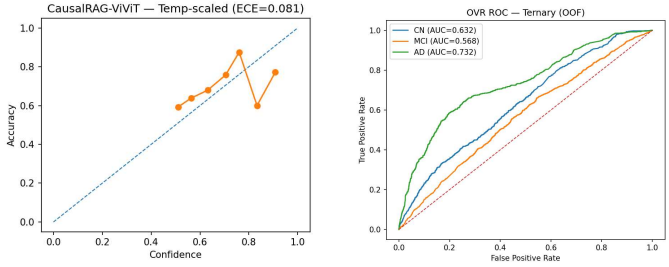
Missing biomarkers are marked as A?, T?. The text includes an *Evidence* line listing each value vs. its cut-point, a short *rationale* (which domain crossed and by how much), and a minimal *counterfactual* change to flip N.

Together with model predictions and calibration summaries, we construct **4,427** organized, NIA-AA aligned reports (one per visit when data permit).

IV. Experimental Results

All models are trained from scratch on a unified, preprocessed *ADNI-1/GO/2/3 cohort* using frozen, subject-wise 5-fold cross-validation. All visits from a given subject are confined to a single fold, ensuring no patient overlap across training, validation, and test sets. We evaluate both discrimination and probability quality using accuracy, macro-averaged F1 (*macro-F1*), area under the ROC curve (*AUC*), and expected calibration error (*ECE*). Per-fold temperature scaling is applied prior to computing ECE, and pooled out-of-fold metrics are reported. In addition, we assess guideline-compliant caption coverage and numeric fidelity for key biomarkers. Baseline models are end-to-end volumetric learners trained on the same cohort; in contrast, CausalRAG-ViViT injects hippocampal, ventricular, and temporal z-scores into the attention mechanism to improve calibration and auditability.

First, we evaluate **binary classification (AD vs. CN)**. Calibration (ECE) and discrimination (F1, AUC, Acc) are shown in Table I. Whereas CausalRAG-ViViT attains the lowest ECE (0.078), over $3\times$ lower than ViT+BiLSTM (0.240), ViT+BiLSTM achieves the best AUC/Acc. One plausible reason is that AD-vs-CN is a simpler decision boundary where slice-sequence aggregation can maximize discrimination by exploiting global patterns across slices. Our ROI/clinical gating regularizes attention toward anatomically motivated regions, improving probability reliability (ECE) but occasionally trading a small amount of raw discrimination. CausalRAG-ViViT is better calibrated, which is beneficial for thresholded decisions and abstention policies where probability quality is important, even if ViT+BiLSTM is marginally stronger in pure discrimination. We next consider the more challenging three-way task (CN/MCI/AD). Table I shows that ViViT-lite has the lowest ECE, whereas CausalRAG-ViViT achieves the highest accuracy (0.519) with competitive F1/AUC. This reflects a calibration–complexity trade-off compared to a lighter ViViT variant, while the causal attention approach yields the strongest overall accuracy across the three classes.



(a) (b)

Fig. 2. (a) Binary reliability curve after temperature scaling. (b) One-vs-rest ROC curves for the tri-class CN/MCI/AD task.

To visualize calibration and separability, Fig. 2(a) shows the binary reliability curve after temperature scaling, and Fig. 2(b) shows the one-vs-rest ROC for the tri-class task. The reliability curve tracks the diagonal closely (consistent with ECE $= 0.078$), and the OvR ROC curves illustrate class separability, matching Table I.

Finally, for captioning compliance and fidelity, we evaluate the generated caption PNGs, which are self-contained clinical cards with a structured report (*Summary, Patient Information, Findings, Biomarker Context, Impression, Disclaimer*), a side-by-side Original/Processed mid-slice, and AT(N) chips on the processed slice. Each report includes an explicit Evidence line (MRI cut-points; CSF cohort thresholds), a brief causal rationale (which domain crossed which cut-point and by how much), and a counterfactual (minimum change to flip N).

On $N = 4{,}427$ reports (Table II), we observe **100%** section coverage, **93.6%** MRI provenance mentions, and **42.1%** CSF line presence; numeric fidelity is **100%** for MRI z-scores, MMSE, and CDR-SB, and Aβ42 matches in **93.9%**. Metrics were computed automatically via a verifier script that parses exported report JSON/text and exact-matches numeric fields to retrieved source-table values (after fixed formatting); mismatches are logged. AT(N) tokens agree with rules in

TABLE I

BINARY AND TRI-CLASS CROSS-VALIDATION (OOF; ECE AFTER PER-FOLD TEMPERATURE SCALING). THE PROPOSED MODEL ACHIEVES THE LOWEST ECE IN THE BINARY TASK AND COMPETITIVE TRI-CLASS PERFORMANCE.

	Binary cross-validation				Tri-Class cross-validation			
Model	**F1**	**AUC**	**Acc**	**ECE**	**F1**	**AUC**	**Acc**	**ECE**
ViT+BiLSTM	**0.711**	**0.771**	**0.712**	0.240	**0.485**	**0.658**	0.510	0.371
ViViT-lite	0.576	0.635	0.599	0.079	0.381	0.567	0.443	**0.124**
Proposed	0.646	0.719	0.674	**0.078**	0.452	0.643	**0.519**	0.276

TABLE II

CAPTION AUDIT (4,427 REPORTS). HIGH COVERAGE AND NUMERIC FIDELITY SUPPORT AUDITABILITY; CSF LINE PRESENCE REFLECTS DATASET AVAILABILITY.

Metric	**Value**
All required sections present	**100.0%**
MRI provenance mentioned	93.6%
CSF values line present	42.1%
Numeric fidelity (MRI z, MMSE, CDR-SB)	**100.0%**
CSF Aβ42 fidelity (where present)	93.9%
AT(N) agreement (A/T/N)	**100.0%**

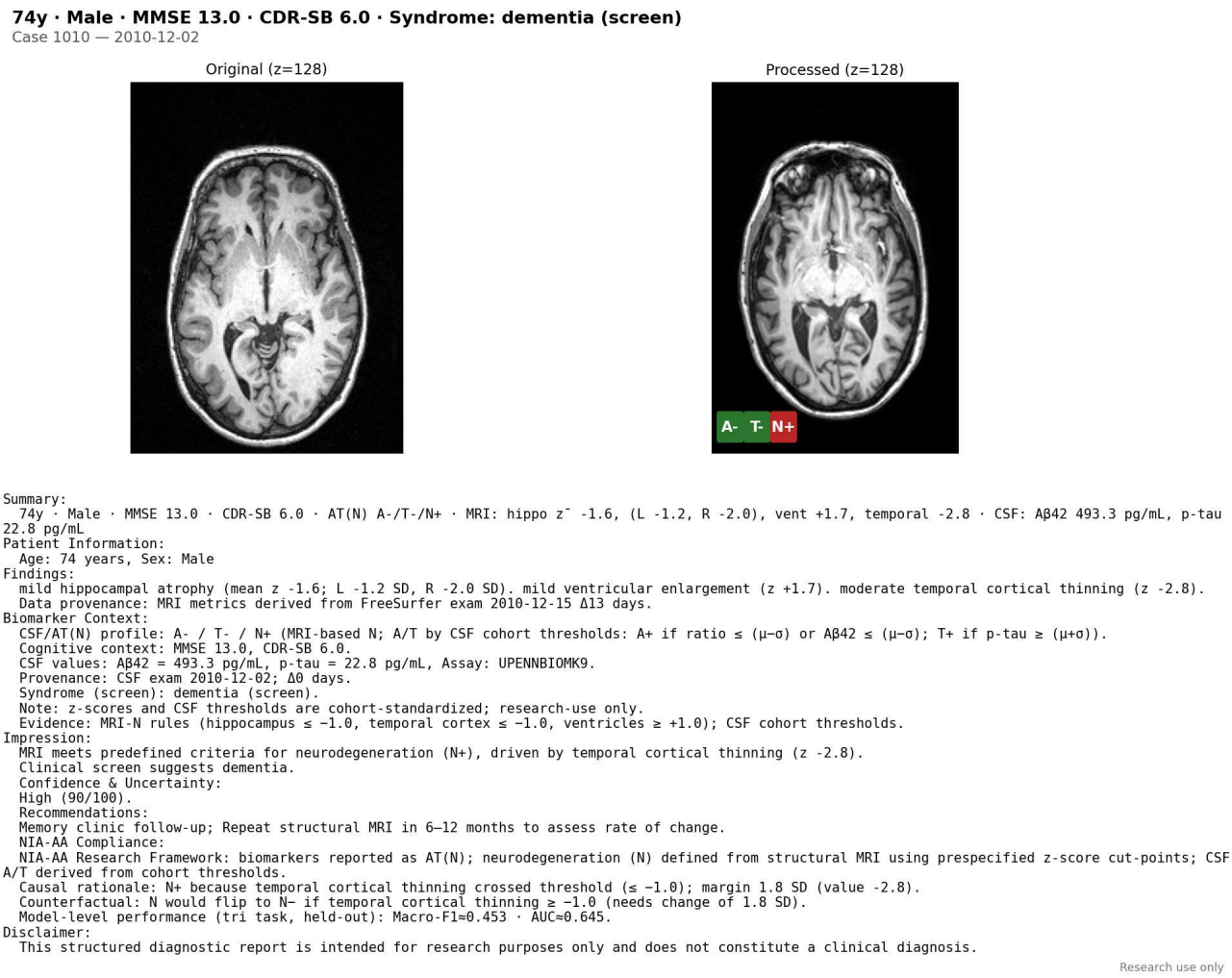


Fig. 3. **Guideline-compliant MRI caption (example dementia screen).** Banner, Original/Processed slice with AT(N) chips, and structured report (truncated).

100% of cases where data are available. The takeaway is that captions are complete, numerically faithful, and largely guideline-consistent, supporting clinical auditability; the lower CSF presence reflects real-world missingness rather than formatting errors.

Overall, CausalRAG-ViViT achieves highest accuracy in the tri-class setting while trading a small amount of raw discrimination in the binary setting for markedly improved calibration. The caption pipeline aligns probabilistic results with reporting guidelines by achieving perfect numeric fidelity along with section coverage on core MRI/clinical fields.

V. DISCUSSION AND CONCLUSION

Our study trained all models from scratch on the same unified ADNI 1/GO/2/3 cohort with subject-wise out-of-fold (OOF) evaluation, enabling a fair comparison of architectures. A primary observation is a calibration–discrimination trade-off: in the binary task (AD vs. CN), ViT+BiLSTM achieves the highest AUC/Acc, whereas CausalRAG-ViViT attains the lowest ECE after per-fold temperature scaling (Table I). This makes probabilities more reliable for thresholded decisions with principled abstention. For example, a high calibrated AD

probability can prioritize referral for confirmatory biomarkers/specialist follow-up, while mid-range probabilities can trigger repeat testing and clinician review rather than an overconfident label. Calibrated probabilities also enable an abstention band ($|p - 0.5| < \delta$) to defer uncertain cases. Because each sentence in the report is generated from retrieved, versioned fields ROI z-scores, MMSE/CDR-SB, available CSF with assay/date provenance, and cohort cut-points the RAG design keeps every claim traceable to its source, simplifying audit and QA. In the more challenging CN/MCI/AD setting, CausalRAG-ViViT yields the highest overall accuracy with competitive F1/AUC, while a lighter ViViT variant achieves the lowest ECE (Table I). Fig. 2 reflects this: the reliability curve closely follows the diagonal (low binary ECE), and the one-vs-rest ROCs show clear separability. Injecting hippocampal/temporal/ventricular z-scores into attention balances discrimination and probability trustworthiness. Beyond scalar metrics, our caption pipeline renders model outputs as auditable artifacts; an example card appears in Fig. 3. Across $N = 4{,}427$ caption PNGs (Table II), we observe 100% section coverage and 100% numeric fidelity for MRI z-scores, MMSE, and CDR-SB; the 42.1% CSF line presence reflects dataset availability rather than formatting gaps. These cards consolidate calibrated predictions and explicit evidence, enabling clear policies, fixed referral thresholds with uncertainty-aware abstention bands ($|p-0.5| < \delta$) and emphasizing deployability over headline accuracy. CausalRAG-AD combines calibrated MRI transformers with retrieval-augmented, NIA-AA aligned reporting on a single ADNI 1/GO/2/3 cohort. With temperature scaling and ROI/clinical gating, CausalRAG-ViViT is competitive on discrimination, achieves the best binary calibration, and attains the highest tri-class accuracy. The AT(N) summaries have exact numerical fidelity and section coverage, and tokens match rule-derived labels when evidence is present. We provide OOF predictions, frozen splits, and caption audits to support impartial comparisons and replication, shifting emphasis to deployability: calibrated decisions and checkable narratives for accountable, real-world use.

Although ADNI spans multiple sites and vendors, external validation is needed since demographics and acquisition may not represent all populations or settings. AT(N) cut-points are cohort-based and CSF availability varies; we therefore report AT(N) only with supporting evidence and flag uncertainty. Reports are research-use, not clinical diagnoses. Next, we plan to add FLAIR, diffusion, PET, APOE, and key EHR fields, and strengthen uncertainty via improved multiclass calibration and simple coverage-guaranteed methods (e.g., conformal prediction). We will move toward clinical integration via a lightweight SMART-on-FHIR/DICOM-SR interface linked to PACS/EHR with human-in-the-loop review, perform external validation (AIBL, OASIS-3, NACC), and run a brief clinical study to assess usability and decision support.

ACKNOWLEDGMENT

This work is supported by the National Science Foundation (NSF) grant (ID. 2131307) "CISE-MSI: DP: IIS: III: Deep Learning-Based Automated Concept and Caption Generation of Medical Images Towards Developing an Effective Decision Support. Data were obtained from ADNI. We thank the ADNI consortium and the developers of FreeSurfer and dcm2niix.

REFERENCES

[1] World Health Organization, "Dementia — Key facts," Fact sheet, 31 Mar 2025. Available: https://www.who.int/news-room/fact-sheets/detail/dementia [Accessed: 8 Sep 2025].

[2] C. R. Jack Jr. *et al.*, "The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods," *Journal of Magnetic Resonance Imaging*, vol. 27, no. 4, pp. 685–691, 2008. doi:10.1002/jmri.21049.

[3] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations (ICLR)*, 2021. Available: https://openreview.net/forum?id=YicbFdNTTy.

[4] T. Akan *et al.*, "Vision Transformers and Bi-LSTM for Alzheimer's Disease Diagnosis from 3D MRI," *arXiv:2401.03132*, 2024. doi:10.48550/arXiv.2401.03132. Available: https://arxiv.org/abs/2401.03132

[5] T. Akan *et al.*, "Leveraging Video Vision Transformer for Alzheimer's Disease Diagnosis from 3D Brain MRI (ViTranZheimer)," *arXiv:2501.15733*, 2025. doi:10.48550/arXiv.2501.15733. Available: https://arxiv.org/abs/2501.15733

[6] A. Arnab *et al.*, "ViViT: A Video Vision Transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 6836–6846. doi:10.1109/ICCV48922.2021.00676. Available: https://doi.org/10.1109/ICCV48922.2021.00676

[7] T. Wu *et al.*, "BiCAL: Bi-directional Contrastive Active Learning for Clinical Report Generation," in *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing (BioNLP)*, Bangkok, Thailand, Aug. 2024, pp. 328–341. doi:10.18653/v1/2024.bionlp-1.25. Available: https://aclanthology.org/2024.bionlp-1.25/

[8] M. Ranjit *et al.*, "Retrieval Augmented Chest X-Ray Report Generation using OpenAI GPT models (CXR-RePaiR-Gen)," *arXiv:2305.03660*, 2023. doi:10.48550/arXiv.2305.03660. Available: https://arxiv.org/abs/2305.03660

[9] S. Liu *et al.*, "Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines," *Journal of the American Medical Informatics Association*, vol. 32, no. 4, pp. 605–615, 2025. doi:10.1093/jamia/ocaf008. PMID: 39812777.

[10] Doshi, R. et al., "Quantitative Evaluation of Large Language Models to Streamline Radiology Report Impressions: A Multimodal Retrospective Analysis," *Radiology*, 2024. doi:10.1148/radiol.231593.

[11] N. J. Dhinagar *et al.*, "Leveraging a Vision-Language Model with Natural Text Supervision for MRI Retrieval, Captioning, Classification, and Visual Question Answering," *bioRxiv*, 2025. doi:10.1101/2025.02.15.638446. PMID: 40027630. PMCID: PMC11870526. Available: https://www.biorxiv.org/content/10.1101/2025.02.15.638446v1

[12] C. R. Jack Jr. *et al.*, "NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease," *Alzheimer's & Dementia*, vol. 14, no. 4, pp. 535–562, 2018. doi:10.1016/j.jalz.2018.02.018.

[13] C. Rorden, "dcm2niix," 2018. Available: https://github.com/rordenlab/dcm2niix

[14] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4ITK: Improved N3 bias correction," *IEEE Transactions on Medical Imaging*, 2010. doi:10.1109/TMI.2010.2046908. PMID: 20378467.

[15] F. Isensee *et al.*, "Automated brain extraction of multisequence MRI using artificial neural networks," *Human Brain Mapping*, 2019. doi:10.1002/hbm.24750. PMID: 31403237.

[16] B. Fischl, "FreeSurfer," *NeuroImage*, 2012. doi:10.1016/j.neuroimage.2012.01.021. PMID: 22248573

[17] C. Guo *et al.*, "On Calibration of Modern Neural Networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, PMLR, vol. 70, pp. 1321–1330, 2017. Available: https://proceedings.mlr.press/v70/guo17a.html