EMAD: Evidence-Centric Grounded Multimodal Diagnosis for Alzheimer's Disease

Qiuhui Chen[1], Xuancheng Yao[2], Zhenglei Zhou[3], Xinyue Hu[2], Yi Hong[2*]
[1]East China University of Science and Technology, [2]Shanghai Jiao Tong University, [3]Tencent
chenqh@ecust.edu.cn, {2212582443, hxy1246475237, yi.hong}@sjtu.edu.cn, rianzhou@tencent.com

1. Implementation Details and Exact Scoring in NIA-AA Diagnostic Reward

NIA-AA Diagnostic Reward ($R_{\text{NIA-AA}}$) Provides comprehensive clinical assessment through a multi-dimensional scoring framework that evaluates diagnostic accuracy against established NIA-AA standards. The reward integrates three core components:

$$R_{\text{NIA-AA}} = 0.4 \cdot R_{\text{cat}} + 0.3 \cdot R_{\text{bio}} + 0.3 \cdot R_{\text{feat}}. \quad (1)$$

Diagnostic Category Alignment (R_{cat}) ensures precise classification into standardized diagnostic categories (CN, MCI, Dementia) through keyword matching and exclusion criteria validation. This component evaluates both the presence of appropriate diagnostic terminology and the absence of contradictory indicators.

Biomarker Consistency Assessment (R_{bio}) quantifies the coverage and contextual accuracy of essential AD biomarkers (Aβ, tTau, pTau). The scoring incorporates both mention frequency and pathological status characterization (normal/abnormal patterns) based on established clinical thresholds.

Clinical Feature Comprehensiveness (R_{feat}) evaluates the depth of cognitive domain analysis across memory, executive function, visuospatial abilities, and language domains. The scoring rewards not only feature inclusion but also detailed characterization within specific subdomains.

This structured approach ensures rigorous adherence to NIA-AA diagnostic protocols while maintaining computational efficiency through weighted component integration.

The NIA-AA diagnostic reward function provides a comprehensive assessment framework for evaluating Alzheimer's disease diagnostic reports generated by our model. This multi-dimensional scoring system ensures clinical accuracy and adherence to established NIA-AA diagnostic standards through three core components with weighted integration:

$$R_{\text{NIA-AA}} = 0.4 \cdot R_{\text{category}} + 0.3 \cdot R_{\text{biomarker}} + 0.3 \cdot R_{\text{feature}} \quad (2)$$

1.1. Diagnostic Category Matching (R_{category})

The diagnostic category component evaluates the accuracy of diagnostic classification through multi-tiered keyword validation. This 40%-weighted component ensures precise alignment with standard diagnostic categories (CN, MCI, Dementia) while penalizing contradictory terminology.

The scoring incorporates inclusion validation and exclusion penalty mechanisms:

$$R_{\text{category}} = \mathbb{I}\text{inclusion} \cdot (1 - \mathbb{I}\text{exclusion}) + R_{\text{staging}} \quad (3)$$

where $\mathbb{I}$inclusion validates presence of category-appropriate keywords, $\mathbb{I}$exclusion penalizes contradictory terminology, and R_{staging} provides additional scoring for dementia stage assessment.

1.2. Biomarker Consistency ($R_{\text{biomarker}}$)

The biomarker consistency component (30% weight) evaluates both coverage and pathological characterization of core AD biomarkers (Aβ, pTau, tTau). The assessment employs clinical importance weighting and status consistency validation.

The scoring formula integrates mention frequency and status accuracy:

$$R_{\text{biomarker}} = \sum_{b \in \mathcal{B}} w_b \cdot (\alpha \cdot \mathbb{I}\text{mention}(b) + \beta \cdot \mathbb{I}\text{status}(b)) \quad (4)$$

where $\mathcal{B} = \text{A}\beta, \text{pTau}, \text{tTau}$ represents the biomarker set, w_b denotes clinical weights ($w_{\text{A}\beta} = 0.4$, $w_{\text{pTau}} = 0.3$, $w_{\text{tTau}} = 0.3$), $\mathbb{I}$mention detects biomarker presence, and $\mathbb{I}$status evaluates pathological status consistency.

Status assessment utilizes pattern recognition for normal/abnormal classification:

*Corresponding author.

$$\mathbb{I}\text{status}(b) = \frac{\sum p \in P_b^{\text{normal}} \mathbb{I}(p) + \sum_{p \in P_b^{\text{abnormal}}} \mathbb{I}(p)}{|P_b^{\text{normal}} \cup P_b^{\text{abnormal}}|} \quad (5)$$

where P_b represents status-indicative patterns for biomarker b.

1.3. Clinical Feature Coverage (R_{feature})

Clinical feature assessment (30% weight) evaluates cognitive domain coverage across memory, executive function, visuospatial abilities, and language domains. The scoring incorporates both breadth of coverage and descriptive specificity with clinical significance weighting.

The comprehensive scoring framework:

$$R_{\text{feature}} = \sum_{f \in \mathcal{F}} w_f \cdot (\gamma \cdot \mathbb{I}\text{domain}(f) + \delta \cdot \mathbb{I}\text{specificity}(f)) \quad (6)$$

where $\mathcal{F} = \text{memory}, \text{executive}, \text{visuospatial}, \text{language}$ represents cognitive domains, w_f denotes clinical significance weights, $\mathbb{I}$domain evaluates primary domain coverage, and $\mathbb{I}$specificity assesses subdomain characterization depth.

Domain-specific weighting reflects clinical importance in AD diagnosis:

$$w_f = \begin{cases} 0.4 & \text{memory} \\ 0.3 & \text{executive function} \\ 0.2 & \text{visuospatial abilities} \\ 0.1 & \text{language} \end{cases} \quad (7)$$

1.4. Text Processing Pipeline

The reward function employs a robust text processing workflow including format sanitization, case normalization, and clinical tokenization. Structured field extraction utilizes regular expression patterns:

$$\text{Diagnosis} = \text{extract}(\text{response}, \langle\text{diagnosis}\rangle.*?\langle/\text{diagnosis}\rangle) \quad (8)$$

$$\text{Reasoning} = \text{extract}(\text{response}, \langle\text{reasoning}\rangle.*?\langle/\text{reasoning}\rangle) \quad (9)$$

This algorithmic framework ensures rigorous adherence to NIA-AA diagnostic protocols while maintaining computational efficiency through weighted component integration. The implementation provides clinically meaningful reward signals that guide the reinforcement learning process toward generating accurate, comprehensive, and logically consistent AD diagnostic reports.

2. AD-MultiSense Dataset Statistics and Preprocessing

2.1. Multimodal Data Collection

To enable MLLMs to perform both physiological understanding and diagnostic reasoning over heterogeneous medical data, we construct a multimodal dataset that conforms to established clinical logic, as shown in Fig. 1. Raw data are collected from the ADNI [7] and AIBL [3] cohorts, covering a wide spectrum of patient characteristics and disease stages. For each subject, we acquire sMRI scans alongside six types of clinical data encompassing demographic, cognitive, and biochemical information. After aligning data across modalities and visit timepoints, we curate a total of 10,378 multimodal samples from 2,619 unique subjects. Each sample reflects a consistent physiological state at a specific visit, enabling clinically valid reasoning over disease progression.

To enhance clinical interpretability, quantitative measurements are systematically converted into standardized textual reports. For sMRI analysis, we calculate age-adjusted z-scores for structural volumes (e.g., hippocampal/ventricular) using population norms, with textual descriptors generated based on established thresholds: bilateral hippocampus atrophy is reported as "mild" ($1 \leq |z| < 1.5$), "moderate" ($1.5 \leq |z| < 2$), "significant" ($2 \leq |z| < 3$) or "profound" ($|z| \geq 3$). Similarly, laboratory data undergoes z-score normalization against age/sex-matched cohorts, though only clinically significant abnormalities ($|z| > 2.0$) are included in final reports. Biomarkers are consistently interpreted with contextual information, and each value is accompanied by reference-based interpretation, e.g., "Amyloid beta: 858.30 pg/mL (normal)." This quantitative-to-textual transformation bridges raw biomarker measurements with clinically meaningful narratives, enabling natural language reasoning about pathological changes while preserving data fidelity.

The vision description generation module transforms quantitative neuroimaging measurements into clinically interpretable natural language descriptions. This transformation employs a multi-step analytical process that contextualizes individual volumetric data within population-based reference distributions. For each brain structure of interest, the system first establishes an age and gender-matched reference cohort derived from cognitively normal subjects. This cohort is stratified into decade-wide age groups (50-59, 60-69, 70-79, 80-89 years) with separate distributions maintained for male and female populations.

Three core metrics are computed to quantify deviations from normative values. The Z-score represents standard deviation units from the reference mean, calculated as

$$Z = (V_{subject} - \mu_{ref})/\sigma_{ref} \quad (10)$$

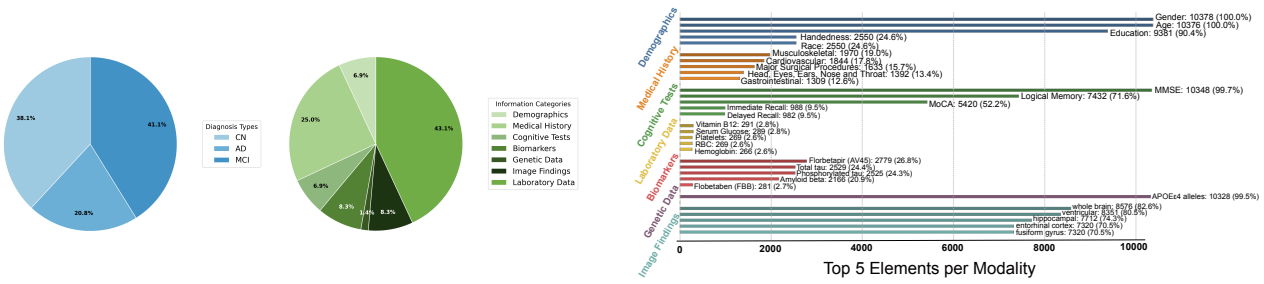


Figure 1. Disease and Modality Distribution of AD-MultiSense Dataset.

where $V_{subject}$ is the observed volume, μ_{ref} is the reference mean, and σ_{ref} is the reference standard deviation. The percentile rank indicates the proportion of healthy individuals with smaller volumes, derived from the cumulative distribution function of the reference population. The percentage difference expresses relative deviation as

$$\Delta\% = (V_{subject} - \mu_{ref})/\mu_{ref} \times 100, \tag{11}$$

providing an intuitive measure of volumetric change.

Clinical severity classifications incorporate structure-specific pathological directionality. For atrophy-sensitive structures including the hippocampus, entorhinal cortex, fusiform gyrus, middle temporal gyrus, and whole brain, we apply the criteria in Table 1:

Table 1. Clinical interpretation of Z-scores for brain structures

Z-score Range	Clinical Interpretation
$Z < -3$	Profound atrophy
$-3 \leq Z < -2$	Significant atrophy
$-2 \leq Z < -1.5$	Moderate atrophy
$-1.5 \leq Z < -1$	Mild atrophy
$-1 \leq Z \leq 1$	Normal volume
$1 < Z \leq 1.5$	Mild enlargement
$1.5 < Z \leq 2$	Moderate enlargement
$2 < Z \leq 3$	Significant enlargement
$Z > 3$	Profound enlargement

These thresholds align with established radiological practice while maintaining statistical rigor.

Natural language generation follows a standardized template that synthesizes these quantitative metrics into clinically actionable interpretations for all six structures. Each description includes four key elements: 1) the absolute volumetric measurement, 2) percentage difference from the reference mean, 3) Z-score with corresponding percentile rank, and 4) clinical severity assessment. The template dynamically adapts terminology based on pathological directionality, using "below" and "atrophy" for cortical structures versus "above" and "enlargement" for ventricles. This approach ensures consistent reporting while maintaining clinical relevance across diverse brain structures.

Table 2 presents representative outputs of the vision description generation system for all six brain structures. These structured interpretations provide clinicians with immediately actionable information by contextualizing quantitative measurements within population norms. The comprehensive coverage of ventricles, hippocampal formation, global brain volume, and temporal lobe structures enables a holistic assessment of neurodegenerative patterns. The framework's modular design permits seamless integration of additional brain regions while maintaining standardized reporting protocols across neuroimaging evaluations.

Figure 3 presents a comparative analysis of six key brain structure volumes across diagnostic groups: cognitively normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease dementia (AD/Dementia). Violin and box plots demonstrate significant volumetric differences in all structures that effectively discriminate between diagnostic categories. Most notably, ventricular volume exhibits progressive enlargement across the CN→MCI→AD continuum, while hippocampal, entorhinal, and mid-temporal volumes show corresponding stepwise reductions. Fusiform and whole brain volumes similarly decrease with disease progression. The distributions reveal three critical patterns: 1) AD patients consistently demonstrate the most pronounced atrophy (or ventricular expansion), 2) MCI subjects exhibit intermediate values with greater distributional overlap with both CN and AD groups, and 3) CN individuals maintain the highest preserved volumes. These z-score distributions provide robust imaging biomarkers that collectively differentiate diagnostic categories, with ventricular and hippocampal measures showing the most distinct group separation.

Figure 2 presents a comprehensive analysis of hippocampal volume Z-scores, normalized to age- and gender-matched cognitively normal references. Panel A shows the overall distribution with clinically significant thresholds at Z = -1 and Z = -2, revealing a right-skewed distribution indicative of hippocampal atrophy in the cohort. The boxplot analysis in Panel B demonstrates progressive Z-score reduction across the diagnostic continuum (CN → MCI → Dementia), with females exhibiting consistently lower Z-scores than males within each diagnostic category (ΔZ = [gender-diff], $p \leq 0.001$).

Table 2. Representative vision descriptions for brain structures

Structure	Generated Description
Ventricles	Ventricular volume measures 42,500 mm^3, 32.5% above the reference mean (32,070 ± 2,850 mm^3). With a Z-score of 3.65 (99.9^{th} percentile), this represents significant enlargement.
Hippocampus	Hippocampal volume measures 2,850 mm^3, 28.2% below the reference mean (3,970 ± 350 mm^3) for this demographic. The Z-score of -3.21 (0.1^{th} percentile) indicates significant atrophy.
Whole Brain	Whole brain volume measures 950,000 mm^3, 8.7% below the reference mean (1,040,000 ± 45,000 mm^3). The Z-score of -2.00 (2.3^{th} percentile) demonstrates mild atrophy.
Entorhinal Cortex	Entorhinal cortex volume is 2,350 mm^3, 35.1% below reference values. The Z-score of -3.02 (0.1^{th} percentile) is consistent with significant atrophy.
Fusiform Gyrus	Fusiform gyrus volume measures 18,600 mm^3, 15.3% below the reference mean (21,970 ± 1,850 mm^3). With a Z-score of -1.82 (3.4^{th} percentile), this suggests mild atrophy.
Middle Temporal Gyrus	Middle temporal gyrus volume measures 17,600 mm^3, 22.7% below the reference mean (22,750 ± 2,100 mm^3). The Z-score of -2.45 (0.7^{th} percentile) demonstrates significant atrophy.

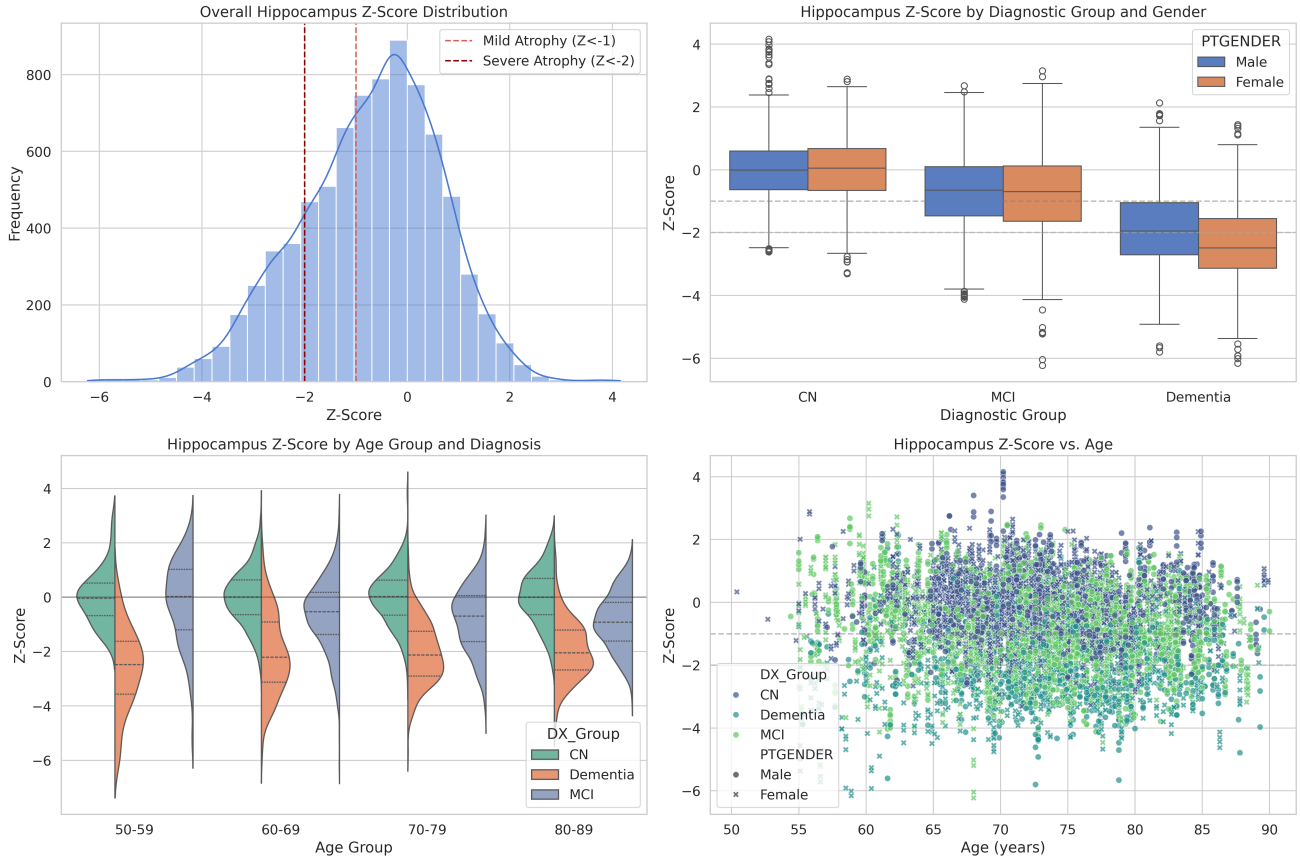


Figure 2. Distribution of hippocampal Z-scores across demographic and clinical dimensions.

Panel C illustrates the interaction between aging and neurodegeneration, where dementia patients show substantially lower Z-scores across all age groups, particularly in the 70-79 cohort. The scatterplot in Panel D confirms the expected age-related decline in hippocampal volumes (r = [correlation-value], p $\leq$ [p-value]), while highlighting the diagnostic separation maintained across the age spectrum. The horizontal reference lines at Z = -1 and Z = -2 provide clinical context for interpreting individual data points.

2.2. Reasoning Generation

Based on these raw data, we construct multimodal QA pairs from disease-level diagnostic reasoning. The process begins by querying the *Thinker* model (DeepSeek-V3) using

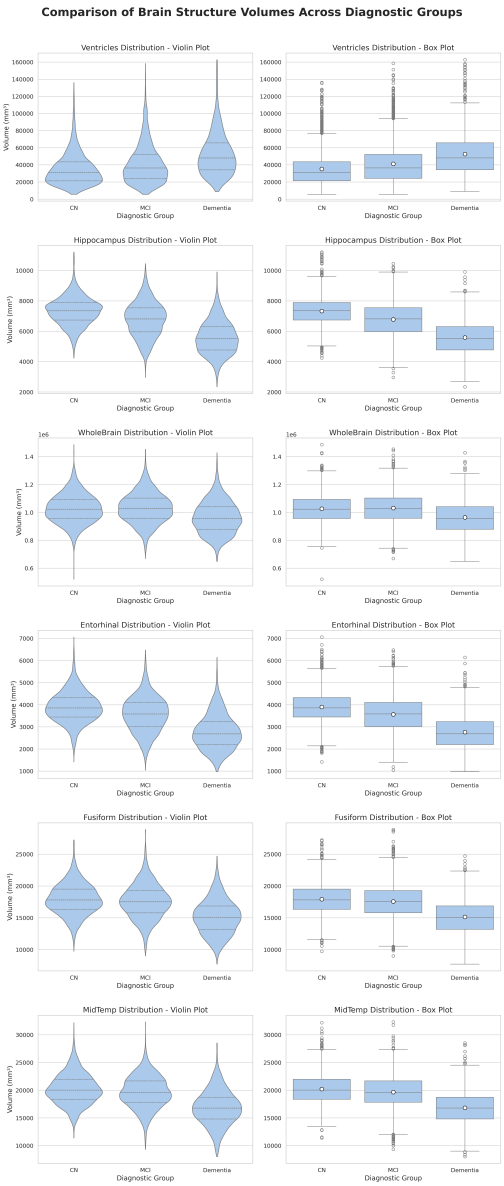


Figure 3. Volumetric distributions of six brain structures across diagnostic groups. Left column: Violin plots showing density distributions and quartiles. Right column: Box plots with white circles indicating means. Structures shown (top to bottom): Ventricles, Hippocampus, WholeBrain, Entorhinal, Fusiform, and MidTemp. CN = Cognitively Normal (n=2732), MCI = Mild Cognitive Impairment (n=3150), Dementia = Alzheimer's Disease Dementia (n=1349). Note progressive ventricular enlargement and hippocampal/entorhinal atrophy across the CN→MCI→AD continuum.

a structured diagnostic prompt template:

```
SYSTEM_PROMPT: "You are an
Alzheimer's specialist.  Analyze
the data and provide:
1.  Reasoning
2.  Final diagnosis:
CN/MCI/Dementia
3.  Confidence level:
High/Medium/Low
Format:
Reasoning:  [analysis]
Diagnosis:  [CN/MCI/Dementia]
Confidence:  [High/Medium/Low]"
```

This is an initial response $\langle R_0, C_0 \rangle = \text{Thinker}(M, P_d)$, where R_0 denotes the reasoning chain, C_0 is the preliminary diagnosis, M represents multimodal inputs (i.e., sMRIs and clinical data), and P_d is the diagnosis prompt.

The *Validator* module evaluates C_0 against ground truth diagnoses. When mismatches occur, the system triggers rethinking cycles: the Thinker regenerates reasoning using refinement prompts (P_r) constructed from explicit NIA-AA criteria dictionaries. These dictionaries map clinical findings to diagnostic rules, enabling targeted feedback. This iterative process continues for up to N cycles (i.e., 2), with random expert sampling providing quality control.

For cases where diagnosis remains incorrect after N iterations, the prompts with correct diagnosis (P_c) is explicitly provided to the Thinker, instructing it to correct its reasoning and conclusion accordingly. The Thinker then produces final reasoning R^F and diagnosis C^F, formatted into training pairs $\langle M \circ P_d, R^F \circ C^F \rangle$ for supervised fine-tuning.

3. Hyperparameters for Optimization

Optimization hyperparameters used during training are listed: Learning Rate $1e{-}4$ for LLaMA and $5e{-}5$ for backbone models. Batch Size 4 per batch. Contrastive Temperature (τ) 0.07 for contrastive learning between text and image features. Distillation Temperature 2.0 for GTX-Distill distillation process. KL Weight (λ_{KL}) 1.0 for the distillation loss. Group Size (G) 4, Clipping Parameter (ϵ) 0.2 and KL Coefficient (β) 0.1 for Executable-Rule GRPO.

The LLaMA 3.2-1B model [4] is used as the text decoder with rank-8 LoRA adapters [5]. The LLaMA model consists of multiple layers, and in this work, we employ LoRA to improve performance while maintaining computational efficiency. A 3D Vision Transformer (ViT) [2] is used to process sMRI images. The model has been adapted to handle 3D volumetric data, which is crucial for working with medical imaging data such as MRI scans. We utilize the Longformer encoder [1] to handle clinical text data. Longformer is known for its efficient handling of long sequences, and it is essential for processing medical and clinical records. For SEA Grounding, we adopt the Segformer3D architecture [6]. This model is particularly effective in segmentation tasks and is adapted to handle 3D data from medical scans. The model uses a 768-dimensional text embedding and a contrastive temperature of $\tau = 0.07$. The

model employs a 4-layer 3D decoder for generating the reconstructed sMRI scans. This decoder is responsible for taking the embeddings generated by the vision and text encoders and reconstructing the target outputs.

4. Cross-cohort Generalization

To evaluate the generalization capability of EMAD across different patient cohorts, we conduct cross-dataset experiments: training on the ADNI dataset and testing on the AIBL dataset. Table 4 summarizes the diagnostic performance of EMAD and baseline methods under this setting. EMAD achieves an accuracy of 85.6% and an AUC of 84.3% on AIBL when trained solely on ADNI, demonstrating robust cross-cohort generalization. This suggests that the evidence-grounded reasoning and multimodal alignment in EMAD help mitigate dataset-specific biases and improve model reliability in real-world clinical settings.

5. Label Efficiency via GTX-Distill

To evaluate the label efficiency of our proposed GTX-Distill framework, we conducted experiments with varying fractions of fully annotated grounding data. As shown in Table 3, we trained teacher models with different amounts of grounding supervision (10%, 25%, 50%, and 100% of the training set) and then distilled this knowledge to student models using GTX-Distill on the full generated reports.

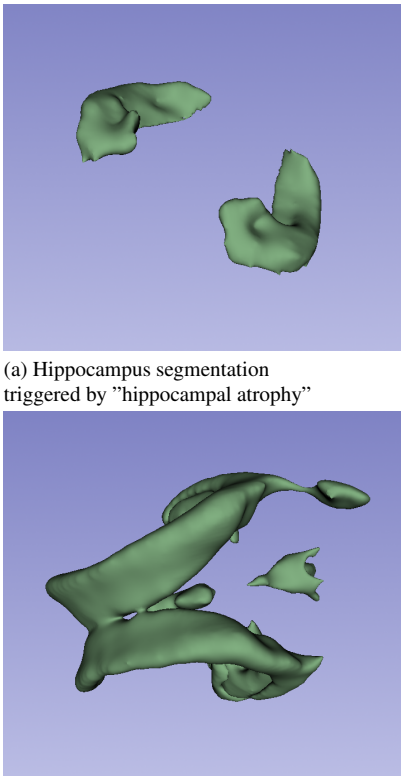
The results demonstrate that GTX-Distill effectively transfers grounding knowledge even with limited annotations. Notably, with only 25% of grounding labels, the student model retains approximately 81% of the teacher's sentence-evidence retrieval performance and 90% of the evidence-anatomy segmentation quality. This highlights the practical value of GTX-Distill in reducing annotation costs while maintaining grounding faithfulness in clinical report generation.

Furthermore, we observed diminishing returns beyond 50% annotation coverage, suggesting that GTX-Distill can achieve near-optimal performance with half the annotation effort required for full supervision. This makes our approach particularly suitable for medical domains where expert annotations are scarce and expensive to obtain.

6. Qualitative Results

We present additional qualitative examples to demonstrate EMAD's capability in generating evidence-grounded diagnostic reports. Figure 5 shows two representative cases comparing EMAD's outputs with ground truth reports.

Figure 4 demonstrates EMAD's evidence-to-anatomy grounding capability, showing how clinical evidence phrases trigger precise 3D segmentation of relevant brain structures.



(a) Hippocampus segmentation triggered by "hippocampal atrophy"

(b) Ventricular segmentation triggered by "ventricular enlargement"

Figure 4. Evidence-conditioned 3D anatomical segmentation. Each clinical evidence phrase activates segmentation of the corresponding brain structure, providing explicit anatomical grounding for report claims.

References

[1] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. 5

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020. 5

[3] Kathryn A Ellis, Ashley I Bush, David Darby, et al. The australian imaging, biomarkers and lifestyle (aibl) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of alzheimer's disease. *International psychogeriatrics*, 21(4):672–687, 2009. 2

[4] Aaron Grattafiori et al. The llama 3 herd of models, 2024. 5

[5] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 5

[6] Shehan Perera, Pouyan Navard, and Alper Yilmaz. Segformer3d: an efficient transformer for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on*

Table 3. Label efficiency analysis of GTX-Distill with varying fractions of grounding annotations. Performance is measured by sentence-evidence R@3 and evidence-anatomy Dice scores.

Grounding Labels	Sentence-Evidence (R@3)			Evidence-Anatomy (Dice)		
	Teacher	Student	% of Full	Teacher	Student	% of Full
10%	0.63	0.60	71.4%	0.72	0.69	84.1%
25%	0.72	0.68	81.0%	0.76	0.74	90.2%
50%	0.79	0.77	91.7%	0.80	0.79	96.3%
100%	0.84	0.82	97.6%	0.82	0.81	98.8%

Table 4. Cross-cohort generalization results (in %) of models trained on ADNI and tested on AIBL. EMAD demonstrates strong generalization, outperforming baselines in diagnostic accuracy and AUC.

Method	ACC	AUC	SEN	SPE
LLaVA-Med	74.8	72.1	70.3	79.5
Med-PaLM-M	77.5	75.6	73.8	81.2
M3D-LaMed	80.2	78.9	76.5	83.7
EMAD (ours)	**85.6**	**84.3**	**82.1**	**88.9**

Case 1: Mild Cognitive Impairment (MCI)

Ground Truth	EMAD
Reasoning: Moderate hippocampal atrophy (-28%), normal ventricular volume. MMSE: 28, MoCA: 24 indicates mild executive dysfunction. APOEe4: 1 allele.	*Reasoning:* Significant hippocampal volume reduction (27.3% below mean) with preserved whole-brain volume. MoCA score of 20 suggests mild cognitive impairment.
Diagnosis: MCI — *Confidence:* Medium	*Diagnosis:* MCI — *Confidence:* High

Evidence Grounding:

- *Hippocampal volume:* 4,724 mm³ (-27.3%)
- *MoCA score:* 20/30

Case 2: Alzheimer's Disease (AD)

Ground Truth	EMAD
Reasoning: Severe bilateral hippocampal atrophy, ventricular enlargement. MMSE: 20, significant memory impairment. Elevated A and pTau biomarkers.	*Reasoning:* Diffuse cortical atrophy with pronounced medial temporal lobe involvement. Hippocampal volume 35% below reference. A PET SUVR: 1.78, consistent with amyloid pathology.
Diagnosis: AD — *Confidence:* High	*Diagnosis:* AD — *Confidence:* High

Evidence Grounding:

- *Hippocampal volume:* 3,892 mm³ (-35.1%)
- *Aβ PET:* 1.78 SUVR
- *MMSE:* 20/30

Figure 5. Qualitative examples comparing EMAD-generated reports with ground truth annotations. EMAD produces clinically coherent reasoning chains while explicitly linking claims to supporting evidence and anatomical structures.

Computer Vision and Pattern Recognition, pages 4981–4988, 2024. 5

[7] Ronald Carl Petersen, Paul S Aisen, Laurel A Beckett, Michael C Donohue, Anthony Collins Gamst, Danielle J Harvey, et al. Alzheimer's disease neuroimaging initiative (adni): clinical characterization. *Neurology*, 74(3):201–209, 2010. 2