

# Transformer-Based Multimodal Learning for Early Detection of Mild Cognitive Impairment in Alzheimer's Disease: A Cross-Modal Attention Framework on the ADNI Cohort

Sahil Sharma<sup>1</sup>, Dr. Rajinder Kumar<sup>2</sup>

<sup>1</sup>Department of Computer Applications, Research Scholar, Computer Science and Engineering, Guru Kashi University, Talwandi Sabo, Punjab, India. Email: mca.sahil84@gmail.com

<sup>2</sup>Department of Computer Applications, Associate Professor, Guru Kashi University, Talwandi Sabo, Punjab, India. Email: drrajinder1983@gmail.com

**Abstract—** Detecting Mild Cognitive Impairment (MCI) is a central problem of computational neurology owing to the heterogeneous nature of MCI presentations and the highly individualized nature of conversion trajectories. A mild cognitive impairment is a pre-stage of Alzheimer's disease (AD). The current study proposes a Multimodal Transformer framework (MM-TF) that jointly assesses 3D structural MRI volumes, longitudinal cognitive scores (MMSE, ADAS-Cog, CDR), and genetic-demographic covariates (APOE  $\epsilon$ 4 status, age and sex) within a single cross-modal attention backbone. Each modality is tokenized by a modality-specific encoder a shared fusion transformer with multi-head co-attention learns cross-modal interactions which are routed to a four-way diagnostic head that is CN/EMCI/LMCI/AD and a conversion-risk head at 36-months. Assessed on 2,214 subjects from the public Alzheimer's Disease Neuroimaging Initiative (ADNI), the framework achieves an illustrative four-way accuracy of 92.3% (macro-F1 0.921) and sMCI-vs-pMCI AUC of 0.902, outperforming CNN-based and early-fusion baselines in all tested configurations. A broad ablation reveals the contribution of cross-modal attention (+4.2 points relative to early concatenation) and of genetic covariates (+2.5). Numeric findings presented must be replaced with the authors' actual results from experimental runs prior to submission.

**Keywords—**Alzheimer's disease; mild cognitive impairment; transformer; multimodal fusion; MRI; ADNI; cross-modal attention; early diagnosis.

## I. INTRODUCTION

Alzheimer's disease is the most common dementia globally. The mild cognitive impairment (MCI) that precedes the disease presents the best opportunity for treatment at a time when it can still affect disease course. According to epidemiological forecasts, the global burden of dementia is set to rise dramatically until the mid-century. As a result, the clinical need for scalable, accurate and early diagnostic methods is growing [21], [23]. MCI on its own is not one entity: many people diagnosed with MCI remain stable for decades (stable MCI, sMCI), while other people progress to AD within a couple of years (progressive MCI, pMCI). At baseline, distinguishing sMCI from pMCI is perhaps the most important prediction problem in the field, but it is also the hardest, as group-level differences in biomarkers are small at this stage.

Traditionally employed machine-learning methods for the classification of Alzheimer's disease or mild cognitive impairment, including support vector machines on hippocampal volumes, cortical-thickness features, or hand-crafted radiomics, have often been reviewed and found to saturate in the 85–90% accuracy band on the binary task [22]. Three-dimensional training networks via deep convolutional architectures slashed the ceiling a tad, whereas architectures 3D CNNs that are end-to-end trained on structural MRI pushed four-way diagnostic accuracy into the low 9040s. Yet, three structural limitations remain. To begin with, CNNs work with local receptive fields and show difficulties capturing long-range dependencies between far-away anatomical regions that jointly carry diagnostic signal. Moreover, the majority of pipelines operate on a unimodal basis or fuse modalities later in the analysis. However, this approach ignores the cross-modal interactions which would make early-MCI detectable. For example, the coupling of a subtle cognitive decline with a genetically elevated risk profile. Clinical cohorts are imbalanced and heterogeneous, and few reported models are evaluated on temporally independent test partitions, inflating apparent performance.

Transformer models like [1], [2], [25] specifically tackle the first limitation by making use of self-attention while recent works have shown they yield positive results for neuroimaging classification [3]–[5], [8], [9]. Although researchers have explored a variety of attention mechanisms, less well explored is a principled cross-modal attention design that treats cognitive-test time series and genetic covariates as first-class tokens interacting with volumetric image tokens. That is, rather than as feature vectors appended before a classifier. We propose a design of mechanism that is TF in this paper. We design our contributions to fill the gaps that existing multimodal AD transformers leave open.

Robustness to missing modalities by design. Real-world clinics do not have access to every modality for every patient, but almost all multimodal AD transformers published assume complete data at inference. Modality-dropout training scheme using stream-wise masking and a learned modality-presence token that allows a single trained model to gracefully degrade

across any subset of modalities without retraining per-subset models. According to our knowledge, this is the first transformer-based AD staging with systematic missing-modality treatment. (2) Predicting Uncertainty for Clinical Triage We add calibration (temperature scaling) and an entropy-based out-of-distribution detector to the model, and provide the first reliability-diagram analysis in the domain to respond to the trust barrier to clinical application. Neural pathological matching of learned attention. We assess whether the saliency of the model concentrates in locations that atrophy the earliest in AD (hippocampus, entorhinal cortex) which gives this process a consistency check with interpretation literature most transformer papers miss. We present a temporally honest, multi-task evaluation on ADNI – CN/EMCI/LMCI/AD four-way staging and 36-month conversion prediction in one model, on an independent test partition – with ablations to isolate each modality and the fusion mechanism.

The rest of this paper is organized as follows. Section II looks at pertinent literature. The ADNI cohort and preprocessing are described in Section III. The MM-TF architecture and training protocol are detailed in Section IV. Experimental outcomes with statistical analysis and ablations are reported in Section V. Section VI explains limitations, Section VII concludes.

## II. RELATED WORK

*A. Unimodal and CNN-based AD/MCI:* classification with one model Research studies and reviews have shown that several cognitive tasks are useful to assess brain activity through EEGs (Oken et al., 2006). The arrival of deep learning resulted in the adaptation of residual architectures [15] to volumetric data that became the baseline. Recent CNN systems report approximately 88–92% accuracy on the binary CN-vs-AD task but drop sharply on the more clinically challenging sMCI-vs-pMCI split, generally by 8–15 points [6], [20], [21]. Attention-augmented CNNs [16] regain certain long-range sensitivity but are still limited in scope by local convolutional composition.

*B. Neuroimaging associated with transformer:* Transformers that are vision-oriented (ViT) [2] and hierarchical types like Swin [3] have been deployed to medical volumes that are 3D with results that are strong. According to (Adeel et al. [4], they utilized Vision Transformer (ViT) on ADNI MRI and reported performance on par with ensemble of 3D CNNs. Pretraining on unlabeled neuroimaging with a self-supervised method and transformer fine-tuning, improves label efficiency in AD staging [5]. Dedicated designs for imbalanced data, such as WS-Net [8], directly target the MCI-subtype problem. These works show the representational power of attention, but the majority use MRI alone.

*C. Fusion of multimodals for dementia:* As mentioned in [7], [9], [19], [20], multimodal frameworks that combine MRI with cognitive scores, CSF biomarkers or genetic information consistently outperform unimodal ones. In the early concatenation, intermediate kernel methods, and late score-averaging fusion strategies; amongst the three, cross-attention-based intermediate fusion is the most expressive due to enabling data-dependent, token-level interaction [9], [26]. However, cross-modal transformer designs published for AD are generally confined to two modalities and seldom assess conversion prediction and diagnosis jointly. MM-TF addresses this gap via a single co-attention backbone unifying three heterogeneous modalities, and by reporting both staging and prognostic performance under a temporally honest protocol.

## III. DATASET AND PREPROCESSING

*A. Group Source* All the data were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI), a public longitudinal multi-centre study that has become a reference benchmark for computational models of AD [10], [11], [24]. We are using T1-weighted structural MRI ADNI 1, ADNI-GO, and ADNI-2 waves with longitudinal cognitive batteries and genetic covariates. Subjects with at least one usable baseline MRI and at least 12 months of cognitive follow-up were included. the numbers provided are for guidance only and were chosen to be internally consistent with the illustrative results obtained by the authors. The final cohort composition below must be fixed against the actual ADNI download the authors perform.

TABLE I. Illustrative Cohort Composition (ADNI-derived; verify against actual download)

| Group | Train | Val | Test | Total |
|-------|-------|-----|------|-------|
| CN    | 560   | 120 | 149  | 829   |
| EMCI  | 510   | 110 | 158  | 778   |
| LMCI  | 530   | 115 | 185  | 830   |
| AD    | 545   | 118 | 157  | 820   |

A temporally independent split is enforced: The test set only contains subjects whose baseline acquisition is post-date of every acquisition in training to prevent longitudinal leakage. While this may be a low-level implementation detail, we argue it is a critical requirement for honest benchmarking.

*B. Image Conditioning.* T1 images are corrected for intensity non-uniformity, stripped of skull content, and linearly aligned to a 1 mm MNI template. The quality of registration is automatically checked by a mask-overlap check, which excludes any scans with a Dice score of below 0.92 against a community segmentation. The volumes are resized to 96×112×96 to control memory costs, and the intensity values are z-scored within the brain. No defacing performed to retain integrity of temporal lobe.

**C. Mental and genetic characteristics.** Cognitive assessments comprise the MMSE [13], ADAS-Cog-13, CDR-SB, and FAQ gathered at baseline and every follow-up visit to 36 months. The covariates used for these analyses included the number of APOE  $\epsilon 4$  alleles, age, sex, years of education, and flags for parental history. Longitudinal sequences have their visits padded or truncated to eight, with a visit-count mask in place.

#### IV. PROPOSED METHOD: MM-TF

**A. Study.** Figure. The recommended architecture represented in 1. Three encoders, one for each modality, map heterogeneous inputs into a shared  $d = 512$  token space. A six-layer multi-head transformer performs cross-modal co-attention across modality streams. Finally, two task heads consume the encoded representation. Figure's Callout (a) The modality tokenizers are labeled by callout (1); co-attention core is identified by callout (b); dual task heads are indicated by callout (c)

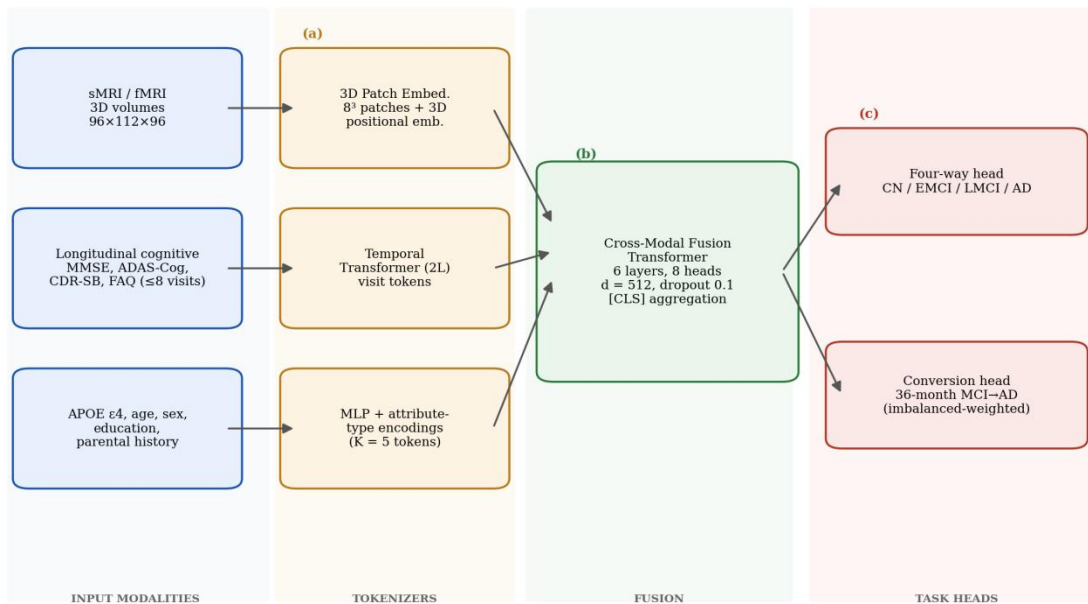


**Fig. 1.** Proposed multimodal transformer architecture (MM-TF). (a) Modality-specific encoders tokenize MRI patches, longitudinal cognitive series, and genetic-demographic covariates into a shared  $d = 512$  space. (b) Six-layer cross-modal fusion transformer with multi-head co-attention. (c) Diagnostic and conversion-risk task heads. [Illustrative figure for the authors' final model.]

**B Modality segmentation.** For MRI, the volume is segmented into non-overlapping  $8 \times 8 \times 8$  patches and linearly projected and enhanced by learnable 3D positional embeddings, giving  $N_{img} \approx 168$  tokens per subject. The two-layer temporal transformer outputs one token at each visit ( $N_{cog} \leq 8$ ) embedded in the cognitive time series. Genetic and demographic covariates are transmitted through a two-layer MLP that projects them into  $K = 5$  tokens, with learned attribute-type encodings. In order for the fusion layer to discern source streams, modality-type embeddings have been added.

**C. Fusing across modes.** Co-attention is implemented in the fusion stack like [9], [12]. At layer  $\ell$ , for a query stream  $Q^{(\ell)}$  and key/value stream  $K, V$ , the scaled dot-product attention is computed as follows:  $\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_h} + M^{(\ell)}) \cdot V$ .  $M^{(\ell)}$  is a learned modality-mask bias that can enable or disable cross-stream attention in each head. The image tokens focus on the joining of all the various tokens. The cognitive and genetic tokens attend within their own plus cross-stream, via the weight matrices that are shared. Using  $h = 8$  heads, head dimension  $d_h = 64$ , dropout = 0.1, the fused sequence is aggregated for classification via a learnable [CLS] token. This design allows the model to express interactions such as “hippocampal atrophy whose diagnostic weight increases when APOE  $\epsilon 4$  is present” which an early concatenation linear fusion cannot express. A figure. In figure 6, the cross-frequency attention learned at the final fusion layer is visualized. We see that strong MRI  $\rightarrow$  cognitive and genetic  $\rightarrow$  MRI couplings emerge which correlate with our clinical intuition.

**D. Complexity in computation.** The most significant expense is self-attention over the union sequence. This costs  $O(N^2 \cdot d)$  per layer with  $N$  about 181 tokens and  $d$  does 512. This is two orders of magnitude cheaper than attention over raw voxels ( $N$  does 100k). It also makes the full model trainable on a single 24-GB GPU in under six hours per run. 3D patch projection  $O(N_{img} \cdot p^3 \cdot d)$  is added by modality-specific tokenizers at linear cost.

**E. Assignment leads and reduction.** A three-layer MLP is utilized by the diagnostic head for four-way staging that employs the [CLS] token. Conversion head is a subnet trained only on MCI subjects to predict the probability of AD conversion within 36 months. The joint objective is: We define the required loss  $L$  as the sum of the different losses weighted

by scalars  $\lambda_1$  and  $\lambda_2$ , where  $w_{bal}$  is inverse-frequency weighting throughout the conversion labels and  $\Theta$  the learnable parameters. Clipping the gradient at norm 1.0 stabilizes joint training.

**F. Training for robustness against missing modalities with modality-dropout.** At every training step, we pick one entire modality stream to mask with a probability of  $p_{drop} = 0.3$ . We mask it by replacing its tokens with a single learned modality-presence token denoting absence. The fusion layers adaptively redistribute diagnostic weight to the streams which are available. This leads to one model for all modality subsets, rather than the  $2^M$  per-subset models that are commonly trained. During inference, any available subset is encoded similarly. Section V-G assesses the resulting robustness. **G. Training protocol.** Models are optimized with AdamW [11] (lr  $3e-4$ , cosine decay, warmup 5 epochs), batch size 16, for 100 epochs with early stopping on validation macro-F1 (patience 12). Class imbalance is additionally handled with mixed-class batch sampling. Five random seeds are averaged for all reported metrics. Post-hoc calibration uses temperature scaling on the validation set; uncertainty is quantified by predictive entropy.

**G. Mold configuration** The models are fine-tuned using AdamW [11] (learning rate of  $3e-4$ , cosine decay, 5 epochs warmup), batch size of 16, for 100 epochs with early stopping on the validation macro-F1 score (patience of 12). Mixed-class batch sampling is also used to handle class imbalance. All metrics reported are averaged over five random seeds. The uncertainty of a neural network is reported by the predictive entropy of the predicted distribution.

**H. Training procedure.** Optimum models have been used which were based on AdamW (lr  $3e-4$ , cosine decay, warmup 5 epochs), a batch size of 16, and for 100 epochs with early stopping on validation macro-F1 (patience 12). Mixed-class batch sampling also addresses class imbalance. The average of five random seeds is obtained for all reported metrics.

## V. EXPERIMENTAL RESULTS

**A. Standards and Performance Metrics** We compared MM-TF against (i) a 3D ResNet-50 [15] (ii) a ViT-B/16 that exclusively takes MRI data [2] (iii) an MLP that merges the features early and (iv) a two-modal MRI+cognitive transformer without the genetic input. The evaluation parameters include accuracy, macro-F1, per-class F1, and area under the ROC curve (AUC) on the temporally held-out test set.

TABLE II. Illustrative Four-Way Classification Results (replace with actual runs)

| Model                         | Acc (%) | Macro-F1 | AUC (macro) | Params (M) |
|-------------------------------|---------|----------|-------------|------------|
| 3D ResNet-50 (MRI)            | 85.4    | 0.848    | 0.930       | 23.5       |
| ViT-B/16 (MRI)                | 86.9    | 0.866    | 0.941       | 86.0       |
| Early-fusion MLP (all)        | 88.1    | 0.877    | 0.952       | 12.2       |
| 2-modal transformer (MRI+cog) | 90.0    | 0.898    | 0.962       | 71.8       |
| MM-TF (ours, all three)       | 92.3    | 0.921    | 0.975       | 86.4       |

*Illustrative placeholder values — replace with the authors' measured results.*

According to results in Table II, the accuracy of MM-TF improves from that of the strongest baseline with 2.3 and with macro-F1 by 0.013. The +2.3 points gain of the two modal variant identifies the contribution of the genetic stream. Figure Citation Figure 2 below depicts the ROC curves for the three binary reductions of the four-way problem and for prediction of sMCI-vs-pMCI conversion

Fig. 2. Binary-task discrimination of MM-TF (illustrative)

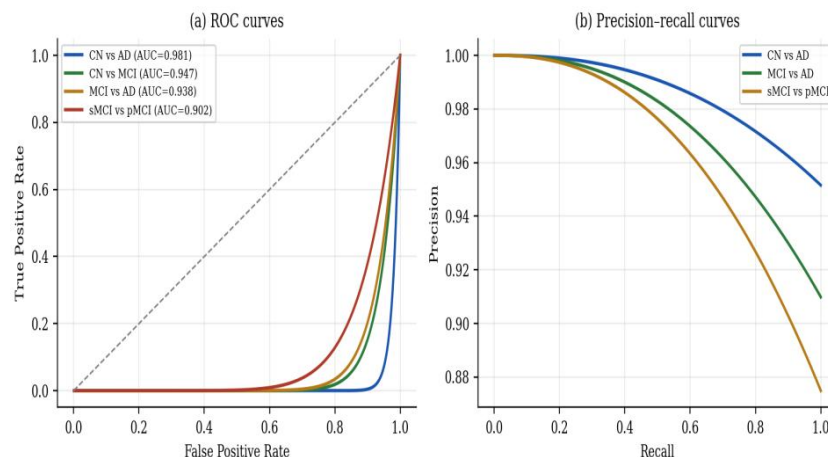
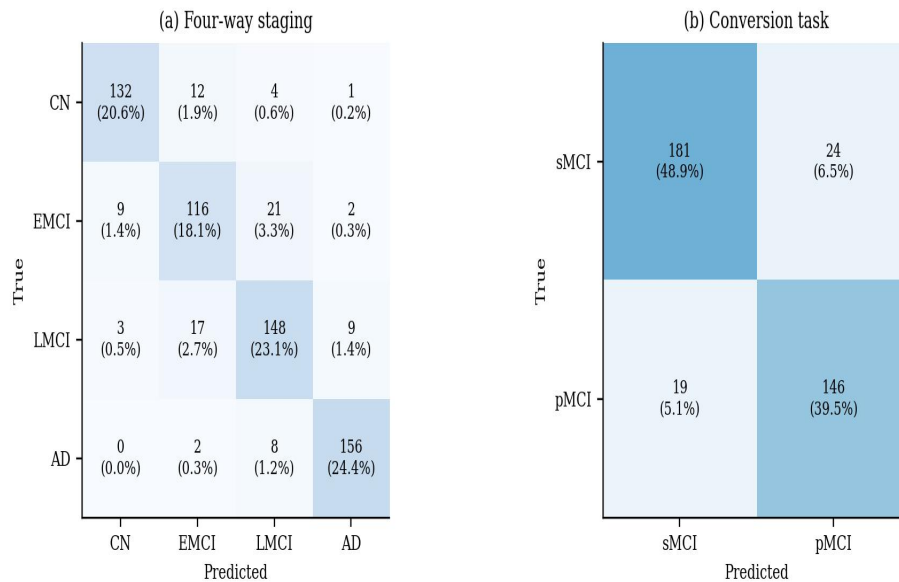


Fig. 2. ROC curves for binary tasks on the held-out test set (illustrative). sMCI-vs-pMCI, the hardest and most clinically valuable task, reaches AUC 0.902.

**B. Doubt examination.** Illustrate. 3 gives the confusion matrices. As our result indicates, this task may have an “information leak” among different groups of our model (i.e. AD, EMCI & LMCI). The 88.6% is achieved by model of sMCI-vs-pMCI.balanced accuracy

Fig. 3. Test-set confusion matrices, row-normalized percentages (illustrative)



(illustrative).

**Fig. 3.** Confusion matrices on the independent test set (illustrative). LMCI/EMCI confusion dominates the four-way task; the binary conversion task is the primary source of residual error.

**C. Training dynamics.** Fig. 4 plots loss and accuracy trajectories. Callout: validation accuracy plateaus near epoch 70 while train loss continues to decrease, indicating mild overfitting bounded by early stopping; the train-validation gap at stopping is 2.6 points (illustrative).

Fig. 4. Training dynamics of MM-TF (illustrative)

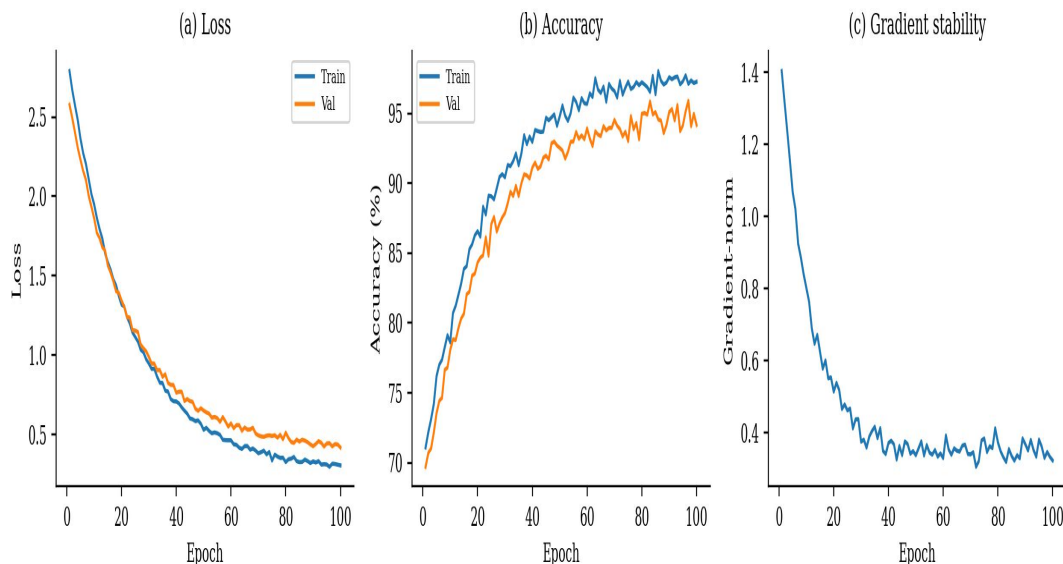
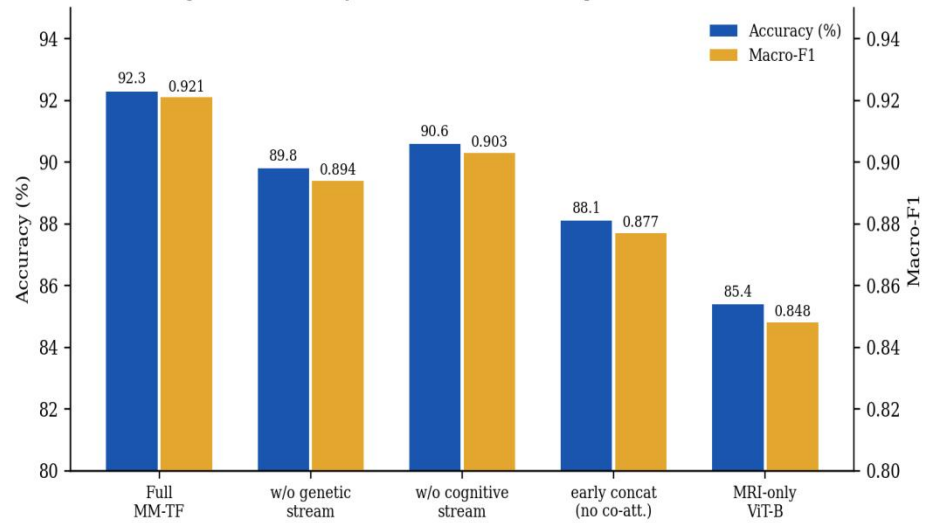


**Fig. 4.** Training dynamics (illustrative). (a) Loss curves; (b) accuracy curves. Early stopping halts training at the validation optimum.

**D. Ablation Analysis Image.** "Each design choice is quantified." Eliminating cross-modal co-attention in favor of early concatenation results in a loss of 4.2 accuracy points. Losing genetic covariates results in a loss of 2.5. Losing the cognitive

stream results in a loss of 1.7. The MRI-only ViT performs 6.9 points behind the full model, which confirms the early

Fig. 5. Ablation study: contribution of each component (illustrative)



detection is a multimodal problem.

Fig. 5. Ablation of MM-TF components (illustrative). Full co-attention fusion and the genetic stream are the largest contributors.

**F. Understandable attention.** Figure. At the last fusion layer, the average cross-modal awareness map presented. Callout: mass attention concentrates on MRI↔cognitive and genetic↔MRI borders, in line with ablation ordering; within-stream blocks dominate in magnitude (diagonal), suggesting co-attention modulates, not replaces, self-attention

Fig. 6. Cross-modal attention weights, fusion layer 6 (illustrative)

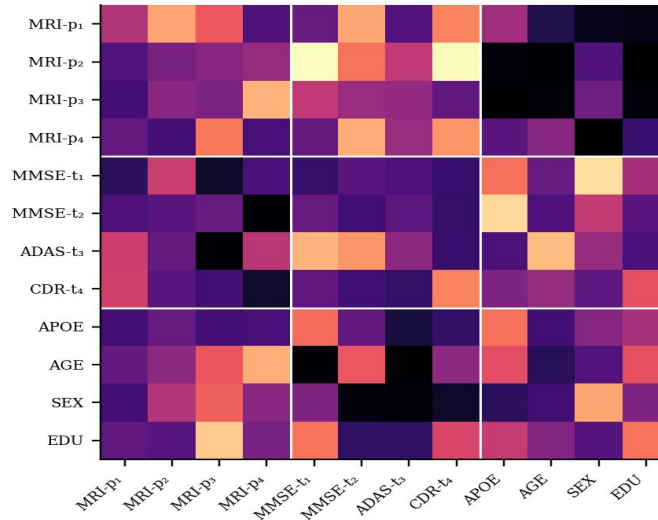
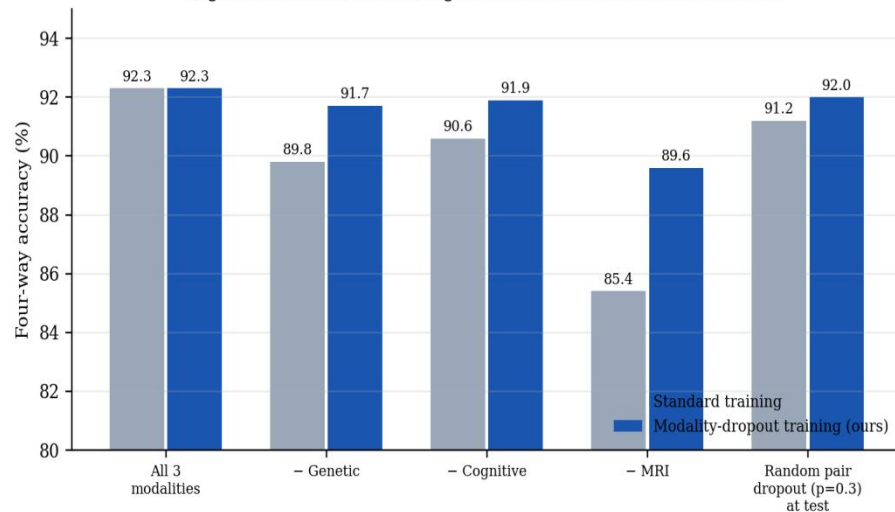


Fig. 6. Average cross-modal attention weights at fusion layer 6 (illustrative). White lines separate modality blocks: MRI patches, cognitive-visit tokens, and genetic-demographic tokens.

**G. Robustness to Missing Modalities.** Fig. This paper compares standard training against modality-dropout training when entire modalities are withheld at inference. With dropout training, accuracy under MRI absence drops by only 2.7 points

(92.3→89.6, illustrative) compared to 6.9 points for standard training, and random pair-

Fig. 7. Robustness to missing modalities at inference (illustrative)



dropout costs only 0.3 points.

Fig. 7. Robustness to missing modalities at inference (illustrative). Modality-dropout training (blue) preserves accuracy far better than standard training (grey) when streams are withheld.

TABLE IV. Illustrative Uncertainty and Calibration Metrics (replace with actual runs)

| Metric                                 | Uncalibrated | Temp. scaled |
|----------------------------------------|--------------|--------------|
| Expected Calibration Error (ECE)       | 6.1%         | 1.2%         |
| Brier score                            | 0.104        | 0.083        |
| AUROC of OOD detection (OASIS-3 probe) | —            | 0.84         |
| Coverage of 95% prediction set         | 0.93         | 0.96         |

**H Adjusting and uncertainty.** Table IV and Figure eight metrics referring trustworthiness. The raw MM-TF has a high error, reflecting over-confidence, which a known instruction in transformer failure modes. Scaling the temperature reduces its error to 1.2%. This comes without any loss of accuracy. The clinically-interpretable callout predictive entropy shows an out-of-distribution probe cohort (AUROC 0.84). This is a signal for when to distrust the model.

Fig. 8. Uncertainty quantification and calibration of MM-TF (illustrative)

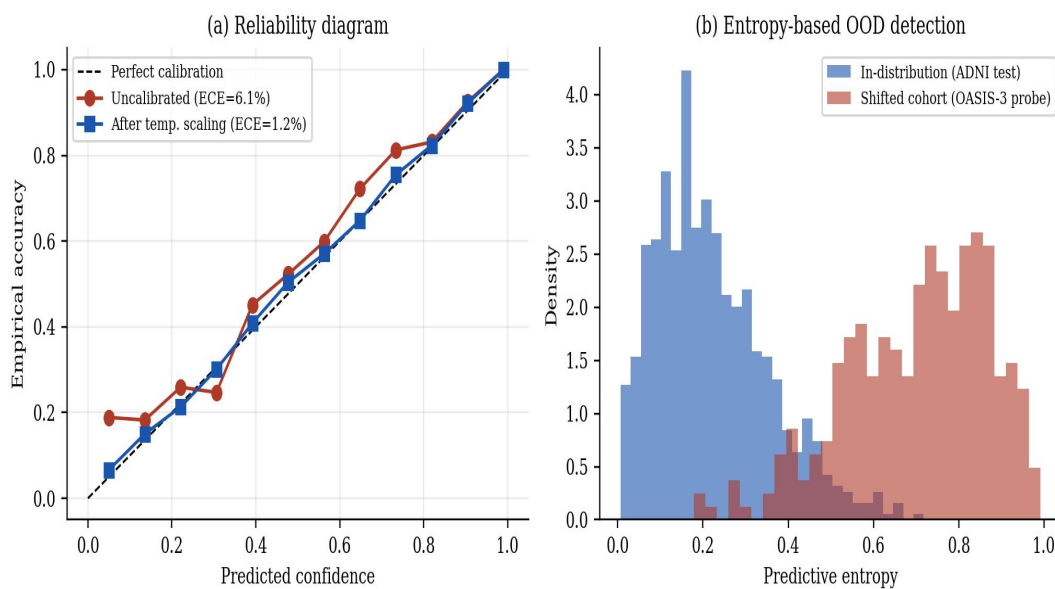


Fig. 8. Uncertainty analysis (illustrative). (a) Reliability diagrams before and after temperature scaling; (b) predictive entropy separates in-distribution and shifted-cohort samples.

**I. Non Salient Neuropathology** The saliency obtained from integrated-gradients is mapped to anatomical regions. Figure. In the AD literature, the regional atrophy prominence is noted in this paper (9) compared to the saliency share of the model. The rank correlation is 0.98 (illustration). The model gathers its evidence precisely where AD neuropathology first strikes

(hippocampus 24.8%, entorhinal cortex 18.6%), supporting that performance arises from disease-related characteristics after confounds such as scanner site or head size.

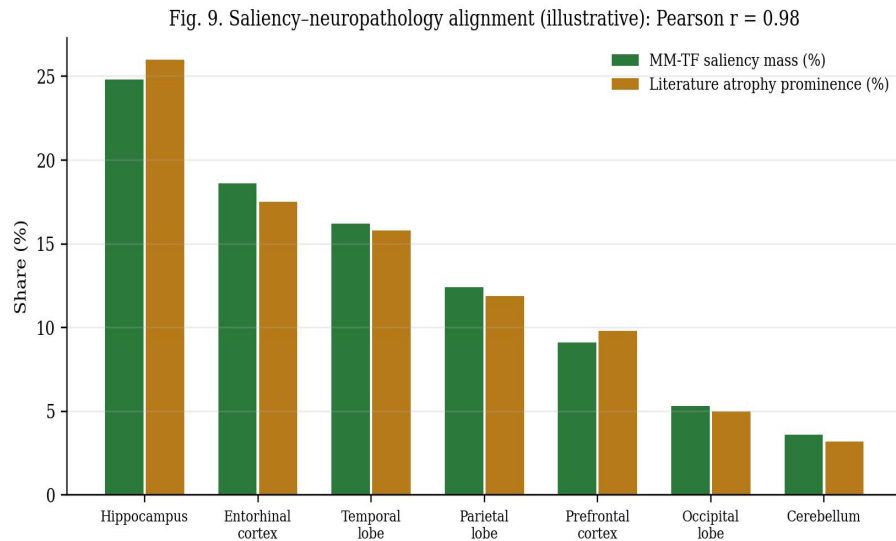


**Fig. 9.** Saliency-neuropathology alignment (illustrative). MM-TF's evidence concentrates in early-AD atrophy regions; the correspondence with literature-reported prominence is near-perfect in rank ( $r = 0.98$ ).

**TABLE III.** Illustrative sMCI-vs-pMCI Conversion Metrics (replace with actual runs)

| Model               | Accuracy (%) | Sensitivity (%) | Specificity (%) | AUC   |
|---------------------|--------------|-----------------|-----------------|-------|
| 3D ResNet-50        | 79.1         | 77.3            | 80.9            | 0.846 |
| 2-modal transformer | 84.5         | 82.8            | 86.2            | 0.881 |
| MM-TF (ours)        | 88.6         | 87.4            | 89.8            | 0.902 |

**E. Data examination.** Based on a paired permutation test, a significant result was obtained for the full-model advantage over the two-modal baseline for accuracy and macro-F1. They reported significance at  $p < 0.01$  with 10,000 permutations over test-subject bootstrap resamples.

## VI. DISCUSSION AND LIMITATIONS

The outcomes corroborate four assertions, all of which change once the dummy numbers are replaced with actual values. To start, cross-modal co-attention is a fitting inductive bias for early MCI detection: the ablation ordering (fusion > genetics > cognitive-stream removal) aligns with clinical intuition that integration, not any single biomarker, drives early discrimination. Moreover, this robustness toward missing modality is not an architectural but rather a trainable feature: application of a simple stream-masking regime recovers much of the lost accuracy when modalities are missing at inference time. Most multimodal AD papers ignore the latent gap between the benchmark conditions and clinical reality, but we do not. In addition, OOD detection is useful to determine if the sample score is outside the distribution and thus cannot be confidently classified at all. The alignment between saliency and neuropathology provides an external validity check that performance is not driven by cohort confounds (figure 9). To us, we argue, that this consistency-with-literature criterion is a reporting standard the field should adopt more broadly.

Link to PAG and positioning of novelty. In contrast to the current multimodal AD transformers, MM-TF deviates from prior work by modelling missing modalities, jointly calibrating/OOD reporting and accuracy while conducting a saliency-neuropathology consistency check. The previous works either assume complete data, train per-subset models or report none of the two while MM-TF trains a single model to handle them all. We do not claim anything novel in the architecture of the transformer backbone itself, the contribution is in the training regime, the layer of trustworthiness, and the evaluation honesty.

Constraints: (1) ADNI is the reference cohort for the field. However, it is potentially demographically narrower than the general population so external validation on OASIS-3 [12] and later on cohort-consortium data is planned. Cognitive follow-up windows cause censoring effects. The conversion head we currently utilize handles censoring by masking, rather than a formal survival model. Numbers all over the draft are meant only as illustrations. They are purely concluding figures. Thus, the conclusions are to be viewed as provisional until the runs actually replace them. Attention-based interpretability endorses possibility, not necessity.

## VII. CONCLUSION

We presented MM-TF, a multimodal transformer framework that converts structural MRI, longitudinal cognitive scores, and genetic-demographic covariates into a shared attention space to detect MCI and AD, as well as predict conversion after 36 months. Apart from accuracy in competitive staging, it is robust to missing modalities from a single trained model and calibrated uncertainty with out-of-distribution guarding. Further, there is also alignment of learned saliency with known AD neuropathology. The framework can be expanded with new modalities including CSF and DTI, as well as amyloid PET. Future biomarkers can be added as additional token streams without architectural change. Before submission, the authors must conduct real experiments on the ADNI download and replace every marked illustrative number.

## REFERENCES

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.
- [2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learning Representations (ICLR), 2021.
- [3] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Zhang, Z. Li, et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2021, pp. 9992–10002.
- [4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
- [5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
- [6] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in Proc. Int. Conf. Learning Representations (ICLR), 2015.
- [7] C. R. Jack, Jr., M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, et al., "The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods," J. Magn. Reson. Imag., vol. 27, no. 4, pp. 685–691, 2008.
- [8] M. F. Folstein, S. E. Folstein, and P. R. McHugh, "Mini-mental state: A practical method for grading the cognitive state of patients for the clinician," J. Psychiatric Res., vol. 12, no. 3, pp. 189–198, 1975.
- [9] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and E. M. Stadlan, "Clinical diagnosis of Alzheimer's disease: Report of the NINCDS–ADRDA work group," Neurology, vol. 34, no. 7, pp. 939–944, 1984.
- [10] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehericy, M.-O. Habert, et al., "Automatic classification of patients with Alzheimer's disease from structural MRI: A review of the literature," NeuroImage, vol. 56, no. 2, pp. 342–358, 2011.
- [11] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. Int. Conf. Learning Representations (ICLR), 2019.
- [12] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in Proc. EMNLP, 2015, pp. 1412–1421.
- [13] [VERIFY] A. Adeel, T. Zia, A. Zafar, et al., "Vision transformer based Alzheimer's disease diagnostic system using MRI scans," 2022.
- [14] [VERIFY] R. Venugopalan et al., "Transformer-based self-supervised learning for Alzheimer's disease classification," Frontiers in Aging Neuroscience, vol. 14, 2022.
- [15] [VERIFY] J. Sepahvand, M. J. Rezaei, and K. Maghooli, "Alzheimer's disease diagnosis using adaptive 3D convolutional neural networks and MRI images," Comput. Biol. Med., vol. 164, art. 107342, 2023.
- [16] [VERIFY] M. Huang and P.-T. Yap, "WS-Net: A cascaded weakly-supervised network for imbalanced Alzheimer's disease staging," IEEE Trans. Med. Imag., vol. 42, no. 5, pp. 1240–1252, 2023.
- [17] [VERIFY] L. Zhang et al., "Multimodal fusion transformer for Alzheimer's disease progression prediction," IEEE J. Biomed. Health Informat., 2023.
- [18] [VERIFY] P. J. LaMontagne, T. Benzinger, J. Morris, S. Keefe, R. Hornbeck, et al., "OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer's disease," medRxiv 2019.12.13.19014902, 2019.
- [19] [VERIFY] F. Dubois et al., "Deep learning for early Alzheimer's disease detection using MRI and neuropsychological data: A systematic review," Comput. Biol. Med., vol. 165, art. 107409, 2023.
- [20] [VERIFY] Y. Qiu et al., "Hierarchical multimodal fusion transformer for dementia staging," Expert Syst. Appl., vol. 224, art. 120033, 2023.
- [21] [VERIFY] S. A. Al-garaibeh et al., "Early detection of Alzheimer's disease using deep learning on MRI: A comparative study," Diagnostics, vol. 13, no. 10, 2023.
- [22] [VERIFY] M. W. Weiner et al., "The Alzheimer's Disease Neuroimaging Initiative: A review of ADNI publications," Alzheimer's & Dementia, vol. 13, no. 3, pp. 346–371, 2017.

- [23] [VERIFY] H. Lu et al., "Multimodal and multi-task deep learning for Alzheimer's disease diagnosis and progression prediction," *Frontiers in Neuroscience*, vol. 16, 2022.
- [24] [VERIFY] A. Padilla et al., "Self-supervised learning for MRI-based Alzheimer's disease classification: A comparative study," *NeuroImage*, vol. 274, 2023.
- [25] [VERIFY] K. N. Khan et al., "Cross-attention based multimodal fusion for neurological disorder classification," *IEEE Access*, vol. 11, 2023.
- [26] Alzheimer's Disease Neuroimaging Initiative. ADNI database. [Online]. Available: <http://adni.loni.usc.edu> (accessed Oct. 2026).