

M²PFN: End-to-End Disentangled Alignment for Generalizable Multimodal In-Context Learning in Alzheimer’s Disease

Lujia Zhong^{1,2}, Shuo Huang^{1,3}, Jianwei Zhang^{1,2}, Xinyu Nie¹, and Yonggang Shi^{1,2,3*}

¹Stevens Neuroimaging and Informatics Institute, Keck School of Medicine,
University of Southern California, Los Angeles, CA 90033, USA

²Ming Hsieh Department of Electrical and Computer Engineering, Viterbi School of Engineering,
University of Southern California, Los Angeles, CA 90089, USA

³Alfred E. Mann Department of Biomedical Engineering, Viterbi School of Engineering,
University of Southern California, Los Angeles, CA 90089, USA

Abstract

While various multimodal methods combining imaging and tabular data for Alzheimer’s disease (AD) diagnosis were proposed, they are often limited in generalization across cohorts. In-context learning (ICL) has demonstrated excellent generalization performances and high flexibility in foundational tabular models such as TabPFN. To extend TabPFN’s ICL to multimodal AD analysis, the main obstacle is that TabPFN is meta-trained on synthetic tabular priors that do not naturally match the statistical structure of image-derived features. We propose M²PFN, an end-to-end framework that turns this tabular foundation model into a multimodal AD predictor. M²PFN (i) performs differentiable inference through TabPFN’s transformer, back-propagating task gradients into 3D-MRI and tabular encoders; (ii) aligns the two modalities into a shared subspace, via disentanglement and a contrastive objective, matched to the ICL engine’s prior; and (iii) folds in a frozen tabular-only prediction through a learnable gated shortcut. Because the ICL engine stays frozen, its in-context mechanism is preserved for test-time generalization, while end-to-end training shapes the encoders into features it can exploit. On ADNI ($n = 2240$, three-class CN/MCI/AD), M²PFN attains 65.55% macro-F1 and 82.21% macro-AUC, surpassing a comprehensive set of unimodal and multimodal baselines. By swapping only the head for a TabPFN regressor, the same architecture regresses baseline MMSE on a 1250-subject sub-cohort to test MAE 1.743, outperforming every multimodal baseline. On two external cohorts (OASIS-3 and SCAN) with no retraining, M²PFN achieves the best AUC and the lowest MMSE MAE across all baselines, and transfers even when the cognitive instrument changes.

Keywords: Alzheimer’s disease, in-context learning, multimodal fusion, prior-data fitted networks, contrastive learning, MRI.

1 Introduction

For Alzheimer’s disease (AD) diagnosis and prognosis, the value of multimodal analysis that combines complimentary information from imaging and non-imaging tabular data has been well demonstrated [1–7]. Among imaging modalities, structural MRI is the most widely available and routinely acquired in clinical practice. It characterizes neurodegeneration through cortical and hippocampal

*Corresponding author: Yonggang Shi (yshi@loni.usc.edu).

atrophy, which are among the most reliable *in vivo* markers of cognitive decline. Non-imaging tabular data, including demographic characteristics, clinical history, genetic factors such as APOE genotype, and fluid biomarkers such as CSF amyloid and tau, provide additional information about disease risk and progression that imaging alone cannot reveal. Despite their benefits, most existing multimodal approaches combine imaging and tabular data through task- and dataset-specific fusion architectures with fixed input structures. Consequently, they typically require retraining or extensive fine-tuning when transferred across cohorts with different imaging protocols, scanner vendors, and population characteristics, and struggle to accommodate variation in the number and type of available tabular features, ranging from more than one hundred variables in Alzheimer’s Disease Neuroimaging Initiative (ADNI) [8,9] to fewer than ten in some external cohorts.

These challenges naturally motivate foundation models, which leverage large-scale pretraining to learn transferable representations that can be adapted across downstream tasks and domains. Such capability is particularly desirable for AD analysis, where both imaging distributions and multimodal feature availability vary considerably across clinical cohorts. However, most existing AD prediction models, whether based on convolutional neural networks [10–12], transformers [13–16], or mixture-of-experts architectures [3,4,17–19], remain task-specific, with model parameters optimized for the distributions observed during training. Consequently, they provide limited support for flexible multimodal integration and robust cross-cohort deployment.

In-context learning (ICL) offers a principled answer to this foundation-model demand. Rather than compressing data into fixed parameters, a pretrained ICL model conditions on a labeled support set supplied at inference time and predicts a query directly through attention over that support, without updating a single weight. Pioneered by large language models [20], ICL has repeatedly shown strong generalization and rapid adaptation to new tasks and distributions purely by changing the support—a capability structurally unavailable to parametric predictors. For tabular data, this paradigm has already yielded a family of foundation models: prior-data fitted networks (PFNs) such as TabPFN [21] and TabFM [22], meta-trained on large families of synthetic priors, deliver strong, well-calibrated predictions on medium to small-sized datasets typical of clinical cohorts, all off the shelf and with the network frozen. Two obstacles, however, keep this tabular foundation model from multimodal AD analysis. First, no ICL foundation model natively ingests medical imaging or, more generally, multiple modalities; a tabular-only ICL engine therefore leaves most of the AD signal unused. Second, TabPFN is meta-trained on large families of synthetic tabular priors sampled from structural causal models; its learned in-context prior therefore reflects the statistical structure of tabular data-generating processes, not that of image-derived features. Naively extracting image embeddings and feeding them as additional tabular columns yields a representation mismatched to what the engine was trained to exploit, making end-to-end training essential to shape the encoder’s output to the engine’s prior. However, TabPFN’s preprocessing pipeline contains non-differentiable operations that fracture gradient flow, so it cannot be coupled to and jointly trained with the learnable encoder this adaptation demands.

We propose M²PFN, an end-to-end framework that repurposes a frozen, off-the-shelf tabular foundation model for multimodal Alzheimer’s disease analysis. M²PFN preserves the frozen TabPFN in-context learning (ICL) engine as the prediction head and learns multimodal representations that are directly optimized for in-context reasoning through end-to-end training. By keeping the ICL engine frozen, M²PFN supports support-set replacement at inference, enabling gradient-free transfer to unseen clinical cohorts without retraining.

Contributions. Our main contributions are summarized as follows:

- We instantiate differentiable prompt tuning through TabPFN by bypassing its non-differentiable preprocessing pipeline and back-propagating gradients through the in-context attention com-

putation.¹

- We introduce a disentangled alignment-based multimodal representation learning strategy that jointly optimizes MRI and tabular features through the frozen ICL engine.
- We achieve state-of-the-art performance on ADNI for both three-class CN/MCI/AD classification and MMSE regression.
- We demonstrate robust zero-retraining transfer to two independent external cohorts (OASIS-3 and SCAN) via support-set replacement, achieving the best performance among all gradient-free adaptation baselines.

2 Related Work

2.1 Multimodal Alzheimer’s Disease Analysis

AD analysis spans two paradigms: categorical classification (e.g., CN/MCI/AD) and regression of continuous cognitive scores such as MoCA and MMSE. Deep learning for both has progressed from unimodal architectures, such as ResNet [10], DenseNet [11], and transformer-based SwinUNETR [14], to multimodal fusion architectures that combine imaging with clinical/tabular features, spanning cross-modal transformers (MulT [16], mmFormer [15]) and mixture-of-experts (MoE) routing across modalities (Flex-MoE [3], MoE++ [18], SwitchGate [17], I2MoE [4], InterpretCC [19]); these consistently show that combining MRI with tabular/clinical data outperforms either modality alone. All of these methods stay within the parametric learning paradigm, compressing the training distribution into fixed weights and cannot reference labeled exemplars at inference or effectively adapt to a new cohort without retraining or finetuning.

2.2 In-Context Learning

In-context learning (ICL) was popularized by large language models, which solve new tasks by conditioning on demonstrations without parameter updates [20], and has since been extended to vision-language models [23]. For tabular data, two main paradigms have emerged: retrieval-based methods that attend to training examples during inference [24, 25], and Prior-Data Fitted Networks (PFNs) [26], which meta-learn Bayesian inference over synthetic task distributions. TabPFN [21, 27] is the most mature PFN-based model, achieving competitive performance on small-to-medium tabular datasets without test-time optimization, while TabFM [22] further scales this paradigm to larger tabular benchmarks. We build upon TabPFN. However, like existing tabular foundation models, it is designed exclusively for structured tabular inputs and relies on a non-differentiable preprocessing pipeline before the transformer. Consequently, multimodal inputs such as 3D MRI cannot be incorporated into the in-context reasoning process, nor can gradients propagate from the ICL engine back to an imaging encoder for end-to-end representation learning.

2.3 Prompt Tuning for ICL Engines

When a large pretrained model cannot be finetuned directly, being expensive to update or easy to collapse on a small dataset, prompt tuning instead optimizes a small set of continuous input

¹The differentiable-input path for TabPFN’s regressor used in this work has been contributed by the authors to the official TabPFN library and merged upstream (<https://github.com/PriorLabs/TabPFN/pull/923>).

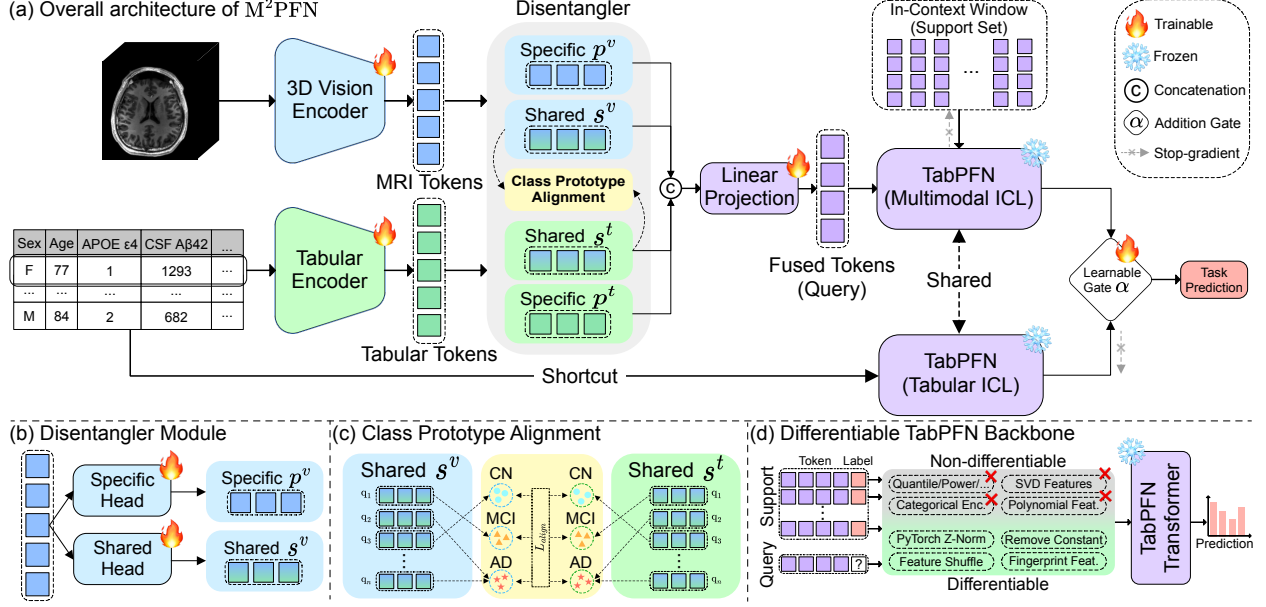


Figure 1: Overview of the proposed M²PFN framework (§3). (a) Overall architecture. (b) Disentangler module for separating modality-specific and shared representations. (c) Class prototype alignment for encouraging cross-modal consistency. (d) Differentiable TabPFN backbone enabling end-to-end gradient propagation through the frozen PFN engine.

vectors to steer the frozen model, spanning soft prompts at the input layer [28], deep/prefix variants that inject tunable vectors at every layer [29, 30], and parameter-efficient adapters such as LoRA [31] that learn a low-rank update to the frozen weights. Prompt tuning a PFN, however, needs a gradient path from the engine’s output back to its input, which its non-differentiable pre-processing stack does not provide. A concurrent multimodal PFN, MMPFN [32], works around this architecturally, inserting a multimodal mixer (a multi-head gated MLP with cross-attention pooling) between TabPFN’s feature encoder and its transformer and lightly finetuning that mixer on frozen, precomputed image embeddings; this avoids re-extracting image features during training but prevents end-to-end adaptation of the imaging encoder. We take the opposite stance on what to adapt: the engine’s weights stay frozen, which prevents a foundation model from collapsing on the small data of a clinical cohort, and task gradients instead flow through TabPFN’s transformer into the imaging and tabular encoders, aligning their representations with the inductive biases of TabPFN.

3 Method

An overview of M²PFN is shown in Fig. 1. We first formulate multimodal in-context prediction (Sec. 3.1), then describe the alignment-based representation learner (Sec. 3.2), the differentiable TabPFN prediction head (Sec. 3.3), and the gated tabular shortcut (Sec. 3.4). The complete training objective is given in Sec. 3.5.

3.1 Problem Formulation

Subject i is represented by $\mathbf{x}_i = (x_i^{\text{img}}, x_i^{\text{tab}})$, where $x_i^{\text{img}} \in \mathbb{R}^{1 \times 128 \times 128 \times 128}$ is a 3D MRI volume and $x_i^{\text{tab}} \in \mathbb{R}^d$ contains continuous and ordinal-encoded categorical variables. The task target is denoted by y_i : for diagnosis, $y_i \in \{0, 1, 2\}$ corresponds to $\{\text{CN}, \text{MCI}, \text{AD}\}$; for cognitive-score regression, $y_i \in \mathbb{R}$ (§4.7). We additionally denote the diagnostic category by $c_i \in \{0, 1, 2\}$ for the prototype-alignment loss below. Thus, $c_i = y_i$ for classification, whereas c_i is an auxiliary diagnosis label for regression. Given a query subject q and a labeled support set $\mathcal{D}_S = \{(\mathbf{x}_j, y_j)\}_{j \in \mathcal{S}}$, both tasks are formulated as

$$p_{\Theta_{\text{PFN}}}(y_q \mid \tilde{h}_q, \{(\tilde{h}_j, y_j)\}_{j \in \mathcal{S}}), \quad (1)$$

where $\tilde{h}_i \in \mathbb{R}^{D'}$ is the learned multimodal representation supplied to TabPFN, defined in Sec. 3.3, and Θ_{PFN} denotes the pretrained TabPFN parameters. For classification, $\hat{y}_q = \arg \max_c p(y_q = c \mid \cdot)$; for regression, $\hat{y}_q$ is the posterior mean of TabPFN’s discretized target distribution. Because the support set enters only as context, $\mathcal{D}_S$ can be replaced at inference without updating the model parameters, enabling the external support-swap evaluation in §4.8. Our inference path is differentiable with respect to $\tilde{h}_q$ and hence the query inputs. The PFN weights Θ_{PFN} remain frozen, while gradients optimize the modality encoders, disentangler, projection, and fusion gate.

3.2 Alignment-Based Multimodal Representation Learning

Per-modality encoders. The visual encoder $E_v(\cdot; \theta_v)$ maps x_i^{img} to $u_i^v \in \mathbb{R}^H$, and the tabular encoder $E_t(\cdot; \theta_t)$ maps x_i^{tab} to $u_i^t \in \mathbb{R}^H$, with $H = 768$ and superscripts v and t denoting visual and tabular modalities. Specifically, E_v is a MONAI [33] 3D ResNet-18 [10] initialized from MedicalNet [34]. Global average pooling produces a 512-dimensional descriptor, followed by Linear+LayerNorm+GELU to obtain u_i^v . The FT-Transformer [35] used for E_t embeds each continuous or categorical variable as a token, prepends a learnable [CLS] token, and applies three width- H transformer blocks with eight attention heads. Its [CLS] output is mapped by Linear+RMSNorm to u_i^t .

Shared/specific disentanglement. For each modality $m \in \{v, t\}$, two parallel heads transform u_i^m into a shared representation $s_i^m = S_m(u_i^m) \in \mathbb{R}^{D_s}$ and a modality-specific representation $p_i^m = P_m(u_i^m) \in \mathbb{R}^{D_p}$. Each head comprises Linear+LayerNorm+GELU, with $D_s = D_p = 576$ (Fig. 1(b)); their parameters are collectively denoted by ϕ . The heads have the same architecture but separate parameters. Their functional roles are induced by the alignment objective: only s_i^v and s_i^t are directly aligned, whereas p_i^v and p_i^t are optimized only through the downstream task loss.

Class-prototype alignment loss. Instead of aligning individual MRI–tabular pairs with an InfoNCE objective [36, 37], we align class prototypes within the current query minibatch $\mathcal{B}$ (Fig. 1(c)). Let $\mathcal{B}_c = \{i \in \mathcal{B} : c_i = c\}$ and $\mathcal{C}_B = \{c : \mathcal{B}_c \neq \emptyset\}$. The modality- m prototype for category c is

$$\bar{s}_c^m = \frac{1}{|\mathcal{B}_c|} \sum_{i \in \mathcal{B}_c} s_i^m, \quad (2)$$

and the alignment loss is

$$\mathcal{L}_{\text{align}} = \frac{1}{|\mathcal{C}_B|} \sum_{c \in \mathcal{C}_B} \left[1 - \cos(\bar{s}_c^v, \bar{s}_c^t) \right], \quad (3)$$

where $\cos(a, b) = a^\top b / (\|a\|_2 \|b\|_2)$. The number of alignment targets is therefore bounded by the number of diagnostic categories rather than the number of subjects, limiting memorization of

individual cross-modal pairs. Minimizing Eq. (3) encourages the shared heads to encode diagnosis-consistent cross-modal information, while the specific heads remain free to retain complementary task-relevant information.

Concatenated representation. Following the ordering in Fig. 1(a), the modality-specific and shared representations are concatenated as

$$h_i^C = [p_i^v; s_i^v; s_i^t; p_i^t] \in \mathbb{R}^{2(D_s+D_p)} = \mathbb{R}^{2304}, \quad (4)$$

where the superscript C denotes concatenation.

3.3 Differentiable ICL Classification and Regression

A learnable Linear+LayerNorm projection

$$\text{Proj}_\psi : \mathbb{R}^{2(D_s+D_p)} \rightarrow \mathbb{R}^{D'}$$

maps the concatenated representation to a compact fused representation:

$$h_i^{\text{fused}} = \text{Proj}_\psi(h_i^C), \quad D' = 512. \quad (5)$$

Given the fused representation h_q^{fused} of query subject q and a cached support context, the multi-modal ICL branch computes

$$z_{\text{mm},q} = \text{TabPFN}\left(h_q^{\text{fused}}, \left\{(\text{sg}(h_j^{\text{fused}}), y_j)\right\}_{j \in \mathcal{S}}; \Theta_{\text{PFN}}\right), \quad (6)$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation applied to the support representations, as illustrated in Fig. 1(a), and Θ_{PFN} denotes the frozen TabPFN parameters. For classification, $z_{\text{mm},q} \in \Delta^{K-1}$ is a probability vector over $K = 3$ diagnostic classes. For regression, the TabPFN classifier is replaced by a TabPFN regressor, and $z_{\text{mm},q} \in \mathbb{R}$ denotes the posterior mean of its discretized predictive distribution in standardized target space. The projection is optimized jointly with the modality encoders and disentanglement heads, allowing the task loss to retain the feature directions most useful to the frozen ICL engine.

Differentiable inference. We bypass TabPFN’s non-differentiable feature preprocessing and retain differentiable PyTorch z-normalization before the transformer, as illustrated in Fig. 1(d). The transformer forward pass remains differentiable even though Θ_{PFN} is frozen. Consequently, the task gradient follows $z_{\text{mm},q} \rightarrow h_q^{\text{fused}} \rightarrow h_q^C \rightarrow \{u_q^v, u_q^t\}$, and updates the trainable front end. No gradient is propagated through the detached support cache. Freezing Θ_{PFN} preserves its meta-learned in-context prior, while adapting its input representation to multimodal AD data, analogous to continuous prompt tuning.

Per-epoch support refresh. At epoch e , the training set is randomly partitioned into a support subset $\mathcal{S}^{(e)}$ and a disjoint query subset $\mathcal{Q}^{(e)}$. Support subjects are encoded once using the current front-end parameters, detached, and reused for all query updates in that epoch. For classification, 1,400 of 1,556 training subjects form the support, leaving 156 queries processed in ten batches of at most 16. For MMSE regression, 859 of 875 subjects form the support, and the remaining 16 queries constitute one ICL episode and one optimizer update. Support caching avoids repeated 3D MRI encoding, although the cached representations become slightly stale as the front end is updated within an epoch.

3.4 Gated TabPFN Shortcut

The multimodal branch uses both MRI and tabular representations. As shown in Fig. 1(a), we add a frozen tabular-only TabPFN shortcut and blend its output with z_{mm} through a learnable scalar gate. The shortcut provides a calibrated tabular reference, while the multimodal branch learns complementary information from MRI and tabular features.

TabPFN shortcut. Let $z_{\text{tab},i} \in \Delta^{C-1}$ denote the frozen tabular-only TabPFN prediction for classification. For validation and test subjects, its context is a random subset of the training data (1,400 of 1,556 subjects). For training subjects, $z_{\text{tab},i}$ is precomputed by stratified K -fold out-of-fold inference with $K = 5$: each held-out fold is predicted using the other $K - 1$ folds as support, so the subject being predicted never appears in its own context. The cache $\{z_{\text{tab},i}\}_{i \in \mathcal{D}_{\text{train}}}$ is generated once per random seed and retrieved by subject index during training. For regression, $z_{\text{tab},i} \in \mathbb{R}$ is the corresponding posterior mean in standardized target space.

Learnable gating. For each subject, the two branches are combined by

$$z_{\text{final},i} = \alpha z_{\text{mm},i} + (1 - \alpha) z_{\text{tab},i}, \quad \alpha = \sigma(\rho), \quad (7)$$

where $\rho \in \mathbb{R}$ is trainable, σ is the sigmoid, and ρ is initialized to zero so that $\alpha = 0.5$. For classification, the three z variables are probability vectors; for regression, they are standardized scalars.

Because the shortcut prediction is frozen and detached,

$$\frac{\partial z_{\text{final},i}}{\partial h_i^{\text{fused}}} = \alpha \frac{\partial z_{\text{mm},i}}{\partial h_i^{\text{fused}}}. \quad (8)$$

Thus, representation gradients pass only through the multimodal branch, whereas ρ is updated by the final task loss and learns the relative contribution of the two predictions.

3.5 Training Objective

Let $\Omega = \{\theta_v, \theta_t, \phi, \psi, \rho\}$ collect the trainable parameters of the two encoders, four disentangler heads, projection, and gate, respectively. For a query minibatch $\mathcal{B}$, we optimize

$$\mathcal{L}(\Omega) = \mathcal{L}_{\text{task}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}}, \quad (9)$$

where

$$\mathcal{L}_{\text{task}} = \begin{cases} -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log[z_{\text{final},i}]_{y_i}, & \text{classification,} \\ \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \text{SmoothL1}_{\beta}(z_{\text{final},i} - y_i^z), & \text{regression,} \end{cases} \quad (10)$$

and $y_i^z = (y_i - \mu_y)/\sigma_y$ is the regression target standardized using training-set statistics. For residual e ,

$$\text{SmoothL1}_{\beta}(e) = \begin{cases} \frac{e^2}{2\beta}, & |e| < \beta, \\ |e| - \frac{\beta}{2}, & \text{otherwise.} \end{cases} \quad (11)$$

All TabPFN parameters Θ_{PFN} remain fixed. We use $\lambda_{\text{align}} = 0.1$ and $\beta = 1$ unless stated otherwise.

4 Experiments

4.1 Dataset and Preprocessing

Experiments are conducted on three public datasets: the Alzheimer’s Disease Neuroimaging Initiative (ADNI) [8,9], the Standardized Centralized Alzheimer’s & Related Dementias Neuroimaging (SCAN) initiative [38], and the Open Access Series of Imaging Studies 3 (OASIS-3) [39]. All T1-weighted MRI volumes are processed with FreeSurfer [40], and the resulting skull-stripped brain images are resampled to $128 \times 128 \times 128$. The accompanying tabular modality comprises 172 features spanning medical history, neurological examination (excluding cognitive tests), demographics, vital signs, APOE genotype, and CSF biomarkers.

The three cohorts differ in tabular completeness. ADNI provides the full 172-feature set, whereas OASIS-3 and SCAN provide only 7 and 27 features, respectively. The seven features shared by both cohorts are sex, age, education, handedness, ethnicity, and the two APOE alleles; SCAN additionally supplies race and marital status, three vital signs, nine per-system medical-history indicators, three neurological-exam signs, and CSF amyloid and tau values derived from a coarse two-level status. Available features are coded into their corresponding ADNI slots, and how a missing value is filled depends on the feature type rather than on the cohort: continuous features are mean-imputed, and categorical features are assigned a reserved index whose embedding is learned with the rest of the model. No architectural change is needed, so the same frozen network consumes all three cohorts.

A subject enters the classification cohort if a FreeSurfer-reconstructed baseline T1 and a baseline diagnosis that maps onto CN/MCI/AD are both available. In ADNI this retains the 2240 subjects that have both, out of 2962 with a baseline diagnosis. In OASIS-3, the label is the clinician diagnosis recorded at the baseline Uniform Data Set visit, which leaves 1163 of 1195 subjects; the 32 dropped either carry no diagnosis or carry a non-AD dementia with no counterpart in a three-way CN/MCI/AD task. In SCAN, the label is the NACC Uniform Data Set diagnosis [41], and the 34 of 1151 subjects coded impaired-but-not-MCI are dropped for the same reason, leaving 1117. The regression cohort is the subset of the classification cohort carrying a valid baseline cognitive test score for ADNI (1250 of 2240) and SCAN (1097 of 1117), whereas for OASIS-3 it comprises all 1195 subjects with an MMSE score, including the 32 that the three-way diagnosis excludes. The target is the Mini-Mental State Examination (MMSE) for ADNI and OASIS-3 and the Montreal Cognitive Assessment (MoCA) for SCAN. ADNI is partitioned into subject-disjoint training, validation, and test sets following a 70/15/15 split (1556/333/351 subjects for classification, 875/185/190 for regression). All models are trained exclusively on ADNI; OASIS-3 and SCAN are held out and used solely as external test sets. Detailed cohort statistics are reported in Table 1, where rows describe the classification cohort.

4.2 Implementation Details

Our model uses the TabPFN-v2.5 checkpoint [21] for both the multimodal in-context head and the tabular shortcut in M²PFN. Per-layer activation recomputation is enabled within the TabPFN backbone to improve memory efficiency. For classification, we use a learning rate of 5×10^{-5} , weight decay 10^{-4} , batch size 16, 175 epochs, and a linear warmup ratio of 0.1. The regression instantiation (§4.7) adopts the same hyperparameter configuration, except that it is trained for 1200 epochs with a batch size of 1. Both instantiations are trained and evaluated with 5 random seeds, and we report the mean $\pm$ standard deviation across seeds; a single seed requires approximately 5.5 hours for classification and 23 hours for regression. Statistical comparisons are performed with a paired one-sided Student’s *t*-test on the per-seed metrics. We report Accuracy, macro F1, macro

Table 1: Characteristics of the three study cohorts, broken down by diagnostic group. Age is mean $\pm$ std; Sex is shown in female/male; APOE $\epsilon 4^+$ is carriers over genotyped subjects.

	n	Age (Year)	Sex (F/M)	APOE $\epsilon 4^+$
ADNI				
CN	1000	71.9 $\pm$ 7.7	616/384	223/726
MCI	770	73.0 $\pm$ 7.9	354/416	258/610
AD	470	74.8 $\pm$ 7.9	217/253	279/413
All	2240	72.9 $\pm$ 7.9	1187/1053	760/1749
OASIS-3				
CN	855	67.0 $\pm$ 9.2	497/358	299/850
MCI	95	71.8 $\pm$ 6.2	49/46	41/94
AD	213	74.0 $\pm$ 7.8	93/120	133/211
All	1163	68.7 $\pm$ 9.2	639/524	473/1155
SCAN				
CN	756	71.9 $\pm$ 8.3	502/254	247/709
MCI	180	74.2 $\pm$ 7.0	87/93	70/152
AD	181	72.8 $\pm$ 8.5	99/82	88/140
All	1117	72.5 $\pm$ 8.2	688/429	405/1001

Table 2: Performance on the ADNI test set. Img/Tab indicates which modalities the method consumes.

Method	Img	Tab	Acc. (%)	F1 (%)	AUC (%)	Sens. (%)	Spec. (%)
<i>Unimodal Baselines</i>							
ResNet [10]	✓		46.91 $\pm$ 3.90	39.27 $\pm$ 3.17	69.70 $\pm$ 2.17	46.55 $\pm$ 3.22	71.96 $\pm$ 2.37
DenseNet [11]	✓		46.06 $\pm$ 4.89	39.43 $\pm$ 4.33	68.55 $\pm$ 2.22	45.39 $\pm$ 3.85	72.20 $\pm$ 3.54
SE-ResNet [12]	✓		46.91 $\pm$ 4.78	43.27 $\pm$ 3.51	62.86 $\pm$ 1.04	44.34 $\pm$ 0.50	71.31 $\pm$ 0.94
SwinUNETR [14]	✓		48.53 $\pm$ 1.65	42.32 $\pm$ 1.34	64.77 $\pm$ 0.96	43.29 $\pm$ 2.85	71.89 $\pm$ 0.51
XGBoost [42]		✓	57.68 $\pm$ 0.26	54.51 $\pm$ 0.39	74.83 $\pm$ 0.23	53.90 $\pm$ 0.36	77.40 $\pm$ 0.09
AutoGluon [43]		✓	60.49 $\pm$ 1.52	58.01 $\pm$ 0.70	77.77 $\pm$ 0.18	57.96 $\pm$ 1.05	79.01 $\pm$ 0.62
TabPFN-v2.5 [21]		✓	61.57 $\pm$ 0.39	59.19 $\pm$ 0.25	78.84 $\pm$ 0.04	58.64 $\pm$ 0.20	78.20 $\pm$ 0.20
<i>Multimodal Baselines</i>							
InterpretCC [19]	✓	✓	61.88 $\pm$ 1.77	61.51 $\pm$ 2.37	80.52 $\pm$ 0.32	63.65 $\pm$ 2.67	80.44 $\pm$ 1.14
MoE++ [18]	✓	✓	60.23 $\pm$ 2.10	60.32 $\pm$ 1.84	79.95 $\pm$ 0.70	61.76 $\pm$ 1.29	79.61 $\pm$ 0.82
SwitchGate [17]	✓	✓	60.23 $\pm$ 2.41	59.43 $\pm$ 3.33	79.71 $\pm$ 0.74	62.12 $\pm$ 1.96	79.55 $\pm$ 1.25
mmFormer [15]	✓	✓	51.34 $\pm$ 3.42	49.84 $\pm$ 4.54	69.68 $\pm$ 4.02	50.76 $\pm$ 4.48	74.91 $\pm$ 1.68
MuT [16]	✓	✓	60.63 $\pm$ 1.18	60.39 $\pm$ 0.85	80.50 $\pm$ 0.75	62.54 $\pm$ 0.95	79.64 $\pm$ 0.53
Flex-MoE [3]	✓	✓	62.91 $\pm$ 1.85	62.26 $\pm$ 1.93	80.13 $\pm$ 0.47	63.94 $\pm$ 1.59	80.73 $\pm$ 1.10
I2MoE [4]	✓	✓	61.42 $\pm$ 5.88	60.49 $\pm$ 6.37	79.23 $\pm$ 0.97	63.14 $\pm$ 3.90	80.36 $\pm$ 2.09
MMPFN [32]	✓	✓	64.50 $\pm$ 0.80	62.85 $\pm$ 0.95	81.14 $\pm$ 0.45	63.33 $\pm$ 0.94	81.04 $\pm$ 0.46
M²PFN (ours)	✓	✓	67.18$\pm$1.13	65.55$\pm$1.26	82.21$\pm$0.54	65.71$\pm$1.44	82.33$\pm$0.64

AUC (one-vs-rest), Sensitivity, and Specificity for classification, and MAE, Pearson r , and R^2 for regression, all computed on the held-out test set.

4.3 Baselines

For baselines, we consider three families of methods: unimodal models operating on the MRI volume, unimodal models operating on the tabular data, and multimodal fusion models. All baselines use the standard public implementation of their respective architectures and are trained and evaluated on the same splits. To ensure a fair comparison on the imaging side, any multimodal baseline whose released code does not provide a 3D image encoder is equipped with the same MedicalNet-pretrained 3D ResNet-18 vision encoder used by M²PFN, fine-tuned end-to-end, while retaining its own tabular encoder. For every baseline, we adopt the default hyperparameters specified in the published code. The full list of baselines is as follows:

- *Unimodal MRI*: ResNet-3D [10], DenseNet-3D [11], SE-ResNet-3D [12], SwinUNETR [14].

Table 3: Component ablation on the ADNI test set. Each row removes one mechanism from the full recipe.

Variant	Acc. (%)	F1 (%)	AUC (%)
Full M ² PFN	67.18$\pm$1.13	65.55$\pm$1.26	82.21$\pm$0.54
w/o Disentangled Alignment (DA)	65.70 $\pm$ 1.18	64.36 $\pm$ 1.72	81.41 $\pm$ 0.70
w/o End-to-end Training (E2E)	65.41 $\pm$ 1.47	64.29 $\pm$ 1.37	81.72 $\pm$ 1.12
w/o DA & E2E	63.93 $\pm$ 1.52	62.33 $\pm$ 1.66	80.77 $\pm$ 1.00
w/o Gated TabPFN Shortcut	62.96 $\pm$ 1.27	61.00 $\pm$ 1.00	79.30 $\pm$ 0.93
w/o ICL (TabPFN $\rightarrow$ MLP)	60.06 $\pm$ 3.42	59.32 $\pm$ 3.74	76.18 $\pm$ 2.13

- *Unimodal tabular*: XGBoost [42], AutoGluon [43], TabPFN-v2.5 [21].
- *Multimodal fusion*: InterpretCC [19], MoE++ [18], SwitchGate [17], mmFormer [15], MulT [16], Flex-MoE [3], I2MoE [4], MMPFN [32].

4.4 Classification on ADNI

We perform a three-way diagnostic classification experiment on the ADNI dataset. Table 2 summarizes the ADNI test-set results. M²PFN achieves the best performance across all evaluation metrics, including accuracy (67.18%), macro-F1 (65.55%), macro-AUC (82.21%), sensitivity (65.71%), and specificity (82.33%). It consistently outperforms both unimodal and multimodal baselines, demonstrating the benefit of integrating MRI with tabular biomarkers. Among the multimodal methods, MMPFN is the strongest competitor, while Flex-MoE and InterpretCC also improve upon tabular-only models but remain behind M²PFN. Notably, both MMPFN and M²PFN use the TabPFN ICL engine; the key difference is that MMPFN relies on frozen image embeddings, whereas M²PFN optimizes the imaging encoder end-to-end through the ICL transformer. The statistically significant improvement over MMPFN therefore highlights the effectiveness of end-to-end prompt tuning rather than in-context learning alone.

4.5 Ablation Study

To assess the contribution of each component of M²PFN to the final performance, we conduct an ablation study, reported in Table 3. The disentangled alignment module consistently improves both predictive performance and training stability, with its removal causing a modest but significant reduction in AUC ($p < 0.05$). In contrast, the architectural designs contribute more substantially. End-to-end optimization enables gradients from the frozen ICL engine to guide multimodal representation learning, and jointly removing end-to-end optimization and representation alignment leads to a much larger performance degradation than removing either component alone, indicating that the two are complementary. The frozen TabPFN shortcut further encourages the multimodal branch to learn complementary information beyond what is already captured by the tabular features, leading to richer multimodal representations and better performance. Finally, replacing the frozen TabPFN heads with trainable MLP classifiers produces the largest performance drop, demonstrating that preserving TabPFN’s pretrained in-context learning capabilities, rather than a conventional parametric classifier, is the largest source of M²PFN’s performance gain.

4.6 Subject-level Interpretability

Within our framework, each query is classified by attending to a labeled support set. The attention assigned to each support subject therefore provides an example-level measure of its contribution to the query prediction. For query q , we obtain the multimodal attention distribution $a_{q,\cdot}^{\text{mm}} \in \Delta^{|S|-1}$

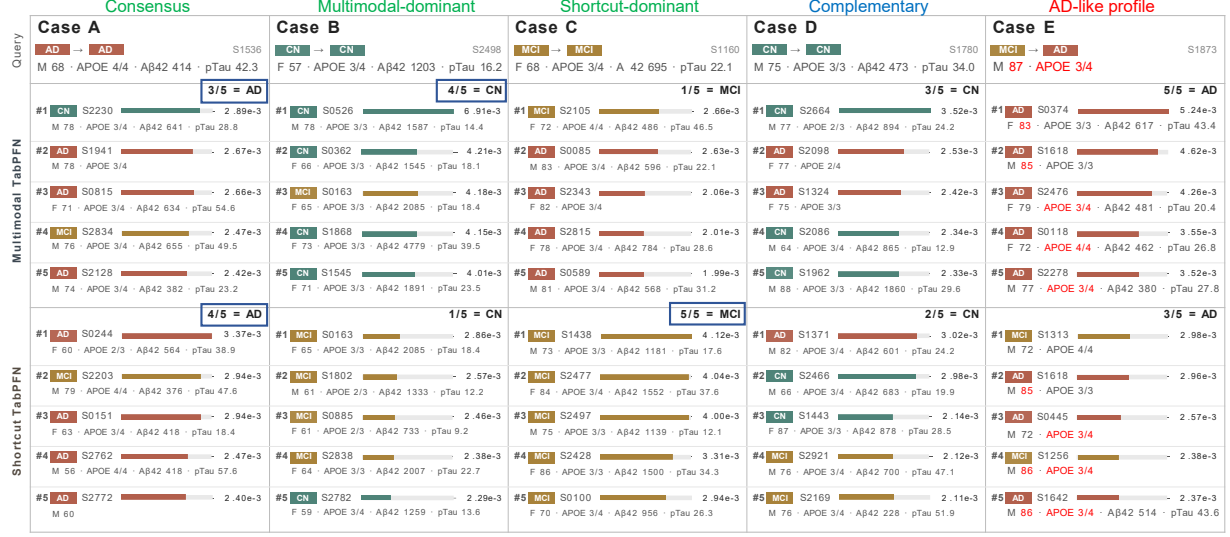


Figure 2: Exemplar reference attribution for five representative ADNI test subjects. Each column corresponds to one query subject. The first row summarizes the ground-truth and predicted diagnosis together with selected clinical variables (sex, age, APOE genotype, $A\beta_{42}$, and pTau; missing values indicate unavailable biomarkers). The two panels show the top-5 support subjects retrieved by the multimodal TabPFN and the shortcut TabPFN, respectively. Bar length represents the normalized attention weight within each query, and colors denote diagnostic class (CN, MCI, and AD). Subject identifiers are anonymized (S_n), and ages above 90 years are reported as 90+ in accordance with the ADNI Data Use Agreement.

by extracting the softmax attention weights from each TabPFN layer and head, averaging them over layers and heads, and renormalizing the resulting vector over the support set $\mathcal{S}$. We similarly obtain $a_{q,\cdot}^{\text{tab}} \in \Delta^{|\mathcal{S}|-1}$ from the tabular-only TabPFN shortcut. Both branches use the same indexed support set $\mathcal{S}$.

Because the final prediction is

$$z_{\text{final},q} = \alpha z_{\text{mm},q} + (1 - \alpha) z_{\text{tab},q}, \quad (12)$$

we define a branch-weighted unified attribution by

$$a_{q,j}^{\text{uni}} = \alpha a_{q,j}^{\text{mm}} + (1 - \alpha) a_{q,j}^{\text{tab}}, \quad j \in \mathcal{S}. \quad (13)$$

Since both branch-specific attention vectors lie in the probability simplex and $\alpha \in [0, 1]$, their convex combination satisfies $a_{q,\cdot}^{\text{uni}} \in \Delta^{|\mathcal{S}|-1}$. Consequently, entropy, top- k , and effective-support-size analyses remain applicable. The value $a_{q,j}^{\text{uni}}$ represents the relative attention-based attribution of support subject j to query q .

Figure 2 visualizes the attribution for five representative test subjects by showing the five most highly attended support subjects from the multimodal and shortcut TabPFN branches. Several representative retrieval patterns emerge. First, Case A illustrates branch consensus: both branches predominantly retrieve AD support subjects and predict AD. Second, Cases B and C show that the two branches provide different evidence across query subjects. In Case B, the multimodal branch predominantly retrieves CN support subjects, whereas the shortcut branch favors MCI subjects; their gated combination correctly predicts CN. In Case C, the shortcut branch predominantly retrieves MCI support subjects and offsets the multimodal branch’s tendency toward AD. Third,

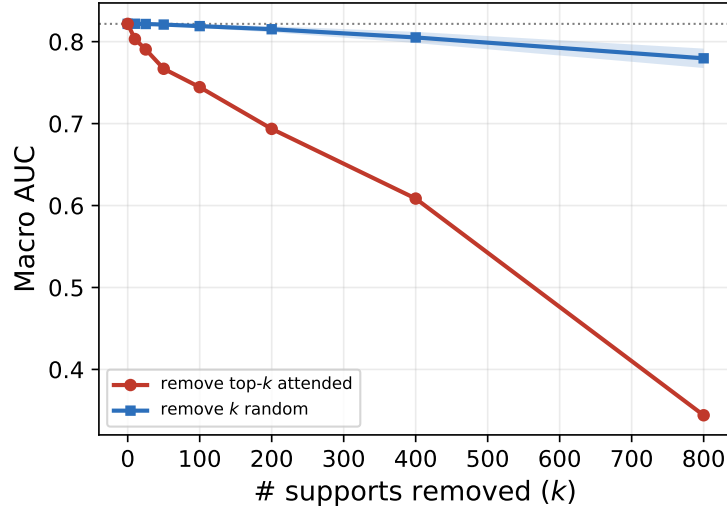


Figure 3: Deletion-curve faithfulness test on all 351 ADNI test queries. Removing the top- k most-attended unified supports (red) collapses macro-AUC far faster than removing k random supports (blue; mean $\pm$ std over 5 draws). Dotted line: the unablated baseline.

Case D illustrates complementary evidence: the multimodal branch retrieves mostly CN support subjects, whereas the shortcut branch retrieves more MCI and AD subjects, and their fusion correctly predicts CN. Finally, Case E presents a representative failure case in which both branches predominantly retrieve AD support subjects for an MCI query, resulting in an AD prediction.

These case studies provide insight into the in-context reasoning of M²PFN. The branch-consensus case suggests that the model predicts by referencing clinically similar support subjects rather than attending uniformly over the context. Cases B–D further indicate that the multimodal and shortcut branches capture complementary information. Although the learned gate α is shared across subjects, the branch predictions and their attended support neighborhoods vary with each query, allowing their fixed weighted combination to integrate different evidence across subjects. The misclassified Case E remains clinically plausible: the query and most of its highly attended support subjects exhibit AD-like clinical profiles, including advanced age and APOE- ϵ 4 positivity. This suggests that the prediction follows clinically meaningful evidence in the support set rather than arbitrary model behavior. Together, these examples show that M²PFN provides subject-level interpretability by exposing the support subjects most strongly associated with each prediction.

In addition to the interpretability experiment, we also validate that this attention reflects causal influence with a leave-one-support-out (LOO) probe. For each query, we drop the top-50 unified-attention support subjects one at a time (plus 50 random controls) and measure the induced change $|\Delta z_{\text{final}}|$ in the predicted-class posterior, denoted as LOO influence. On a class-stratified 30-query subset, the ratio of mean top-50 influence to mean control is 4.52 ± 1.29 (median 4.38), with 100% of queries above unity, and the per-query Spearman correlation between attention and LOO influence is 0.639 (median 0.647), indicating that highly attended support subjects are also those whose removal most strongly perturbs the prediction. We further validate the attention ranking with a standard deletion-curve faithfulness test on all 351 queries (Fig. 3), showing that ablating the top- k most-attended supports degrades task performance far faster than ablating k random supports (mean over 5 draws). Removing the top-800 collapses macro-AUC from 0.822 to 0.344, whereas the same number of supports removed at random leaves macro-AUC slightly degraded (0.780), with

Table 4: Test-set regression of baseline MMSE on the 1,250-subject MMSE sub-cohort of ADNI.

Method	MAE	Pearson r	R^2
InterpretCC [19]	1.925 $\pm$ 0.039	0.492 $\pm$ 0.018	0.225 $\pm$ 0.021
MoE++ [18]	1.873 $\pm$ 0.024	0.540 $\pm$ 0.017	0.277 $\pm$ 0.022
SwitchGate [17]	1.980 $\pm$ 0.095	0.514 $\pm$ 0.037	0.117 $\pm$ 0.101
mmFormer [15]	2.159 $\pm$ 0.057	0.275 $\pm$ 0.059	0.021 $\pm$ 0.061
MuT [16]	1.868 $\pm$ 0.042	0.556 $\pm$ 0.015	0.254 $\pm$ 0.033
Flex-MoE [3]	1.899 $\pm$ 0.053	0.548 $\pm$ 0.031	0.256 $\pm$ 0.049
I2MoE [4]	1.949 $\pm$ 0.025	0.526 $\pm$ 0.006	0.194 $\pm$ 0.014
MMPFN [32]	1.784 $\pm$ 0.041	0.601 $\pm$ 0.014	0.287 $\pm$ 0.057
M²PFN (ours)	1.743$\pm$0.014	0.630$\pm$0.007	0.343$\pm$0.006

the gap widening monotonically in k . Attention therefore focuses each prediction on a small subset of support examples that carries the most informative causal evidence.

4.7 Generalization to Cognitive Score Regression

To evaluate whether the proposed framework generalizes beyond categorical diagnosis, we instantiate the same architecture for continuous MMSE regression by replacing (i) the TabPFN classifier head with a TabPFN regressor head that emits a scalar mean over the bin-distributional output of TabPFN, and (ii) the cross-entropy task loss with a smooth- L_1 (Huber) loss on standardized targets. The disentangler, alignment loss, shortcut branch, and end-to-end training are identical to the classification setup. The target is the baseline Mini-Mental State Examination total score (MMSE; integer scale 0–30, with lower scores indicating greater cognitive impairment), the most widely used cognitive screen in AD research and a standard endpoint for monitoring disease progression. All metrics are reported on the original raw clinical scale (MMSE scores).

Table 4 summarizes the MMSE regression results on ADNI. M²PFN achieves the best performance across all evaluation metrics, attaining the lowest MAE together with the highest Pearson correlation and coefficient of determination. It consistently outperforms both conventional multimodal fusion methods and the recent ICL-based MMPFN, while exhibiting lower variance across random seeds. Since both MMPFN and M²PFN are built upon the same TabPFN-based ICL paradigm, the improvement can be attributed to our end-to-end optimization through the frozen ICL transformer and the proposed disentangled alignment-based multimodal representation learning, rather than to in-context learning alone. These results further demonstrate that the proposed framework generalizes beyond categorical diagnosis and can effectively support continuous cognitive score prediction with only a task-specific prediction head.

4.8 External Validation on OASIS-3 and SCAN

To test generalization beyond ADNI, we evaluate the *frozen* ADNI checkpoints on **OASIS-3** [39] and **SCAN** with *no gradient updates*, in two modes: *zero-shot*, where external subjects are predicted from the ADNI support set, and *support-swap*, where that support set is replaced by available labeled samples of the external cohort. All baselines and our model receive the same labeled sample as support in each setting and is scored on the same held-out test subjects. A parametric baseline cannot accept labels as input, so we use them for the strongest correction available to its output: a class-prior correction for diagnosis [44], and an affine rescaling for regression. The two ICL-based methods can use them as the support set.

Table 5: External CN/MCI/AD diagnosis on OASIS-3 and SCAN, as one-vs-rest AUC per class and macro-averaged.

Method	OASIS-3 (CN/MCI/AD = 855/95/213)				SCAN (756/180/181)			
	CN	one-vs-rest AUC			CN	one-vs-rest AUC		
		MCI	AD	macro		MCI	AD	macro
InterpretCC [19]	0.780 $\pm$.015	0.463 $\pm$.024	0.807 $\pm$.015	0.684 $\pm$.012	0.772 $\pm$.010	0.450 $\pm$.009	0.851 $\pm$.009	0.691 $\pm$.008
MoE++ [18]	0.758 $\pm$.021	0.450 $\pm$.021	0.794 $\pm$.020	0.668 $\pm$.020	0.783 $\pm$.018	0.451 $\pm$.062	0.851 $\pm$.021	0.695 $\pm$.019
SwitchGate [17]	0.776 $\pm$.030	0.432 $\pm$.043	0.801 $\pm$.018	0.670 $\pm$.019	0.797 $\pm$.015	0.394 $\pm$.019	0.856$\pm$.015	0.683 $\pm$.012
mmFormer [15]	0.735 $\pm$.024	0.554 $\pm$.033	0.758 $\pm$.012	0.682 $\pm$.015	0.697 $\pm$.029	0.515 $\pm$.026	0.747 $\pm$.047	0.653 $\pm$.020
MuT [16]	0.776 $\pm$.013	0.427 $\pm$.041	<u>0.809$\pm$.012</u>	0.671 $\pm$.014	0.765 $\pm$.023	0.457 $\pm$.046	0.850 $\pm$.019	0.691 $\pm$.025
Flex-MoE [3]	0.778 $\pm$.011	0.437 $\pm$.048	0.806 $\pm$.014	0.674 $\pm$.013	0.779 $\pm$.025	0.429 $\pm$.048	0.852 $\pm$.012	0.687 $\pm$.018
I2MoE [4]	<u>0.787$\pm$.007</u>	0.466 $\pm$.107	0.774 $\pm$.042	0.676 $\pm$.044	0.780 $\pm$.026	0.404 $\pm$.018	0.841 $\pm$.014	0.675 $\pm$.014
MMPFN [32] (zero-shot)	0.569 $\pm$.048	0.529 $\pm$.021	0.565 $\pm$.062	0.554 $\pm$.042	0.478 $\pm$.029	0.441 $\pm$.023	0.495 $\pm$.031	0.471 $\pm$.024
MMPFN [32] (support-swap)	0.768 $\pm$.005	0.605 $\pm$.016	0.791 $\pm$.007	0.721 $\pm$.006	0.819 $\pm$.004	0.735$\pm$.008	0.817 $\pm$.009	0.790 $\pm$.004
M ² PFN (zero-shot)	0.783 $\pm$.010	0.520 $\pm$.078	0.807 $\pm$.011	0.703 $\pm$.022	0.762 $\pm$.026	0.527 $\pm$.074	0.810 $\pm$.018	0.700 $\pm$.034
M²PFN (support-swap)	0.797$\pm$.007	0.607$\pm$.019	0.822$\pm$.007	0.742$\pm$.010	0.837$\pm$.006	<u>0.734$\pm$.009</u>	0.853 $\pm$.008	0.808$\pm$.003

Table 6: External regression on OASIS-3 (MMSE) and SCAN (MoCA).

Method	OASIS-3 (MMSE, same instrument as ADNI)			SCAN (MoCA, cross-instrument)		
	MAE $\downarrow$	r $\uparrow$	R^2 $\uparrow$	MAE $\downarrow$	r $\uparrow$	R^2 $\uparrow$
InterpretCC [19]	1.60 $\pm$.03	0.480 $\pm$.022	0.225 $\pm$.016	3.54 $\pm$.10	0.514 $\pm$.037	0.252 $\pm$.035
MoE++ [18]	1.60 $\pm$.05	0.480 $\pm$.027	<u>0.227$\pm$.022</u>	3.57 $\pm$.12	0.503 $\pm$.050	0.242 $\pm$.047
SwitchGate [17]	1.65 $\pm$.06	0.434 $\pm$.076	0.186 $\pm$.056	3.98 $\pm$.08	0.230 $\pm$.127	0.061 $\pm$.039
mmFormer [15]	1.74 $\pm$.07	0.310 $\pm$.098	0.097 $\pm$.053	3.95 $\pm$.18	0.216 $\pm$.098	0.049 $\pm$.050
MuT [16]	1.59 $\pm$.04	<u>0.489$\pm$.033</u>	0.234$\pm$.027	3.63 $\pm$.23	0.475 $\pm$.072	0.218 $\pm$.075
Flex-MoE [3]	1.64 $\pm$.07	0.456 $\pm$.061	0.205 $\pm$.050	3.80 $\pm$.22	0.363 $\pm$.102	0.135 $\pm$.072
I2MoE [4]	1.83 $\pm$.03	0.107 $\pm$.052	0.007 $\pm$.017	4.07 $\pm$.05	0.060 $\pm$.092	0.006 $\pm$.015
MMPFN [32] (zero-shot)	1.79 $\pm$.06	0.151 $\pm$.136	0.026 $\pm$.040	4.05 $\pm$.10	0.152 $\pm$.059	0.021 $\pm$.019
MMPFN [32] (support-swap)	<u>1.50$\pm$.02</u>	0.437 $\pm$.039	0.156 $\pm$.024	<u>3.48$\pm$.15</u>	0.526 $\pm$.020	0.272 $\pm$.021
M ² PFN (zero-shot)	1.61 $\pm$.01	0.475 $\pm$.010	0.224 $\pm$.009	3.75 $\pm$.11	0.371 $\pm$.038	0.134 $\pm$.032
M²PFN (support-swap)	1.47$\pm$.03	0.500$\pm$.038	0.212 $\pm$.031	3.16$\pm$.12	0.643$\pm$.027	0.386$\pm$.032

Diagnostic transfer. As shown in Table 5, all methods maintain reasonable discrimination between CN and AD, whereas distinguishing the minority MCI class remains the primary challenge. With support-set replacement, M²PFN achieves the highest MCI AUC and the best macro-AUC among parametric learning-based methods on both external cohorts, demonstrating the effectiveness of in-context adaptation for cross-cohort diagnosis. Compared with MMPFN, which is also built on TabPFN, M²PFN maintains substantially stronger zero-shot performance before support replacement and benefits further from support-set adaptation, resulting in significantly higher macro-AUC; this indicates that end-to-end multimodal representation learning and in-context learning are complementary for robust cross-cohort diagnosis.

Cognitive-score transfer. As shown in Table 6, M²PFN achieves the lowest MAE on both OASIS-3 and SCAN, significantly outperforming all baselines. The two cohorts present different transfer scenarios: OASIS-3 shares the same cognitive assessment (MMSE) as ADNI, whereas SCAN requires cross-instrument transfer to MoCA. Consequently, performance differences on OASIS-3 are relatively modest, with several strong baselines achieving comparable correlations. In contrast, SCAN highlights the advantage of in-context adaptation: while rescaling can reduce prediction error, it cannot improve Pearson correlation because the ranking of predictions remains unchanged. By replacing the support set with MoCA-labeled examples, M²PFN substantially improves the correlation and clearly outperforms all competing methods, including the ICL-based MMPFN. These results indicate that effective cross-instrument transfer relies not only on support-set adaptation, but also on learning transferable multimodal representations that remain compatible

with the ICL engine.

5 Conclusion

We presented M²PFN, an end-to-end multimodal framework that repurposes a frozen TabPFN as the prediction engine for Alzheimer’s disease analysis by enabling end-to-end optimization through its in-context learning (ICL) transformer. Rather than fine-tuning the foundation model, M²PFN keeps the ICL engine frozen and instead learns multimodal representations that are directly optimized for in-context reasoning through differentiable prompt tuning. A disentangled alignment-based multimodal representation learner and a gated tabular shortcut further bridge heterogeneous MRI and tabular biomarkers while encouraging complementary information learning. Extensive experiments demonstrate that M²PFN achieves state-of-the-art performance on both three-class AD diagnosis and MMSE regression on ADNI, consistently outperforming strong multimodal baselines. More importantly, preserving the frozen ICL engine enables gradient-free support-set replacement at inference, allowing the same model to transfer effectively to independent clinical cohorts without retraining while maintaining superior performance on both classification and cognitive-score prediction tasks. These results demonstrate that combining multimodal representation learning with in-context learning provides an effective paradigm for robust cross-cohort clinical prediction, and highlight the potential of foundation-model-based in-context learning as an alternative to task-specific finetuning for multimodal medical AI.

Acknowledgments

This work was supported by the National Institutes of Health (NIH) under grants R01EB022744, R01AG077578, R01AG064584, S10OD032285, U19AG078109, and P30AG066530. This work used data from the Alzheimer’s Disease Neuroimaging Initiative (ADNI), OASIS-3, and the Standardized Centralized Alzheimer’s and Related Dementias Neuroimaging (SCAN) initiative linked to the National Alzheimer’s Coordinating Center (NACC). The NACC database is funded by NIA/NIH Grant U24 AG072122, and the SCAN initiative by NIA/NIH Grant U24 AG067418; NACC and SCAN data are contributed by the NIA-funded Alzheimer’s Disease Research Centers (ADRCs).

References

- [1] Sebastian Pölsterl, Tom Nuno Wolf, and Christian Wachinger. Combining 3D image and tabular data via the dynamic affine feature map transform. In *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. (MICCAI)*, pages 688–698, 2021. doi: 10.1007/978-3-030-87240-3_66.
- [2] Shangran Qiu, Matthew I. Miller, Prajakta S. Joshi, Joyce C. Lee, Chonghua Xue, Yunruo Ni, et al. Multimodal deep learning for Alzheimer’s disease dementia assessment. *Nature Commun.*, 13(1):3404, June 2022. doi: 10.1038/s41467-022-31037-5.
- [3] Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. Flex-MoE: Modeling arbitrary modality combination via the flexible mixture-of-experts. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 37, pages 98782–98805, 2024.

- [4] Jiayi Xin, Sukwon Yun, Jie Peng, Inyoung Choi, Jenna L. Ballard, Tianlong Chen, and Qi Long. I2MoE: Interpretable multimodal interaction-aware mixture-of-experts. In *Proc. 42nd Int. Conf. Mach. Learn. (ICML)*, volume 267 of *Proc. Mach. Learn. Res.*, pages 68870–68888, 2025.
- [5] Shuo Huang, Lujia Zhong, and Yonggang Shi. Multistage alignment and fusion for multimodal multiclass Alzheimer’s disease diagnosis. In *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. (MICCAI)*, pages 375–385, 2025. doi: 10.1007/978-3-032-05182-0_37.
- [6] Guangqian Yang, Kangrui Du, Zhihan Yang, Ye Du, Eva Yi Wah Cheung, Yongping Zheng, Mo Yang, Zoe Kourtzi, Carola-Bibiane Schönlieb, and Shujun Wang. ADFound: A foundation model for diagnosis and prognosis of Alzheimer’s disease. *IEEE J. Biomed. Health Inform.*, 29(11):8395–8408, November 2025. doi: 10.1109/JBHI.2025.3576436.
- [7] Guangqian Yang, Ye Du, Xiaowei Hu, Weiyang Shi, Mo Yang, and Shujun Wang. Taming arbitrary modality missingness and imbalance: A unified graph-MoE framework for Alzheimer’s disease diagnosis. *Med. Image Anal.*, 113:104222, September 2026. doi: 10.1016/j.media.2026.104222.
- [8] Susanne G. Mueller, Michael W. Weiner, Leon J. Thal, Ronald C. Petersen, Clifford Jack, William Jagust, John Q. Trojanowski, Arthur W. Toga, and Laurel Beckett. The Alzheimer’s Disease Neuroimaging Initiative. *Neuroimaging Clin. N. Amer.*, 15(4):869–877, November 2005. doi: 10.1016/j.nic.2005.09.008.
- [9] Clifford R. Jack, Jr., Matt A. Bernstein, Nick C. Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J. Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The Alzheimer’s Disease Neuroimaging Initiative (ADNI): MRI methods. *J. Magn. Reson. Imag.*, 27(4):685–691, April 2008. doi: 10.1002/jmri.21049.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.
- [11] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2261–2269, 2017. doi: 10.1109/CVPR.2017.243.
- [12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 7132–7141, 2018. doi: 10.1109/CVPR.2018.00745.
- [13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
- [14] Yufan He, Vishwesh Nath, Dong Yang, Yucheng Tang, Andriy Myronenko, and Daguang Xu. SwinUNETR-V2: Stronger Swin transformers with stagewise convolutions for 3D medical image segmentation. In *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. (MICCAI)*, pages 416–426, 2023. doi: 10.1007/978-3-031-43901-8_40.

- [15] Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmFormer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. (MICCAI)*, pages 107–117, 2022. doi: 10.1007/978-3-031-16443-9_11.
- [16] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, pages 6558–6569, 2019. doi: 10.18653/v1/P19-1656.
- [17] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23(120):1–39, 2022.
- [18] Peng Jin, Bo Zhu, Li Yuan, and Shuicheng Yan. MoE++: Accelerating mixture-of-experts methods with zero-computation experts. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2025.
- [19] Vinitra Swamy, Syrielle Montariol, Julian Blackwell, Jibril Frej, Martin Jaggi, and Tanja Käser. Intrinsic user-centric interpretability through global mixture of experts. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2025.
- [20] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 33, pages 1877–1901, 2020.
- [21] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeyer, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, January 2025. doi: 10.1038/s41586-024-08328-6.
- [22] Weihao Kong and Abhimanyu Das. Introducing TabFM: A zero-shot foundation model for tabular data. Google Research Blog, June 2026. Accessed: Jul. 2026.
- [23] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 35, pages 23716–23736, 2022.
- [24] Jannik Kossen, Neil Band, Clare Lyle, Aidan N. Gomez, Tom Rainforth, and Yarin Gal. Self-attention between datapoints: Going beyond individual input-output pairs in deep learning. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 34, pages 28742–28756, 2021.
- [25] Yury Gorishniy, Ivan Rubachev, Nikolay Kartashev, Daniil Shlenskii, Akim Kotelnikov, and Artem Babenko. TabR: Tabular deep learning meets nearest neighbors. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.
- [26] Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do Bayesian inference. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.
- [27] Noah Hollmann, Samuel Müller, Katharina Eggersperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.

- [28] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, pages 3045–3059, 2021. doi: 10.18653/v1/2021.emnlp-main.243.
- [29] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process. (ACL-IJCNLP)*, pages 4582–4597, 2021. doi: 10.18653/v1/2021.acl-long.353.
- [30] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. GPT understands, too. *AI Open*, 5:208–215, 2024. doi: 10.1016/j.aiopen.2023.08.012.
- [31] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.
- [32] Wall Kim, Chaeyoung Song, and Hanul Kim. MultiModalPFN: Extending prior-data fitted networks for multimodal tabular learning. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 30357–30367, 2026.
- [33] M. Jorge Cardoso et al. MONAI: An open-source framework for deep learning in healthcare. arXiv:2211.02701, 2022.
- [34] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3D: Transfer learning for 3D medical image analysis. arXiv:1904.00625, 2019.
- [35] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 34, pages 18932–18943, 2021.
- [36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. Int. Conf. Mach. Learn. (ICML)*, pages 8748–8763, 2021.
- [37] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 34, pages 9694–9705, 2021.
- [38] National Institute on Aging. SCAN: Standardized Centralized Alzheimer’s and Related Dementias Neuroimaging, 2020. NIA-funded initiative U24 AG067418; data accessed through the NACC Data Platform. Accessed: Jul. 2026.
- [39] Pamela J. LaMontagne, Tammie L. S. Benzinger, John C. Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G. Vlassenko, et al. OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. *medRxiv*, December 2019. Preprint.
- [40] Bruce Fischl. FreeSurfer. *NeuroImage*, 62(2):774–781, August 2012. doi: 10.1016/j.neuroimage.2012.01.021.

- [41] Duane L. Beekly, Erin M. Ramos, William W. Lee, Woodrow D. Deitrich, Mary E. Jacka, Joylee Wu, Janene L. Hubbard, Thomas D. Koepsell, John C. Morris, Walter A. Kukull, et al. The National Alzheimer’s Coordinating Center (NACC) database: The Uniform Data Set. *Alzheimer Dis. Assoc. Disord.*, 21(3):249–258, July–September 2007. doi: 10.1097/WAD.0b013e318142774e.
- [42] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, pages 785–794, 2016. doi: 10.1145/2939672.2939785.
- [43] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. AutoGluon-Tabular: Robust and accurate AutoML for structured data. arXiv:2003.06505, 2020.
- [44] Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Comput.*, 14(1):21–41, January 2002. doi: 10.1162/089976602753284446.