RPMF: RELIABILITY-AWARE PROGRESSIVE MULTIMODAL FUSION FRAMEWORK FOR MULTIMODAL BRAIN DISEASE DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal learning has emerged as a promising paradigm for brain disease diagnosis by integrating complementary information from heterogeneous clinical modalities. However, the diagnostic reliability of different modalities varies substantially across individual subjects, suggesting that multimodal data should be integrated adaptively according to their reliability. Nevertheless, many existing multimodal methods estimate modality importance only implicitly and directly fuse all modalities in a single step, limiting their ability to prioritize reliable modalities while suppressing noisy or conflicting information. To address these limitations, we propose a Reliability-aware Progressive Multimodal Fusion (RPMF) framework for brain disease diagnosis. Specifically, RPMF first learns modality-specific representations and explicitly estimates subject-specific modality reliability using an entropy-based uncertainty formulation, allowing the model to quantify the diagnostic confidence of each modality. To characterize heterogeneous disease patterns, each modality is further represented by a set of latent prototypes, from which cross-modal compatibility is evaluated through prototype-level relation modeling. Guided by both reliability and compatibility, multimodal information is progressively integrated in a reliability-aware manner, where the most reliable modality serves as the initial anchor and subsequent modalities are sequentially incorporated according to the reliability and compatibility with the current fused representation. This progressive strategy enables the model to adaptively exploit useful modalities that are compatible with the current multimodal representation while suppressing noisy or conflicting information. Extensive experiments on multiple brain disease datasets demonstrate that RPMF consistently outperforms state-of-the-art methods.

1 INTRODUCTION

Brain diseases, such as Alzheimer's disease (AD) (Jack Jr et al., 2024) and autism spectrum disorder (ASD) (Lord et al., 2020), are complex neurological and neurodevelopmental conditions that affect millions of individuals worldwide. Accurate and early diagnosis is critical for timely intervention, personalized treatment, and improved long-term outcomes (Liss et al., 2021). In clinical practice, diagnosis of brain diseases typically relies on integrating multimodal information, including magnetic resonance imaging (MRI), positron emission tomography (PET), behavioral or cognitive assessments, genetic profiles, and demographic data (Shoeibi et al., 2023). Each modality captures complementary yet incomplete aspects of disease status, motivating multimodal learning as a promising approach to brain disease diagnosis (Mirzaei et al., 2026).

Recent years have witnessed remarkable progress in multimodal learning for brain disease analysis (Zhang et al., 2025). Early studies primarily relied on handcrafted features together with feature-level or decision-level fusion strategies (Mirzaei et al., 2026). More recently, deep neural networks have enabled end-to-end multimodal representation learning through modality-specific encoders combined with early, late, or hybrid fusion architectures. Transformer-based models (Cai et al., 2023; Feng et al., 2025), graph neural networks (Zheng et al., 2022; Wan et al., 2025; Pei et al., 2025), and cross-modal attention mechanisms (Tang et al., 2024) have further improved the modeling of interactions across heterogeneous modalities. In parallel, adaptive fusion techniques,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

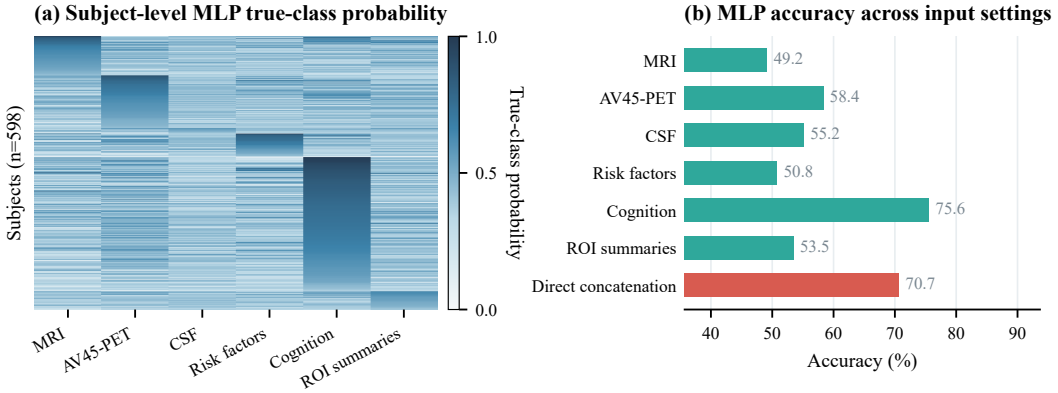


Figure 1: **(a)** Subject-level probabilities assigned to the ground-truth class by individual unimodal classifiers. Each row represents a subject, each column represents a modality, and darker colors indicate higher probabilities assigned to the ground-truth class. **(b)** Classification accuracy obtained by the same MLP using individual modalities or their direct feature concatenation. Bars and error bars denote the mean and standard deviation over ten cross-validation folds, respectively.

including gating networks, mixture-of-experts, and dynamic attention mechanisms, have been proposed to automatically assign different importance to multimodal information, leading to competitive diagnostic performance (Golovanevsky et al., 2022; Tan et al., 2025). However, these methods do not necessarily address two fundamental questions: *which modalities provide reliable diagnostic evidence for a given subject*, and *how should additional modalities be incorporated into the existing representation*?

First, modality reliability can vary substantially across subjects, making it difficult to establish a universal modality importance ranking (Dyrba et al., 2015). As illustrated in Fig. 1(a), the probabilities assigned to the true disease class by individual unimodal classifiers vary considerably across subjects, and the modality assigning the highest probability differs from one subject to another. Fig. 1(b) reveals substantial differences in the overall performance of individual modalities. Together, these observations motivate subject-specific modality reliability estimation rather than relying solely on a globally fixed modality importance ranking. However, most existing approaches estimate modality importance implicitly through attention or gating mechanisms, where the learned weights primarily reflect feature interactions optimized for the downstream objective rather than explicitly characterizing modality reliability (Gapp et al., 2026). Consequently, unreliable modalities may still receive disproportionately large weights, potentially distorting the final diagnosis.

Second, existing multimodal methods predominantly adopt one-shot fusion, in which all modalities are jointly aggregated through feature concatenation, attention, or cross-modal interaction modules (Nawaz et al., 2026). As shown in Fig. 1(b), directly concatenating all modalities does not necessarily provide a clear advantage over the strongest unimodal predictor. This suggests that simply combining all available modalities may introduce redundant or conflicting information rather than consistently improving the performance of brain disease diagnosis. Moreover, one-shot fusion tightly couples heterogeneous modalities throughout the integration process, making it difficult to selectively prioritize reliable modalities or suppress inconsistent information. An effective fusion strategy should therefore determine not only how much each modality contributes but also when it should be incorporated. This progressive reasoning process is also consistent with clinical diagnosis, where physicians first establish an initial hypothesis from the most reliable modality and then selectively incorporate additional examinations to refine their decision (Higgs et al., 2001).

To address these challenges, we propose a Reliability-aware Progressive Multimodal Fusion (RPMF) framework for brain disease diagnosis. Our approach explicitly estimates subject-specific modality reliability via an entropy-based uncertainty formulation, enabling adaptive assessment of each modality's contribution to diagnosis. To characterize heterogeneous disease patterns within each modality, the framework further employs structured latent prototypes for fine-grained representation learning and cross-modal interaction modeling. Based on these representations, cross-modal compatibility is evaluated to determine whether different modalities provide consistent and comple-

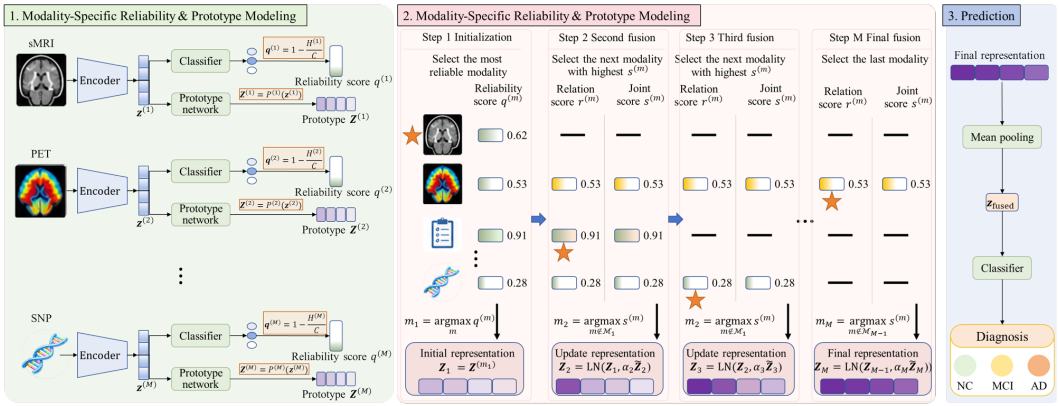


Figure 2: Overview of the proposed Reliability-aware Progressive Multimodal Fusion (RPMF) framework. **(1)** Modality-specific encoders produce latent representations, from which subject-specific reliability scores and prototype sets are derived. **(2)** RPMF initializes fusion with the most reliable modality and progressively selects subsequent modalities according to a joint reliability–relation score, followed by adaptive prototype-level fusion. **(3)** The final fused prototypes are aggregated for multimodal disease diagnosis.

mentary diagnostic modality. Multimodal information is then integrated progressively. The fusion process begins with the most reliable modality and sequentially incorporates additional modalities according to their reliability and compatibility with the current fused representation, thereby suppressing noisy or conflicting modalities and enabling more effective multimodal reasoning. Extensive experiments on multiple brain disease datasets demonstrate that the proposed framework consistently outperforms existing multimodal approaches. The main contributions are as follows.

- We propose a reliability-aware multimodal fusion framework that explicitly estimates subject-specific modality reliability from predictive uncertainty, providing an explicit reliability signal to guide multimodal integration rather than relying solely on implicit modality weighting.
- We develop a progressive fusion strategy in which modalities are sequentially incorporated according to both subject-specific reliability and their compatibility with the evolving fused representation, allowing more reliable and contextually compatible modalities to exert greater influence during integration.
- We conduct extensive experiments on multiple brain disease datasets, demonstrating consistent improvements over state-of-the-art multimodal methods.

2 Method

2.1 Model Motivation and Overview

Given a subject with multimodal observations $\mathcal{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}\}$, where $\mathbf{x}^{(m)} \in \mathbb{R}^{d_m}$ denotes the input of the m-th modality, our goal is to learn a multimodal classifier that effectively integrates complementary information from different modalities to predict the disease label y.

Brain disease diagnosis is inherently a multimodal reasoning process, where clinicians jointly consider neuroimaging biomarkers, demographic factors, and clinical assessments. However, the diagnostic value of each modality varies across individuals and disease stages. Moreover, clinical decision-making usually follows a progressive integration strategy, where more reliable modalities are considered first and additional modalities are incorporated to refine the diagnostic hypothesis (Atri et al., 2025). Inspired by this process, we propose RPMF, a reliability-aware progressive multimodal fusion framework that models subject-specific modality reliability and progressively integrates multimodal information for brain disease diagnosis. As illustrated in Fig. 2, RPMF first extracts modality-specific representations using independent encoders. Then, an uncertainty-based

reliability estimation module evaluates the contribution of each modality, while prototype-based cross-modal modeling captures complementary and consistent disease-related information. Finally, the learned reliability and cross-modal relationships jointly guide progressive fusion to generate a unified representation for diagnosis.

2.2 Reliability-Aware Representation Learning

Multimodal brain disease diagnosis involves heterogeneous modalities that encode distinct biological and clinical information. Beyond such cross-modal heterogeneity, the diagnostic utility of the same modality can also vary substantially across subjects due to acquisition noise and inter-subject variability. Consequently, effective multimodal modeling requires not only modality-specific representations, but also subject-specific estimates of how reliable each modality is for the current prediction. Given the input $\mathbf{x}^{(m)}$ of modality m, we first employ a modality-specific encoder $E^{(m)}(\cdot)$ to obtain the embedding $\mathbf{z}^{(m)} \in \mathbb{R}^D$: $\mathbf{z}^{(m)} = E^{(m)}(\mathbf{x}^{(m)})$.

We then estimate subject-specific modality reliability from the uncertainty of its unimodal prediction. A modality that provides informative diagnostic evidence for a subject should induce a relatively concentrated predictive distribution, whereas an ambiguous modality is more likely to yield a diffuse distribution. Specifically, given the representation $\mathbf{z}^{(m)}$, a lightweight modality-specific classifier produces the class logits $\mathbf{f}^{(m)} \in \mathbb{R}^C$. The corresponding class probabilities are computed as

$$p_c^{(m)} = \frac{\exp(f_c^{(m)}/T)}{\sum_{j=1}^{C} \exp(f_j^{(m)}/T)}, \tag{1}$$

where T is a temperature parameter controlling the sharpness of the predictive distribution. We then quantify the predictive uncertainty of the m-th modality using Shannon entropy:

$$H^{(m)} = -\sum_{c=1}^{C} p_c^{(m)} \log p_c^{(m)}. \tag{2}$$

We characterize subject-specific modality reliability according to the certainty of its unimodal prediction. A modality that provides more discriminative diagnostic evidence for a subject is expected to produce a more concentrated predictive distribution, whereas less informative evidence generally leads to greater predictive uncertainty. We therefore quantify the instantaneous predictive certainty using normalized entropy and define

$$q^{(m)} = 1 - \frac{H^{(m)}}{\log C}. \tag{3}$$

Accordingly, $q^{(m)} \in [0, 1]$, where a larger value indicates a more concentrated and confident modality-specific prediction, while a smaller value corresponds to a more uncertain prediction. This reliability score provides a subject-adaptive estimate of the predictive informativeness of each modality and is subsequently used to measure its contribution during multimodal fusion.

The entropy-based reliability score is estimated from the current prediction and may fluctuate during optimization. To make it better reflect the modality-specific prediction difficulty of individual samples, we regularize the estimated reliability using historical unimodal prediction losses. Specifically, samples that are consistently easier to predict under a given modality are encouraged to receive higher reliability scores. For modality m, we maintain a historical prediction difficulty $h_i^{(m)}$ for each training sample i by accumulating its sample-wise cross-entropy loss

$$h_i^{(m)} \leftarrow h_i^{(m)} + \ell_{\text{ce}} y_i, \mathbf{f}_i^{(m)}, \tag{4}$$

where $\mathbf{f}_i^{(m)}$ denotes the unimodal class logits produced by the modality-specific classifier for sample i under modality m, and y_i is the corresponding ground-truth label. The accumulated difficulty is min-max normalized within each modality:

$$\tilde{h}_i^{(m)} = \frac{h_i^{(m)} - \min_j h_j^{(m)}}{\max_j h_j^{(m)} - \min_j h_j^{(m)}}. \tag{5}$$

A smaller $\tilde{h}_i^{(m)}$ indicates that sample i has been relatively easier to predict under modality m throughout training. For a sample pair (i, j), we define the historical ordering and its corresponding difference as

$$\gamma_{ij}^{(m)} = \mathrm{sign}(\tilde{h}_i^{(m)} - \tilde{h}_j^{(m)}), \qquad \delta_{ij}^{(m)} = |\tilde{h}_i^{(m)} - \tilde{h}_j^{(m)}|. \tag{6}$$

We then impose a pairwise consistency constraint on their reliability scores:

$$\mathcal{L}_{\mathrm{rank}}^{(m)} = \frac{1}{|\mathcal{P}^{(m)}|} \sum_{(i,j)\in\mathcal{P}^{(m)}} [\gamma_{ij}^{(m)}(q_i^{(m)} - q_j^{(m)}) - \delta_{ij}^{(m)}]_+, \tag{7}$$

where $[a]_+ = \max(0, a)$ and $\mathcal{P}^{(m)}$ denotes the set of sample pairs. This regularization discourages reliability assignments that are substantially inconsistent with the historical prediction difficulty, thereby incorporating training-time prediction behavior into the instantaneous uncertainty-based reliability estimate.

2.3 Prototype-Based Cross-Modal Relation Modeling

While modality-level representations provide a compact summary of disease-related information, brain diseases often involve multiple heterogeneous pathological patterns that may coexist within the same modality. For example, neuroimaging data may exhibit distinct structural or functional abnormalities across different brain regions, whereas cognitive or behavioral assessments capture impairments from different functional perspectives (Brown et al., 2025). Compressing such diverse information into a single global representation may obscure fine-grained disease-related characteristics that are important for multimodal interaction modeling.

To preserve such heterogeneous patterns, we transform each modality representation into a set of latent prototypes through multiple learnable projection functions: $\mathbf{Z}^{(m)} = P^{(m)}(\mathbf{z}^{(m)}) \in \mathbb{R}^{K\times D}$, where $P^{(m)}(\cdot) = \{P_1^{(m)}(\cdot), \ldots, P_K^{(m)}(\cdot)\}$ consists of K lightweight MLPs. Specifically, the k-th prototype is obtained as $\mathbf{z}_k^{(m)} = P_k^{(m)}(\mathbf{z}^{(m)})$. The resulting prototype set $\mathbf{Z}^{(m)} = [\mathbf{z}_1^{(m)}, \ldots, \mathbf{z}_K^{(m)}]$ captures multiple latent sub-patterns within the same modality, providing a richer representation than a single global embedding for subsequent cross-modal interaction modeling.

Building upon the prototype representations, we introduce a cross-modal compatibility modeling mechanism. Instead of measuring similarity between global modality embeddings, we evaluate whether latent disease-related patterns from different modalities provide mutually supportive information under the current diagnostic context. This design is motivated by the observation that different modalities often contain partially overlapping yet complementary information (Zhang et al., 2025), making holistic embedding similarity insufficient for reliable multimodal reasoning. Given two prototype sets $\mathbf{Z}^{(a)}$ and $\mathbf{Z}^{(b)}$, the matching distance is defined through bidirectional nearest-neighbor matching:

$$d(\mathbf{Z}^{(a)}, \mathbf{Z}^{(b)}) = \frac{1}{K}\sum_{i=1}^{K} \min_j \|\mathbf{z}_i^{(a)} - \mathbf{z}_j^{(b)}\|_2, \tag{8}$$

This formulation evaluates whether each latent pattern in one modality can be supported by at least one corresponding pattern in another modality. Compared with global similarity measures, the proposed matching mechanism is more suitable for heterogeneous multimodal data, where modalities may only partially align while still providing clinically coherent information. To ensure symmetric comparison, we compute the matching distance in both directions and define the cross-modal relation score as

$$r^{(a,b)} = -(d(\mathbf{Z}^{(a)}, \mathbf{Z}^{(b)}) + d(\mathbf{Z}^{(b)}, \mathbf{Z}^{(a)})), \tag{9}$$

A higher relation score indicates stronger alignment and consistency.

2.4 Progressive Multimodal Fusion

In clinical practice, the information of different modalities is rarely integrated all at once. Instead, clinicians typically anchor their judgment on the most reliable modality and then progressively incorporate additional modalities that are both consistent with and compatible to the existing diagnostic hypothesis (Atri et al., 2025). Motivated by this process, rather than treating all modalities equally or

integrating them simultaneously, we progressively incorporate modalities according to their subject-specific reliability and their relation to the currently fused representation. This design allows the fusion process to prioritize more reliable modalities while accounting for cross-modal compatibility during subsequent integration. The fusion process is initialized by selecting the modality with the highest reliability score, which serves as the most trustworthy diagnostic anchor:

$$m_1 = \arg\max_m q^{(m)}, \quad \mathbf{Z}_1 = \mathbf{Z}^{(m_1)}. \tag{10}$$

where $\mathbf{Z}_1$ serves as the initial fused prototype representation. At each subsequent fusion stage t, we evaluate the remaining modalities based on a joint score that balances reliability and relation with the current fused representation. Specifically, the joint selection score of modality m is defined as

$$s^{(m)} = \beta q^{(m)} + \eta r^{(m)}, \tag{11}$$

where $q^{(m)}$ reflects the subject-specific reliability, and $r^{(m)}$ measures its cross-modal compatibility with the already integrated representation. The coefficients β and η control the relative importance of these two factors. This formulation favors modalities that are both individually reliable and compatible to the current context, while discouraging the integration of noisy or irrelevant modalities. The modality with the highest score among the remaining candidates is selected at stage t:

$$m_t = \arg\max_{m \notin \mathcal{M}_{t-1}} s^{(m)}, \tag{12}$$

where $\mathcal{M}_{t-1}$ denotes the set of modalities already integrated. The selected modality is then incorporated into the fused representation through a prototype-level cross-attention module:

$$\tilde{\mathbf{Z}}_t = \mathrm{CrossAttn}(\mathbf{Z}_{t-1}, \mathbf{Z}^{(m_t)}), \tag{13}$$

where the current fused prototypes $\mathbf{Z}_{t-1}$ serve as queries, while the prototypes of the selected modality provide the keys and values. This operation enables each fused prototype to selectively retrieve information from the newly introduced modality instead of directly concatenating the two representations. To adaptively control the influence of the newly integrated modality, we compute a fusion strength:

$$\alpha_t = \mathrm{softplus}(\beta q^{(m_t)} + \eta r^{(m_t)}), \tag{14}$$

The joint score α_t increases when the modality is both reliable and consistent with the fused modality. The fused representation is subsequently updated through a gated residual operation:

$$\mathbf{Z}_t = \mathrm{LayerNorm}(\mathbf{Z}_{t-1} + \alpha_t \tilde{\mathbf{Z}}_t). \tag{15}$$

Consequently, a selected modality with a larger joint score exerts a stronger influence on the updated representation, whereas its contribution is attenuated when the score is low. Repeating this process over the remaining modalities yields the final fused prototype representation $\mathbf{Z}_M$.

2.5 Training Objective and Final Prediction

After progressive fusion, we aggregate the final prototype representation $\mathbf{Z}_M$ by mean pooling as $\mathbf{z}_{\text{fuse}} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{Z}_{M,k}$. The resulting representation is fed into a classification head $C(\cdot)$ to obtain the final logits $\mathbf{f} = C(\mathbf{z}_{\text{fuse}})$ and the primary classification objective is

$$\mathcal{L}_{\text{cls}} = \ell_{\text{ce}}(y, \mathbf{f}), \tag{16}$$

where y denotes the ground-truth label and $\ell_{\text{ce}}(\cdot,\cdot)$ is the cross-entropy loss. To promote discriminative learning within each modality, the modality-specific representation $\mathbf{z}^{(m)}$ and unimodal logits $\mathbf{f}^{(m)}$ are further supervised by an auxiliary classification objective:

$$\mathcal{L}_{\text{aux}}^{(m)} = \ell_{\text{ce}}(y, \mathbf{f}^{(m)}). \tag{17}$$

The overall training objective combines the final classification loss, unimodal auxiliary supervision, and the reliability regularization introduced in Eq. (7):

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_1 \frac{1}{M}\sum_{m=1}^{M} \mathcal{L}_{\text{aux}}^{(m)} + \lambda_2 \frac{1}{M}\sum_{m=1}^{M} \mathcal{L}_{\text{rank}}^{(m)}, \tag{18}$$

where λ_1 and λ_2 control the contributions of different terms.

3 EXPERIMENTS

3.1 DATASETS

Experiments are conducted on two widely used multimodal medical datasets for brain disease analysis, namely the TADPOLE dataset (Marinescu et al., 2018) and the ABIDE dataset (Abraham et al., 2017). Their detailed statistics are summarized in Table 4. TADPOLE is derived from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and includes six modalities, including MRI, PET, CSF biomarkers, risk factors, cognitive assessments, and ROI summaries. A total of 598 subjects are used, including 209 Normal Control (NC), 315 Mild Cognitive Impairment (MCI), and 74 Alzheimer's Disease (AD) subjects. ABIDE is a large-scale multi-site dataset for autism research, containing resting-state fMRI data and phenotypic information. ABIDE contains 871 subjects, including 468 Normal Control (NC) and 403 Autism Spectrum Disorder (ASD) subjects. Each subject contains four modalities, including demographic features, automated anatomical quality assessment metrics, automated functional quality assessment metrics and functional connectivity networks. To ensure consistency with prior studies, we use the dataset preprocessed with the Configurable Pipeline for the Analysis of Connectomes (C-PAC) following the workflow described in (Parisot et al., 2017; Zheng et al., 2022).

3.2 IMPLEMENTATION DETAILS

We conduct experiments on the TADPOLE dataset using a three-way classification protocol among NC, MCI, and AD. Experiments on the ABIDE dataset follow a binary classification setting for differentiating NC and ASD. For multi-class classification, we report accuracy (ACC) and area under the ROC curve (AUC), where AUC is computed using the one-vs-rest strategy and macro-averaged across all classes. For binary classification, we report ACC, AUC, sensitivity (SEN), and specificity (SPE). We conduct all experiments using 10-fold stratified cross-validation. In each fold, 10% of the training data are further held out as a validation set for model selection, while the outer held-out fold is used exclusively for testing. The average performance over the ten test folds is reported. The proposed model is implemented in PyTorch and trained on a single NVIDIA RTX 3090 GPU with 24 GB memory. The model is optimized using the Adam optimizer with a learning rate of 1×10^{-3}, a batch size of 256, and trained for 1000 epochs, with early stopping applied according to the validation performance using a patience of 50 epochs. The temperature parameter in reliability estimation is set to $T = 1$. The detailed implementation settings and hyperparameter settings are provided in the supplementary material. We compare the proposed method with several representative baselines, including MLP, MMGL (Zheng et al., 2022), MD (Han et al., 2022), QMF (Zhang et al., 2023), COMM (Dufumier et al., 2025), CMoB (Wang et al., 2025), HierSSL (Qu et al., 2025), SPromptGL (Wan et al., 2025), and MMFL (Dong et al., 2026).

3.3 OVERALL PERFORMANCE

As shown in Table 1, RPMF achieves the best results across all reported metrics on both datasets. On TADPOLE, RPMF obtains 89.47% ACC and 95.98% AUC, outperforming the strongest competing results from MMFL by 4.70 and 1.72 percentage points, respectively. Compared with QMF, which explicitly considers modality quality during multimodal fusion, RPMF further improves both ACC and AUC, suggesting that reliability estimation can benefit from being coupled with relation-aware progressive evidence integration rather than being used solely for modality reweighting. On ABIDE, RPMF achieves 85.65% ACC and 91.77% AUC, compared with 84.38% ACC from QMF and 91.38% AUC from CMoB. It also achieves the highest SEN and SPE simultaneously, indicating a more balanced performance between the two classes. Overall, the consistent gains across two datasets support the effectiveness of RPMF in brain disease diagnosis.

3.4 ABLATION STUDY

3.4.1 DISSECTING THE ROLE OF EACH COMPONENT

We conduct ablation experiments to evaluate the contributions of the key components. Specifically, we construct six variants: (i) **w/o Reliability**, which removes reliability guidance from modality selection and fusion; (ii) **w/o Relation**, which removes cross-modal relation guidance; (iii) **w/o**

Table 1: Performance comparison on TADPOLE and ABIDE. Results are reported as mean ± standard deviation (%) over 10-fold cross-validation. The best results are shown in **bold**, and the second-best results are underlined.

	TADPOLE		ABIDE			
Method	**ACC**	**AUC**	**ACC**	**AUC**	**SEN**	**SPE**
MLP	69.56±5.43	84.78±4.48	82.43±4.96	89.32±3.89	84.81±5.23	79.66±7.54
MMGL	80.94±7.32	88.87±4.53	83.81±5.61	90.26±5.23	84.78±10.68	82.63±6.62
MD	81.74±7.97	92.60±4.16	83.58±6.20	90.13±5.41	85.19±9.55	81.62±5.38
QMF	78.59±4.78	91.45±2.93	84.38±6.80	90.02±5.96	86.93±6.97	81.38±8.77
COMM	78.90±5.53	92.79±2.84	82.66±4.60	89.06±5.27	82.88±6.25	82.42±7.22
CMoB	76.92±6.15	90.43±3.97	83.91±6.65	91.38±4.17	85.85±7.90	81.60±12.50
HierSSL	83.12±6.28	90.30±4.82	84.27±5.50	90.79±5.11	85.41±9.04	82.86±8.97
SPromptGL	78.25±4.36	90.02±3.47	84.03±5.63	89.91±4.01	86.27±8.24	81.37±7.21
MMFL	84.77±5.58	94.26±2.05	83.23±6.60	88.23±6.73	85.67±8.76	80.34±9.45
RPMF (Ours)	**89.47±2.58**	**95.98±1.69**	**85.65±3.54**	**91.77±3.55**	**86.94±6.43**	**84.10±6.03**

Prototype, which removes prototype-based representation and directly performs progressive fusion on encoder-level modality vectors, with cross-modal relations measured by negative Euclidean distance; (iv) **Weighted fusion**, which replaces the proposed progressive fusion with one-shot reliability-weighted averaging; (v) **w/o** $\mathcal{L}_{\text{aux}}$ and (vi) **w/o** $\mathcal{L}_{\text{rank}}$, which remove the corresponding loss terms.

Table 2: Ablation study results (%).

	TADPOLE		ABIDE			
Variant	**ACC**	**AUC**	**ACC**	**AUC**	**SEN**	**SPE**
RPMF	**89.47**	**95.98**	**85.65**	**91.77**	86.94	**84.10**
w/o Reliability	70.89	85.61	67.49	71.26	**90.93**	39.88
w/o Relation	79.27	90.38	74.96	84.47	86.31	61.73
w/o Prototype	83.11	93.44	81.85	89.19	86.29	76.66
Weighted fusion	80.27	92.49	82.66	91.27	82.67	82.62
w/o $\mathcal{L}_{\text{aux}}$	69.45	83.88	73.12	77.37	79.06	66.23
w/o $\mathcal{L}_{\text{rank}}$	85.45	93.70	80.13	85.40	87.35	71.79

As shown in Table 2, removing either reliability or relation guidance substantially degrades performance on both datasets, demonstrating their complementary roles in modality selection and fusion. The particularly large degradation caused by removing reliability guidance highlights the importance of subject-specific reliability estimation. Removing the prototype-based representation also consistently reduces performance. Compared with the full model, ACC decreases from 89.47% to 83.11% on TADPOLE and from 85.65% to 81.85% on ABIDE, while AUC drops from 95.98% to 93.44% and from 91.77% to 89.19%, respectively. These results indicate that representing each modality with multiple latent prototypes provides more informative structure for relation modeling. Replacing progressive fusion with one-shot reliability-weighted fusion also reduces performance, suggesting that sequentially incorporating multimodal evidence offers additional benefits over direct weighted aggregation. Finally, removing either $\mathcal{L}_{\text{aux}}$ or $\mathcal{L}_{\text{rank}}$ leads to clear performance degradation, confirming that both objectives contribute to effective reliability learning and model optimization. Overall, the ablation results support the complementary roles of reliability estimation, prototype-based relation modeling, and progressive multimodal fusion in the proposed framework.

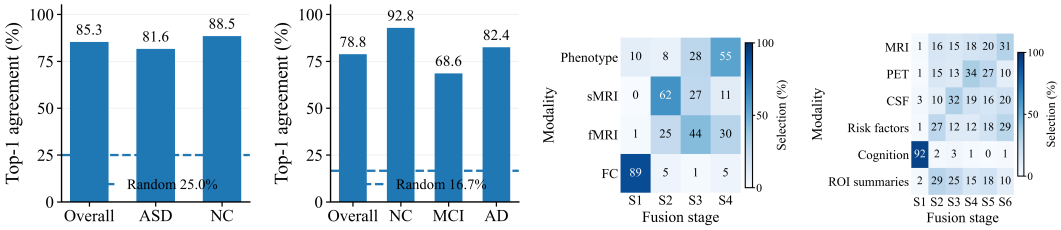
3.4.2 Analysis of Reliability–Prediction Consistency

We first evaluate whether the learned modality reliability is consistent with empirical unimodal predictive quality. Fig. 3(a)-(b) reports the Top-1 Agreement, i.e., whether the modality assigned the highest reliability is also the modality yielding the lowest unimodal classification loss. The proposed reliability estimator achieves 85.30% agreement on ABIDE and 78.76% on TADPOLE, substantially exceeding the corresponding random baselines of 25.00% and 16.67%.

Table 3 reports the within-subject Spearman rank correlation between learned reliability and unimodal predictive quality. The mean correlations reach 0.470 on ABIDE and 0.462 on TADPOLE,

Table 3: Consistency between learned modality reliability and empirical unimodal predictive quality.

Dataset	Class	N	Top-1 (%)	Random (%)	Mean ρ	Median ρ
ABIDE	Overall	871	**85.30**	25.00	0.470	0.400
	ASD	403	81.64	25.00	0.303	0.200
	NC	468	88.46	25.00	0.613	0.800
TADPOLE	Overall	598	**78.76**	16.67	0.462	0.600
	NC	209	92.82	16.67	0.797	0.886
	MCI	315	68.57	16.67	0.346	0.429
	AD	74	82.43	16.67	0.010	-0.057



(a) ABIDE (b) TADPOLE (c) ABIDE (d) TADPOLE

Figure 3: Analysis of learned modality reliability and progressive modality selection. (a)–(b) Top-1 Agreement between the modality with the highest learned reliability and the modality with the best empirical unimodal predictive quality. Dashed lines indicate random-selection baselines. (c)–(d) Stage-wise modality selection distributions, where each cell reports the percentage of subjects selecting a particular modality at the corresponding fusion stage.

showing that the learned reliability generally preserves the relative predictive quality of different modalities. Class-specific results reveal that this correspondence varies across disease groups. The AD group achieves a high Top-1 Agreement of 82.43% despite a relatively low full-ranking correlation. This result suggests that the learned reliability is particularly effective at identifying the most informative modality, while the relative ordering of the remaining modalities may be less pronounced. This is consistent with RPMF, where reliability determines the initial anchor, whereas subsequent modality integration is jointly guided by reliability and cross-modal relations.

3.4.3 Analysis of Progressive Modality Selection

We next investigate how the learned reliability is translated into actual modality selection during progressive fusion. Fig. 3(c)-(d) visualizes the stage-wise selection frequencies, where each cell denotes the proportion of subjects for which a modality is selected at a particular fusion stage. The first stage is determined by subject-specific reliability, whereas subsequent stages additionally incorporate cross-modal relation information with respect to the current fused representation. The initial selection is dominated by FC on ABIDE and cognition on TADPOLE, accounting for 88.86% and 92.31% of subjects, respectively. Nevertheless, alternative modalities are selected for subsets of subjects, reflecting subject-dependent variation in the learned reliability. More importantly, the stage-wise heatmap reveals how different modalities are progressively introduced after the initial evidence has been established, providing a visualization of the progressive fusion mechanism.

4 Conclusion

In this paper, we present RPMF, a reliability-aware progressive multimodal fusion framework for brain disease diagnosis. RPMF integrates multimodal evidence progressively according to subject-specific modality reliability and prototype-level cross-modal relations, allowing the fusion process to prioritize informative and compatible information. Experiments on TADPOLE and ABIDE demonstrate consistent improvements over representative multimodal methods, while the ablation and modality-selection analyses further support the effectiveness of the proposed components.

AI USE STATEMENT

In this work, we used generative AI tools for design or provide feedback on research methodology or experiments, implement methods, assist with translation, support qualitative and thematic data analysis. We have not used generative AI tools for help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, clean and reformat dataset, interpret results. Generate synthetic data sets, assist in the writing of proofs, propose or refine hypotheses are not applicable to this work. Additionally, we used generative AI tools for edit a research paper to improve readability, identify relevant literature. We have reviewed all AI-assisted work. All AI-assisted methodological suggestions and code were manually reviewed and verified by the authors, and suggested literature was checked against the original publications. All experimental results and conclusions were independently validated by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed cross-validation protocols, hyperparameter search spaces, and final configurations in Appendix B-C. All baselines are evaluated using identical data splits and preprocessing pipelines. Code will be released upon publication.

REFERENCES

Alexandre Abraham, Michael P Milham, Adriana Di Martino, R Cameron Craddock, Dimitris Samaras, Bertrand Thirion, and Gael Varoquaux. Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example. *NeuroImage*, 147:736–745, 2017.

Alireza Atri, Bradford C Dickerson, Carolyn Clevenger, Jason Karlawish, David Knopman, Pei-Jung Lin, Mary Norman, Chiadi Onyike, Mary Sano, Susan Scanland, et al. Alzheimer's association clinical practice guideline for the diagnostic evaluation, testing, counseling, and disclosure of suspected alzheimer's disease and related disorders (detecd-adrd): Executive summary of recommendations for primary care. *Alzheimer's & Dementia*, 21(6):e14333, 2025.

Jesse A Brown, Alex J Lee, Kristen Fernhoff, Taylor Pistone, Lorenzo Pasquini, Amy B Wise, Adam M Staffaroni, Maria Luisa Mandelli, Suzee E Lee, Adam L Boxer, et al. Functional network collapse in neurodegenerative disease. *Nature communications*, 16(1):10273, 2025.

Hongjie Cai, Yue Gao, and Manhua Liu. Graph transformer geometric learning of brain networks using multimodal mr images for brain age estimation. *IEEE Transactions on Medical Imaging*, 42(2):456–466, 2023.

Aimei Dong, Yongxing Cai, Long Wang, Jingyuan Xu, Guohua Lv, and Guixin Zhao. Multimodal multi-graph fusion learning for alzheimer's disease diagnosis. *IEEE Transactions on Multimedia*, 28:628–642, 2026.

Benoit Dufumier, Javiera Castillo Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *International Conference on Learning Representations*, volume 2025, pp. 5408–5432, 2025.

Martin Dyrba, Michel Grothe, Thomas Kirste, and Stefan J Teipel. Multimodal analysis of functional and structural disconnection in a lzheimer's disease using multiple kernel svm. *Human brain mapping*, 36(6):2118–2131, 2015.

Jingxi Feng, Heming Xu, Junhao Cai, Yujie Chang, Dong Zhang, Shaoyi Du, and Juan Wang. Cross-modal brain graph transformer via function-structure connectivity network for brain disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 247–256. Springer, 2025.

Christian Gapp, Elias Tappeiner, Martin Welk, Karl Fritscher, Elke R Gizewski, and Rainer Schubert. What are you looking at? modality contribution in multimodal medical deep learning. *International Journal of Computer Assisted Radiology and Surgery*, 21(3):507–517, 2026.

Michal Golovanevsky, Carsten Eickhoff, and Ritambhara Singh. Multimodal attention-based deep learning for alzheimer's disease diagnosis. *Journal of the American Medical Informatics Association*, 29(12):2014–2022, 2022.

Zongbo Han, Fan Yang, Junzhou Huang, Changqing Zhang, and Jianhua Yao. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20675–20685. IEEE, 2022.

Joy Higgs, Alastair Burn, and Mark Jones. Integrating clinical reasoning and evidence-based practice. *AACN Advanced Critical Care*, 12(4):482–490, 2001.

Clifford R Jack Jr, J Scott Andrews, Thomas G Beach, Teresa Buracchio, Billy Dunn, Ana Graf, Oskar Hansson, Carole Ho, William Jagust, Eric McDade, et al. Revised criteria for diagnosis and staging of alzheimer's disease: Alzheimer's association workgroup. *Alzheimer's & Dementia*, 20(8):5143–5169, 2024.

JL Liss, S Seleri Assunção, Jeffrey Cummings, A Atri, DS Geldmacher, SF Candela, DP Devanand, HM Fillit, J Susman, J Mintzer, et al. Practical recommendations for timely, accurate diagnosis of symptomatic alzheimer's disease (mci and dementia) in primary care: a review and synthesis. *Journal of internal medicine*, 290(2):310–334, 2021.

Catherine Lord, Traolach S Brugha, Tony Charman, James Cusack, Guillaume Dumas, Thomas Frazier, Emily JH Jones, Rebecca M Jones, Andrew Pickles, Matthew W State, et al. Autism spectrum disorder. *Nature reviews Disease primers*, 6(1):5, 2020.

Razvan V Marinescu, Neil P Oxtoby, Alexandra L Young, Esther E Bron, Arthur W Toga, Michael W Weiner, Frederik Barkhof, Nick C Fox, Stefan Klein, Daniel C Alexander, et al. Tadpole challenge: prediction of longitudinal evolution in alzheimer's disease. *arXiv preprint arXiv:1805.03909*, 2018.

Golrokh Mirzaei, Aaditya Gupta, and Hojjat Adeli. Data fusion of medical imaging in neurological disorders. *Reviews in the Neurosciences*, 37(1):43–60, 2026.

Asif Nawaz, Afaf Edinat, Muhammad Rizwan Rashid Rana, Tariq Ali, Ghulam Mustafa, Sidra Tahir, and Seung Won Lee. The role of multimodality in clinical disease diagnosis: advances, challenges, and opportunities. *Frontiers in Public Health*, 14:1788454, 2026.

Sarah Parisot, Sofia Ira Ktena, Enzo Ferrante, Matthew Lee, Ricardo Guerrerro Moreno, Ben Glocker, and Daniel Rueckert. Spectral graph convolutions for population-based disease prediction. In *International conference on medical image computing and computer-assisted intervention*, pp. 177–185. Springer, 2017.

Shengbing Pei, Jiajun Ma, Zhao Lv, Chao Zhang, and Jihong Guan. Community-aware graph transformer for brain disorder identification. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 4191–4199, 2025.

Zhe Qu, Taihua Chen, Xin Zhou, Fanglin Zhu, Wei Guo, Yonghui Xu, Yixin Zhang, and Lizhen Cui. Multi-modal disease prediction with hierarchical self-supervised learning. *IEEE Journal of Biomedical and Health Informatics*, 29(10):7057–7067, 2025.

Afshin Shoeibi, Marjane Khodatars, Mahboobeh Jafari, Navid Ghassemi, Parisa Moridian, Roohallah Alizadehsani, Sai Ho Ling, Abbas Khosravi, Hamid Alinejad-Rokny, H.K. Lam, Matthew Fuller-Tyszkiewicz, U. Rajendra Acharya, Donovan Anderson, Yudong Zhang, and Juan Manuel Gorriz. Diagnosis of brain diseases in fusion of neuroimaging modalities using deep learning: A review. *Information Fusion*, 93:85–117, 2023. ISSN 1566-2535.

Wenfeng Tan, Hong Yang, Shanshan Ye, Ji Zhang, Peng Zhang, and Zhi Qiao. Text-graph alignment based on llms instruction learning and wavelet graph transformers for brain disease analysis. *Neurocomputing*, 654:131355, 2025. ISSN 0925-2312.

Chaosheng Tang, Mengbo Xi, Junding Sun, Shuihua Wang, Yudong Zhang, Alzheimer's Disease Neuroimaging Initiative, et al. Macfnet: Detection of alzheimer's disease via multiscale attention and cross-enhancement fusion network. *Computer Methods and Programs in Biomedicine*, 254:108259, 2024.

Xixi Wan, Bo Jiang, Shihao Li, and Aihua Zheng. Spromptgl: Semantic prompt guided graph learning for multi-modal brain disease. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 616–626. Springer, 2025.

Jun Wang, Fuyuan Cao, Zhixin Xue, Xingwang Zhao, and Jiye Liang. Cmob: Modality valuation via causal effect for balanced multimodal learning. *Advances in Neural Information Processing Systems*, 38:82585–82607, 2025.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *International Conference on Machine Learning*, 2023.

Rong Zhang, Jinhua Sheng, Qiao Zhang, Junmei Wang, and Binbing Wang. A review of multimodal fusion–based deep learning for alzheimer's disease. *Neuroscience*, 576:80–95, 2025.

Shuai Zheng, Zhenfeng Zhu, Zhizhe Liu, Zhenyu Guo, Yang Liu, Yuchen Yang, and Yao Zhao. Multi-modal graph learning for disease prediction. *IEEE Transactions on Medical Imaging*, 41 (9):2207–2216, 2022.

A Limitations and Future Work

Despite the encouraging results, several aspects of RPMF can be further improved. First, RPMF currently adopts a greedy progressive fusion strategy, where the next modality is selected according to their reliability and relation to the current fused representation. This sequential design provides an interpretable fusion trajectory and avoids integrating heterogeneous modalities in a single step, but each selection decision is made locally and all available modalities are eventually incorporated. Consequently, the current framework does not explicitly determine whether a newly introduced modality provides sufficient additional information to justify further fusion. Future work could formulate progressive multimodal integration as a sequential decision process, in which the modality order and fusion policy are jointly optimized. In particular, a stopping mechanism could estimate the incremental utility of each newly incorporated modality and terminate the fusion process when the expected information gain becomes sufficiently small. This extension may further reduce the influence of redundant or conflicting modalities while improving computational efficiency and interpretability.

Second, the reliability and relation terms in the joint selection score are derived from different mechanisms and naturally exhibit different numerical scales. Their relative contributions are therefore controlled by manually selected coefficients, which may lead to dataset-dependent configurations and require additional tuning. One promising direction is to first transform reliability and relation scores into a comparable normalized space, and then replace the globally fixed coefficients with a lightweight gating network. Such a network could dynamically estimate their relative importance from the current subject-specific reliability profile and the evolving fused representation, enabling the balance between individual modality quality and cross-modal compatibility to vary across subjects and fusion stages.

Third, the current experiments are conducted on two benchmark datasets with a fixed set of input modalities, whereas real-world clinical data are typically more heterogeneous. Different subjects may undergo different examinations, and substantial distribution shifts can arise from acquisition protocols, scanner platforms, demographic composition, and clinical sites. These factors may alter both modality reliability and cross-modal relationships. We therefore plan to extend RPMF to incomplete-modality settings by introducing mask-aware reliability and relation estimation, such that modality selection and fusion are performed only over the observations available for each subject. The resulting framework could dynamically construct subject-specific fusion trajectories of different lengths and compositions. In parallel, evaluating RPMF under leave-site-out, cross-site, and independent external-cohort protocols will be important for assessing whether the learned reliability and progressive fusion strategies remain robust under clinically relevant distribution shifts.

Table 4: Demographic and clinical characteristics of subjects in the TADPOLE and ABIDE datasets. Continuous variables are reported as mean ± standard deviation.

TADPOLE				
Group	Gender (F/M)	Age	MMSE	MoCA
NC	114/95	72.81 ± 5.96	29.13 ± 1.11	25.93 ± 2.45
MCI	144/171	70.87 ± 7.19	28.14 ± 1.70	23.53 ± 3.10
AD	30/44	73.29 ± 7.97	22.82 ± 2.93	16.86 ± 5.06
ABIDE				
Group	Gender (F/M)	Age	Eyes Open/Closed	
NC	90/378	16.84 ± 7.23	321/147	
ASD	54/349	17.07 ± 7.95	288/115	

B Dataset Characteristics

Table 4 summarizes the demographic and clinical characteristics of the subjects included in TADPOLE and ABIDE. The TADPOLE cohort consists of 209 NC, 315 MCI, and 74 AD subjects. The three diagnostic groups have broadly comparable age distributions, while cognitive scores show the expected differences across disease stages, with the AD group exhibiting substantially lower MMSE and MoCA scores than the NC and MCI groups. The ABIDE cohort contains 468 NC and 403 ASD subjects, with similar mean ages across the two groups. These statistics provide an overview of the demographic and clinical composition of the datasets used in our experiments.

C Additional Experimental Details

We employ stratified 10-fold cross-validation to preserve the class distribution across folds. In each fold, one fold is held out as the test set, while 10% of the remaining training samples are further separated as a validation set for model selection and early stopping. The same data splits are used for all compared methods and ablation variants to ensure a fair comparison.

For all baselines, we use the publicly released implementation whenever available. All methods are evaluated using the same 10-fold cross-validation splits and preprocessing pipeline. Hyperparameters of the baseline methods are either set according to the recommended values reported in the corresponding papers or selected using the same validation protocol as our method.

For all experiments, we use a latent feature dimension of $D = 16$ and $K = 4$ prototypes. These two architectural parameters are fixed across datasets rather than selected through hyperparameter optimization. The temperature used for reliability estimation is fixed to $T = 1$. The progressive fusion module uses a single attention head, with a residual scale of 0.1. Dropout is set to 0.3, and weight decay is fixed to 3.6×10^{-3}. We optimize the model using Adam with a batch size of 256 for training. The validation and test batch sizes are set to 128. The maximum number of training epochs is 1000, and early stopping is applied according to validation accuracy with a patience of 50 epochs.

The remaining hyperparameters are selected using a staged pairwise search strategy. The candidate ranges for the two coefficients controlling reliability and relation guidance are $\beta, \eta \in \{10^{-4}, 10^{-3}, \cdots, 10^{3}, 10^{4}\}$, while the two loss weights are selected from $\lambda_1, \lambda_2 \in \{0.1, 0.2, \cdots, 0.9, 1.0\}$. Hyperparameters are optimized in pairs. At each stage, all combinations of the two currently selected hyperparameters are evaluated while the remaining parameters are fixed to their current best values. The configuration achieving the highest accuracy is retained before proceeding to the next pair. After all parameter pairs have been traversed, the complete sequence is repeated for a second round using the configuration obtained in the first round. This second pass allows parameters selected at earlier stages to be revisited after the remaining parameters have been updated. The final configuration for ABIDE is $\beta = 1 \times 10^{-4}$, $\eta = 1 \times 10^{-2}$, $\lambda_1 = 0.6$, and $\lambda_2 = 0.1$. For TADPOLE, the selected values are $\beta = 1.0$, $\eta = 10^{4}$, $\lambda_1 = 0.2$, and $\lambda_2 = 0.1$.

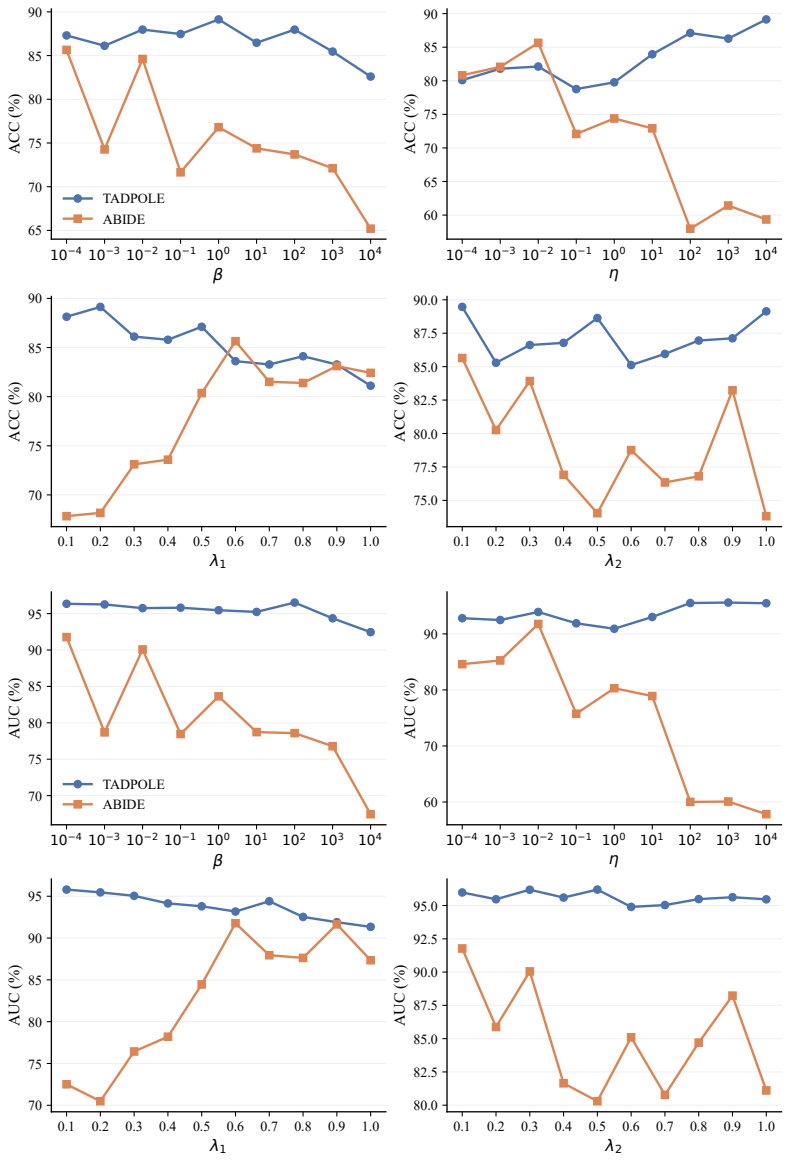


Figure 4: Hyperparameter sensitivity analysis on TADPOLE and ABIDE. The top and bottom panels show ACC and AUC, respectively. For each hyperparameter, the remaining hyperparameters are fixed.

D Hyperparameter Sensitivity Analysis

We further investigate the sensitivity of β, η, λ_1, and λ_2, as shown in Fig. 4. For β, TADPOLE maintains competitive performance over a wide range and degrades noticeably only at excessively large values. In contrast, ABIDE favors relatively small β, with performance gradually deteriorating as the weight increases. A similar dataset-dependent behavior is observed for η: relatively large values are favorable on TADPOLE, while ABIDE achieves its best performance around $\eta = 10^{-2}$ and deteriorates considerably when η becomes large. For the loss weights, TADPOLE favors relatively small λ_1, while ABIDE reaches its best performance at a moderate value around $\lambda_1 = 0.6$. Compared with the other hyperparameters, λ_2 produces relatively smaller performance variations on both datasets, indicating that the model is less sensitive to this parameter.

Overall, the model exhibits different levels of sensitivity across the two datasets. TADPOLE shows relatively stable performance over a broad range of hyperparameter values, whereas ABIDE is more sensitive to their variation. The relatively higher sensitivity observed on ABIDE may be related to its greater data heterogeneity and more complex cross-modal relationships, which can make the learned reliability and relation estimates more dependent on the strength of the corresponding regularization terms. An interesting direction for future work is therefore to replace globally fixed weighting coefficients with data-adaptive or training-aware weighting strategies, allowing the contribution of different objectives to be automatically adjusted according to dataset characteristics and modality reliability.