# FMD-IMHF: FOUNDATION MODEL DISTILLATION AND HYPERGRAPH FUSION FOR NEURODEGENERATIVE DISEASE DIAGNOSIS WITH INCOMPLETE MULTIMODAL DATA

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Accurate diagnosis of neurodegenerative diseases such as Alzheimer's disease (AD) and Parkinson's disease (PD) requires integrating complementary information from multiple modalities, including structural MRI (sMRI), PET, and clinical data. However, a limited number of labeled samples and frequently missing modalities in clinical data pose challenges to representation learning and multimodal fusion. Medical foundation models offer powerful pretrained representations for modality-specific learning, but jointly using independently pretrained models incurs substantial computational costs and leads to misaligned representation spaces. Meanwhile, existing incomplete multimodal learning approaches remain limited in exploiting complementary information from available modalities to compensate for missing information. To address these challenges, we propose FMD-IMHF, a framework that distills modality-specific foundation models into lightweight encoders and aligns their representations in a shared diagnostic semantic space. A patient-modality hypergraph further models cross-modal complementarity and inter-patient similarity, enabling incomplete multimodal fusion without reconstructing missing modalities. Experiments on ADNI demonstrate the effectiveness of FMD-IMHF for AD diagnosis and MCI progression prediction, as well as its robustness under varying missing-modality rates. Evaluations on AIBL, SCAN, and PPMI further demonstrate its generalizability. Distillation reduces parameter counts, inference time, and peak GPU memory, while ablation studies validate the key components of FMD-IMHF.

## 1 INTRODUCTION

Neurodegenerative diseases such as Alzheimer's disease (AD) and Parkinson's disease (PD) are often characterized by subtle early symptoms and slow progression (Knopman et al., 2021). Accurate diagnosis therefore typically requires integrating complementary information on brain structure, metabolic function, and clinical phenotypes to capture subtle disease-related abnormalities from different perspectives. Structural magnetic resonance imaging (sMRI), fluorodeoxyglucose positron emission tomography (FDG-PET), and clinical data provide disease-related evidence from these different perspectives, enabling a more comprehensive characterization of the disease (Knopman et al., 2021; Acosta et al., 2022).

Rapid development in deep learning technologies has significantly improved multimodal disease diagnosis. However, the clinical application of these technologies remains constrained by several challenges. For example, the limited availability of clinical samples and heterogeneous data distributions makes it difficult to learn robust and generalizable disease representations (Liu et al., 2023b). Additionally, the frequent absence of one or more modalities in patient samples limits the effectiveness of multimodal fusion methods that rely on complete inputs (Yao et al., 2024).

Medical foundation models provide a promising approach for obtaining robust disease representations from limited clinical samples by leveraging knowledge acquired through large-scale pretraining (Sun et al., 2025; Lin et al., 2026; Hollmann et al., 2025). However, in multimodal settings,



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.

jointly deploying multiple foundation models incurs substantial computational and memory overhead. Moreover, because these models are pretrained independently, their representation spaces are not aligned, making cross-modal integration challenging. Knowledge distillation can transfer discriminative knowledge from foundation models to lightweight encoders, thereby reducing model complexity (Yang et al., 2025). But independent intra-modal distillation still leaves cross-modal representation space misalignment untackled. This motivates integrating knowledge transfer with cross-modal semantic alignment to learn lightweight, discriminative, and semantically consistent multimodal representations for effective fusion.

Meanwhile, subsequent multimodal fusion remains challenging when one or more modalities are missing. Existing methods mainly rely on flexible fusion (Liu et al., 2017; 2023b) or missing-modality generation (Chen et al., 2023; Wang et al., 2024). The former directly integrates available modalities but has limited ability to compensate for missing information. The latter reconstructs missing modalities or features but may introduce unreliable generated information into downstream classification. Importantly, available multimodal data contain relational information that can potentially compensate for missing information without explicit reconstruction. Such information arises from both cross-modal relationships within individual patients and similarity relationships across patients within each modality. Exploiting these complementary relationships therefore provides a natural alternative to explicit missing-modality generation.

To address these challenges, we propose FMD-IMHF (Foundation Model Distillation and Incomplete Multimodal Hypergraph Fusion), a multimodal learning framework for neurodegenerative disease diagnosis with incomplete multimodal data. In this framework, a semantically aligned foundation model distillation method first transfers discriminative knowledge from modality-specific foundation models to lightweight student encoders while aligning their representations in a shared diagnostic semantic space. A patient-modality hypergraph is then constructed to jointly model cross-modal complementarity within patients and inter-patient similarity within each modality. At inference time, available multimodal data are encoded by lightweight student encoders and incorporated into the patient-modality hypergraph. Higher-order propagation aggregates information from related patient-modality nodes to compensate for missing information, enabling prediction without reconstructing or generating missing modalities.

We evaluate FMD-IMHF on ADNI for AD diagnosis and MCI progression prediction using sMRI, FDG-PET, and clinical data, obtaining bACC scores of 0.966 and 0.793, respectively. We further assess robustness under varying missing-modality rates and generalization to AIBL, SCAN, and PPMI. For MCI progression prediction, FMD-IMHF achieves bACC scores of 0.823 and 0.808 on AIBL and SCAN without target-domain retraining; after retraining on PPMI, it achieves 0.773 for PD progression prediction. Relative to their respective modality-specific teachers, the distilled student encoders reduce parameter counts by 35.1%–93.0%, per-modality inference time by 56.7%–88.9%, and peak GPU memory by 85.1%–92.8%. Ablation studies further demonstrate the effectiveness of cross-modal semantic alignment and patient–modality hypergraph modeling.

## 2 Related Works

### 2.1 Foundation Models and Multimodal Fusion

Foundation models provide transferable representations through large-scale pretraining and have been increasingly applied to medical imaging and clinical tasks (Ma et al., 2024; Toma et al., 2023; Li et al., 2025). In brain imaging, modality-specific foundation models have been developed for different imaging modalities, including BrainMVP (Rui et al., 2025), BrainLM (Caro et al., 2024), and Brain-JEPA (Dong et al., 2024) for MRI or fMRI, and SegAnyPET (Zhang et al., 2025) for 3D PET imaging. These studies demonstrate the potential of foundation models for downstream medical representation learning.

However, modality-specific foundation models are typically pretrained independently, resulting in heterogeneous representation spaces and high computational overhead when used jointly for multimodal fusion (Girdhar et al., 2023; Gu et al., 2026). Existing studies have explored cross-model alignment, such as modality-aware Q-Former (Gu et al., 2026), but still limited to full-scale large foundation models. How to retain the diagnostic knowledge of modality-specific foundation models

while optimising computational overhead and achieving cross-modal semantic alignment remains insufficiently explored.

### 2.2 Incomplete Multimodal Learning in Medical Applications

Missing modalities in patient data are common in clinical practice due to examination costs, acquisition constraints, and data availability. Existing methods mainly follow two directions. Flexible-modality approaches directly fuse available modalities, such as 3MT and HEALNet (Liu et al., 2023b; Hemker et al., 2024), but provide limited compensation for missing information. Modality reconstruction methods instead recover missing images or features from available modalities, such as DFTD, M3AE, M3FeCon, and DrFuse (Chen et al., 2023; Liu et al., 2023a; Zeng et al., 2024; Yao et al., 2024), but may introduce reconstruction errors. Recently, developments in structural relation modeling approaches have provided another direction for incomplete multimodal learning. For example, MUSE models patient–modality associations using a bipartite graph (Wu et al., 2024), while Graph-SLC exploits similarity among subjects (Ou et al., 2024). These studies demonstrate the value of relational information but model different types of relationships separately. In contrast, our approach jointly captures intra-patient cross-modal complementarity and within-modality inter-patient similarity in a unified hypergraph.

## 3 Method

As illustrated in Fig. 1, FMD-IMHF consists of two stages: semantically aligned foundation-model distillation and incomplete multimodal hypergraph fusion. First, modality-specific foundation models are adapted to the target task and distilled into lightweight student models, while the student representations are simultaneously aligned in a shared diagnostic semantic space. Second, a patient–modality hypergraph is constructed from the aligned representations to jointly model intra-patient cross-modal complementarity and inter-patient similarity within each modality, enabling higher-order information propagation under incomplete multimodal settings. During inference, available modalities are encoded by student models and incorporated into the hypergraph, where higher-order propagation aggregates information from related patient–modality nodes to compensate for missing information without reconstructing or generating missing modalities.

### 3.1 Foundation Model Task Adaptation

We employ a modality-specific pretrained foundation model for each modality and adapt it to the target diagnostic tasks using LoRA (Hu et al., 2022). This lightweight adaptation introduces only a small number of trainable parameters while preserving most pretrained knowledge. Each foundation model is independently optimized on the training samples available for its corresponding modality using a cross-entropy classification objective. The resulting features are mapped to a unified 512-dimensional space through modality-specific pooling and projection layers. After adaptation, all foundation models are frozen and serve as teachers, providing task-specific representations and predictive distributions for subsequent distillation.

### 3.2 Semantically Aligned Multi-Foundation-Model Distillation

To reduce the computational overhead of jointly deploying multiple foundation models and address cross-modal semantic misalignment, we propose a semantically aligned multi-foundation-model distillation method that combines intra-modal knowledge transfer with cross-modal semantic alignment. Lightweight student models are jointly optimized through teacher supervision and cross-modal semantic alignment in a shared diagnostic semantic space, thereby improving cross-modal consistency while preserving discriminative structure.

**Intra-modal Knowledge Distillation.** For each modality $m \in \mathcal{M}$, the task-adapted foundation model is used as the teacher, and a lightweight student model is constructed accordingly. MRI and PET employ separate 3D ResNet-10 students, while clinical data are converted into text and distilled from Qwen2.5-7B to Qwen2.5-0.5B, with the resulting representations projected to 512 dimensions. For patient $i$, the teacher and student representations are

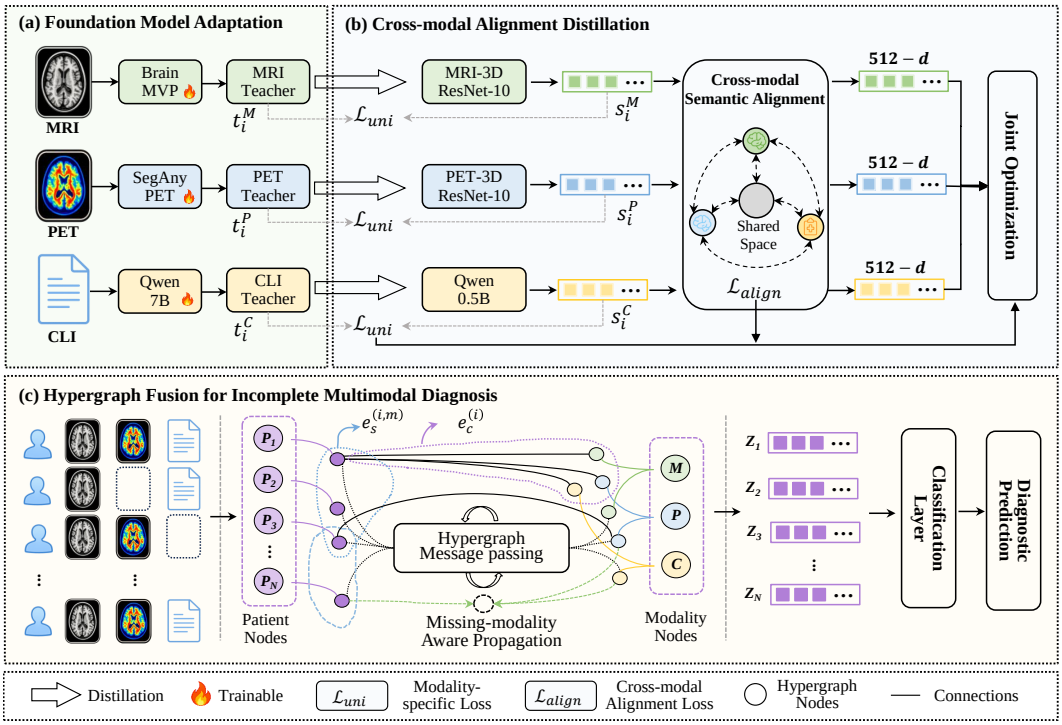


Figure 1: Overall framework of FMD-IMHF. (a) Modality-specific foundation model adaptation. (b) Cross-modal alignment distillation to lightweight students. (c) Patient–modality hypergraph fusion with higher-order message passing for diagnosis.

$$t_i^{(m)} = T_m(x_i^{(m)}), \qquad s_i^{(m)} = S_m(x_i^{(m)}), \tag{1}$$

where $t_i^{(m)}, s_i^{(m)} \in \mathbb{R}^{512}$. The teacher is frozen during distillation, and the student is jointly supervised by the ground-truth labels, teacher predictions, and teacher representations:

$$\mathcal{L}_{\text{uni}} = \sum_{m \in \mathcal{M}} \left( \lambda_{\text{cls}} \mathcal{L}_{\text{cls}}^{(m)} + \lambda_{\text{KD}} \mathcal{L}_{\text{KD}}^{(m)} + \lambda_{\text{feat}} \mathcal{L}_{\text{feat}}^{(m)} \right). \tag{2}$$

where $\mathcal{L}_{\text{cls}}$ is the cross-entropy loss with ground-truth labels, $\mathcal{L}_{\text{KD}}$ is the temperature-scaled KL-divergence from the teacher to the student predictive distribution (Zhao et al., 2022), and $\mathcal{L}_{\text{feat}}$ is the cosine-distance loss between their representations. All losses are computed only on samples with the corresponding modality. For the clinical modality, $\mathcal{L}_{\text{KD}}$ is computed over diagnostic class probabilities rather than token-level predictions.

**Cross-Modal Semantic Alignment.** Intra-modal knowledge distillation preserves modality-specific discriminative information but does not explicitly align student representations across modalities. We therefore introduce cross-modal semantic alignment to learn a shared diagnostic semantic structure (Girdhar et al., 2023).

Specifically, the student models are supervised by the corresponding teacher models using the intra-modal distillation loss, while cross-modal constraints are simultaneously imposed on their projected 512-dimensional representations. Representations from different available modalities of the same patient are pulled closer to improve cross-modal semantic consistency, while representations from samples with different diagnostic labels are pushed apart to preserve discriminative structure.

For patient $i$, positive pairs are constructed from the available modality-specific student representations $s_i^{(m)}$, and cross-modal consistency is encouraged by maximizing their feature similarity:

$$\mathcal{L}_{\text{pos}} = \frac{1}{|\mathcal{S}^+|} \sum_{(i,m,n)\in\mathcal{S}^+} \left[1 - \text{sim}\left(s_i^{(m)}, s_i^{(n)}\right)\right], \tag{3}$$

where $\mathcal{S}^+$ denotes the set of positive pairs formed by different available modalities ($m \neq n$), and $\text{sim}(\cdot,\cdot)$ denotes cosine similarity. To prevent representations from different diagnostic classes from becoming overly close in the shared space, negative pairs are constructed from samples with different labels and constrained only when their similarity exceeds a predefined margin:

$$\mathcal{L}_{\text{neg}} = \frac{1}{|\mathcal{S}^-|} \sum_{(i,m,j,n)\in\mathcal{S}^-} \max\left(0, \text{sim}\left(s_i^{(m)}, s_j^{(n)}\right) - \delta\right). \tag{4}$$

where $\mathcal{S}^-$ contains pairs of available student representations from patients with different diagnostic labels and different modalities ($m \neq n$), and $\delta$ is the similarity margin. The cross-modal semantic alignment loss is then defined as $\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{pos}} + \lambda_{\text{neg}}\mathcal{L}_{\text{neg}}$, where $\lambda_{\text{neg}}$ balances the positive and negative constraints. Under missing-modality settings, only available representations participate in pair construction, without generating or imputing missing modalities.

The overall objective of this stage is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{uni}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}$, where $\mathcal{L}_{\text{uni}}$ transfers modality-specific knowledge from the teacher models to the corresponding student models, while $\mathcal{L}_{\text{align}}$ regularizes the semantic relationships among student representations across modalities. Joint optimization enables the student models to preserve modality-specific discriminative knowledge while learning semantically consistent cross-modal representations.

### 3.3 Patient–Modality Hypergraph Construction

**Patient–Modality Hypergraph Construction.** We construct a patient–modality hypergraph in which each available patient–modality pair is represented as a node. Two types of hyperedges are introduced to model cross-modal complementarity within patients and representation similarity across patients within each modality, enabling incomplete multimodal fusion through higher-order information propagation without reconstructing missing modalities (Chien et al., 2022).

**Node Definition.** Let $s_i^m \in \mathbb{R}^d$ denote the aligned student representation of patient $i$ for modality $m$, where $d = 512$ and $m \in \mathcal{M} = \{M, P, C\}$ correspond to MRI, PET, and clinical data, respectively. We define a modality-availability indicator $a_i^m \in \{0, 1\}$, where $a_i^m = 1$ indicates that modality $m$ is available for patient $i$. Each available patient–modality pair is represented as a node $v_i^m$, with its initial node feature defined in Sec. 3.4. The node set is defined as $\mathcal{V} = \{v_i^m \mid a_i^m = 1, \forall i \in \{1, \ldots, N\}, m \in \mathcal{M}\}$.

**Hyperedge Construction.** To jointly model within-modality patient similarity and cross-modal complementarity, we construct two types of hyperedges, $\mathcal{E} = \mathcal{E}_{\text{s}} \cup \mathcal{E}_{\text{c}}$. For intra-modality hyperedges $\mathcal{E}_{\text{s}}$, each node $v_i^m$ is connected to its $k$ nearest neighbors according to cosine similarity in the same modality representation space: $e_{\text{s}}^{(i,m)} = \{v_i^m\} \cup \{v_j^m | j \in \mathcal{N}_k^m(i)\}$, where $\mathcal{N}_k^m(i)$ denotes the $k$ most similar patients with available modality $m$. For validation and test subjects, nearest neighbors are retrieved only from the training set, and no direct kNN connections are constructed among evaluation subjects. If fewer than $k$ candidates are available, all candidates are included. These hyperedges aggregate information from patients with similar representations within the same modality. For inter-modality hyperedges $\mathcal{E}_{\text{c}}$, all available modality nodes of the same patient are connected: $e_{\text{c}}^{(i)} = \{v_i^m | a_i^m = 1, m \in \mathcal{M}\}$.

These hyperedges aggregate complementary information across the available modalities of each patient and naturally adapt to different modality-availability patterns. When only one modality is available, the cross-modal hyperedge degenerates to a singleton, while the patient can still receive information from similar patients through the within-modality hyperedges.

### 3.4 Hypergraph-Based Incomplete Multimodal Learning

Based on the patient–modality hypergraph, we perform hypergraph message passing to integrate information across patients and modalities. For each available node $v_i^m$, the student representation

$s_i^m$ is concatenated with its corresponding student predictive probability vector $p_i^m$ to form the initial node feature: $x_i^m = [s_i^m; p_i^m]$, and all node features constitute $X^{(0)}$. Let $H$ denote the node–hyperedge incidence matrix. We adopt the normalized hypergraph propagation operator (Feng et al., 2019):

$$A_{\mathcal{H}} = D_v^{-\frac{1}{2}} H D_e^{-1} H^\top D_v^{-\frac{1}{2}}, \tag{5}$$

where $D_v$ and $D_e$ are the node- and hyperedge-degree matrices, respectively. The first propagation layer applies the same transformation without a residual connection, while the subsequent layers are defined as

$$X^{(l+1)} = X^{(l)} + \mathrm{Dropout}\left(\mathrm{ELU}\left(\mathrm{Linear}_l(A_{\mathcal{H}} X^{(l)})\right)\right). \tag{6}$$

This operator first aggregates node information into hyperedges through $H^\top$ and then propagates the aggregated information back to associated nodes through $H$. Within-modality hyperedges propagate information among similar patients, while cross-modal hyperedges integrate the available modalities of the same patient. We employ three propagation layers with residual connections in the second and third layers, allowing cross-modal information aggregated within a patient to further propagate to similar patients and thereby provide indirect information compensation under missing-modality settings.

Given the final propagated node representation $h_i^m$ obtained after the three propagation layers, we aggregate the available modalities of patient $i$ using attention pooling with modality-specific biases:

$$\alpha_i^m = \frac{\exp(g(h_i^m) + b_m)}{\sum_{n \in \mathcal{M}_i} \exp(g(h_i^n) + b_n)}, \qquad z_i = \sum_{m \in \mathcal{M}_i} \alpha_i^m h_i^m, \tag{7}$$

where $\mathcal{M}_i = \{m : a_i^m = 1\}$, $g(\cdot)$ is a learnable scoring function, and $b_m$ is a modality-specific bias. The attention weights are normalized only over the available modalities, allowing the pooling process to adapt to different missing-modality patterns. Finally, $z_i$ is fed into a classification head for diagnostic prediction. During this stage, the student models are frozen, and only the hypergraph propagation, pooling, and classification modules are optimized.

# 4 Experiments

## 4.1 Data and Preprocessing

We conduct main experiments on ADNI using T1-weighted MRI, FDG-PET, and clinical data for AD diagnosis (AD vs. NC) and MCI progression prediction (sMCI vs. pMCI). AIBL and SCAN are used for cross-domain evaluation, while PPMI is used for PD extension. Following prior studies (Dong et al., 2026; Gu et al., 2026), MCI and PD subjects are categorized as stable or progressive according to 36-month follow-up. MRI and PET volumes undergo preprocessing and are resampled to $128^3$. Clinical variables are organized into unified text templates for language-model encoding, while diagnosis-related variables, such as MMSE and CDR-SB are excluded to avoid potential label leakage. Detailed statistics and preprocessing procedures are provided in Appendices A.

## 4.2 Experimental Settings

For all ADNI experiments, we use subject-level stratified five-fold cross-validation, with each fold using 20% of subjects for testing and the remaining subjects split 8:2 for training and validation. We evaluate complete, naturally incomplete, and simulated missing-modality settings. The complete setting uses subjects with MRI, PET, and clinical data, while the naturally incomplete setting preserves the original ADNI modality availability. For simulated missingness, MRI, PET, and clinical modalities are independently dropped at test time, with at least one modality retained per subject. Missing modalities are neither reconstructed nor imputed.

Table 1: Comparison Results under Complete and Naturally Incomplete Multimodal Settings on ADNI.

**(a) Complete multimodal setting**

| Method | AD vs. NC | | sMCI vs. pMCI | |
|---|---|---|---|---|
| | AUC | bACC | AUC | bACC |
| MRI-only | 0.936±0.015 | 0.868±0.025 | 0.774±0.053 | 0.722±0.057 |
| PET-only | 0.968±0.016 | 0.901±0.024 | 0.830±0.022 | 0.761±0.038 |
| Clinical-only | 0.806±0.028 | 0.721±0.040 | 0.717±0.042 | 0.633±0.035 |
| TMC | 0.815±0.116 | 0.733±0.116 | 0.655±0.031 | 0.620±0.026 |
| MMIN | 0.733±0.054 | 0.651±0.093 | 0.666±0.029 | 0.624±0.028 |
| ShaSpec | 0.736±0.032 | 0.660±0.038 | 0.660±0.027 | 0.605±0.049 |
| DFTD | 0.762±0.089 | 0.720±0.068 | 0.644±0.030 | 0.611±0.057 |
| 3MT | 0.993±0.013 | 0.967±0.032 | 0.850±0.033 | 0.748±0.049 |
| DrFuse | 0.735±0.041 | 0.697±0.053 | 0.656±0.037 | 0.626±0.054 |
| HEALNet | 0.736±0.040 | 0.677±0.061 | 0.636±0.030 | 0.616±0.056 |
| Flex-MoE | 0.775±0.062 | 0.702±0.073 | 0.659±0.037 | 0.596±0.047 |
| IMDL | 0.704±0.064 | 0.646±0.066 | 0.645±0.028 | 0.596±0.026 |
| Q-Former | 0.969±0.015 | 0.886±0.041 | 0.839±0.040 | **0.752±0.039** |
| Ours | **0.996±0.007** | **0.976±0.011** | **0.850±0.031** | 0.750±0.044 |

**(b) Naturally incomplete multimodal setting**

| Method | AD vs. NC | | sMCI vs. pMCI | |
|---|---|---|---|---|
| | AUC | bACC | AUC | bACC |
| TMC | 0.916±0.014 | 0.838±0.018 | 0.705±0.039 | 0.633±0.042 |
| MMIN | 0.810±0.033 | 0.734±0.019 | 0.695±0.062 | 0.638±0.065 |
| ShaSpec | 0.851±0.080 | 0.775±0.075 | 0.711±0.066 | 0.644±0.037 |
| DFTD | 0.841±0.055 | 0.762±0.051 | 0.703±0.042 | 0.623±0.021 |
| 3MT | **0.990±0.004** | 0.959±0.010 | 0.828±0.021 | 0.754±0.038 |
| DrFuse | 0.863±0.077 | 0.783±0.074 | 0.712±0.069 | 0.629±0.047 |
| HEALNet | 0.790±0.022 | 0.713±0.025 | 0.706±0.028 | 0.635±0.022 |
| Flex-MoE | 0.806±0.026 | 0.717±0.032 | 0.718±0.065 | 0.651±0.051 |
| IMDL | 0.799±0.042 | 0.729±0.049 | 0.680±0.042 | 0.635±0.054 |
| Q-Former | 0.956±0.011 | 0.885±0.031 | 0.854±0.032 | 0.761±0.048 |
| Ours | **0.990±0.006** | **0.966±0.013** | **0.863±0.033** | **0.793±0.026** |

Table 2: Cross-domain generalization and disease extension results.

| Method | AIBL (sMCI vs. pMCI) | | | | SCAN (sMCI vs. pMCI) | | | | PPMI (sPD vs. pPD) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | bACC | SEN | SPE | ACC | bACC | SEN | SPE | ACC | bACC | SEN | SPE |
| MRI-only | 0.739 | 0.735 | 0.700 | 0.769 | 0.722 | 0.720 | 0.714 | 0.726 | 0.649 | 0.629 | 0.577 | 0.680 |
| PET-only | 0.783 | 0.785 | 0.800 | 0.769 | 0.722 | 0.710 | 0.679 | 0.742 | - | - | - | - |
| Clinical-only | 0.652 | 0.658 | 0.700 | 0.615 | 0.589 | 0.584 | 0.571 | 0.597 | 0.601 | 0.594 | 0.577 | 0.611 |
| Q-Former | 0.783 | 0.785 | 0.800 | 0.769 | **0.789** | 0.798 | 0.821 | **0.774** | 0.770 | 0.742 | 0.673 | 0.811 |
| Ours | **0.826** | **0.823** | **0.800** | **0.846** | **0.789** | **0.808** | **0.857** | 0.758 | **0.796** | **0.773** | **0.720** | **0.827** |

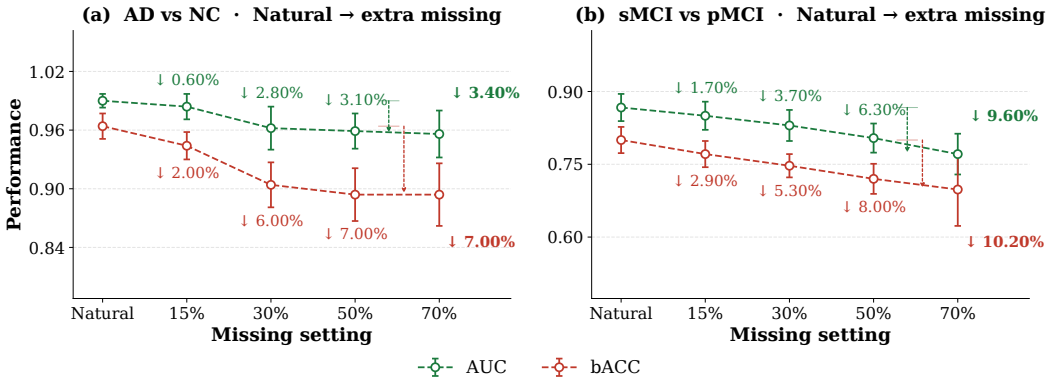


Figure 2: Performance of FMD-IMHF under different missing-modality rates on AD diagnosis and MCI progression prediction.

For MRI and FDG-PET, the UniFormer-S encoder of BrainMVP (Rui et al., 2025) and the ViT-B encoder of SegAnyPET (Zhang et al., 2025) are used as modality-specific teachers and distilled into separate 3D ResNet-10 student encoders (He et al., 2016). For clinical data, Qwen2.5-7B (Qwen et al., 2024) is used as the teacher and distilled into Qwen2.5-0.5B. All modalities produce 512-dimensional representations for cross-modal learning and subsequent patient–modality hypergraph fusion. During hypergraph fusion, the student models are frozen, and only the hypergraph, aggregation, and classification modules are optimized. Further details on evaluation metrics and hyperparameter settings are provided in Appendix B.

### 4.3 Main Results

**(1) Comparison on ADNI.** We evaluate AD vs. NC and sMCI vs. pMCI using the same stratified 5-fold splits under both complete three-modality and naturally incomplete settings, reporting mean ± standard deviation across folds. Unimodal baselines use the distilled student models before cross-modal alignment. As shown in Table 1, we compare FMD-IMHF with representative multimodal methods, including TMC (Han et al., 2021), MMIN (Zhao et al., 2021), ShaSpec (Wang et al., 2023), DFTD (Chen et al., 2023), 3MT (Liu et al., 2023b), DrFuse (Yao et al., 2024), HEALNet (Hemker et al., 2024), Flex-MoE (Yun et al., 2024), IMDL (Han et al., 2026), and Q-Former (Gu et al.,

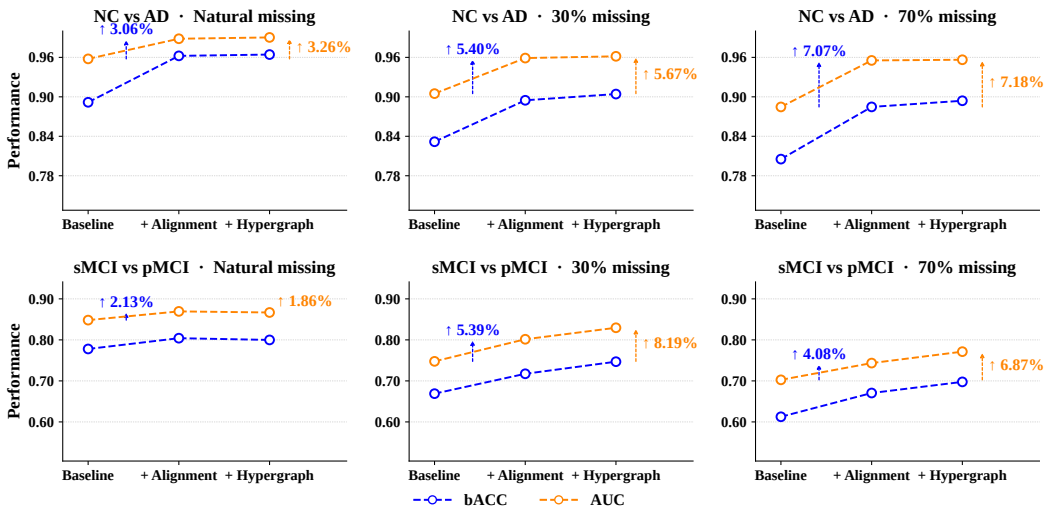


Figure 3: Ablation analysis of key components in FMD-IMHF under natural and simulated missing-modality settings on AD diagnosis and MCI progression prediction.

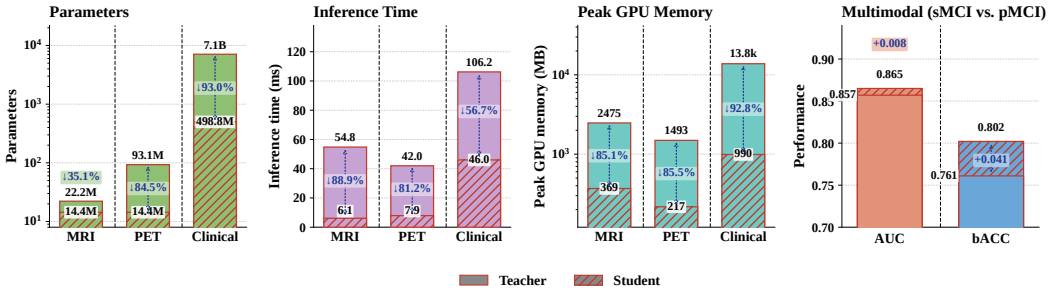


Figure 4: Computational efficiency before and after foundation-model distillation.

2026). All compared methods are retrained based on their official implementations using the same subject splits and available input modalities. Full results with AUC, bACC, ACC, SEN, and SPE are provided in Appendix C.2.

**(2) Missing-Modality Settings.** To evaluate robustness to modality missingness, we randomly remove available modality nodes from naturally incomplete ADNI data with probabilities of 15%, 30%, 50%, and 70%, while retaining at least one modality per subject. The model is trained under natural missingness and directly evaluated at all missing rates without retraining. The training graph and kNN neighbor bank remain fixed, and evaluation subjects are connected only to training subjects. As shown in Fig. 2, FMD-IMHF maintains robust performance as modality missingness increases, evaluated by AUC and bACC.

**(3) Cross-Domain Generalization and Disease Extension.** To evaluate cross-domain generalization, we train FMD-IMHF on ADNI and directly evaluate it on the independent AIBL and SCAN cohorts for sMCI vs. pMCI prediction without target-domain retraining. To assess disease-level extensibility, we further retrain the same framework on the PPMI dataset using the available MRI and clinical modalities for PD progression prediction. The detailed results are summarized in Table 2.

### 4.4 Ablation and Further Analysis

**(1) Effectiveness of Key Components.** To assess the effectiveness of the key components in FMD-IMHF, we compare three progressive variants: (1) **Baseline**, which concatenates the distilled modality-specific representations for classification; (2) **+ Alignment**, which introduces cross-modal

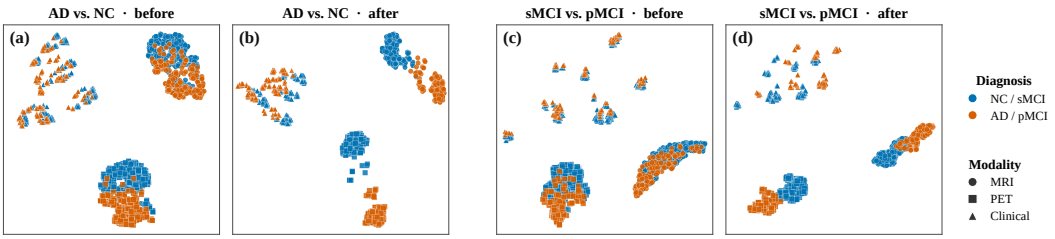


Figure 5: Visualization of multimodal feature distributions before and after cross-modal semantic alignment.

semantic alignment; and (3) **+ Hypergraph**, which further incorporates patient–modality hypergraph fusion, forming the full FMD-IMHF. We further evaluate all variants under different missing-modality settings. As shown in Fig. 3, semantic alignment consistently improves the distilled baseline, while hypergraph fusion provides further benefits particularly under higher missing-modality rates, demonstrating their complementary roles in multimodal representation learning and incomplete multimodal fusion. Further analyses of hyperedge components and the $k$ sensitivity are presented in Appendix C.1 and Appendix C.3.

**(2) Analysis of Distillation Efficiency.** Compared with directly deploying multiple modality-specific foundation models, FMD-IMHF reduces deployment cost through knowledge distillation. As shown in Fig. 4, the MRI and PET teachers are compressed from 22.2M and 93.1M to 14.4M parameters, respectively, while the clinical model is reduced from 7B to 498.8M, substantially lowering inference time and peak GPU memory while maintaining comparable predictive performance. In contrast, Q-Former (Gu et al., 2026) jointly deploy multiple foundation models, including LLaMA2-13B and BrainMVP, together with a 63.82M-parameter Q-Former for alignment. Thus, FMD-IMHF reduces the computational overhead of multimodal deployment.

**(3) Effect of Cross-Modal Semantic Alignment.** We use t-SNE to visualize the 512-dimensional multimodal representations before and after alignment. Fig. 5 shows more consistent cross-modal structures and clearer class separation after alignment. This agrees with the ablation results and supports the effectiveness of cross-modal semantic alignment.

**(4) Modality Ablation.**

We remove each modality at test time to assess its contribution to multimodal prediction. As shown in Fig. 6, MRI contributes most to AD diagnosis, whereas PET has the largest impact on MCI progression prediction. Clinical data provide smaller but consistent complementary benefits, likely because only demographic and genetic variables are used, while disease-related cognitive scores such as MMSE are excluded to avoid information leakage.

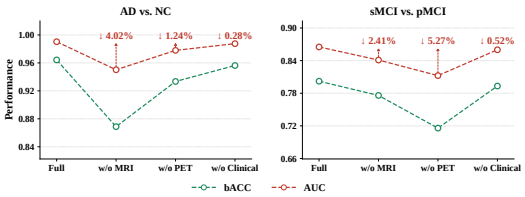


Figure 6: Modality ablation results.

## 5 Conclusion

We propose FMD-IMHF for incomplete multimodal neurodegenerative disease diagnosis. It distills modality-specific pretrained knowledge into lightweight, semantically aligned student models and constructs a patient–modality hypergraph to capture cross-modal complementarity and inter-patient similarity without reconstructing missing modalities. Experiments on AD diagnosis, MCI progression prediction, and additional PD evaluation demonstrate effective multimodal fusion, robustness to missing modalities, reduced deployment cost, and cross-domain generalization. However, the framework depends on pretrained models and the availability of clinical multimodal data. Future work will explore more efficient adaptation strategies and broader clinical validation.

## AI USE STATEMENT

In this study, we used generative AI tools as auxiliary tools to support various aspects of our research process. Specifically, we used them to provide feedback on research methodology and experiments, including discussions on methodology, experimental design, and potential improvements. We also used generative AI tools to assist with implementing research methods, including code writing, code modification, and debugging. In addition, we used them for translation and language refinement. We did not use generative AI tools for generating synthetic datasets, cleaning and reformatting datasets, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of mathematical proofs, developing theoretical models or conceptual frameworks, proposing or refining research hypotheses, or automatically conducting experiments and generating experimental results. Additionally, we used generative AI tools for supplementary purposes, including summarizing or analyzing existing literature, brainstorming research ideas and experimental designs, sourcing and searching for relevant information, and editing the manuscript to improve readability and clarity.

We used generative AI tools only as auxiliary tools. All research decisions, implementations, experiments, and conclusions were made and verified by us. We reviewed all AI-assisted content, including text, code, and analytical suggestions. AI-assisted text and translations were carefully revised to ensure accuracy and academic quality, AI-assisted code was tested and verified to ensure correctness, and AI-assisted methodological suggestions and literature information were manually evaluated based on relevant literature and experimental evidence. We take full responsibility for the final content of this work.

## ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. No human or animal experiments were conducted in this study. All datasets were publicly available and utilized in accordance with their respective guidelines, with appropriate measures taken to ensure privacy protection. We are committed to upholding ethical standards and promoting the responsible use of AI.

## REPRODUCIBILITY STATEMENT

The cohorts used in this work, including ADNI (`https://adni.loni.usc.edu/`), AIBL (`https://aibl.csiro.au/`), SCAN (`https://scan.naccdata.org/`) for Alzheimer's disease-related tasks, and PPMI (`https://www.ppmi-info.org/`) for Parkinson's disease progression prediction, are publicly available from their respective official sources. Detailed data selection criteria and preprocessing procedures are provided in the Appendix. The model architecture, algorithmic details, data splits, training configurations, and hyperparameters are described in the paper and Appendix. The implementation code is available at an anonymous repository: `https://anonymous.4open.science/r/ICLR_code-B819/`.

## REFERENCES

Jesse N. Acosta, Guido J. Falcone, Pranav Rajpurkar, and Eric J. Topol. Multimodal biomedical ai. *Nature Medicine*, 28(9):1773–1784, 2022. doi: 10.1038/s41591-022-01981-2.

Jose O. Caro, Andre H. de Oliveira Fonseca, Saba A. Rizvi, Matteo Rosati, Caroline L. Averill, Jennifer L. Cross, Pankaj Mittal, Enrico Zappala, Rakesh M. Dhodapkar, Chadi Abdallah, and David van Dijk. Brainlm: A foundation model for brain activity recordings. In *International Conference on Learning Representations (ICLR)*, 2024.

Yu Chen, Yifei Pan, Yan Xia, and Yixuan Yuan. Disentangle first, then distill: A unified framework for missing modality imputation and alzheimer's disease diagnosis. *IEEE Transactions on Medical Imaging*, 42(12):3566–3578, 2023. doi: 10.1109/TMI.2023.3295489.

Eli Chien, Chao Pan, Jane Peng, and Olgica Milenkovic. You are allset: A multiset learning framework for hypergraph neural networks. In *International Conference on Learning Representations (ICLR)*, 2022.

Zengjie Dong, Hong Liu, Xinting Ge, Hongkuan Zhang, Zhengxu Li, Yafan Chen, and Wenhao Li. Brain region-centered multimodal hypergraph fusion for mci conversion prediction. *IEEE Journal of Biomedical and Health Informatics*, pp. 1–14, 2026. doi: 10.1109/JBHI.2026.3715131.

Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Su Xian Chong, Fang Ji, Nathanael Ren Jie Tong, Christopher Li Hsian Chen, and Juan Helen Zhou. Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 86048–86073. Curran Associates, Inc., 2024. doi: 10.52202/079017-2732. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9c3828adf1500f5de3c56f6550dfe43c-Paper-Conference.pdf.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3558–3565, 2019.

Rohit Girdhar, Alaa El-Nouby, Zhuang Liu, Mannat Singh, Kalyan V. Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190, 2023.

Wentao Gu, Yuquan Li, Xinyang Jiang, Zilong Wang, Dongsheng Li, Zehui Li, Zijian Dong, and Cai Zhao. Joint adaptation of uni-modal foundation models for multi-modal alzheimer's disease diagnosis. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 14667–14693, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/186094ef879dcbd5ef6879782916c309-Paper-Conference.pdf.

Kangfu Han, Dan Hu, Fenqiang Zhao, Tianming Liu, Feng Yang, and Gang Li. Incomplete multimodal disentanglement learning with application to alzheimer's disease diagnosis. *IEEE Transactions on Medical Imaging*, 45(2):632–645, 2026. doi: 10.1109/TMI.2025.3604361.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted Multi-View Classification. In *International Conference on Learning Representations*, 2021. URL https://mlanthology.org/iclr/2021/han2021iclr-trusted/.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Katharina Hemker, Nikola Simidjievski, and Mateja Jamnik. Healnet: Multimodal fusion for heterogeneous biomedical data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 64479–64498, 2024. doi: 10.52202/079017-2057.

Noah Hollmann, Sebastian Müller, Lukas Purucker, Anish Krishnakumar, Michael Körfer, Sean B. Hoo, Robin T. Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025. doi: 10.1038/s41586-024-08328-6.

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 *International Conference on Learning Representations (ICLR)*, 2022.

David S. Knopman, Hélène Amieva, Ronald C. Petersen, Gaël Chételat, David M. Holtzman, Bradley T. Hyman, Ralph A. Nixon, and David T. Jones. Alzheimer disease. *Nature Reviews Disease Primers*, 7(1):Article 33, 2021. doi: 10.1038/s41572-021-00269-y.

Wenhao Li, Hongkuan Zhang, Hongwei Zhang, Zhengxu Li, Zengjie Dong, Yafan Chen, Niranjan Bidargaddi, and Hong Liu. Refine medical diagnosis using generation augmented retrieval and clinical practice guidelines. *IEEE Journal of Biomedical and Health Informatics*, pp. 1–14, 2025. doi: 10.1109/JBHI.2025.3641931.

Yu-Nong Scarlett Lin, Shansong Wang, Mojtaba Safari, and Xiaofeng Yang. Beyond natural images: A dual-stream dinov3 framework for pet/ct segmentation. In Yuankai Huo, Mingchen Gao, Chang-Fu Kuo, Yueming Jin, and Ruining Deng (eds.), *Proceedings of The 9th International Conference on Medical Imaging with Deep Learning*, volume 315 of *Proceedings of Machine Learning Research*, pp. 2780–2794. PMLR, 08–10 Jul 2026. URL https://proceedings.mlr.press/v315/lin26b.html.

Hongyu Liu, Dong Wei, Dong Lu, Jing Sun, Lin Wang, and Yefeng Zheng. M3ae: Multimodal representation learning for brain tumor segmentation with missing modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1657–1665, 2023a. doi: 10.1609/aaai.v37i2.25253.

Lin Liu, Shuai Liu, Liang Zhang, X. V. To, Feras Nasrallah, and Siddharth S. Chandra. Cascaded multi-modal mixing transformers for alzheimer's disease classification with incomplete data. *NeuroImage*, 277:120267, 2023b. doi: 10.1016/j.neuroimage.2023.120267.

Miao Liu, Jie Zhang, Pew-Thian Yap, and Dinggang Shen. View-aligned hypergraph learning for alzheimer's disease diagnosis with incomplete multi-modality data. *Medical Image Analysis*, 36: 123–134, 2017. doi: 10.1016/j.media.2016.11.002.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. doi: 10.1038/s41467-024-44824-z.

Zhen Ou, Cheng Jiang, Yifan Liu, Yuxin Zhang, Zhiming Cui, and Dinggang Shen. A graph-embedded latent space learning and clustering framework for incomplete multimodal multiclass alzheimer's disease diagnosis. In *Medical Image Computing and Computer Assisted Intervention (MICCAI 2024)*, volume 15007 of *Lecture Notes in Computer Science*, pp. 45–55, Cham, Switzerland, 2024. Springer. doi: 10.1007/978-3-031-72104-5_5.

Qwen, An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Shaohao Rui, Lingzhi Chen, Zhenyu Tang, Lilong Wang, Mianxin Liu, Shaoting Zhang, and Xiaosong Wang. Multi-modal vision pre-training for medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5164–5174, June 2025.

Y. Sun, L. Wang, G. Li, W. Lin, and L. Wang. A foundation model for enhancing magnetic resonance images and downstream segmentation, registration and diagnostic tasks. *Nature Biomedical Engineering*, 9(4):521–538, 2025. doi: 10.1038/s41551-024-01283-7.

Alexandru Toma, Patrick R. Lawler, Jimmy Ba, Ramesh G. Krishnan, Benjamin R. Rubin, and Bo Wang. Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*, 2023. doi: 10.48550/arXiv.2305.12031.

Chen Wang, Shuang Piao, Zhiyong Huang, Qian Gao, Jie Zhang, Yang Li, and Hongming Shan. Joint learning framework of cross-modal synthesis and diagnosis for alzheimer's disease by mining underlying shared modality information. *Medical Image Analysis*, 91:103032, 2024. doi: 10.1016/j.media.2023.103032.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15878–15887, 2023. doi: 10.1109/CVPR52729.2023.01524.

Zhen Wu, Anisha Dadu, Nicholas J. Tustison, Brian B. Avants, Michael A. Nalls, Jian Sun, and Faraz Faghri. Multimodal patient representation learning with missing modalities and labels. In *International Conference on Learning Representations (ICLR)*, 2024.

Xiaoyu Yang, Shuo Wang, Lukas Luking, and Ying Chen. Enabling visual foundation models to teach compact students via mixture of distillation. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 11145–11153, 2025. doi: 10.24963/ijcai.2025/1281.

Wei Yao, Kai Yin, William K. Cheung, Jing Liu, and Jun Qin. Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16416–16424, 2024. doi: 10.1609/aaai.v38i15.29578.

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 *Advances in Neural Information Processing Systems*, volume 37, pp. 98782–98805, 2024. doi: 10.52202/079017-3135.

Zhilin Zeng, Zelin Peng, Xiaokang Yang, and Wei Shen. Missing as Masking: Arbitrary Cross-modal Feature Reconstruction for Incomplete Multimodal Brain Tumor Segmentation . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15008. Springer Nature Switzerland, October 2024.

Yichi Zhang, Le Xue, Wenbo Zhang, Lanlan Li, Yuchen Liu, Chen Jiang, Yuan Cheng, and Yuan Qi. Seganypet: Universal promptable segmentation from positron emission tomography images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21107–21116, October 2025.

Borui Zhao, Quan Cui, Ren Song, Yiyu Qiu, and Jinjun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11953–11962, 2022.

Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2608–2618. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.203.

# A Dataset Preparation

## A.1 Cohort Construction

For Alzheimer's disease (AD) diagnosis, we constructed the cohort from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. Subjects with baseline diagnoses of Alzheimer's disease (AD) and normal controls (NC) were selected for binary classification. For mild cognitive impairment (MCI) progression prediction, subjects were selected from the ADNI dataset, with the Australian Imaging, Biomarkers and Lifestyle (AIBL) study and the Standardized Centralized Alzheimer's & Related Dementias Neuroimaging (SCAN) cohort used for external validation. MCI progression labels were defined based on longitudinal diagnosis changes: subjects converting to AD within 36 months after baseline were labeled as progressive MCI (pMCI), whereas those remaining stable for at least 36 months were labeled as stable MCI (sMCI). Subjects with insufficient follow-up, baseline dementia, or inconsistent diagnostic trajectories were excluded. For Parkinson's disease (PD) progression prediction, the Parkinson's Progression Markers Initiative (PPMI) dataset was used. Subjects were categorized into stable PD (sPD) and progressive PD (pPD) groups according to disease progression during the 36-month follow-up period.

Following these criteria, task-specific cohorts were constructed for subsequent experiments. Detailed sample sizes and demographic characteristics are provided in Table A1.

Table A1: Demographic and clinical characteristics of subjects in different cohorts.

**(a) ADNI Cohort**

| **Variable** | **NC** ($n = 1084$) | **AD** ($n = 461$) | **sMCI** ($n = 536$) | **pMCI** ($n = 253$) |
|---|---|---|---|---|
| Age | $70.98 \pm 7.14$ | $74.83 \pm 7.87$ | $71.99 \pm 7.53$ | $73.63 \pm 7.28$ |
| Sex (M/F) | 434 / 650 | 256 / 205 | 314 / 222 | 152 / 101 |
| Education | $16.44 \pm 2.55$ | $15.28 \pm 2.92$ | $16.13 \pm 2.69$ | $15.92 \pm 2.72$ |
| MMSE | $29.03 \pm 1.18$ | $23.06 \pm 2.16$ | $28.01 \pm 1.70$ | $26.75 \pm 1.75$ |
| CDR-SB | $0.04 \pm 0.15$ | $4.38 \pm 1.64$ | $1.28 \pm 0.72$ | $2.05 \pm 1.01$ |

**(b) AIBL Cohort**

| **Variable** | **sMCI** ($n = 13$) | **pMCI** ($n = 10$) |
|---|---|---|
| Age | $77.31 \pm 6.01$ | $74.30 \pm 6.34$ |
| Sex (M/F) | 8 / 5 | 7 / 3 |
| MMSE | $28.00 \pm 1.58$ | $26.30 \pm 1.70$ |
| CDR-SB | $0.46 \pm 0.14$ | $0.50 \pm 0.00$ |

**(c) SCAN Cohort**

| **Variable** | **sMCI** ($n = 62$) | **pMCI** ($n = 28$) |
|---|---|---|
| Age | $66.89 \pm 8.26$ | $68.71 \pm 7.51$ |
| Sex (M/F) | 47 / 15 | 14 / 14 |
| FAQ | $1.87 \pm 4.93$ | $3.35 \pm 5.41$ |
| CDR-SB | $0.94 \pm 0.87$ | $1.55 \pm 1.13$ |

**(d) PPMI Cohort**

| **Variable** | **sPD** ($n = 248$) | **pPD** ($n = 100$) |
|---|---|---|
| Age | $60.81 \pm 9.51$ | $68.22 \pm 7.55$ |
| Sex (M/F) | 146 / 102 | 65 / 35 |
| Education | $16.64 \pm 3.31$ | $16.37 \pm 2.97$ |
| MoCA | $27.15 \pm 2.57$ | $26.48 \pm 2.81$ |

## A.2 Data Preprocessing

**T1-weighted MRI preprocessing.** T1-weighted structural MRI (sMRI) scans were used as the anatomical imaging modality. We selected the official ADNI-provided MPRAGE-based images. The MRI preprocessing pipeline included N4 bias field correction, reorientation to RAS coordinate

space, skull stripping using SynthStrip, spatial normalization to the MNI152NLin2009cAsym template, and intensity normalization. Specifically, voxel intensities within the brain mask were clipped using the 5th and 95th percentiles and then normalized to the range $[0, 1]$. Finally, all MRI volumes were resized to $128 \times 128 \times 128$ for model input.

**FDG-PET preprocessing.** FDG-PET images were used as the metabolic imaging modality. The images were converted to NIfTI format and reoriented to RAS space. To ensure temporal consistency with MRI, each PET scan was matched with the nearest available T1-weighted MRI scan, and subjects with MRI-PET acquisition intervals exceeding 90 days were excluded. PET images were first registered to the corresponding T1 MRI space and subsequently transformed into the MNI space using the T1-to-MNI transformation. For intensity normalization, FDG-PET images were converted into SUVR-like representations using positive voxels within the MNI brain mask as the reference region, followed by clipping at the 99.5th percentile. The resulting images were further normalized using 1st–99th percentile intensity clipping and min-max normalization to the range of $[0, 1]$. Finally, all PET volumes were resized to $128 \times 128 \times 128$ for model input.

**Clinical tabular data preprocessing.** Clinical features included demographic and genetic information, such as age, sex, education years, and APOE genotype. To avoid potential data leakage, diagnosis-related variables such as MMSE, CDR-SB, and FAQ were excluded. The selected features were directly converted into textual descriptions and fed into the Qwen model for representation learning without additional preprocessing. Specifically, the retained variables were serialized into a fixed-order English paragraph, where missing entries were omitted. The generated text descriptions were encoded by Qwen2.5-0.5B, and the hidden state of the last non-padding token was projected to a 512-dimensional embedding as the clinical modality representation.

# B Experimental Implementation Details

## B.1 Experimental Settings and Evaluation Metrics

For primary evaluation, the AD cohort includes 1,545 subjects (461 AD and 1,084 NC), with 545 subjects having complete MRI, PET, and clinical modalities. The MCI cohort includes 789 subjects (253 pMCI and 536 sMCI), including 615 subjects with complete multimodal data. Results on complete multimodal subsets are reported only as supplementary comparisons.

Hypergraph fusion was evaluated using subject-level stratified five-fold cross-validation with a fixed seed of 42. In each fold, 20% of subjects were held out for testing, and the remaining subjects were split into training and validation sets with an 8:2 ratio. The decision threshold was determined on the validation set by maximizing balanced accuracy and applied to the test set. We report mean and standard deviation across five folds for AUC, balanced accuracy (bACC), accuracy (ACC), sensitivity (SEN), and specificity (SPE). Balanced accuracy is calculated as $\mathrm{bACC} = (\mathrm{SEN} + \mathrm{SPE})/2$, which provides a more reliable evaluation under class-imbalanced conditions by assigning equal importance to positive and negative classes.

Intra-modal MRI, PET, and clinical encoders followed the same subject-level splits as hypergraph evaluation. Foundation model adaptation and student distillation were performed using only training subjects in each fold, while validation and test subjects remained unseen during training. After modality-specific distillation, all encoders were frozen before hypergraph fusion.

## B.2 Model Training and Hyperparameters

The whole framework consists of four training stages. The first three stages generate frozen 512-dimensional modality representations, while only the hypergraph fusion module is optimized in the final stage.

**Modality-Specific Foundation Model Adaptation and Distillation.** The MRI teacher is initialized from BrainMVP with a UniFormer-S backbone, and the PET teacher is initialized from SegAnyPET with a ViT-B backbone. All teacher models are adapted to the downstream tasks using LoRA. The modality-specific student encoders adopt 3D ResNet-10 with a base width of 64 and output 512-dimensional representations. During distillation, the teacher models are frozen, and the student encoders are optimized using classification loss, response distillation loss, and feature alignment

loss. The distillation temperature is set to $T = 4$, with $\lambda_{\text{feat}} = 0.5$. For AD vs. NC, the weights of the classification and response distillation losses are set to 0.4 and 0.6, respectively. For sMCI vs. pMCI and sPD vs. pPD, the corresponding weights are set to 0.3 and 0.7. The student encoders are trained using AdamW for up to 40 epochs with early stopping. The clinical branch uses Qwen2.5-7B as the teacher and Qwen2.5-0.5B as the student. The student model is adapted with LoRA ($r = 32, \alpha = 32$) and trained through probability-level distillation. The resulting 512-dimensional clinical representation is used for subsequent cross-modal semantic alignment. During hypergraph fusion, all student encoders are frozen, and only the hypergraph fusion module is optimized.

**Cross-modal semantic alignment.** The student encoders are jointly optimized during semantic alignment. For AD vs. NC, alignment is performed in the 512-dimensional representation space, with $\lambda_{\text{align}} = 0.25$ and negative-pair weight $\gamma = 0.5$. The distillation temperature is set to 3, with $\lambda_{\text{kd}} = 0.5$ and $\lambda_{\text{feat}} = 0.4$. Training uses learning rates of $3 \times 10^{-4}$ for the student encoders and $1 \times 10^{-3}$ for the projection layers, with early stopping. For sMCI vs. pMCI and sPD vs. pPD, semantic alignment is also conducted in the 512-dimensional space. The patient-level alignment term is introduced with $\lambda_{\text{align}} = 0.15$. Training uses a learning rate of $2 \times 10^{-4}$ with early stopping.

**Hypergraph fusion.** Each hypergraph node is constructed by concatenating the 512-dimensional modality feature and its two-dimensional softmax prediction. Two types of hyperedges are constructed: intra-modality hyperedges connect $k = 8$ nearest neighbors within the same modality based on cosine similarity, and intra-patient hyperedges connect available modalities from the same subject. Hypergraph propagation is implemented using uniform hypergraph convolution with a hidden dimension of 256, three layers, and dropout of 0.35. The hypergraph module is trained using AdamW for 50 epochs. The validation criterion is the average of AUC and balanced accuracy. No additional modality masking is introduced during training, and missing modalities are only derived from naturally incomplete samples.

# C Additional Experimental Results

## C.1 Ablation Study of Hyperedge

To evaluate the contribution of different hyperedge types, we perform hyperedge ablation experiments on the ADNI dataset using the proposed patient–modality hypergraph. Specifically, we compare three settings: (1) the full hypergraph with both intra-modality patient similarity hyperedges and intra-patient cross-modal hyperedges; (2) removing intra-patient cross-modal hyperedges while retaining only intra-modality patient similarity hyperedges; and (3) removing intra-modality patient similarity hyperedges while retaining only intra-patient cross-modal hyperedges.

Table A2 reports the results under naturally incomplete multimodal settings. The results demonstrate that both types of hyperedges contribute to the final performance, as removing either component consistently degrades the model performance. These findings validate the effectiveness of jointly modeling intra-patient cross-modal relationships and intra-modality patient similarities for incomplete multimodal fusion.

Table A2: Ablation study of hyperedge under naturally incomplete multimodal settings.

(a) AD vs. NC

| Setting | AUC | ACC | bACC | SEN | SPE |
|---|---|---|---|---|---|
| Full hypergraph | 0.990±0.006 | 0.966±0.016 | 0.966±0.013 | 0.965±0.012 | 0.966±0.022 |
| w/o cross-modal hyperedges | 0.985±0.007 | 0.939±0.019 | 0.939±0.018 | 0.937±0.026 | 0.940±0.024 |
| w/o patient similarity hyperedges | 0.983±0.007 | 0.937±0.011 | 0.931±0.012 | 0.915±0.018 | 0.946±0.012 |

(b) sMCI vs. pMCI

| Setting | AUC | ACC | bACC | SEN | SPE |
|---|---|---|---|---|---|
| Full hypergraph | 0.863±0.033 | 0.778±0.041 | 0.793±0.026 | 0.834±0.032 | 0.752±0.070 |
| w/o cross-modal hyperedges | 0.850±0.025 | 0.786±0.036 | 0.781±0.044 | 0.766±0.073 | 0.795±0.032 |
| w/o patient similarity hyperedges | 0.834±0.027 | 0.750±0.044 | 0.751±0.035 | 0.755±0.050 | 0.748±0.069 |

### C.2 Complete Comparison Results on ADNI

The complete comparison results, including ACC, SEN, and SPE, are provided in Tables A3 and A4.

Table A3: Comparison Results for AD vs. NC on the ADNI Dataset under Complete and Naturally Incomplete Multimodal Settings.

**(a) Complete multimodal setting**

| Method | AUC | bACC | ACC | SEN | SPE |
|---|---|---|---|---|---|
| MRI-only | 0.936±0.015 | 0.868±0.025 | 0.868±0.021 | 0.870±0.045 | 0.867±0.023 |
| PET-only | 0.968±0.016 | 0.901±0.024 | 0.901±0.023 | 0.911±0.041 | 0.892±0.037 |
| Clinical-only | 0.806±0.028 | 0.721±0.040 | 0.741±0.037 | 0.670±0.066 | 0.771±0.045 |
| TMC (ICLR 2021) | 0.815±0.116 | 0.733±0.116 | 0.732±0.109 | 0.736±0.179 | 0.730±0.071 |
| MMIN (ACL 2021) | 0.733±0.054 | 0.651±0.093 | 0.653±0.100 | 0.641±0.219 | 0.662±0.229 |
| ShaSpec (CVPR 2023) | 0.736±0.032 | 0.660±0.038 | 0.655±0.055 | 0.700±0.146 | 0.620±0.186 |
| DFTD (TMI 2023) | 0.762±0.089 | 0.720±0.068 | 0.723±0.077 | 0.705±0.055 | 0.736±0.134 |
| 3MT (NeuroImage 2023) | 0.993±0.013 | 0.967±0.032 | 0.965±0.033 | 0.983±0.028 | 0.952±0.043 |
| DrFuse (AAAI 2024) | 0.735±0.041 | 0.697±0.053 | 0.705±0.056 | 0.645±0.097 | 0.749±0.111 |
| HEALNet (NeurIPS 2024) | 0.736±0.040 | 0.677±0.061 | 0.692±0.062 | 0.572±0.064 | 0.781±0.078 |
| Flex-MoE (NeurIPS 2024) | 0.775±0.062 | 0.702±0.073 | 0.695±0.086 | 0.755±0.134 | 0.649±0.200 |
| IMDL (TMI 2026) | 0.704±0.064 | 0.646±0.066 | 0.657±0.056 | 0.573±0.181 | 0.720±0.099 |
| Q-Former (ICLR 2026) | 0.969±0.015 | 0.886±0.041 | 0.888±0.039 | 0.868±0.089 | 0.904±0.062 |
| Ours | **0.996±0.007** | **0.976±0.011** | **0.974±0.012** | **0.991±0.012** | **0.961±0.018** |

**(b) Naturally incomplete multimodal setting**

| Method | AUC | bACC | ACC | SEN | SPE |
|---|---|---|---|---|---|
| TMC (ICLR 2021) | 0.916±0.014 | 0.838±0.018 | 0.856±0.017 | 0.792±0.077 | 0.884±0.051 |
| MMIN (ACL 2021) | 0.810±0.033 | 0.734±0.019 | 0.748±0.037 | 0.700±0.068 | 0.767±0.077 |
| ShaSpec (CVPR 2023) | 0.851±0.080 | 0.775±0.075 | 0.775±0.091 | 0.774±0.045 | 0.776±0.115 |
| DFTD (TMI 2023) | 0.841±0.055 | 0.762±0.051 | 0.768±0.057 | 0.748±0.089 | 0.776±0.084 |
| 3MT (NeuroImage 2023) | **0.990±0.004** | 0.959±0.010 | 0.961±0.011 | 0.954±0.025 | 0.963±0.020 |
| DrFuse (AAAI 2024) | 0.863±0.077 | 0.783±0.074 | 0.801±0.072 | 0.738±0.087 | 0.827±0.073 |
| HEALNet (NeurIPS 2024) | 0.790±0.022 | 0.713±0.025 | 0.737±0.038 | 0.655±0.081 | 0.771±0.078 |
| Flex-MoE (NeurIPS 2024) | 0.806±0.026 | 0.717±0.032 | 0.748±0.026 | 0.640±0.091 | 0.794±0.053 |
| IMDL (TMI 2026) | 0.799±0.042 | 0.729±0.049 | 0.714±0.072 | 0.766±0.059 | 0.692±0.112 |
| Q-Former (ICLR 2026) | 0.956±0.011 | 0.885±0.031 | 0.883±0.039 | 0.889±0.030 | 0.881±0.053 |
| Ours | **0.990±0.006** | **0.966±0.013** | **0.966±0.016** | **0.965±0.012** | **0.966±0.022** |

Table A4: Comparison Results for sMCI vs. pMCI on the ADNI Dataset under Complete and Naturally Incomplete Multimodal Settings.

**(a) Complete multimodal setting**

| Method | AUC | bACC | ACC | SEN | SPE |
|---|---|---|---|---|---|
| MRI-only | 0.774±0.053 | 0.722±0.057 | 0.718±0.049 | 0.733±0.094 | 0.711±0.047 |
| PET-only | 0.830±0.022 | 0.761±0.038 | 0.776±0.060 | 0.734±0.083 | 0.788±0.089 |
| Clinical-only | 0.717±0.042 | 0.633±0.035 | 0.635±0.075 | 0.628±0.155 | 0.638±0.172 |
| TMC (ICLR 2021) | 0.655±0.031 | 0.620±0.026 | 0.659±0.031 | 0.545±0.106 | 0.695±0.070 |
| MMIN (ACL 2021) | 0.666±0.029 | 0.624±0.028 | 0.655±0.045 | 0.564±0.127 | 0.684±0.094 |
| ShaSpec (CVPR 2023) | 0.660±0.027 | 0.605±0.049 | 0.616±0.035 | 0.585±0.168 | 0.626±0.087 |
| DFTD (TMI 2023) | 0.644±0.030 | 0.611±0.057 | 0.636±0.049 | 0.564±0.213 | 0.658±0.121 |
| 3MT (NeuroImage 2023) | **0.850±0.033** | 0.748±0.049 | 0.738±0.106 | **0.768±0.117** | 0.728±0.170 |
| DrFuse (AAAI 2024) | 0.656±0.037 | 0.626±0.054 | 0.641±0.050 | 0.598±0.129 | 0.654±0.079 |
| HEALNet (NeurIPS 2024) | 0.636±0.030 | 0.616±0.056 | 0.629±0.037 | 0.592±0.164 | 0.641±0.079 |
| Flex-MoE (NeurIPS 2024) | 0.659±0.037 | 0.596±0.047 | 0.633±0.059 | 0.524±0.195 | 0.667±0.130 |
| IMDL (TMI 2026) | 0.645±0.028 | 0.596±0.026 | 0.637±0.100 | 0.517±0.219 | 0.675±0.199 |
| Q-Former (ICLR 2026) | 0.839±0.040 | **0.752±0.039** | **0.779±0.057** | 0.700±0.089 | **0.803±0.090** |
| Ours | **0.850±0.031** | 0.750±0.044 | 0.751±0.089 | 0.748±0.082 | 0.752±0.137 |

**(b) Naturally incomplete multimodal setting**

| Method | AUC | bACC | ACC | SEN | SPE |
|---|---|---|---|---|---|
| TMC (ICLR 2021) | 0.705±0.039 | 0.633±0.042 | 0.634±0.091 | 0.632±0.179 | 0.634±0.208 |
| MMIN (ACL 2021) | 0.695±0.062 | 0.638±0.065 | 0.640±0.071 | 0.633±0.123 | 0.644±0.113 |
| ShaSpec (CVPR 2023) | 0.711±0.066 | 0.644±0.037 | 0.649±0.028 | 0.632±0.162 | 0.657±0.103 |
| DFTD (TMI 2023) | 0.703±0.042 | 0.623±0.021 | 0.603±0.042 | 0.676±0.108 | 0.569±0.107 |
| 3MT (NeuroImage 2023) | 0.828±0.021 | 0.754±0.038 | 0.719±0.081 | 0.823±0.061 | 0.686±0.125 |
| DrFuse (AAAI 2024) | 0.712±0.069 | 0.629±0.047 | 0.602±0.049 | 0.704±0.156 | 0.554±0.117 |
| HEALNet (NeurIPS 2024) | 0.706±0.028 | 0.635±0.022 | 0.624±0.029 | 0.668±0.087 | 0.603±0.073 |
| Flex-MoE (NeurIPS 2024) | 0.718±0.065 | 0.651±0.051 | 0.635±0.031 | 0.696±0.131 | 0.606±0.052 |
| IMDL (TMI 2026) | 0.680±0.042 | 0.635±0.054 | 0.650±0.048 | 0.593±0.172 | 0.677±0.114 |
| Q-Former (ICLR 2026) | 0.854±0.032 | 0.761±0.048 | 0.768±0.030 | 0.743±0.171 | **0.780±0.095** |
| Ours | **0.863±0.033** | **0.793±0.026** | **0.778±0.041** | **0.834±0.032** | 0.752±0.070 |

### C.3 Sensitivity Analysis of the kNN Hyperparameter

To investigate the effect of the neighborhood size $k$ in constructing intra-modality patient similarity hyperedges, we conduct a sensitivity analysis on the ADNI dataset by varying $k$ while keeping all other experimental settings unchanged. The evaluated values of $k$ range from 4 to 12, and the results are reported in Fig. 7.

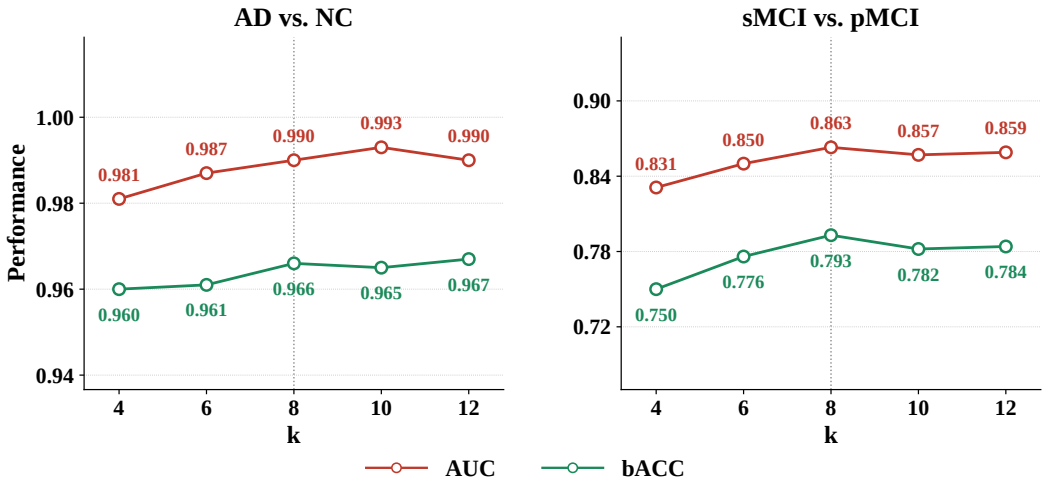


Figure 7: Sensitivity analysis of the kNN neighborhood size $k$ for constructing intra-modality patient similarity hyperedges. The red and green curves denote AUC and balanced accuracy (bACC).

For AD vs. NC, the model achieves consistently high performance across different values of $k$, with AUC remaining above 0.98 and only minor variations observed. Although $k = 10$ achieves the highest AUC, the performance differences among different settings are relatively small, indicating the robustness of the hypergraph construction to the neighborhood size. For sMCI vs. pMCI, the performance shows slightly larger variations with different $k$ values. The highest AUC is obtained at $k = 8$, while larger neighborhoods do not provide further improvement. Considering both classification tasks and maintaining a consistent hypergraph construction strategy, we select $k = 8$ as the default neighborhood size.