Hierarchical Information Aggregation for Incomplete Multimodal Alzheimer's Disease Diagnosis

Chengliang Liu[1, 2, 3], Yuanxi Que[1], Qihao Xu[3], Yabo Liu[4], Jie Wen[3], Jinghua Wang[3], Xiaoling Luo[1]*

[1]College of Computer Science and Software Engineering, Shenzhen University
[2]Laboratory for Artificial Intelligence in Design, The Hong Kong Polytechnic University
[3]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
[4]College of Artificial Intelligence, Ocean University of China
liucl1996@163.com, {queyuanxi, xqh51199597, xiaolingluoo}@outlook.com,
yaboliu.ug@gmail.com, jiewen_pr@126.com, wangjh2012@foxmail.com

Abstract

Alzheimer's Disease (AD) poses a significant health threat to the aging population, underscoring the critical need for early diagnosis to delay disease progression and improve patient quality of life. Recent advances in heterogeneous multimodal artificial intelligence (AI) have facilitated comprehensive joint diagnosis, yet practical clinical scenarios frequently encounter incomplete modalities due to factors like high acquisition costs or radiation risks. Moreover, traditional convolution-based architecture face inherent limitations in capturing long-range dependencies and handling heterogeneous medical data efficiently. To address these challenges, in our proposed heterogeneous multimodal diagnostic framework (HAD), we develop a multi-view Hilbert curve-based Mamba block and a hierarchical spatial feature extraction module to simultaneously capture local spatial features and global dependencies, effectively alleviating spatial discontinuities introduced by voxel serialization. Furthermore, to balance semantic consistency and modal specificity, we build a unified mutual information learning objective in the heterogeneous multimodal embedding space, which maintains effective learning of modality-specific information to avoid modality collapse caused by model preference. Extensive experiments demonstrate that our HAD significantly outperforms state-of-the-art methods in various modality-missing scenarios, providing an efficient and reliable solution for early-stage AD diagnosis.

1 Introduction

AD is a progressive neurodegenerative disorder characterized by cognitive impairment, gradual loss of memory, and a decline in self-care abilities as its primary clinical manifestations [1, 2, 3]. Due to the lack of effective cures, AD poses a significant threat to the health of the elderly population, severely impacting the quality of life of patients and their families and imposing a heavy medical burden on society [4]. Mild Cognitive Impairment (MCI) is considered a precursor stage of AD, marked by mild cognitive decline without a noticeable impact on daily functional abilities. Early diagnosis and intervention during this stage are critical for delaying disease progression and improving patients' quality of life. In recent years, with the rapid increase of the type of multimodal data, researchers have been able to better understand and diagnose early-stage AD from multiple perspectives, providing more comprehensive and objective decision-making support for clinical diagnosis and treatment [5, 6].

[*] Corresponding author: Xiaoling Luo (email: xiaolingluoo@outlook.com).

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

These advances also lay the groundwork for the application of multimodal Artificial Intelligence (AI) in joint AD diagnosis.

However, in practical clinical settings, the collection of multimodal data is often hampered by issues such as radiation risks, high costs, and unexpected patient withdrawal, leading to the frequent problem of missing modalities. This makes it difficult to obtain complete multimodal datasets [7, 8]. To address missing data, existing studies often discard cases with incomplete modalities and rely solely on data with complete modalities for analysis [9]. This approach reduces the subject scale, thereby limiting the performance of models. To overcome this issue, researchers have proposed various multimodal learning frameworks based on strategies such as subspace learning, knowledge distillation, and missing data imputation [10, 11]. For instance, data imputation methods use generative models like Generative Adversarial Networks (GANs) and Autoencoders to fill in missing modalities, thereby expanding the training dataset. However, due to challenges in ensuring the quality of imputed data, such methods often introduce redundant or even misleading information, resulting in decreased performance. Furthermore, the heterogeneity of multimodal data in AD, which ranges from three-dimensional (3D) image data to 1D biomarker data, presents additional challenges. Effectively integrating heterogeneous multimodal data and mining their shared semantic information under conditions of missing modalities remains a key difficulty in AI-assisted AD diagnosis.

Currently, mainstream multimodal diagnostic methods are typically based on 3D Convolutional Neural Networks (CNNs) or Transformer architectures [12, 13]. However, 3D CNNs often struggle to effectively capture long-range spatial dependencies and are constrained by large parameter sizes and high computational costs [14]. While Transformer-based models can model long-range dependencies, their computational complexity increases quadratically with the input data dimensions. This is particularly problematic when dealing with high-dimensional multimodal 3D medical image data, where computational inefficiency becomes a significant bottleneck, severely limiting their practical clinical applications. Recently, Structured State Space Models (SSMs), exemplified by Mamba, have gained attention for their efficient information extraction capabilities and linear computational complexity. For example, Liu et al. [15] proposed VMamba with a 2D Selective Scan module (SS2D) that bridges the ordered nature of 1D selective scan and the non-sequential structure of 2D visual data. Xing et al. [16] developed SegMamba, which captures remote dependencies in the entire 3D voxel at multi-scale. In the context of 3D multimodal medical image, exploring an efficient feature extraction module based on SSMs holds the potential to enhance diagnostic performance while reducing model complexity, thereby better meeting the demands of clinical applications.

To address the aforementioned challenges, this paper proposes a novel heterogeneous multimodal diagnostic framework for AD, named HAD. The framework is capable of processing multimodal data that includes Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), Cerebrospinal Fluid (CSF), and Clinical Assessment Data (CAD) with arbitrary modality missing, providing a flexible solution for early-stage AD diagnosis. One the one hand, to address the challenges of long-range dependencies and high computational complexity in 3D brain image data, we develop a hierarchical spatial feature extraction module. Building upon existing work [17], we adopt the same Hilbert curves for space-filling transformations while maintaining their core concept of "locality-preserving property of space-filling curves". However, our proposed HSFE module introduces two key innovations: (1) A hierarchical architecture based on the fractal theory of 3D Hilbert curves enables multi-scale information fusion through multi-level recursive structures; (2) A multi-directional scanning mechanism (incorporating axial rotation and mirror transformations) enhances complementary spatial information capture. This module integrates traditional convolutional operation with efficient state space model, enabling the effective capture of both shallow features and global dependencies in 3D image data. On the other hand, to enhance the consistency of discriminative capabilities across different modalities for AD diagnosis, we propose an optimization objective based on maximizing mutual information between multimodal joint semantic features and shallow features in the semantic embedding space. This strategy ensures that the framework effectively integrates heterogeneous information from various modalities, improving diagnostic performance in scenarios with incomplete multimodal datasets. Our main contributions are summarized as follows:

- We propose a heterogeneous multimodal AD diagnosis framework capable of handling arbitrary incomplete modalities. Unlike existing methods primarily tailored for 3D brain image data, our framework can effectively process heterogeneous multimodal data with significant informational and structural differences.

- To effectively handle complex 3D MRI and PET image data, we propose the multi-view Hilbert curve-based Mamba block (HMamba), along with a hierarchical spatial feature extraction strategy built upon HMamba. These modules alleviate discontinuities introduced by spatial voxel serialization and unify long-sequence modeling with local feature extraction across multiple scales.
- We propose a multimodal semantic representation learning framework, which establishes a goal of maximizing mutual information between modality-specific features and semantic labels, simultaneously considering modality-specific information extraction and consistency representation learning.

2 Preliminary

2.1 Problem Definition

Given a multimodal AD diagnosis dataset $\mathcal{D} = (\mathbf{x}, \mathbf{y})$ with n subjects, and each subject x consists of heterogeneous multimodal data (e.g., structural MRI, PET, CSF, and CAD), denoted as $x = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$, where m denotes the total number of modalities. Noted that we use x^{img} to indicate the MRI or PET data. The corresponding diagnostic label is denoted as y, representing the clinical status of the subject (e.g., cognitively normal (CN), MCI, and AD). In practical clinical scenarios, some modalities might be missing for certain subjects due to various reasons, thus we let $\mathcal{V}$ denote the set of available modalities and $|\mathcal{V}| \leq m$. The objective of multimodal AD diagnosis is thus to train a neural network model capable of accurately predicting the clinical label $\mathbf{y}$ using any available subset of modalities, even when some modalities are missing during inference.

2.2 State Space Modals

State Space Models (SSMs) are classical linear time invariant systems widely used in control theory and signal processing, characterized by their linear complexity and effectiveness in modeling sequential data. Recently, SSMs have gained renewed attention in deep learning due to their ability to efficiently capture long-range dependencies. Gu et al. [18] first introduced the HiPPO framework, providing a theoretically optimal approach to represent continuous-time state-space models by high-order polynomial projections. Subsequently, Gu et al. proposed the Structured State Space Sequence model (S4) [19] that introduces discretization and convolutional representation for parallel training, demonstrating superior performance in processing time series data. More recently, Mamba [20] introduced selective structured state spaces, simplifying the architecture and improving parallelism. Various vision Mamba architectures [15, 21] further extended the state-space modeling paradigm from sequences to two-dimensional image data, effectively addressing the quadratic complexity issue inherent in vision Transformers.

2.2.1 Hilbert Curve for 3D Brain Image Data

Existing Vision Mamba methods typically convert structured 2D or 3D visual data into 1D sequences through serialization approaches such as bi-directional scanning, cross scanning, or continuous scanning [22, 23, 24]. However, these simple scanning strategies inevitably disrupt the inherent spatial relationships, causing spatially adjacent pixels or voxels to become distant from each other in the serialized sequence. Such spatial discontinuity significantly impairs the model's ability to capture local structural information and long-range spatial dependencies, thereby limiting diagnostic performance in medical imaging tasks. To alleviate this issue, we propose adopting the 3D Hilbert curve for scanning brain image data. Unlike traditional scanning approaches [25, 26, 27], the Hilbert curve is a continuous, space-filling fractal curve known for its excellent locality-preserving property . Specifically, the Hilbert curve mapping ensures that voxels spatially close in the 3D space remain close in the serialized 1D representation to a large extent, thus minimizing spatial distortion and aiding the model in effectively capturing local and global contextual information from structured medical image data.

The 3D Hilbert curve can be generated iteratively, with each iteration referred to as the curve's *order* N. At each iteration, the curve recursively subdivides the original 3D cubic space into $2^N \times 2^N \times 2^N$ sub-cubes, as shown in Fig. 5. The Hilbert curve of order N is thus constructed by connecting the Hilbert curves of order $(N - 1)$ from eight smaller sub-cubes through rotations and reflections,

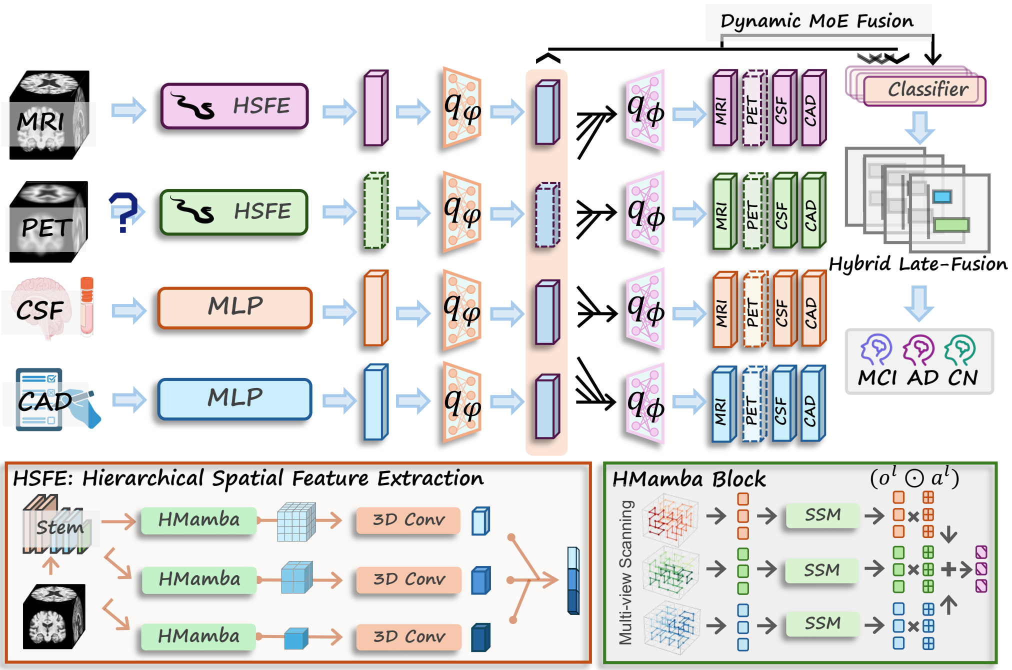


Figure 1: The schematic diagram of our HAD. It contains 2 main parts, a heterogeneous modality-specific shallow feature extraction consisting of HSFE module and MLP (left half) and high-level multimodal semantic coding (right half). "q_φ" and "q_ϕ" denote the modality-specific encoder and cross-modal decoder, respectively; "3D Conv" denotes the 3D residual convolution module.

preserving the locality and continuity of the space-filling curve. Formally, given a voxel coordinate (x_c, y_c, z_c) within a cubic voxel grid of size 2^N, the 3D Hilbert curve defines a mapping $\mathcal{H}_N(\cdot)$ from the 3D coordinate to a 1D sequence index h:

$$h = \mathcal{H}_N(x_c, y_c, z_c), \quad h \in \{0, 1, \ldots, 8^N - 1\}. \tag{1}$$

An appropriate Hilbert curve order N is selected based on the input image data resolution.

3 Method

3.1 HMamba: Hilbert Curve-Based Mamba Block

Multi-View Spatial Scanning. As illustrated in Fig. 5, spatially adjacent voxels located at the boundaries between neighboring sub-cubes might become relatively distant within the serialized sequence due to the intrinsic fractal structure of the Hilbert curve. This spatial tearing phenomenon potentially leads to the loss of critical adjacency information, adversely affecting the model's ability to capture fine-grained spatial dependencies. To bridge this gap and further alleviate the spatial discontinuity issue, we propose utilizing multiple Hilbert curves oriented along different spatial axes, complementing the conventional approach of scanning with a Hilbert curve along a single orientation. In other words, we expect that multiple Hilbert curves collectively maximize the coverage of adjacency edges among voxels, effectively preserving comprehensive spatial context within serialized data. A higher coverage rate indicates better preservation of spatial adjacency information and thus facilitates more effective modeling of spatial context (please refer to Appendix A.3 for the definition of *coverage rate*).

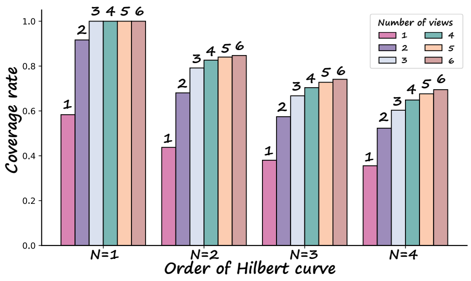


Figure 2: Coverage rate vs. order of Hilbert curve under different view numbers.

As we know, starting from a fixed vertex, six independent curves can be drawn. Therefore, in Fig. 2, we illustrate the relationship between the number of scanning views and the coverage rate. At lower orders, three-view scanning is sufficient to fully cover all adjacent edges. However, as image resolution increases, achieving complete coverage becomes more challenging. Furthermore, increasing the number of scanning curves should be considered with caution due to the additional computational overhead. Overall, for input $x^{img} \in \mathbb{R}^{d_i \times d_i \times d_i}$, L multi-view scanning layers are denoted as $\{\mathcal{F}_S^l : x^{img} \in \mathbb{R}^{d_i \times d_i \times d_i} \to \mathbb{R}^{d_i^3}\}_{l=1}^{L}$.

Multi-View Dynamic Fusion-Based Mamba Block. Upon serializing the input 3D image data, we employ the SSMs to capture long-range dependencies across the resulting long sequences, which originate from continuous-time linear dynamical systems, mapping an input $x_t \in \mathbb{R}$ to an output $o_t \in \mathbb{R}$ via a hidden state $h_t \in \mathbb{R}^{d_e}$ as follows:

$$h_t = Ah_{t-1} + Bx_t, \quad o_t = Ch_t, \tag{2}$$

where $A \in \mathbb{R}^{d_e \times d_e}$, $B \in \mathbb{R}^{d_e \times 1}$, and $C \in \mathbb{R}^{1 \times d_e}$ are learnable parameters.

For discrete sequence modeling, the continuous parameters (A, B) are discretized using zero-order hold (ZOH) with a step size Δ: $\overline{A} = e^{\Delta A}, \quad \overline{B} = A^{-1}(e^{\Delta A} - I)B$. The resulting discrete-time SSM is formulated as:

$$h_t = \overline{A}h_{t-1} + \overline{B}x_t, \quad o_t = Ch_t. \tag{3}$$

In practice, model outputs are efficiently computed through convolution:

$$o = x * H, \quad H = (C\overline{B}, C\overline{AB}, \dots, C\overline{A}^{M-1}\overline{B}), \tag{4}$$

where $M = d_i^3$ is the input sequence length, and $H \in \mathbb{R}^M$ is the structured convolution kernel. Since sequences derived from different views correspond to spatially misaligned voxel indices, a reverse indexing operation is subsequently applied to map the multi-view serialized features back to the original 3D voxel grid structure, i.e., $\bar{\mathcal{F}}_S^l : o \in \mathbb{R}^{d_i^3} \to \mathbb{R}^{d_i \times d_i \times d_i}$. Furthermore, due to the distinct scanning views, the SSMs applied to each serialized sequence capture complementary spatial dependencies. To effectively aggregate these multi-view features, we propose a voxel-wise dynamic fusion strategy. Formally, given the encoded feature tensors from L distinct scanning views, denoted as $\{o^l\}_{l=1}^{L}$, we assign a learnable spatially adaptive weighting tensor $a^l \in \mathbb{R}^{d_i \times d_i \times d_i}$ for each view. Each element of a^l dynamically balances the importance of each voxel from the l-th view. The fused voxel-wise feature map $\hat{o}$ is thus obtained as a weighted combination: $\hat{o} = \sum_{l=1}^{L} a^l \odot o^l$, where $\odot$ denotes the voxel-wise (element-wise) multiplication. The weighting tensors a^l are optimized during training, enabling the model to dynamically emphasize the most informative features from different scanning views for each voxel individually. This proposed voxel-wise dynamic multi-view fusion effectively integrates complementary spatial context captured by multi-view SSMs, significantly enhancing the model's representation capability for 3D medical image data.

3.2 Hierarchical Spatial Feature Extraction

As described in Section 3.1, our proposed multi-view Hilbert curve-based scanning approach significantly mitigates the spatial tearing issue. However, at the high imaging resolutions (i.e., larger Hilbert curve orders), the multi-view scanning strategy may fail to resolve all spatial discontinuities, as complete adjacency edge coverage becomes increasingly challenging. Motivated by the hierarchical receptive field scaling property inherent in multi-scale CNNs, we propose a hierarchical spatial feature extraction (HSFE) module that serializes and processes 3D image data at multi-scale resolutions. At each hierarchical level, the resolution of the input 3D image data is progressively reduced by the downsampling operation. Specifically, given an original 3D image with dimensions $2^N \times 2^N \times 2^N$, we iteratively construct K lower-resolution images by a series of downsampling modules consisting of convolutional layer and MaxPool layer: $\{\hat{x}|_{k+1} = \text{Down}(\hat{x}|_k)\}_{k=0}^{K-1}$, where $\hat{x}|_0 = \text{Stem}(x^{img})$ and Stem module is to expand channels and reduce dimensions. At k-th level, we utilize HMamba module with $N - k$ order Hilbert curve and 3D residual convolution block to model long-sequence dependency and local spatial information:

$$g|_k = \mathcal{C}^k(\mathcal{M}^k(\hat{x}|_k)), \quad k \in [0, K-1], \tag{5}$$

where $g|_k$ denotes the k-th level output. $\mathcal{C}^k$ and $\mathcal{M}^k$ mean the corresponding 3D residual block and HMamba module, respectively. Then, we simply concatenate all K outputs and perform max-pooling to obtain the final output of the HSFE module:

$$g = \text{MaxPool}(\text{Concat}(g|_0, g|_1, \ldots, g|_{K-1})). \tag{6}$$

This hierarchical strategy naturally alleviates the spatial discontinuity issues that may arise from high-order Hilbert curves, as the receptive field expands significantly at deeper layers. Consequently, our hierarchical state space architecture effectively compresses spatial features at various scales, progressively enhancing the information density of the learned representations.

3.3 Multimodal Semantic Representation Learning

Heterogeneous multimodal data inherently exhibit a modality gap, with different modalities contributing distinct perspectives to AD diagnosis. To preserve modality-specific characteristics, we avoid inter-modality information interactions in the initial stage instead of conducting modality-specific feature extraction (i.e., HSFE module for 3D image data and Multi-Layer Perceptrons (MLPs) for CSF and CAD). This is to map the heterogeneous multimodal data into a unified feature space, facilitating further cross-modal alignment [28, 29]. In general, we aim for the learned features to preserve semantic information as fully as possible, while simultaneously striving to achieve semantic consistency across multiple modalities. Therefore, we propose the following mutual information maximization objective:

$$\max I(\mathbf{g}; \mathbf{y}) + \alpha I(\mathbf{g}; \mathbf{z}), \tag{7}$$

where $\mathbf{g} = \{\mathbf{g}^{(v)}\}_{v \in \mathcal{V}}$ and random variable $\mathbf{g}^{(v)}$ corresponds to the modality-specific feature of v-th modality. $\mathbf{z}$ represents the cross-modal joint semantic representation and α is the balanced parameter. Eq. (7) consists of two parts: the first term focus on learning discriminative information from labels, while the second term aims to align modality-specific representations with cross-modal semantic representations. For the first term of Eq. (7), the equivalent objective is as follows:

$$\max I(\mathbf{g}; \mathbf{y}) \Leftrightarrow \min H(P, Q), \tag{8}$$

where $P \sim p(\mathbf{y}|\mathbf{g})$ and $Q \sim q(\mathbf{y}|\mathbf{g})$ denote the real distribution of $\mathbf{y}$ and predicted distribution, respectively, and $H(P, Q)$ is the cross entropy. Specifically, we employ a dynamic Mixture-of-Experts (MoE) fusion to get the typical joint posterior of latent representation $\mathbf{z}$, i.e., $q_\varphi(\mathbf{z}|\mathbf{g}) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \omega^{(v)} q_{\varphi_v}(\mathbf{z}|\mathbf{g}^{(v)})$, where $\omega^{(v)} = \frac{e^{\eta^v/\tau}}{\sum_{v \in \mathcal{V}} e^{\eta^v/\tau}}$ (τ: the temperature parameter) is calculated by the modality-specific learnable parameters $\{\eta^1, \eta^2, ..., \eta^m\}$, and then inference the prediction probability by parameterized neural networks $q_\theta(\mathbf{y}|\mathbf{z})$. This corresponds to constructing the probabilistic graph model: $\mathbf{g} \rightarrow \mathbf{z} \rightarrow \mathbf{y}$. Together with the second term, this approach simultaneously ensures semantic consistency across multimodal representations and facilitates semantic learning of disease categories. However, it ignores the inherent heterogeneity among different modalities, which can hinder the effective exploration of multimodal complementary information. In addition, due to discrepancy in information and data structure, the network commonly has obvious fitting preference for certain modalities, which can easily lead to training collapse issue for hard-fitting modalities (see further analysis in Section 4.3). Thus, we propose to add a direct inference of $\mathbf{y}$ from modality observations, i.e., $\mathbf{g} \rightarrow \mathbf{y}$. To be specific, we introduce the modality-specific conditional distribution into the final prediction distribution as follows:

$$q(\mathbf{y}|\mathbf{g}) =: \frac{1}{2} q_\theta(\mathbf{y}|\mathbf{z}) + \frac{1}{2} \sum_{v \in \mathcal{V}} \omega^{(v)} q_{\theta_v}(\mathbf{y}|\mathbf{g}^{(v)}), s.t., \sum_{v \in \mathcal{V}} \omega^{(v)} = 1, \tag{9}$$

Finally, given the established prediction distribution in Eq. (9), cross entropy minimization objective given in Eq. (8) can be expressed as the cross entropy loss function $\mathcal{L}_{ce} = \text{CrossEntropy}(y, \hat{y})$, where $\hat{y}$ denotes the joint prediction probability. By introducing dynamic weighting factors, the importance across different modalities can be effectively balanced. Note that we reverse $\omega^{(v)} = \frac{e^{-\eta^v/\tau}}{\sum_{v \in \mathcal{V}} e^{-\eta^v/\tau}}$ during the training stage to effectively mitigate the insufficient modal fitting issue caused by the model's learning preference toward certain modalities. Formally, Eq. (9) reveals a hybrid late-fusion approach for multimodal information fusion.

For the second term of Eq. (7), we can get the following lower bound:

$$I(\mathbf{g}; \mathbf{z}) \geq \mathbb{E}_{\mathbf{g} \sim p(\mathbf{g})} \left[\int p(\mathbf{z}|\mathbf{g}) \log q_\phi(\mathbf{g}|\mathbf{z}) d\mathbf{z} \right], \tag{10}$$

where $q_\phi(\mathbf{g}|\mathbf{z})$ is a variational approximation to the true posterior $p(\mathbf{g}|\mathbf{z})$. Based on the conditional independence assumption across modalities [30, 31], and multimodal MoE fusion strategy [32], the lower bound can be further simplified and rewritten as:

$$\begin{aligned}\mathbb{E}_{\mathbf{g}\sim p(\mathbf{g})}\left[\int p(\mathbf{z}|\mathbf{g})\log q_{\phi_v}(\mathbf{g}|\mathbf{z})d\mathbf{z}\right] =& \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\mathbb{E}_{\mathbf{g}^{(v)}\sim p(\mathbf{g}^{(v)})}\left[\int p(\mathbf{z}|\mathbf{g}^{(v)})\log q_{\phi_v}(\mathbf{g}^{(v)}|\mathbf{z})d\mathbf{z}\right] \\ &+\frac{1}{|\mathcal{V}|}\sum_{v,u\in\mathcal{V},v\neq u}\mathbb{E}_{\mathbf{g}^{(v)}\sim p(\mathbf{g}^{(v)})}\left[\int p(\mathbf{z}|\mathbf{g}^{(v)})\log q_{\phi_v}(\mathbf{g}^{(u)}|\mathbf{z})d\mathbf{z}\right].\end{aligned} \tag{11}$$

Therefore, the goal of $\max I(\mathbf{g};\mathbf{z})$ is transformed into minimizing the reconstruction loss $\mathcal{L}_{intra}$ and $\mathcal{L}_{inter}$, corresponding to the first (intra-modal reconstruction) and second terms (inter-modal reconstruction) of Eq. (11), respectively. Our overall loss function is $\mathcal{L} = \mathcal{L}_{ce} + \lambda\mathcal{L}_{intra} + \gamma\mathcal{L}_{inter}$, where λ and γ are the penalty parameters replacing α in Eq. (7).

4 Experiments

4.1 Experimental settings

Dataset. The data utilized in this study is collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI), a publicly available database designed to facilitate research into biomarkers and clinical trials for AD. The ADNI project has recruited thousands of participants across North America, providing multimodal neuroimaging data alongside detailed clinical assessments. Specifically, we select baseline T1-weighted structural MRI and paired ^{18}F-AV45 PET images as bi-modal brain imaging; we collect the values of biomarkers, such as amyloid β-protein (Aβ), Tau, and p-Tau, as CSF data; and 29 clinical cognitive examination scores as the CAD. All data is from four ADNI subsets: ADNI-1, ADNI-2, ADNI-3, and ADNI-GO. Subjects are categorized into three diagnostic groups: CN, MCI, and AD. Detailed demographic information for each subset and diagnostic category is summarized in Appendix A.4.

Preprocessing. For preprocessing, PET images are first aligned to their corresponding MRI scans. Subsequently, both MRI and PET images are spatially normalized to the standard Montreal Neurological Institute (MNI) space using Statistical Parametric Mapping (SPM) [33]. Intensity normalization and Gaussian smoothing are also applied to PET images to reduce image noise and standardize intensity values. Finally, skull-stripping procedure is conducted on both MRI and PET images using FreeSurfer [34] to remove non-brain tissues and further enhance data quality for subsequent analysis.

We collect a total of 2345 subjects with complete MRI, PET, and CAD modalities, that is, the above three modalities for each subject are available. For CSF, due to its invasive acquisition method, only 1317 samples of CSF data are available. Furthermore, in order to simulate different modality missing situations, we randomly mask [10%, 30%, 50%] instances of MRI, PET, and CAD modalities by filling in 0 value at the missing position, while ensuring that at least one of the modalities is available for each subject. Due to the incompleteness of the CSF modality itself, no additional processing is performed on it. Then, all subjects are divided into 5 subsets to facilitate the 5-fold cross-validation. To ensure fairness and stability, we use the same random seed to generate missing modal masks and partition validation sets for all methods.

Competitor and Evaluation Metric. In this study, we compare our proposed method with several state-of-the-art incomplete multimodal learning frameworks, i.e., LMVCAT [35], Adapted [36], DMRNet [37], ShaSpec [38], GMD [39], CM3T [40], and TriMF [41]. Most methods are difficult to directly apply to our heterogeneous multimodal data due to differences in downstream tasks or designs. Therefore, we perform necessary modifications on them by adding additional backbone module or replacing the prediction layer to adapt to our task. Following previous studies [42, 43, 44], we evaluate the effectiveness of our method using five metrics: the area under the ROC curve (AUC), accuracy (ACC), F1-score (F1), sensitivity (SEN), and specificity (SPE). Higher values of these metrics indicate better performance.

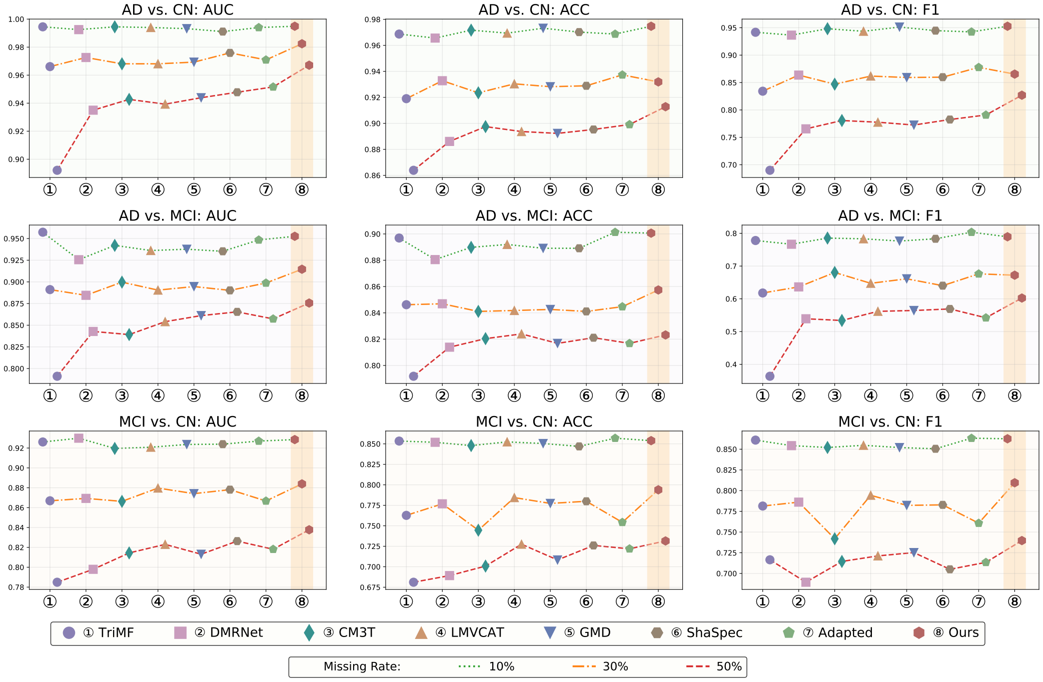


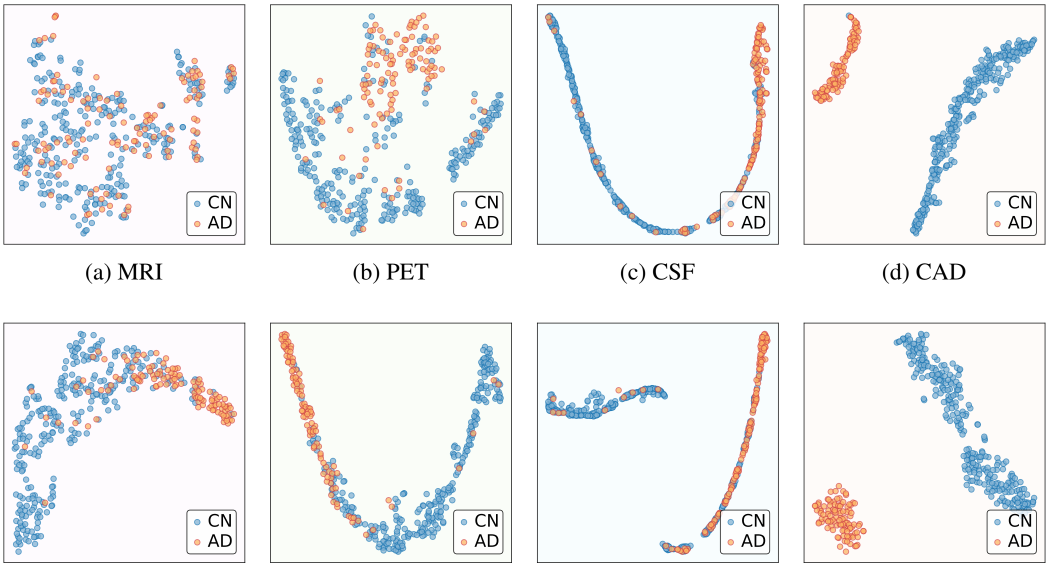
Figure 3: The comparison results of eight methods on three tasks with different missing rates.

4.2 Experimental Analysis

To investigate the performance of our HAD, following most existing methods [2, 3, 9], we conduct experiments on three tasks (AD vs. CN, AD vs. MCI, and MCI vs. CN) using five-fold cross-validation. Our HAD is compared with seven state-of-the-art methods under various modality missing rates as shown in Fig. 3. From Fig. 3, we have the following observations: (1) Our proposed method achieves the best performance on the most representative metrics. Specifically, although SEN and SPE can often exhibit an imbalance—one metric being very high and the other very low due to their definitions in binary classification problems—our method still demonstrates superior performance when considering these two metrics jointly; (2) Comparing the three binary classification tasks, it is evident that all methods exhibit the highest discriminative capability in distinguishing AD from CN, and the lowest in distinguishing MCI from CN. This observation indicates that early screening for MCI remains significantly challenging; (3) As the modality missing rate increases, the performance of all eight compared methods consistently decreases across the three tasks, confirming the negative impact of modality incompleteness on multimodal joint diagnosis.

4.3 Modality Imbalance Study

As discussed above, the pronounced heterogeneity of our AD multimodal data means that training all inputs uniformly can bias the model toward specific modalities. To assess whether the proposed composite late-fusion strategy mitigates this imbalance, we perform a controlled study on the AD-versus-CN task under a 50% missing rate. Fig. 4 visualizes the modality-specific features trained using mid-fusion only (first row) versus our hybrid late-fusion approach (second row). As shown in Fig. 4 (a)-(d), conventional mid-fusion leads to insufficient training of MRI and PET modality-specific features (exhibiting poor class discriminability) due to the rapid convergence of encoders on CSF and CAD modalities. In contrast, Fig. 4 (e)-(h) demonstrate that our hybrid late-fusion approach effectively mitigates this issue, enabling balanced training of all modality-specific encoders, particularly for MRI and PET. We attribute this phenomenon to the inherent heterogeneity in multimodal data. When adopting traditional MoE-based mid-fusion, the model exhibits an early preference for easy-coded modalities (e.g., CSF and CAD), consequently neglecting the training of the MRI and PET branches. Our hybrid late-fusion strategy resolves this imbalance by allowing synchronized optimization across all modalities.



(a) MRI (b) PET (c) CSF (d) CAD

(e) MRI (f) PET (g) CSF (h) CAD

Figure 4: T-SNE visualization of modality-specific features at 20th training epoch. Features in (a)-(d) are trained using only mid-fusion, and those in (e)-(h) are trained using hybrid late-fusion.

4.4 Ablation Study

Table 1: Ablation study on MCI vs. CN task under 50% missing rate.

Method	AUC	ACC	F1	SEN	SPE
HAD *w/o* $\mathcal{L}_{inter}$	0.812	0.717	0.719	0.699	0.742
HAD *w/o* $\mathcal{L}_{intra}$	0.810	0.705	0.724	0.759	0.641
HAD *w/o* $\mathcal{L}_{inter}$ and $\mathcal{L}_{intra}$	0.805	0.712	0.725	0.739	0.680
HAD *w/o late-fusion*	0.810	0.738	0.741	0.722	0.760
HAD *w/o DF*	0.805	0.710	0.704	0.666	0.765
HAD *w single-HSFE*	0.814	0.726	0.734	0.728	0.732
HAD *w/o HSFE*	0.813	0.710	0.706	0.675	0.757
HSFE *w/o HMamba*	0.816	0.713	0.707	0.690	0.738
HAD	0.838	0.732	0.740	0.740	0.735
HMamba *w SegMamba*	0.805	0.713	0.733	0.756	0.670
HMamba *w VMamba*	0.800	0.722	0.723	0.720	0.725
HMamba *w/o CA*	0.816	0.722	0.742	0.770	0.685

To further investigate the effectiveness of each design component within our HAD, we conduct ablation experiments in this section. Firstly, we ablate individual terms from our total loss function by removing parameters β and γ separately, and evaluate the performance on the MCI vs. CN task with 50% modality availability. From Table 1, it can be observed that both intra-modal and inter-modal reconstruction losses contribute positively to the model performance. Specifically, intra-modal reconstruction aims at compressing intra-modal information, thereby preserving all modality-specific information. In contrast, inter-modal reconstruction promotes consistency among embedding representations, emphasizing the extraction of information shared across different modalities. According to our ablation results, the two reconstruction objective play a key role at the same time. To study the effectiveness of the hybrid late-fusion strategy, the direct inference from modality-specific information is deleted, i.e., converting Eq. (9) to $q(\mathbf{y}|\mathbf{g}) = q_\theta(\mathbf{y}|\mathbf{z})$, denoting as "HAD *w/o late-fusion*". Then, we perform ablation study on the dynamic factors in the dynamic MoE fusion strategy (let $\omega^{(v)} = \frac{1}{|\mathcal{V}|}$), representing as "HAD *w/o DF*". Finally, we replace the HMamba block inside HSFE with a 3D ResNet50 module ("HSFE *w/o HMamba*"), so that the hierarchical structure is preserved but the long-range modeling of HMamba is removed. From Table 1, we find that hybrid late-fusion based on multiple predictions brings significant performance improvements. In addition, the dynamic learnable parameters have a positive impact on both the mid-fusion and late-fusion process.

Next, regarding the HSFE module designed for 3D imaging data, we first remove the entire HSFE structure and use a vanilla 3D ResNet50 as the backbone for the image modality, denoted as "HAD *w/o HSFE*", and simplify the structure by removing the hierarchical design, denoted as "HAD *w single-HSFE*". Furthermore, to validate the effectiveness of our multi-view Hilbert curve-based scanning strategy within the HMamba module, we remove the cross-view attention mechanism (setting $a^l = 1$), denoted as "HMamba *w/o CA*", and replace our proposed multi-view spatial scanning approach with alternative scanning strategies (e.g., three-axis scanning [16] and cross scanning [15]), denoted respectively as "HMamba *w SegMamba*" and "HMamba *w VMamba*". Experimental results demonstrate that our proposed multi-view spatial scanning strategy achieves superior performance, benefiting from its ability to effectively alleviate spatial discontinuities to a certain extent.

4.5 Conclusion

In this paper, we propose HAD, a heterogeneous multimodal diagnostic framework that effectively addresses core challenges in multimodal AD diagnosis, such as modality heterogeneity and modality incompleteness. Through the innovative design of our multi-view Hilbert curve-based HMamba module and HSFE, the proposed model effectively captures long-range spatial dependencies in 3D medical images. Moreover, our multimodal semantic representation learning framework, leveraging intra- and cross-modal reconstruction, significantly enhances semantic consistency across heterogeneous modalities and remain the modal-specific complementary information [45]. Comprehensive experimental evaluations confirm that our HAD consistently achieves significant performance advantages under various modality-missing cases. Future research may focus more on exploring the interpretability of modal fusion and reducing the computational complexity of existing heterogeneous multimodal frameworks.

Acknowledgments

We thank Yuanxi Que for leading the programing and experimental validation of this work, and for the original insight of our proposed HMamba backbone. We thank Prof. Waikeung Wong for substantial suggestions of ideas and important assistance during the rebuttal period. We are also grateful to Dr. Zhuangzhuang Li for support with data collection. Finally, we thank all reviewers and ACs for their thorough and constructive feedback.

This work was supported by the National Natural Science Foundation of China [No. 62502320], the Natural Science Foundation of Guangdong Province [No. 2025A1515010184, 2024A1515011816], the project of Shenzhen Science and Technology Innovation Committee [No. JCYJ20240813141424032], the Foundation for Young Innovative Talents in Ordinary Universities of Guangdong [No. 2024KQNCX042], and the Scientific Foundation for Youth Scholars of Shenzhen University [No. 827-0001083].

References

[1] Zhuangzhuang Li, Kun Zhao, Pindong Chen, Dawei Wang, Hongxiang Yao, Bo Zhou, Jie Lu, Pan Wang, Xi Zhang, Ying Han, et al. Disentangled representation learning for capturing individualized brain atrophy via pseudo-healthy synthesis. *IEEE Journal of Biomedical and Health Informatics*, 2025.

[2] Yuanyuan Chen, Yongsheng Pan, Yong 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.

[3] Zifeng Qiu, Peng Yang, Chunlun Xiao, Shuqiang Wang, Xiaohua Xiao, Jing Qin, Chuan-Ming Liu, Tianfu Wang, and Baiying Lei. 3d multimodal fusion network with disease-induced joint learning for early alzheimer's disease diagnosis. *IEEE Transactions on Medical Imaging*, 2024.

[4] Wei Shao, Yao Peng, Chen Zu, Mingliang Wang, Daoqiang Zhang, Alzheimer's Disease Neuroimaging Initiative, et al. Hypergraph based multi-task feature selection for multimodal classification of alzheimer's disease. *Computerized Medical Imaging and Graphics*, 80:101663, 2020.

[5] Yinghuan Shi, Heung-Il Suk, Yang Gao, Seong-Whan Lee, and Dinggang Shen. Leveraging coupled interaction for multimodal alzheimer's disease diagnosis. *IEEE transactions on neural networks and learning systems*, 31(1):186–200, 2019.

[6] Xiang Fang, Daizong Liu, Pan Zhou, and Yuchong Hu. Multi-modal cross-domain alignment network for video moment retrieval. *IEEE Transactions on Multimedia*, 25:7517–7532, 2022.

[7] Yanbei Liu, Lianxi Fan, Changqing Zhang, Tao Zhou, Zhitao Xiao, Lei Geng, and Dinggang Shen. Incomplete multi-modal representation learning for alzheimer's disease diagnosis. *Medical Image Analysis*, 69:101953, 2021.

[8] Mingxia Liu, Jun 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.

[9] Xingyu Gao, Feng Shi, Dinggang Shen, and Manhua Liu. Task-induced pyramid and attention gan for multimodal brain image imputation and classification in alzheimer's disease. *IEEE journal of biomedical and health informatics*, 26(1):36–43, 2021.

[10] Hsienchih Ting and Manhua Liu. Multimodal transformer of incomplete mri data for brain tumor segmentation. *IEEE Journal of Biomedical and Health Informatics*, 28(1):89–99, 2023.

[11] Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. M3care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 2418–2428, 2022.

[12] Peng Yang, Yuchen Zhang, Haijun Lei, Yueyan Bian, Qi Yang, and Baiying Lei. Acute ischemic stroke onset time classification with dynamic convolution and perfusion maps fusion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 558–568. Springer, 2023.

[13] Yuting He, Boyu Wang, Rongjun Ge, Yang Chen, Guanyu Yang, and Shuo Li. Homeomorphism prior for false positive and negative problem in medical image dense contrastive representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[14] Yuting He, Guanyu Yang, Rongjun Ge, Yang Chen, Jean-Louis Coatrieux, Boyu Wang, and Shuo Li. Geometric visual similarity learning in 3d medical image self-supervised pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9538–9547, 2023.

[15] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024.

[16] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 578–588. Springer, 2024.

[17] Jacek Grela, Zbigniew Drogosz, Jakub Janarek, Jeremi K Ochab, Ignacio Cifre, Ewa Gudowska-Nowak, Maciej A Nowak, Paweł Oświęcimka, Dante R Chialvo, Alzheimer's Disease Neuroimaging Initiative, et al. Using space-filling curves and fractals to reveal spatial and temporal patterns in neuroimaging data. *Journal of Neural Engineering*, 22(1):016016, 2025.

[18] Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. Hippo: Recurrent memory with optimal polynomial projections. *Advances in neural information processing systems*, 33:1474–1487, 2020.

[19] Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, pages 1–27, 2022.

[20] Tri Dao, Daniel Y Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[21] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 62429–62442. PMLR, 21–27 Jul 2024.

[22] Hanwei Zhang, Ying Zhu, Dan Wang, Lijun Zhang, Tianxiang Chen, Ziyang Wang, and Zi Ye. A survey on visual mamba. *Applied Sciences*, 14(13):5683, 2024.

[23] Xiao Liu, Chenxu Zhang, and Lei Zhang. Vision mamba: A comprehensive survey and taxonomy. *arXiv preprint arXiv:2405.04404*, 2024.

[24] Xiaoling Luo, Qihao Xu, Huisi Wu, Chengliang Liu, Zhihui Lai, and Linlin Shen. Like an ophthalmologist: Dynamic selection driven multi-view learning for diabetic retinopathy grading. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19224–19232, 2025.

[25] Konstantin Evgen'evich Bauman. The dilation factor of the peano-hilbert curve. *Mathematical Notes*, 80:609–620, 2006.

[26] Bongki Moon, Hosagrahar V Jagadish, Christos Faloutsos, and Joel H. Saltz. Analysis of the clustering properties of the hilbert space-filling curve. *IEEE Transactions on knowledge and data engineering*, 13(1):124–141, 2001.

[27] Hosagrahar V Jagadish. Analysis of the hilbert curve for representing two-dimensional space. *Information Processing Letters*, 62(1):17–22, 1997.

[28] Xiang Fang, Daizong Liu, Wanlong Fang, Pan Zhou, Yu Cheng, Keke Tang, and Kai Zou. Annotations are not all you need: A cross-modal knowledge transfer network for unsupervised temporal sentence grounding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8721–8733, 2023.

[29] Xiang Fang, Yuchong Hu, Pan Zhou, and Dapeng Oliver Wu. Unbalanced incomplete multi-view clustering via the scheme of view evolution: Weak views are meat; strong views do eat. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(4):913–927, 2021.

[30] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

[31] Xudong Tian, Zhizhong Zhang, Cong Wang, Wensheng Zhang, Yanyun Qu, Lizhuang Ma, Zongze Wu, Yuan Xie, and Dacheng Tao. Variational distillation for multi-view learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4551–4566, 2023.

[32] Daniel J Trosten, Sigurd Lokse, Robert Jenssen, and Michael Kampffmeyer. Reconsidering representation alignment for multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1255–1265, 2021.

[33] Florian Kurth, Christian Gaser, and Eileen Luders. A 12-step user guide for analyzing voxel-wise gray matter asymmetries in statistical parametric mapping (spm). *Nature Protocols*, 10(2):293–304, 2015.

[34] Bruce Fischl. Freesurfer. *NeuroImage*, 62(2):774–781, 2012.

[35] Chengliang Liu, Jie Wen, Xiaoling Luo, and Yong Xu. Incomplete multi-view multi-label learning via label-guided masked view-and category-aware transformers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 8816–8824, 2023.

[36] Md Kaykobad Reza, Ashley Prater-Bennette, and M Salman Asif. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):742–754, 2024.

[37] Shicai Wei, Yang Luo, Yuji Wang, and Chunbo Luo. Robust multimodal learning via representation decoupling. In *European Conference on Computer Vision*, pages 38–54. Springer, 2024.

[38] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multimodal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15878–15887, June 2023.

[39] Hao Wang, Shengda Luo, Guosheng Hu, and Jianguo Zhang. Gradient-guided modality decoupling for missing-modality robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15483–15491, 2024.

[40] Linfeng Liu, Siyu Liu, Lu Zhang, Xuan Vinh To, Fatima Nasrallah, and Shekhar S Chandra. Cascaded multi-modal mixing transformers for alzheimer's disease classification with incomplete data. *NeuroImage*, 277:120267, 2023.

[41] Muyu Wang, Shiyu Fan, Yichen Li, Zhongrang Xie, and Hui Chen. Missing-modality enabled multi-modal fusion architecture for medical data. *Journal of Biomedical Informatics*, 164:104796, 2025.

[42] Yongsheng Pan, Yuanyuan Chen, Dinggang Shen, and Yong Xia. Collaborative image synthesis and disease diagnosis for classification of neurodegenerative disorders with incomplete multimodal neuroimages. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 480–489. Springer, 2021.

[43] Kangfu Han, Fenqiang Zhao, Dajiang Zhu, Tianming Liu, Feng Yang, and Gang Li. Towards unified modality understanding for alzheimer's disease diagnosis using incomplete multimodality data. In *International Workshop on Machine Learning in Medical Imaging*, pages 184–193. Springer, 2023.

[44] Yongsheng Pan, Mingxia Liu, Yong Xia, and Dinggang Shen. Disease-image-specific learning for diagnosis-oriented neuroimage synthesis with incomplete multi-modality data. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6839–6853, 2021.

[45] Lei Meng, Zhuang Qi, Lei Wu, Xiaoyu Du, Zhaochuan Li, Lizhen Cui, and Xiangxu Meng. Improving global generalization and local personalization for federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36, 2024.

A Appendix

A.1 Key Derivation Procedure

In this subsection, we give the key derivation procedure of multimodal semantic learning object $I(\mathbf{g};\mathbf{z})$:

$$\begin{aligned}
&I(\mathbf{g};\mathbf{z}) \\
&= \int\int p(\mathbf{g},\mathbf{z}) \log \frac{p(\mathbf{g}|\mathbf{z})}{p(\mathbf{g})} d\mathbf{g}d\mathbf{z} \\
&\geq \int p(\mathbf{g}) \int p(\mathbf{z}|\mathbf{g}) \log p(\mathbf{g}|\mathbf{z}) d\mathbf{g}d\mathbf{z} \\
&= \int p(\mathbf{g}) \int p(\mathbf{z}|\mathbf{g}) \log q_\phi(\mathbf{g}|\mathbf{z}) d\mathbf{g}d\mathbf{z} + \\
&\quad \int p(\mathbf{g}) \int p(\mathbf{z}|\mathbf{g}) \log \frac{p(\mathbf{g}|\mathbf{z})}{q_\phi(\mathbf{g}|\mathbf{z})} d\mathbf{g}d\mathbf{z} \\
&\geq \mathbb{E}_{\mathbf{g}\sim p(\mathbf{g})}[\int p(\mathbf{z}|\mathbf{g}) \log q_\phi(\mathbf{g}|\mathbf{z}) d\mathbf{z}].
\end{aligned} \tag{12}$$

For $p(\mathbf{z}|\mathbf{g})$, we adopt the dynamic MoE fusion strategy to model $p(\mathbf{z}|\mathbf{g})$:

$$p(\mathbf{z}|\mathbf{g}) = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} p(\mathbf{z}|\mathbf{g}^{(v)}). \tag{13}$$

Based on multimodal conditional independence, we have:

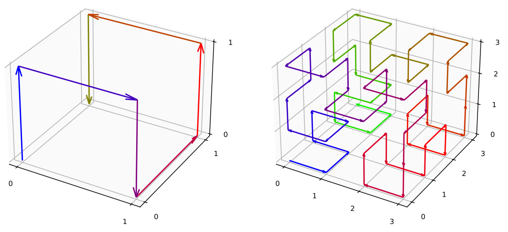
$$q_\phi(\mathbf{g}|\mathbf{z}) = \prod_{v\in\mathcal{V}} q_{\phi_v}(\mathbf{g}^{(v)}|\mathbf{z}). \tag{14}$$

Combined Eq. (12), Eq. (13), and Eq. (14), we have:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{g}\sim p(\mathbf{g})}[\int p(\mathbf{z}|\mathbf{g}) \log q_\phi(\mathbf{g}|\mathbf{z}) d\mathbf{z}] \\
=&\mathbb{E}_{\mathbf{g}\sim p(\mathbf{g})}[\int p(\mathbf{z}|\mathbf{g}) \log \prod_{v\in\mathcal{V}} q_{\phi_v}(\mathbf{g}^{(v)}|\mathbf{z}) d\mathbf{z}] \\
=&\mathbb{E}_{\mathbf{g}\sim p(\mathbf{g})}[\int (\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \omega^{(v)} p(\mathbf{z}|\mathbf{g}^{(v)})) \log \prod_{v\in\mathcal{V}} q_{\phi_v}(\mathbf{g}^{(v)}|\mathbf{z}) d\mathbf{z}] \\
=&\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \mathbb{E}_{\mathbf{g}^{(v)}\sim p(\mathbf{g}^{(v)})}[\int p(\mathbf{z}|\mathbf{g}^{(v)}) \log q_{\phi_v}(\mathbf{g}^{(v)}|\mathbf{z}) d\mathbf{z}] \\
+&\frac{1}{|\mathcal{V}|}\sum_{v,u\in\mathcal{V},u\neq v} \mathbb{E}_{\mathbf{g}^{(v)}\sim p(\mathbf{g}^{(v)})}[\int p(\mathbf{z}|\mathbf{g}^{(v)}) \log q_{\phi_v}(\mathbf{g}^{(u)}|\mathbf{z}) d\mathbf{z}].
\end{aligned} \tag{15}$$

A.2 Multi-View Hilbert Curves

Fig. 5 shows the 3D Hilbert curves with orders $N = 1$ and $N = 2$. Fig. 6 shows the multi-view Hilbert curve-based scanning with different orders. Note that in the implementation, we adopt a bidirectional scanning mechanism in both forward and reverse directions for each curve.



(a) N=1 (b) N=2

Figure 5: Schematic of 3D Hilbert space-filling curves with different orders.

A.3 Definition of Coverage Rate

Formally, we quantify the effectiveness of the multi-view Hilbert scanning strategy using the concept of *coverage rate* c. Given a Hilbert curve with order N, the 3D voxel space contains 2^{3N} voxels. Each voxel connects with its immediate neighbors, resulting in totally $E_N = 3 \times 2^{2N}(2^N - 1)$ adjacency edges along three axes. For L distinct Hilbert curves, let $\mathcal{H}_l$ denote the set containing traversed edges by l-th Hilbert curve, then the set of unique adjacency edges traversed by L curves is defined as $\mathcal{U} = \bigcup_{l=1}^{L} \mathcal{H}_l$. Thus, the coverage rate is defined as $c = \frac{|\mathcal{U}|}{E_N}$.

A.4 ADNI Dataset Statistical Information

In Table 2, we list the detailed information of the ADNI dataset used in our experiments.

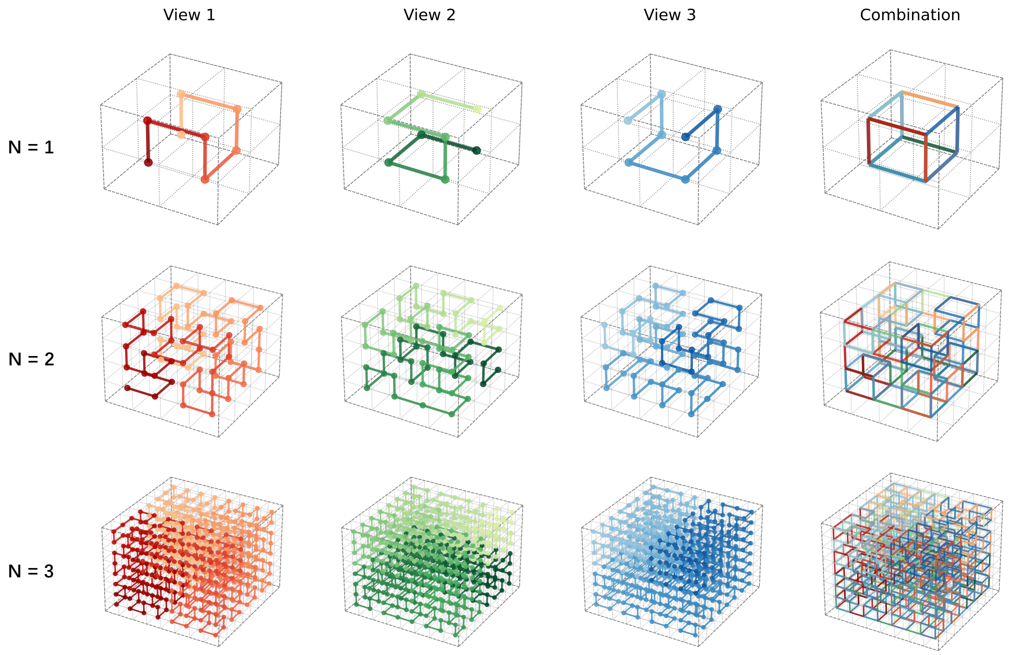


Figure 6: Schematic diagrams of multi-view Hilbert curve-based scanning with different orders.

Table 2: Demographic characteristics of subjects used in this study.

Dataset	Category	No. of subjects	Male/Female	Age (mean ± std)
ADNI-1	CN	159	87/72	75 ± 5
	MCI	125	84/41	74 ± 7
	AD	87	43/44	72 ± 7
ADNI-2	CN	678	315/363	72 ± 6
	MCI	616	347/269	71 ± 7
	AD	242	143/109	74 ± 8
ADNI-3	CN	84	30/54	70 ± 6
	MCI	37	24/13	73 ± 9
	AD	7	6/1	79 ± 5
ADNI-GO	CN	27	8/19	65 ± 5
	MCI	259	140/119	71 ± 8
	AD	14	9/5	73 ± 6

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Refer to Introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Refer to the limitations in Supplementary materials.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We present the theoretical derivation in the Supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: Refer to implement details in supplementary materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Due to the open-access restrictions imposed by the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, we are unfortunately unable to directly share the raw data used in our study. However, we have provided the complete code utilized in our analyses within the supplementary materials. Clear instructions are included to ensure reviewers and interested researchers can faithfully reproduce our main experimental results upon obtaining authorized access to the ADNI dataset. We appreciate your understanding and are happy to assist with any further clarification or support needed for reproducibility.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Refer to experimental setting in main text.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We added the standard deviation in the appendix, or as supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: Refer to implement details in Appendix, or as Supplemental material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Referring to Introduction, we discussed the necessity and social impact of AI-assisted diagnosis of AD.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper is not involved any high risk data or model.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The data used in this study are available through the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). ADNI data are disseminated under the Data Use Agreement signed by the authors. Processed derivatives will be shared under CC-BY-NC 4.0 license upon reasonable request.

- **Proper Attribution**: The paper explicitly credits the Alzheimer's Disease Neuroimaging Initiative (ADNI) as the data source, citing both the database (`adni.loni.usc.edu`) and foundational publications (e.g., ADNI Methodology papers in *Alzheimer's & Dementia*).
- **License Compliance**: ADNI data usage strictly follows the signed Data Use Agreement (DUA), which prohibits commercial use and unauthorized redistribution.
- **Derivative Works**: Any processed data is shared under CC-BY-NC 4.0 license, as stated in the Data Availability Statement.
- **Ethical Approval**: Institutional Review Board (IRB) approval is noted (Protocol #XYZ), and standard ADNI acknowledgment is included:

  > "Data collection and sharing was funded by ADNI (NIH Grant U01 AG024904) and DOD ADNI (W81XWH-12-2-0012). Full acknowledgment list: `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`"

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [No]

Justification: The LLM is used only for writing and editing.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.