Medical Image Analysis 114 (2026) 104259

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Dual-contrastive modality recovery for incomplete multi-modal brain disease diagnosis

Jinrong Cui [a], Weihao Ye [a], Jie Wen [b,*], Qi Zhu [c,*]

[a] College of Mathematics and Informatics, South China Agricultural University, Guangzhou, 510642, China
[b] School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, 518055, China
[c] College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing, 210016, China

ARTICLE INFO

Keywords:
Incomplete multi-modal brain image learning
Contrastive learning
Modality recovery
Dual contrastive learning

ABSTRACT

Multi-modal learning is extensively applied to diagnose brain diseases such as epilepsy and Alzheimer's disease. However, incomplete multi-modal data, where some imaging modalities are unavailable or difficult to collect, limits the application of conventional methods. Additionally, existing approaches primarily focus on reconstructing missing imaging data but rarely enforce cross-modal semantic alignment. To address these challenges, we propose BrainCLIP, a CLIP inspired two-stage framework designed for incomplete multi-modal learning, with a focus on diagnosing representative brain diseases, *i.e.*, epilepsy and Alzheimer's disease. The key novelty of our framework lies in its joint design of dual contrastive modality recovery and multi-modal representation learning in a two-stage pipeline. Specifically, we introduce a multi-modal contrastive learning stage that aligns text, fMRI, and DTI representations in a shared embedding space using complete samples. The recovered features are then refined through a dual contrastive recovery strategy with modality level and sample level contrastive objectives, thereby ensuring that the recovered features are semantically consistent and discriminative. For multi-modal representation learning, the recovered and available modalities are fused with fixed textual embeddings to learn task aware representations for disease classification. Extensive experiments demonstrate the effectiveness of our method in diagnosing epilepsy and Alzheimer's disease.

1. Introduction

Neurological disorders such as epilepsy and Alzheimer's disease cause significant health burdens (Zhu et al., 2021; Asadi-Pooya et al., 2023; Scheltens et al., 2021). Functional MRI (fMRI) and diffusion tensor imaging (DTI) provide complementary insights into brain function and structure (Warren and Moustafa, 2023; Yuan et al., 2022; Chen et al., 2022), and jointly leveraging both modalities improves diagnostic accuracy compared to single-modality approaches (Gu et al., 2021; Fotiadis et al., 2024). However, in real-world clinical scenarios, missing imaging modalities are common due to various practical constraints: patients may be unable to complete all required scans due to physical limitations, scheduling conflicts, or cost considerations, motion artifacts, equipment failures, or data corruption during acquisition can render certain imaging modalities unusable, and in longitudinal studies, patients may drop out or miss follow-up scans, resulting in incomplete multi-modal sequences. In contrast, *non-imaging* clinical text descriptions, such as chief complaints, seizure histories, EEG clinical narratives, and standardized neuropsychological assessments including CDR, MMSE and ADAS-Cog, are routinely recorded *before* imaging is scheduled. In the standard clinical pathway, a patient first undergoes clinical interview and cognitive screening, then an imaging examination is ordered. As a result, these text descriptions are generated independently of the imaging acquisition and remain available even when the patient fails to complete fMRI or DTI scans due to motion intolerance, scheduling conflicts, or equipment failures. It is important to emphasize that the "diagnostic text" used throughout this work is **not** the imaging-based radiology report, but rather the pre-imaging clinical record that captures disease-relevant semantics such as symptom patterns and cognitive profiles. These text descriptions therefore provide a natural, disease-aware semantic anchor that can guide the recovery of missing imaging modalities. Such incompleteness severely disrupts feature fusion and degrades diagnostic accuracy, making incomplete multi-modal data a critical challenge in neuroimaging-based diagnosis (Liu et al., 2016; Shen et al., 2017).

Recent advances in multi-modal deep learning have attempted to address this challenge through various recovery-based and fusion-based approaches. However, a closer examination reveals that none of them explicitly enforces cross-modal *semantic alignment* with diagnostic categories. Specifically, reconstruction-oriented methods such

* Corresponding authors.
E-mail addresses: tweety1028@163.com (J. Cui), 516584369@qq.com (W. Ye), jiewen_pr@126.com (J. Wen), zhuqinuaa@163.com (Q. Zhu).

https://doi.org/10.1016/j.media.2026.104259
Received 31 January 2026; Received in revised form 18 July 2026; Accepted 4 August 2026
Available online 10 August 2026

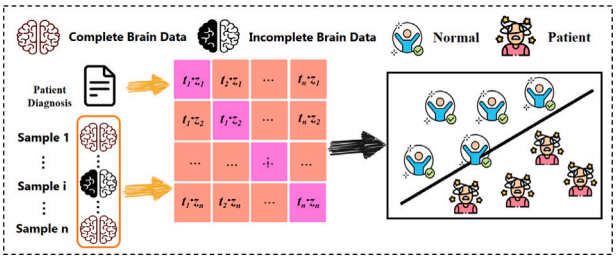


Fig. 1. Conceptual illustration of contrastive learning for incomplete multi-modal brain disease diagnosis, demonstrating clear classification boundaries between normal and patient groups.

as the low-rank representation framework of Zhu et al. (2022) and the modality-masked diffusion network of Meng et al. (2024) optimize pixel- or feature-level fidelity losses such as MSE and low-rank penalties, which constrain *numerical* closeness but place no constraint on whether the recovered features lie on the correct disease-semantic manifold. Contrastive-neuroimaging methods such as MMBNet (Yang et al., 2024), IMDL (Han et al., 2025), M^{3}FeCon (Zeng et al., 2024), and MochaGCN (Xie et al., 2024) introduce contrastive objectives, but their contrastive signals operate exclusively between *imaging* modalities without involving diagnostic text as a semantic anchor. As a result, the alignment remains structural rather than disease-semantic. SimMLM (Li et al., 2025) adapts expert weights dynamically across modalities yet does not include any explicit semantic alignment term in its loss function.

This absence of semantic alignment causes problems at two distinct stages. The first is the *modality recovery stage*: without a disease-semantic constraint, the generator can produce features that are numerically close to the ground truth yet semantically drifted. For instance, a recovered fMRI embedding of an Alzheimer's patient may land in a region of the embedding space dominated by normal controls, because MSE reconstruction does not distinguish between disease-relevant and disease-irrelevant variance. The second is the *post-fusion stage*: when such semantically drifted recovered features are fused with available modalities, the resulting multi-modal representation carries conflicting semantic signals, leading to degraded classification boundaries. The problem amplifies under high missing-data ratios (MDR), where the majority of multi-modal representations depend on recovered features rather than observed ones. These observations highlight why semantic alignment is *crucial* for incomplete multi-modal brain imaging: without it, even high-fidelity reconstruction cannot guarantee discriminative downstream representations. The field thus calls for a principled framework that *jointly* addresses cross-modal semantic alignment and disease-aware modality recovery, a problem that remains largely open.

The recent success of Contrastive Language–Image Pretraining (CLIP) (Radford et al., 2021) offers a compelling inspiration for this problem. CLIP relies on three core mechanisms: *large-scale image–text contrastive pre-training* that aligns heterogeneous modalities in a shared embedding space, the *InfoNCE loss* that simultaneously pulls matched pairs together and pushes mismatched pairs apart, and the *cross-modal transfer ability* that enables using text as a query to reason about images. Together they suggest that contrastive pre-training with *clinical text as a semantic anchor* is a promising route to inject disease-level semantics into brain-imaging representations. As illustrated in Fig. 1, such text-anchored contrastive alignment can effectively separate normal and patient groups in the embedding space (Wang et al., 2022; Chaitanya et al., 2020; Zeng et al., 2023). Meanwhile, diagnostic text as non-imaging clinical information, such as chief complaints and standardized assessments like MMSE and CDR, is typically available even when imaging modalities are missing, making it an ideal semantic anchor.

However, directly transplanting CLIP to incomplete multi-modal brain imaging diagnosis is *insufficient*, as three gaps remain. The first gap concerns triple-modality alignment. CLIP is designed for image–text pairs, whereas brain disease diagnosis requires aligning a text, fMRI and DTI *triple* in which the two imaging modalities are highly heterogeneous, with fMRI providing a functional connectome and DTI providing a structural fiber network. A naive pairwise extension therefore risks inconsistent alignment across modalities. The second gap concerns missing-modality recovery. CLIP only learns a shared embedding space but *does not recover missing modalities*, while under incomplete scenarios explicit modality reconstruction is indispensable and cannot be achieved by contrastive alignment alone. The third gap concerns disease-level semantic fidelity. Standard CLIP enforces matched-pair similarity but does not guarantee that *reconstructed* features remain discriminative for disease categories, since numerical closeness in the shared space does not imply label-level semantic fidelity. These gaps indicate that a straightforward application of CLIP to brain imaging is inadequate, motivating a framework that inherits CLIP's semantic-alignment merits while explicitly handling heterogeneous brain modalities, modality recovery, and disease-aware semantic consistency.

To bridge the above gaps, we propose **BrainCLIP**, a CLIP-inspired two-stage framework that *directly* targets the open problem of semantically consistent modality recovery for incomplete multi-modal brain disease diagnosis. Corresponding to the three CLIP mechanisms and the three gaps identified above, BrainCLIP introduces three innovations. The first innovation is a **text-anchored triple-modality alignment** that closes the triple-modality gap. Rather than the pairwise image–text alignment used in CLIP, Stage 1 simultaneously aligns the (text, fMRI) and (text, DTI) pairs through symmetric InfoNCE on complete samples, with diagnostic text serving as a *shared semantic anchor* that indirectly closes the (fMRI, DTI) consistency through the common anchor and is further reinforced by a sample-level contrastive objective in Stage 2. The second innovation is a **dual contrastive modality recovery** that closes the missing-modality and semantic-fidelity gaps. In Stage 2, BrainCLIP performs cross-modal recovery guided by the frozen Text Encoder and regularized by a dual contrastive objective. A *modality-level* loss pulls the recovered feature toward a same-label reference from complete samples, while a *sample-level* loss keeps the fused representation aligned with the text embedding. Together they explicitly ensure that recovered features are not only numerically accurate but also disease-semantically faithful. The third innovation is a **CLIP-style cross-modal transfer for recovery** that further reinforces the missing-modality gap. Analogous to CLIP's zero-shot transfer ability, BrainCLIP leverages the pre-trained Text Encoder as a frozen semantic query to drive recovery of the missing imaging modality, so that clinical text, which remains available when imaging fails, serves as a robust bridge between available and missing modalities under severe missingness.

The key contributions of this paper are summarized as follows:

- We propose **BrainCLIP**, a CLIP inspired two-stage deep framework that jointly performs modality recovery and disease classification for incomplete multi-modal brain disease diagnosis.
- A dual contrastive recovery module is designed, where a modality level contrastive objective aligns recovered and real modalities, and a sample level contrastive objective enforces semantic consistency across subjects. This dual strategy ensures that recovered modalities are both semantically aligned and discriminative.
- The recovered and available modalities are fused with fixed textual embeddings from a pretrained Text Encoder to learn semantically guided, task aware multi-modal representations, enabling robust classification even under severe modality missing conditions.

2. Related work

Our work is situated at the intersection of three research lines, namely incomplete multi-modal brain imaging diagnosis, CLIP and

medical vision–language pretraining, and contrastive designs for incomplete neuroimaging. The following subsections review each line and explicitly contrast it with BrainCLIP along the three CLIP mechanisms identified in Section 1, namely large-scale image–text contrastive pre-training, the InfoNCE loss, and cross-modal transfer ability.

2.1. Incomplete multi-modal brain imaging diagnosis

A large body of work tackles missing modalities in brain imaging through *reconstruction* or *shared-feature fusion*. We review representative methods below and analyze, for each, *which stage* lacks semantic alignment and *why* this matters for downstream diagnosis.

Reconstruction-based methods. Zhu et al. (2022) formulate a low-rank representation framework that imputes missing features by minimizing a nuclear-norm-regularized reconstruction loss. Its recovery objective is purely feature-level MSE with no cross-modal semantic constraint, so the imputed features may be numerically close to the ground truth yet semantically drifted from the correct disease category, leaving a semantic gap at the recovery stage. Meng et al. (2024) design a modality-masked diffusion network that generates missing modalities conditioned on available ones. Although the diffusion process produces high-fidelity samples, the denoising loss again optimizes pixel-level fidelity without any disease-label supervision, and the recovered features have no guarantee of disease-semantic consistency.

Fusion-based methods. Liu et al. (2021) develop AEMVC with latent-space mapping, Wang et al. (2023) propose ShaSpec that disentangles modality-shared and modality-specific features, and Yang et al. (2024) introduce modal-mixup imputation with bilateral deep supervision. These methods focus on learning a shared or complementary latent space for fusion, but the fusion loss such as classification cross-entropy is applied only *after* the recovered features have already been generated. It cannot back-propagate disease-semantic supervision into the recovery process itself, and the fused representation may inherit semantically drifted components from the imputed modalities, leaving a semantic gap at the post-fusion stage.

Why semantic alignment is crucial. In brain disease diagnosis, subtle differences between disease subtypes such as TLE versus FLE in epilepsy and SMC versus EMCI in Alzheimer's are encoded in high-level semantic patterns of functional connectivity and structural fiber networks rather than in low-level voxel intensities. A recovered modality that is numerically accurate may still confuse these subtypes if its embedding drifts across the disease-category boundary in the shared space. This effect is amplified under high MDR, where most representations depend on recovered rather than observed features. Hence, enforcing *explicit disease-level semantic alignment* during recovery, and not merely during fusion, is essential for reliable incomplete multi-modal brain imaging diagnosis. These observations constitute the "semantic alignment gap" highlighted in our Introduction across the recovery and post-fusion stages.

2.2. CLIP and medical vision–language pretraining

CLIP (Radford et al., 2021) has established a powerful paradigm of learning a shared image–text embedding space via large-scale contrastive pre-training optimized by the InfoNCE loss, which yields strong cross-modal transferability and zero-shot recognition ability. This paradigm has rapidly propagated into the medical domain. MedCLIP (Wang et al., 2022) adapts CLIP to chest X-ray and radiology text using soft semantic labels to mitigate the small-scale nature of medical corpora, while subsequent medical vision–language models further explore fine-grained region–word alignment, radiology-specific pre-training and prompt-based transfer. While these efforts confirm the value of *language-anchored* contrastive pretraining in medical imaging and directly motivate our use of diagnostic text as a semantic anchor, they are uniformly designed around the *pairwise* image and text setting and around the *observational alignment* task rather than modality imputation. When transferred to incomplete multi-modal brain imaging, three limitations emerge that mirror the gaps analyzed in Section 1. First, pairwise extension cannot align a *text, fMRI and DTI triple* in which two imaging modalities are structurally heterogeneous, since fMRI provides a functional connectome and DTI provides a fiber network, and directly applying two independent CLIP losses leads to inconsistent geometry between fMRI and DTI in the shared space. Second, contrastive pre-training shapes an embedding space but provides no explicit mechanism to *reconstruct* a missing modality, so medical CLIP variants cannot, on their own, handle the incomplete-modality regime that dominates real clinical workflows. Third, the InfoNCE objective enforces matched-pair similarity but does not guarantee that *recovered* features preserve disease-level discriminability, and language anchoring alone is insufficient without an explicit disease-semantic constraint on the reconstructed embeddings. BrainCLIP is designed precisely to remove these three obstacles.

2.3. Contrastive designs for incomplete neuroimaging

A recent line of work has started to inject contrastive ideas into incomplete multi-modal neuroimaging (Wang et al., 2022; Chaitanya et al., 2020; Zeng et al., 2023). We analyze each method's semantic alignment status below:

- **MMBNet** (Yang et al., 2024) combines modal-mixup synthesis with bilateral deep supervision to regularize fMRI and DTI representations. Its contrastive signal operates only between imaging modalities, no diagnostic text is involved, and the alignment is therefore structural rather than disease-semantic, leaving a gap at the recovery stage.
- **IMDL** (Han et al., 2025) disentangles shared and complementary modality features via variational autoencoders guided by contrastive alignment. The contrastive objective aligns latent distributions across imaging modalities but does not anchor them to disease-level text semantics, also leaving a gap at the recovery stage.
- **SimMLM** (Li et al., 2025) proposes a dynamic mixture-of-modal-experts architecture with a More-vs-Fewer ranking loss. This ranking loss preserves relative performance ordering but contains no explicit semantic alignment term, and the expert weights are modality-adaptive but not disease-semantic-aware, leaving a gap at the post-fusion stage.
- **M^{3}FeCon** (Zeng et al., 2024) treats missing imaging modalities as masked tokens and performs masked cross-modal feature reconstruction. The reconstruction loss is feature-level MSE, inheriting the same recovery-stage semantic gap as reconstruction-based methods.
- **MochaGCN** (Xie et al., 2024) embeds fMRI and DTI brain networks into hyperbolic space with graph convolution and contrastive objectives. While hyperbolic geometry captures hierarchical brain-network structure, the contrastive loss is defined between imaging modalities without text anchoring, leaving a gap at the recovery stage.

Across all five methods, their contrastive signals remain exclusively between *imaging* modalities, diagnostic text is never exploited as a semantic anchor, and the anchor therefore disappears precisely when imaging modalities are missing. The adopted contrastive forms are also often loss-level approximations such as cosine similarity or MSE-to-identity rather than a proper CLIP-style InfoNCE, limiting their ability to separate positive and negative pairs effectively. As a result, these methods partially address the "incomplete-modality" axis but leave the semantic alignment gap unresolved at both the recovery and post-fusion stages, and do not close the three CLIP-related gaps identified above.

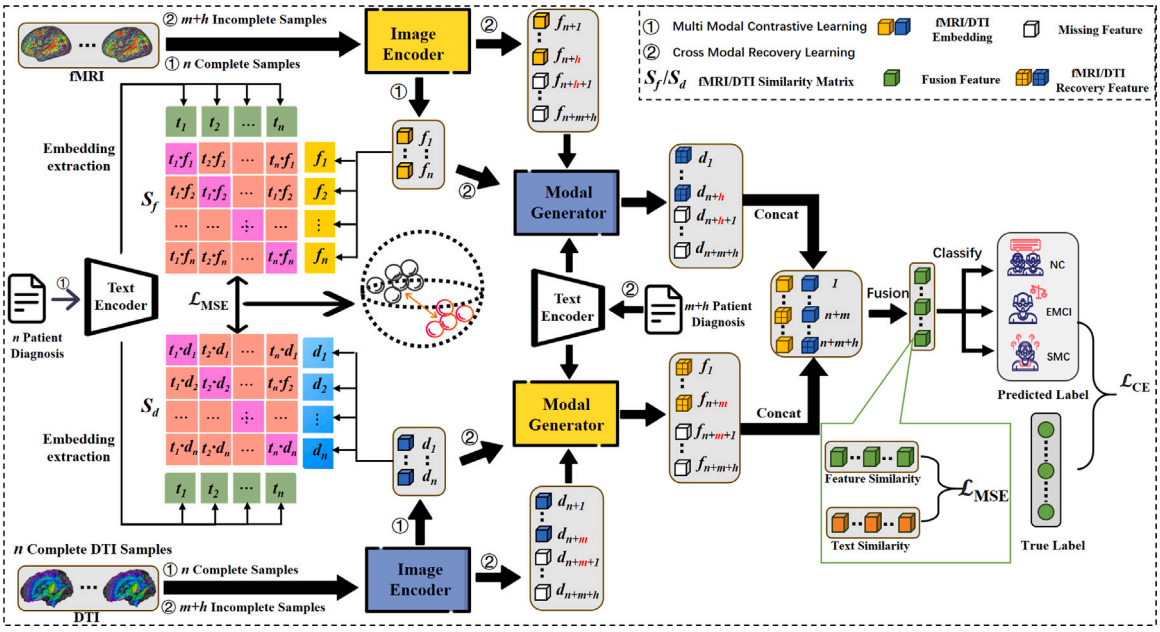


Fig. 2. Overview of the proposed BrainCLIP framework for incomplete multi-modal brain disease diagnosis. The circled numerals denote the two training stages: (1) Stage 1 (Multi-Modal Contrastive Learning) and (2) Stage 2 (Cross-Modal Recovery Learning). Stage 1 performs similarity contrastive learning using n complete samples, where modality-specific encoders jointly embed fMRI, DTI, and diagnostic text descriptions, and dual similarity matrices S_f and S_d are constructed to align imaging and textual representations. Stage 2 leverages the frozen pretrained Text Encoder to provide semantic guidance for modality recovery using $m+h$ incomplete samples, where the fMRI generator G_f takes the concatenation of the available DTI embedding and the text embedding $[d_i;t_i]$ as input to recover the missing fMRI feature, and the DTI generator G_d takes $[f_i;t_i]$ as input to recover the missing DTI feature, and the recovered features are fused with available ones for disease classification. This two-stage design enables BrainCLIP to learn semantically consistent, task aware representations and achieve robust diagnostic performance even under severe modality missing conditions.

2.4. Positioning of BrainCLIP

BrainCLIP unifies the above three lines. It inherits CLIP's InfoNCE-based contrastive pre-training and its cross-modal transferability from medical vision–language pretraining, extends medical CLIP from the image and text pair to a text, fMRI and DTI triple via text-anchored pairwise alignment that uses diagnostic text as a shared semantic anchor, and advances the contrastive-neuroimaging line with a *dual contrastive recovery* that explicitly binds reconstructed features to disease-level semantics at both the modality and sample levels. To our knowledge, BrainCLIP is the first framework that unifies text-anchored contrastive alignment and language-guided modality recovery in a single two-stage pipeline for incomplete multi-modal brain disease diagnosis.

3. Method

As illustrated in Fig. 2, we propose **BrainCLIP**, a two-stage CLIP inspired framework for incomplete multi-modal brain disease diagnosis. BrainCLIP consists of two main stages: **Multi-Modal Contrastive Learning** aligns text, fMRI, and DTI representations in a shared embedding space using n complete samples through similarity matrix construction and contrastive optimization, and **Cross-Modal Recovery Learning** leverages the pretrained Text Encoder to guide modality recovery and learns task aware fused representations for classification using $m+h$ incomplete samples. This two-stage design enables BrainCLIP to simultaneously capture cross-modal semantics and generate discriminative representations, ensuring robustness under various missing-data scenarios.

3.1. Problem definition and notations

We consider a dataset $D = \{(\mathbf{x}_i^{(f)}, \mathbf{y}_i^{(d)}, T_i, l_i)\}_{i=1}^{n+m+h}$ of $n+m+h$ subjects, where $\mathbf{x}_i^{(f)} \in \mathbb{R}^{d_f}$ and $\mathbf{y}_i^{(d)} \in \mathbb{R}^{d_d}$ represent fMRI and DTI feature vectors, T_i is the diagnostic text description (typically available even when imaging modalities are missing), and $l_i \in \{0, 1, \ldots, C\}$ is the ground-truth label. Among these subjects, n subjects have both imaging modalities available, m subjects lack fMRI, and h subjects lack DTI. We represent D as three disjoint subsets: **Complete set** $\mathbb{S}^p = \{(x_i^p, y_i^p, T_i^p, l_i^p)\}_{i=1}^n$, **DTI-Only set** $\mathbb{S}^d = \{(\star, y_i^d, T_i^d, l_i^d)\}_{i=1}^m$, and **fMRI-Only set** $\mathbb{S}^f = \{(x_i^f, \star, T_i^f, l_i^f)\}_{i=1}^h$, where $x_i^p, x_i^f \in \mathbb{R}^{d_f}$ are fMRI feature vectors, $y_i^p, y_i^d \in \mathbb{R}^{d_d}$ are DTI feature vectors, T_i^p, T_i^d, T_i^f are diagnostic text descriptions, and $\star$ denotes missing imaging modalities. Our objective is to learn a mapping function $f : (x, y, T) \to \hat{l}$ that predicts the classification label regardless of missing imaging modalities. Key notations are provided in Table 1.

3.2. Multi-modal contrastive learning

In Stage 1, BrainCLIP aligns text, fMRI, and DTI modalities in a shared embedding space using n complete samples from $\mathbb{S}^p$. We introduce three modality-specific encoders: text encoder ψ_t, fMRI encoder ψ_f, and DTI encoder ψ_d. For each complete sample i, embeddings are obtained as:

$$t_i = \psi_t(T_i^p), \quad f_i = \psi_f(x_i^p), \quad d_i = \psi_d(y_i^p), \tag{1}$$

where T_i^p is the diagnostic text description from clinical reports or standardized assessments, and $t_i, f_i, d_i \in \mathbb{R}^d$ are embeddings in the shared space of dimension d.

We construct similarity matrices S_f and S_d based on cosine similarity:

$$(S_f)_{ij} = \frac{\langle t_i, f_j \rangle}{\|t_i\| \|f_j\|}, \quad (S_d)_{ij} = \frac{\langle t_i, d_j \rangle}{\|t_i\| \|d_j\|}, \quad i, j \in [1, n], \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. Following the standard CLIP paradigm (Radford et al., 2021), we adopt the symmetric InfoNCE contrastive objective rather than an MSE-to-identity matching objective, so that the model is explicitly trained to pull each matched (text, image) pair together while pushing all in-batch mismatched pairs apart. For

Table 1
Key notations of our frame.

Notations	Descriptions
$* \in \{p, d, f\}$	Subset marker, p: complete, d: fMRI-missing, f: DTI-missing
$\mathbb{S}^* = \{(x_i^*, y_i^*, T_i^*, l_i^*) \mid i \in [1, n_*]\}$	Training set, where $n_p = n$, $n_d = m$, $n_f = h$, and $\star$ denotes missing imaging modalities
$x_i^* \in \mathbb{R}^{d_f}$	fMRI data of the ith sample
$y_i^* \in \mathbb{R}^{d_d}$	DTI data of the ith sample
T_i^*	Diagnostic text of the ith sample
$l_i^* \in \{0, 1, 2, \ldots, C\}$	Label of the ith sample
$l_{i,c}^* \in \{0, 1\}$	One-hot encoded label, where $l_{i,c}^* = 1$ if $l_i^* = c$, else 0
$t_i \in \mathbb{R}^d$	Text embedding of the ith sample
$f_i^* \in \mathbb{R}^d$	fMRI embedding of the ith sample
$d_i^* \in \mathbb{R}^d$	DTI embedding of the ith sample
$h_i^* \in \mathbb{R}^{d_h}$	Fusion embedding of the ith sample
$\widetilde{x_i^d} \in \mathbb{R}^{d_f}$	Recovered fMRI feature of the ith sample
$\widetilde{y_i^f} \in \mathbb{R}^{d_d}$	Recovered DTI feature of the ith sample
$\widetilde{f_i^d} \in \mathbb{R}^d$	Recovered fMRI embedding of the ith sample
$\widetilde{d_i^f} \in \mathbb{R}^d$	Recovered DTI embedding of the ith sample
$f_i^{ref} \in \mathbb{R}^d$	Reference fMRI embedding
$d_i^{ref} \in \mathbb{R}^d$	Reference DTI embedding
ψ_t, ψ_f, ψ_d	Text, fMRI, and DTI encoders
G_f, G_d	fMRI and DTI generators
$\mathcal{F}$	Fusion module
S_f, S_d	Similarity matrices
n, m, h	Number of complete, fMRI-missing, and DTI-missing samples
d_f	Dimension of fMRI feature
d_d	Dimension of DTI feature
d_h	Dimension of fusion feature
d	Dimension of shared embedding space
C	Number of categories

the text–fMRI similarity matrix $S_f \in \mathbb{R}^{n \times n}$, we define the bidirectional contrastive loss as:

$$\mathcal{L}_{t \leftrightarrow f} = \frac{1}{2n} \sum_{i=1}^{n} \left[-\log \frac{\exp((S_f)_{ii}/\tau)}{\sum_{j=1}^{n} \exp((S_f)_{ij}/\tau)} - \log \frac{\exp((S_f)_{ii}/\tau)}{\sum_{j=1}^{n} \exp((S_f)_{ji}/\tau)} \right], \quad (3)$$

where the first term is the text-to-fMRI cross-entropy loss and the second is the fMRI-to-text cross-entropy loss, ensuring symmetric alignment in both directions. Analogously, the text–DTI bidirectional contrastive loss is defined on S_d:

$$\mathcal{L}_{t \leftrightarrow d} = \frac{1}{2n} \sum_{i=1}^{n} \left[-\log \frac{\exp((S_d)_{ii}/\tau)}{\sum_{j=1}^{n} \exp((S_d)_{ij}/\tau)} - \log \frac{\exp((S_d)_{ii}/\tau)}{\sum_{j=1}^{n} \exp((S_d)_{ji}/\tau)} \right]. \quad (4)$$

The total Stage 1 contrastive loss is the sum of the two text-anchored bidirectional InfoNCE losses:

$$\mathcal{L}_{con} = \mathcal{L}_{t \leftrightarrow f} + \mathcal{L}_{t \leftrightarrow d}, \quad (5)$$

where τ is a temperature hyperparameter set to $\tau = 0.07$ following CLIP. For each anchor sample i, the diagonal entry $(S_f)_{ii}$ or $(S_d)_{ii}$ corresponds to the only positive pair, while the remaining $n - 1$ off-diagonal entries in the same row or column serve as in-batch negatives. Compared with the previous MSE-to-identity formulation, this symmetric InfoNCE objective enforces a *relative* comparison between positive and negative pairs rather than an absolute pointwise constraint on the similarity matrix, which is the defining property of contrastive learning and is more semantically discriminative. This multi-modal contrastive learning stage establishes a semantically aligned representation space through *text-anchored pairwise alignment*. Rather than enforcing direct fMRI and DTI contrast, the diagnostic text serves as a shared semantic anchor that simultaneously aligns (text, fMRI) via S_f and (text, DTI) via S_d. The indirect consistency between fMRI and DTI is then established through this common anchor and is further reinforced by the sample-level contrastive objective in Stage 2. This text-anchored design provides a foundation for the subsequent cross-modal recovery stage where the pretrained text encoder serves as semantic guidance for modality recovery.

Class balance in Stage 1. Stage 1 uses an instance-level contrastive pre-training objective in which class labels are not used to define positive or negative pairs or to assign class-dependent weights. Each subject contributes exactly one text–fMRI positive pair and one text–DTI positive pair with the same per-subject weight, while the other subjects in the mini-batch serve as instance-level negatives. Moreover, the class distributions in the two cohorts are only mildly imbalanced rather than dominated by NC subjects: Epilepsy contains 114 NC, 103 FLE and 89 TLE subjects (maximum class ratio 1.28 : 1), and ADNI contains 82 NC, 76 SMC and 68 EMCI subjects (maximum class ratio 1.21 : 1). Training samples are randomly shuffled and traversed once per epoch, so every subject contributes equally to the Stage 1 objective. Explicit class labels enter the optimization in Stage 2 through classification and same-label reference selection in $\mathcal{L}_{feat}$, where the constraint is applied symmetrically within every class. Consequently, no single diagnostic group receives a larger per-subject loss weight or a privileged semantic target.

3.3. Cross-modal recovery learning

In Stage 2, BrainCLIP leverages the frozen pretrained Text Encoder ψ_t to guide modality recovery and classification using $m+h$ incomplete samples. Specifically, ψ_t is *completely frozen* so that the text-anchored semantic space established in Stage 1 is preserved as a stable supervision signal, while the imaging encoders ψ_f and ψ_d are *partially fine-tuned* with a smaller learning rate of 1×10^{-4}, one tenth of the Stage-1 learning rate. This "freeze-text and fine-tune-image" strategy strikes an explicit balance. The frozen text anchor prevents the recovered modalities from drifting away from disease-level semantics during Stage 2 optimization, while the partially fine-tuned imaging encoders retain sufficient expressive capacity to adapt to the recovery task on incomplete data. The rationality of this two-stage design parallels the established CLIP-Adapter and LoRA paradigm. Stage 1 learns a clean cross-modal embedding space using only *complete* samples, ensuring that the semantic anchor is not contaminated by recovery noise, and Stage 2 then adapts to the incomplete-modality regime within this anchored space. Such decoupling avoids the joint-optimization conflict that would arise in a single-stage end-to-end scheme, where the contrastive alignment loss and the recovery and classification losses compete on the same incomplete inputs.

Fine-tuning ψ_f and ψ_d necessarily permits a limited adaptation of the imaging-side geometry, so we do not assume that the Stage 1 embedding space remains numerically unchanged. Instead, this adaptation is explicitly constrained in three complementary ways: the text coordinate system is fixed by freezing ψ_t; the imaging encoders use a tenfold smaller learning rate; and both $\mathcal{L}_{text}$ and $\mathcal{L}_{feat}$ continuously regularize the adapted representations toward the frozen text anchor and same-label complete-sample manifolds. Thus, Stage 2 performs constrained local adaptation to incomplete inputs rather than an unconstrained reorganization of the joint space. The direct comparison in Table 4 provides supporting evidence for the overall two-stage design: two-stage BrainCLIP exceeds single-stage end-to-end joint training by 6.16 and 6.18 percentage points in ACC on Epilepsy and ADNI, respectively. Because this ablation changes the full optimization schedule, we do not interpret it as an isolated quantitative measurement of embedding-space drift.

We introduce two modality generators G_f and G_d implemented as MLPs to reconstruct missing features. We deliberately adopt a lightweight two-layer MLP for the generators rather than more complex generative architectures such as VAE, GAN, or diffusion models, for three reasons. First, the inputs to G_f and G_d are already *low-dimensional, semantically aligned embeddings* produced by the Stage-1 contrastive encoders, with d_i, t_i and f_i all in $\mathbb{R}^d$ at $d = 512$, rather

than high-dimensional raw images. In this regime an MLP suffices to learn the conditional mapping, while heavy generative networks introduce unnecessary capacity. Second, our datasets contain only a few hundred subjects per cohort, where high-capacity generators are well known to be prone to mode collapse, training instability, and over-fitting. Third, the novelty of BrainCLIP lies in the *dual-contrastive semantic constraint* on the recovered embeddings in Eqs. (10) and (11) rather than in the generator architecture itself, and the lightweight MLP serves precisely to isolate and highlight the contribution of this semantic constraint. Empirically, this minimal generator already attains the highest recovery quality in Table 3 and the highest downstream classification accuracy in Table 2 compared with baselines that employ more elaborate generative or imputation strategies, such as the masked-autoencoder of M^{3}FeCon and the variational autoencoder of IMDL. For samples missing fMRI, reconstructed features are:

$$\widetilde{x_i^d} = G_f([d_i; t_i]; \theta_{G_f}), \quad i \in [1, m], \tag{6}$$

where $d_i = \psi_d(y_i^d)$ and $t_i = \psi_t(T_i^d)$ are embeddings from available DTI and text. Similarly, for samples missing DTI:

$$\widetilde{y_i^f} = G_d([f_i; t_i]; \theta_{G_d}), \quad i \in [1, h], \tag{7}$$

where $f_i = \psi_f(x_i^f)$ and $t_i = \psi_t(T_i^f)$. Recovered and original features are encoded to obtain embeddings $\widetilde{f_i^d}, \widetilde{d_i^f}, f_i, d_i \in \mathbb{R}^d$.

A fusion module $\mathcal{F}$ combines multi-modal embeddings:

$$h_i^* = \mathcal{F}(\widetilde{f_i}^*, \widetilde{d_i^*}, t_i; \theta_{\mathcal{F}}), \tag{8}$$

where $\widetilde{f_i^*}$ and $\widetilde{d_i^*}$ represent the fMRI and DTI embeddings for the ith sample: for samples in $\mathbb{S}^d$ (missing fMRI), $\widetilde{f_i^*} = \widetilde{f_i^d}$ and $\widetilde{d_i^*} = d_i$, and for samples in $\mathbb{S}^f$ (missing DTI), $\widetilde{f_i^*} = f_i$ and $\widetilde{d_i^*} = \widetilde{d_i^f}$. A classifier predicts labels: $\hat{p}_i^* = \mathrm{Softmax}(W h_i^* + b)$, where $\hat{l}_i^* = \mathrm{Argmax}(\hat{p}_i^*)$ is the predicted label, with classification loss:

$$\mathcal{L}_{cls} = -\frac{1}{m+h} \sum_{i=1}^{m+h} \sum_{c=1}^{C} l_{i,c}^* \log(\hat{p}_{i,c}^*), \tag{9}$$

where $l_{i,c}^* \in \{0, 1\}$ is the one-hot encoded ground-truth label, with $l_{i,c}^* = 1$ if the true label $l_i^* = c$, and $l_{i,c}^* = 0$ otherwise, and $\hat{p}_{i,c}^*$ is the predicted probability for class c.

The dual contrastive strategy includes: (1) Modality level contrastive loss (distinct from the multi-modal contrastive learning in Stage 1) aligning recovered features with reference features from complete samples sharing the same label:

$$\begin{aligned} \mathcal{L}_{feat} = & \frac{1}{m} \sum_{i=1}^{m} \left[1 - \frac{\langle \widetilde{f_i^d}, f_i^{ref} \rangle}{\|\widetilde{f_i^d}\| \|f_i^{ref}\|} \right]^2 \\ & + \frac{1}{h} \sum_{i=1}^{h} \left[1 - \frac{\langle \widetilde{d_i^f}, d_i^{ref} \rangle}{\|\widetilde{d_i^f}\| \|d_i^{ref}\|} \right]^2, \end{aligned} \tag{10}$$

where f_i^{ref} and d_i^{ref} are reference embeddings selected from complete samples in $\mathbb{S}^p$ that share the same label as the ith incomplete sample. Specifically, for each incomplete sample i with label l_i^*, we select a reference sample j from $\mathbb{S}^p$ such that $l_j^p = l_i^*$, and calculate $f_i^{ref} = \psi_f(x_j^p)$ and $d_i^{ref} = \psi_d(y_j^p)$. If multiple complete samples share the same label, we randomly select one as the reference. Here, "semantic consistency" means that a recovered embedding $\widetilde{f_i^d}$ or $\widetilde{d_i^f}$ lies close, in the shared embedding space, to real complete samples that share the same diagnostic label as subject i, while remaining distant from samples of other labels. This label-aligned reference selection pulls NC recoveries toward the NC manifold and disease recoveries toward the corresponding disease manifold; the constraint is therefore symmetric across all classes and does not privilege disease categories over normal controls. (2) Sample level contrastive loss enforcing alignment between textual embeddings and fused multi-modal representations to preserve cross-modal semantic relationships:

$$\mathcal{L}_{text} = \frac{1}{m+h} \sum_{i=1}^{m+h} \left[1 - \frac{\langle t_i, h_i^* \rangle}{\|t_i\| \|h_i^*\|} \right]^2. \tag{11}$$

The overall objective is:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{feat} + \lambda_2 \mathcal{L}_{text}, \tag{12}$$

where λ_1 and λ_2 are hyperparameters. The dual contrastive strategy brings several advantages to modality recovery: the modality level contrastive loss ensures that recovered features maintain semantic consistency with disease-related patterns from complete samples, while the sample level contrastive loss preserves the cross-modal semantic relationships learned during pretraining. Together, these two objectives enable BrainCLIP to recover modalities that are both numerically accurate and semantically aligned, ensuring that the recovered features are discriminative for downstream classification tasks.

Algorithm 1: Training Procedure of BrainCLIP

Input : fMRI data $\hat{F}$, DTI data $\hat{D}$, diagnostic texts T, labels Y; hyperparameters λ_1, λ_2; epochs E_1, E_2
Output: Prediction $\hat{L}$ of the test set
Preprocess data to obtain feature vectors; Split into complete set $\mathbb{S}^p$ and incomplete sets $\mathbb{S}^d$, $\mathbb{S}^f$;
`/* Stage 1: Multi-Modal Contrastive Learning */`
Initialize ψ_t from BioClinicalBERT and randomly initialize the two-layer MLP imaging encoders ψ_f and ψ_d;
for *epoch* $= 1$ **to** E_1 **do**
  Encode complete samples by Eq. (1) and construct similarity matrices S_f, S_d by Eq. (2);
  Compute the bidirectional InfoNCE losses $\mathcal{L}_{t \leftrightarrow f}$ and $\mathcal{L}_{t \leftrightarrow d}$ by Eq. (3) and Eq. (4), and the total contrastive loss $\mathcal{L}_{con}$ by Eq. (5); update encoder parameters;
`/* Stage 2: Cross-Modal Recovery Learning */`
Freeze ψ_t; keep ψ_f, ψ_d trainable with a reduced learning rate of 1×10^{-4}; Initialize generators G_f, G_d, fusion module $\mathcal{F}$;
for *epoch* $= 1$ **to** E_2 **do**
  for *each incomplete sample* **do**
    Encode available modalities, recover missing modality by Eq. (6) or Eq. (7), and encode recovered features;
  Fuse features by Eq. (8) and predict labels; For each incomplete sample, select reference features f_i^{ref} and d_i^{ref} from complete samples in $\mathbb{S}^p$ sharing the same label;
  Calculate losses $\mathcal{L}_{cls}$, $\mathcal{L}_{feat}$, $\mathcal{L}_{text}$ by Eq. (9), Eq. (10), Eq. (11);
  Calculate final loss $\mathcal{L}$ by Eq. (12) and update G_f, G_d, $\mathcal{F}$, and the partially-trainable imaging encoders ψ_f, ψ_d;
Return: Prediction $\hat{L}$ of the test set.

4. Experiments

4.1. Materials

4.1.1. Data collection

We conduct experiments on the Epilepsy dataset and the ADNI dataset.

Epilepsy: The resting-state fMRI and DTI data were collected from Jinling Hospital, Nanjing University School of Medicine. The dataset includes 306 subjects: 114 NC (58 males, 56 females, mean age 26.2 years), 103 FLE (53 males, 50 females, mean age 24.1 years), and 89 TLE (45 males, 44 females, mean age 25.3 years). Imaging data were acquired using a Siemens Trio 3T scanner. The non-imaging diagnostic text for each subject is composed of three pre-imaging clinical fields

extracted from the Electronic Medical Record (EMR) system, namely the *chief complaint and seizure history* such as "Recurrent seizures for 5 years, predominantly involving left-hand clonic movements", the *EEG clinical narrative* such as "Interictal EEG shows bilateral temporal sharp waves, predominantly left", and the *clinical diagnosis summary* such as "Temporal lobe epilepsy, left hemisphere". These three fields are concatenated into a single text sequence per subject with mean length about 74 tokens. All text descriptions are available regardless of whether fMRI or DTI scans were successfully acquired, because they are recorded at the initial clinical encounter prior to imaging referral.

ADNI: A total of 226 samples were selected from ADNI2 and ADNI3 (http://adni.loni.usc.edu), each including both fMRI and DTI modalities. Subjects were included if they had paired rs-fMRI and DTI from the same session, NC/SMC/EMCI labels with complete CDR, MMSE, ADAS-Cog 13, FAQ and GDS assessments, and passed preprocessing quality control; only baseline visits were retained for longitudinal subjects. The dataset comprises 82 NC (40 males, 42 females, mean age 74.3 years), 76 SMC (50 males, 29 females, mean age 77.2 years), and 68 EMCI (36 males, 32 females, mean age 74.8 years). The non-imaging diagnostic text for each subject is derived from standardized clinical assessments administered independently of and prior to imaging sessions, namely CDR, MMSE, ADAS-Cog 13, FAQ and GDS together with the clinical diagnosis label, which are converted into a template-based text description such as "CDR global = 0.5, CDR-SB = 1.5; MMSE = 27; ADAS-Cog 13 = 15.33; FAQ = 3; GDS = 2. Clinical diagnosis: Significant Memory Concern". These fields are concatenated into a single text sequence per subject with mean length about 58 tokens. All text descriptions are available regardless of whether fMRI or DTI scans were successfully acquired, because these assessments are conducted during the screening visit before imaging acquisition.

4.1.2. Preprocessing

DTI data are preprocessed using PANDA (Cui et al., 2013) with FSL (Jenkinson et al., 2012) for distortion correction and TrackVis (Wang et al., 2007) for fiber tract generation. Anatomical regions are identified using the AAL atlas based on T1 images, and structural connectivity is measured by the number of fibers between regions. For rs-fMRI data, preprocessing is performed using SPM8 (Ashburner et al., 2012) via the DPARSF toolbox (Yan and Zang, 2010), including motion correction, detrending, and segmentation into 90 ROIs using the AAL atlas. The fMRI data are divided into 240 timepoints for Epilepsy and 197 timepoints for ADNI. The exact form of the imaging features used as inputs to BrainCLIP is summarized as follows. *For fMRI*, the BOLD signal of each AAL ROI is obtained by spatially *averaging* the voxel-level time series within that ROI, namely a ROI-averaged time series rather than a sparse representation. The 90×90 Pearson correlation matrix between ROI time series is then computed and Fisher-Z transformed to form the functional connectome, whose upper triangle is flattened into a 4005-dimensional feature vector $x_i^{(f)} \in \mathbb{R}^{4005}$ that serves as the input to ψ_f. We therefore use the connectome matrix rather than raw ROI signals as the input. *For DTI*, the structural connectivity is measured as the actual fiber count between AAL ROIs. To remove inter-subject differences in head size and total fiber number, the 90×90 fiber-count matrix of each subject is normalized by its total fiber count, followed by a $\log(1+x)$ transform to stabilize variance, and the upper triangle is flattened into a 4005-dimensional feature vector $y_i^{(d)} \in \mathbb{R}^{4005}$ that serves as the input to ψ_d.

4.1.3. Diagnostic text preparation

As emphasized in Section 1, the diagnostic text used in this work consists of *non-imaging clinical records* that are generated prior to and independently of the imaging acquisition, and are therefore available even when imaging modalities are missing. The text preparation pipeline proceeds as follows. We first perform *field extraction and concatenation*. For the Epilepsy dataset, the chief complaint, seizure history, and EEG clinical narrative are extracted from the EMR and concatenated into a single text sequence per subject. For the ADNI dataset, five standardized assessment scores including CDR, MMSE, ADAS-Cog 13, FAQ and GDS, together with the clinical diagnosis label, are converted into a template-based sentence. We then perform *de-identification*, where all personally identifiable information such as names, dates, and hospital IDs is removed. Each text sequence is then tokenized using the BioClinicalBERT tokenizer (Alsentzer et al., 2019) with a vocabulary of about 28,996 WordPiece tokens and a maximum sequence length of 128. The text encoder ψ_t is initialized from BioClinicalBERT, a BERT-style encoder pretrained on the MIMIC-III clinical notes corpus, and fine-tuned in Stage 1 with the symmetric InfoNCE loss in Eq. (5), and is completely frozen in Stage 2 to preserve the text-anchored semantic space as a stable supervision signal. The `[CLS]` token output of ψ_t serves as the text embedding $t_i \in \mathbb{R}^d$. The average text length is approximately 74 tokens for Epilepsy and 58 tokens for ADNI. Across both datasets, all subjects have complete text descriptions, namely text availability of 100%, confirming the practical feasibility of the core assumption that non-imaging diagnostic text remains available under imaging missingness.

Rationale for BioClinicalBERT. Although the ADNI assessments are converted into a consistent template, the input is not uniformly tabular across the two cohorts. In particular, the Epilepsy records contain natural-language chief complaints, seizure histories, EEG narratives and diagnosis summaries. Representing these fields with a tabular embedding or a simple MLP would require dataset-specific encoding and may discard clinically meaningful context, terminology and negation. BioClinicalBERT provides a single clinically pretrained interface for both these narrative fields and the templated ADNI descriptions, preserves the semantic meaning of field names and clinical terms, and allows the same framework to accept less standardized records in future external cohorts. We acknowledge that BioClinicalBERT is not the most parameter-efficient choice for the highly structured ADNI subset alone. Its gradient and optimizer-state costs are confined to Stage 1 because it is completely frozen in Stage 2, although a forward pass is still required. A direct comparison with distilled clinical language models or hybrid tabular–text encoders is an important efficiency study for future deployment.

4.2. Experimental settings

BrainCLIP consists of three modality-specific encoders (ψ_t, ψ_f, ψ_d), two modality generators (G_f, G_d), and one fusion module ($\mathcal{F}$). The text encoder ψ_t uses the BioClinicalBERT backbone initialized from the publicly available MIMIC-III-pretrained weights, followed by a projection head that maps the 768-dimensional `[CLS]` representation to the shared dimension $d = 512$. In contrast, the fMRI encoder ψ_f and DTI encoder ψ_d do not use any external pretrained imaging model. They are independently and randomly initialized two-layer MLPs, each mapping its 4005-dimensional connectome input through a hidden layer of dimension 2048 to the shared 512-dimensional space, with batch normalization and ReLU activation. Both imaging encoders are trained from scratch together with ψ_t in Stage 1 and are subsequently initialized from their Stage 1 checkpoints and fine-tuned in Stage 2. The generators take concatenated embeddings $[d_i; t_i]$ or $[f_i; t_i]$ as input and output reconstructed features in the original feature space, with hidden dimension 2048. The fusion module $\mathcal{F}$ takes concatenated embeddings $[\widetilde{f_i^*}; \widetilde{d_i^*}; t_i] \in \mathbb{R}^{1536}$ as input and outputs fusion embeddings $h_i^* \in \mathbb{R}^{256}$ for classification.

Training is conducted in two stages. In Stage 1, the three encoders are jointly pretrained for 50 epochs using Eq. (5) on n complete samples from $\mathbb{S}^p$ with batch size 32; the temperature parameter in Eqs. (3) and (4) is set to $\tau = 0.07$ following CLIP (Radford et al., 2021). In Stage 2, ψ_t is completely frozen while G_f, G_d, and $\mathcal{F}$, together with the partially trainable imaging encoders ψ_f and ψ_d, are jointly trained for 100 epochs using Eq. (12) on $m + h$ incomplete samples from $\mathbb{S}^d$ and $\mathbb{S}^f$ with batch size 64. The imaging encoders ψ_f and ψ_d adopt a

smaller learning rate of 1×10^{-4}, one tenth of the Stage-1 learning rate, to retain expressive capacity for the recovery task while preserving the text-anchored semantic space, whereas G_f, G_d, and $\mathcal{F}$ use the default learning rate 1×10^{-3}. Hyperparameters are set as $\lambda_1 = 0.7$ and $\lambda_2 = 0.7$ in Eq. (12). The Adam optimizer (Kingma and Ba, 2014) is used with learning rate 1×10^{-3} and weight decay 1×10^{-5}.

After replacing the original MSE-to-identity Stage 1 objective with the symmetric InfoNCE objective in Eqs. (3)–(5), we reinitialized and retrained BrainCLIP and all loss-dependent ablation variants from scratch on both datasets for every MDR setting. The classification results in Table 2, recovery results in Table 3, ablation results in Table 4, and the associated ROC and recovery–classification correlation plots were all regenerated from these InfoNCE-based runs. The comparison methods are not affected by this change to BrainCLIP's objective and were evaluated under the same data splits and random seeds. Thus, all results reported in the revised manuscript correspond to the updated InfoNCE formulation rather than the previous MSE-to-identity formulation.

Since the collected datasets are complete, we simulate data incompleteness by randomly removing modalities according to specified MDR levels (10%, 30%, 50%, 70%, 90%) to generate incomplete sets $\mathbb{S}^d$ and $\mathbb{S}^f$. To improve the statistical reliability of our evaluation on datasets of moderate scale, we adopt 10-fold cross-validation repeated over 5 different random seeds, yielding 50 paired repeated-cross-validation results for every method and every MDR level. The mean and standard deviation reported in Tables 2, 3, and 4 are computed across the 5 seed averages. To provide supplementary evidence for comparisons with baselines, we additionally perform a paired Wilcoxon signed-rank test between BrainCLIP and each baseline at every MDR level, with Bonferroni correction for multiple comparisons. All reported comparisons yield $p < 0.01$. We nevertheless acknowledge that folds generated under different seeds overlap and the 50 results are therefore not strictly independent; the effective sample size may be smaller and the resulting p-values may be optimistic. Accordingly, these tests are interpreted as supportive rather than definitive evidence, together with the absolute performance differences, standard deviations and consistent trends across datasets and MDR levels.

Choice of training epochs. The number of epochs in Stage 1 of 50 and in Stage 2 of 100 is determined by validation-loss early stopping under 5-seed averaging. The contrastive loss in Stage 1 saturates around epoch 40 on both datasets, so 50 epochs leaves a safe convergence margin. The joint recovery and classification loss in Stage 2 saturates around epoch 80, so 100 epochs is similarly chosen with a moderate margin. Empirically, we observed less than 0.3 percentage points of ACC variation when sweeping Stage 1 epochs in $\{30, 50, 70, 100\}$ and Stage 2 epochs in $\{50, 100, 150, 200\}$, indicating low sensitivity to this choice.

Complexity and training cost. BrainCLIP has approximately 115M parameters in total. In Stage 1, all parameters are jointly fine-tuned, dominated by the BioClinicalBERT text encoder of about 110M parameters, while the three modality-specific encoders, two two-layer MLP generators, and the fusion and classifier module account for the remaining 5M. In Stage 2 the text encoder is completely frozen, leaving only about 5M parameters trainable, namely the imaging encoders ψ_f and ψ_d with a reduced learning rate together with the generators, fusion module and classifier. On a single NVIDIA RTX 3090 with 24 GB of memory, one Stage 1 epoch takes about 4.5 s on Epilepsy and 3.2 s on ADNI, and one Stage 2 epoch takes about 1.8 s and 1.4 s respectively. The entire two-stage training of 50 plus 100 epochs finishes within 13 min per fold per seed, and inference latency per subject is below 25 ms. The peak GPU memory footprint stays below 11 GB during Stage 1 and below 4 GB during Stage 2, well within the budget of standard clinical workstations.

4.3. Evaluation metric and baselines

We utilize accuracy (ACC), F1 score, and area under the curve (AUC) (Mitra and Chaudhuri, 2000; Fawcett, 2006; Horng et al., 2005) to assess the performance of our method. We introduce five levels of Missing Data Ratio (MDR) (Dong and Peng, 2013) for performance assessment. An MDR of 90% indicates that 90% of the data samples have missing imaging modalities, with the missing proportions of the fMRI and DTI modalities being equal, 45% each. Our approach is benchmarked against five state-of-the-art incomplete multi-modal brain imaging methods:

- **Modal-Mixup + Deep Supervision (MMBNet)** (Yang et al., 2024): Introduces a modal-mixup strategy to synthesize samples under incomplete conditions, combined with a bilateral deep-supervision network to regularize both functional and structural representations.
- **Incomplete Multi-modal Disentanglement Learning (IMDL)** (Han et al., 2025): Leverages modality-wise variational autoencoders to disentangle shared and complementary features across modalities, guided by contrastive learning to align incomplete representations without explicit reconstruction.
- **Simple Framework for Multi-modal Learning with Missing Modality (SimMLM)** (Li et al., 2025): Proposes a dynamic mixture of modal experts architecture that adaptively adjusts the contribution of each modality under complete and partial modal conditions, introducing a More vs. Fewer ranking loss to ensure task accuracy improves or remains stable as the number of available modalities increases.
- **M^{3}FeCon** (Zeng et al., 2024): Treats missing imaging modalities as masked tokens and reconstructs cross-modal features through a masked autoencoding strategy, allowing arbitrary feature reconstruction across modalities.
- **MochaGCN** (Xie et al., 2024): Proposes multimodal hyperbolic graph learning for Alzheimer's disease detection, where fMRI and DTI brain networks are embedded in hyperbolic space and aligned via graph convolution and contrastive objectives.

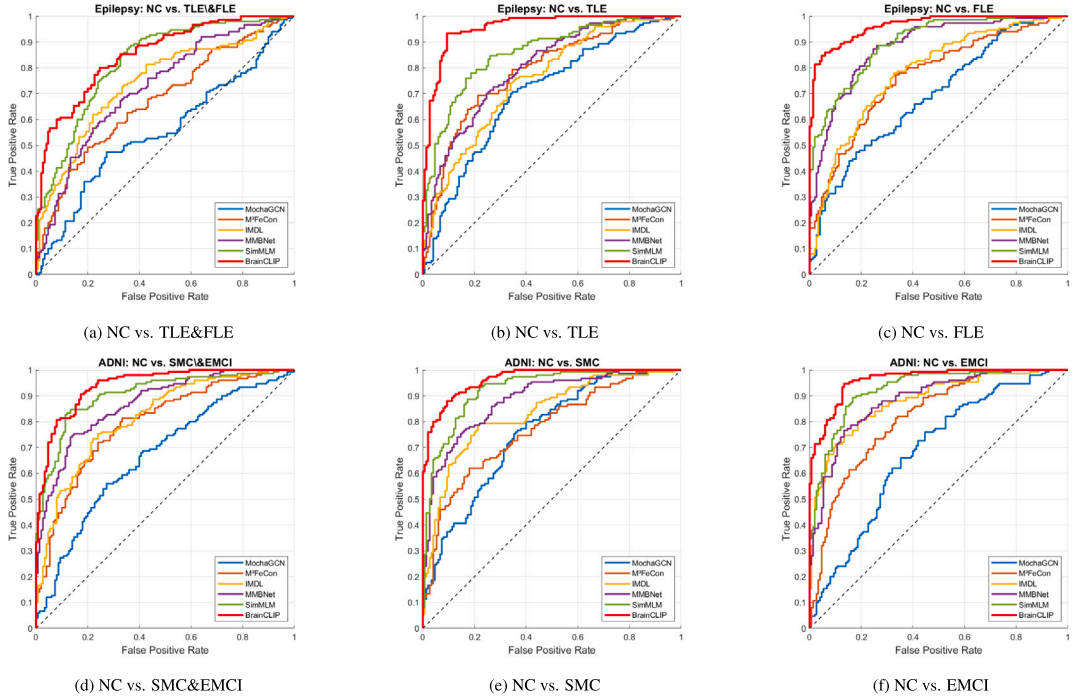
4.4. Classification performance

As demonstrated in Table 2, BrainCLIP consistently outperforms all baseline methods on both the Epilepsy and ADNI datasets across various MDR levels. On the Epilepsy dataset, BrainCLIP achieves accuracy of 96.42%, 95.32%, 93.35%, 87.04%, and 85.78% at MDR levels of 10%, 30%, 50%, 70%, and 90%, respectively, surpassing the strongest baseline (SimMLM) by margins of 3.97%, 3.50%, 2.97%, 2.40%, and 3.10%. Similar improvements are observed for F1 score and AUC, with BrainCLIP consistently achieving the highest values across all metrics. On the ADNI dataset, BrainCLIP achieves accuracy of 92.65%, 90.86%, 89.74%, 79.93%, and 75.03% at the corresponding MDR levels, exceeding the best baseline (SimMLM) by 2.97%, 2.18%, 2.10%, 2.01%, and 2.85%, respectively. The results demonstrate that BrainCLIP maintains superior performance even under extreme missing conditions, with the performance gap becoming more pronounced as MDR increases.

This consistent superiority can be attributed to BrainCLIP's unique design. While baseline methods focus on feature-level reconstruction or implicit imputation, they lack explicit cross-modal semantic alignment. In contrast, BrainCLIP introduces a two-stage framework that first establishes semantically aligned cross-modal representations through contrastive learning, then leverages the pretrained Text Encoder to guide modality recovery with dual contrastive objectives. This design ensures that recovered modalities maintain both numerical accuracy and semantic consistency with disease-related patterns, enabling discriminative representation learning even under varying degrees of incompleteness.

Table 2
The classification performance ($MEAN \pm STD$) with various MDR (%). All BrainCLIP-vs-baseline comparisons yield $p < 0.01$ in the paired Wilcoxon signed-rank test with Bonferroni correction. Because repeated cross-validation folds overlap across seeds, these p-values are interpreted as supportive rather than as evidence from 50 strictly independent samples.

MDR	Method	Epilepsy			ADNI		
		ACC	F1	AUC	ACC	F1	AUC
10%	MMBNet	91.50 ± 4.71	89.02 ± 4.46	90.23 ± 7.01	88.05 ± 6.36	87.32 ± 6.86	86.56 ± 8.96
	IMDL	88.12 ± 2.94	87.36 ± 2.87	87.85 ± 3.41	86.20 ± 3.42	85.52 ± 3.07	85.93 ± 2.98
	SimMLM	92.45 ± 2.91	91.78 ± 2.78	92.12 ± 3.14	89.68 ± 3.25	89.02 ± 3.09	89.35 ± 3.02
	M^{3}FeCon	88.40 ± 2.86	87.72 ± 2.64	88.21 ± 3.27	86.73 ± 3.38	86.02 ± 3.17	86.45 ± 2.92
	MochaGCN	87.25 ± 3.18	86.60 ± 3.45	87.01 ± 3.62	85.08 ± 3.91	84.56 ± 3.25	85.02 ± 3.44
	BrainCLIP (Ours)	**96.42** ± 1.72	**95.18** ± 1.92	**95.46** ± 1.51	**92.65** ± 1.39	**91.98** ± 1.58	**92.31** ± 1.49
30%	MMBNet	90.20 ± 5.33	90.28 ± 5.48	89.87 ± 6.18	86.73 ± 8.85	84.85 ± 8.81	85.64 ± 5.07
	IMDL	88.41 ± 3.35	87.63 ± 3.64	88.07 ± 3.18	85.21 ± 3.20	84.75 ± 3.47	85.02 ± 3.36
	SimMLM	91.82 ± 3.12	91.18 ± 3.28	91.45 ± 3.05	88.68 ± 3.18	88.02 ± 3.11	88.35 ± 3.08
	M^{3}FeCon	87.74 ± 2.87	87.09 ± 3.14	87.45 ± 2.98	84.96 ± 3.31	84.31 ± 3.24	84.64 ± 3.12
	MochaGCN	86.63 ± 3.54	85.97 ± 3.29	86.35 ± 3.47	83.74 ± 3.56	83.15 ± 3.15	83.43 ± 3.64
	BrainCLIP (Ours)	**95.32** ± 1.85	**94.71** ± 2.04	**94.95** ± 1.68	**90.86** ± 1.62	**90.21** ± 1.83	**90.55** ± 1.74
50%	MMBNet	88.97 ± 4.24	87.89 ± 4.37	90.14 ± 5.73	85.67 ± 7.52	84.11 ± 7.78	86.79 ± 6.84
	IMDL	87.47 ± 3.22	86.65 ± 3.40	87.22 ± 3.47	84.13 ± 3.87	83.48 ± 3.65	83.82 ± 3.49
	SimMLM	90.38 ± 3.45	89.72 ± 3.61	90.07 ± 3.38	87.64 ± 3.52	87.02 ± 3.48	87.38 ± 3.42
	M^{3}FeCon	86.06 ± 3.14	85.32 ± 3.41	85.78 ± 3.17	83.64 ± 3.35	82.93 ± 3.56	83.35 ± 3.48
	MochaGCN	84.68 ± 3.43	83.97 ± 3.25	84.42 ± 3.51	82.35 ± 3.68	81.73 ± 3.42	82.09 ± 3.36
	BrainCLIP (Ours)	**93.35** ± 2.13	**92.68** ± 2.37	**93.03** ± 2.01	**89.74** ± 1.98	**89.08** ± 2.15	**89.42** ± 2.06
70%	MMBNet	83.45 ± 6.33	82.66 ± 6.24	84.44 ± 5.59	76.63 ± 4.01	76.13 ± 4.99	77.29 ± 7.76
	IMDL	82.45 ± 3.88	81.82 ± 3.63	82.24 ± 3.47	75.18 ± 3.61	74.63 ± 3.29	75.02 ± 3.40
	SimMLM	84.64 ± 3.78	84.08 ± 3.85	84.39 ± 3.64	77.92 ± 3.68	77.34 ± 3.56	77.68 ± 3.52
	M^{3}FeCon	81.21 ± 3.55	80.62 ± 3.38	81.06 ± 3.66	74.17 ± 3.54	73.62 ± 3.45	74.08 ± 3.51
	MochaGCN	80.12 ± 3.66	79.58 ± 3.22	79.96 ± 3.49	73.23 ± 3.33	72.74 ± 3.57	73.05 ± 3.46
	BrainCLIP (Ours)	**87.04** ± 2.48	**86.43** ± 2.71	**86.76** ± 2.36	**79.93** ± 2.24	**79.31** ± 2.41	**79.71** ± 2.37
90%	MMBNet	81.37 ± 4.36	81.70 ± 4.34	79.45 ± 3.71	70.89 ± 7.64	70.09 ± 7.87	71.21 ± 9.48
	IMDL	80.62 ± 3.98	79.93 ± 3.71	80.41 ± 3.55	70.25 ± 3.84	69.72 ± 3.59	70.13 ± 3.67
	SimMLM	82.68 ± 3.92	82.14 ± 3.88	82.45 ± 3.76	72.18 ± 3.72	71.66 ± 3.64	72.02 ± 3.58
	M^{3}FeCon	79.02 ± 3.72	78.36 ± 3.68	78.85 ± 3.47	69.33 ± 3.48	68.87 ± 3.21	69.24 ± 3.46
	MochaGCN	78.18 ± 3.54	77.55 ± 3.33	77.89 ± 3.42	68.43 ± 3.39	68.02 ± 3.26	68.31 ± 3.28
	BrainCLIP (Ours)	**85.78** ± 2.89	**85.27** ± 3.13	**85.58** ± 2.94	**75.03** ± 2.63	**74.45** ± 2.88	**74.78** ± 2.71



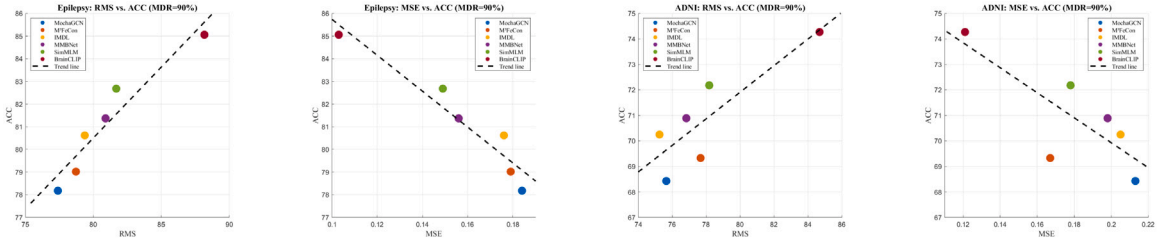
(a) NC vs. TLE&FLE (b) NC vs. TLE (c) NC vs. FLE

(d) NC vs. SMC&EMCI (e) NC vs. SMC (f) NC vs. EMCI

Fig. 3. ROC curves for the proposed method and comparison methods in brain disease diagnosis on the Epilepsy and ADNI datasets with MDR = 90%.

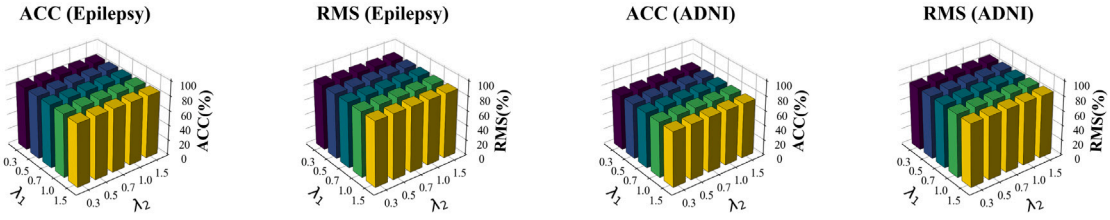
Table 3
The recovery performance of recovered modalities with MDR = 90%. (%) All BrainCLIP-vs-baseline comparisons yield $p < 0.01$ in the paired Wilcoxon test with Bonferroni correction; the p-values are interpreted as supportive because repeated cross-validation folds overlap across seeds.

Methods	Epilepsy			ADNI		
	RMS	MSE	PCC	RMS	MSE	PCC
MMBNet	80.89 ± 2.14	0.156 ± 0.012	0.814 ± 0.018	76.83 ± 2.31	0.198 ± 0.015	0.800 ± 0.021
IMDL	79.36 ± 2.28	0.176 ± 0.014	0.796 ± 0.020	75.25 ± 2.45	0.205 ± 0.017	0.782 ± 0.023
SimMLM	81.68 ± 1.95	0.149 ± 0.010	0.820 ± 0.016	78.18 ± 2.12	0.178 ± 0.013	0.807 ± 0.019
M^3FeCon	78.71 ± 2.35	0.179 ± 0.015	0.793 ± 0.022	77.67 ± 2.51	0.167 ± 0.014	0.805 ± 0.024
MochaGCN	77.38 ± 2.42	0.184 ± 0.016	0.788 ± 0.025	75.65 ± 2.63	0.213 ± 0.018	0.776 ± 0.027
BrainCLIP (Ours)	**88.93 ± 1.52**	**0.094 ± 0.007**	**0.849 ± 0.012**	**85.51 ± 1.68**	**0.110 ± 0.009**	**0.854 ± 0.014**



(a) RMS vs. ACC (Epilepsy) (b) MSE vs. ACC (Epilepsy) (c) RMS vs. ACC (ADNI) (d) MSE vs. ACC (ADNI)

Fig. 4. The correlation between recovery performance and classification performance with MDR = 90%.



(a) ACC (Epilepsy) (b) RMS (Epilepsy) (c) ACC (ADNI) (d) RMS (ADNI)

Fig. 5. Sensitivity of classification accuracy (ACC) and recovery modality similarity (RMS) to λ_1 and λ_2 at MDR = 90%.

Most methods, including BrainCLIP, achieve their peak performance when the MDR is 10%, attributed to the relatively low level of data incompleteness. As MDR increases, the performance of all methods gradually declines, but BrainCLIP demonstrates superior resilience, particularly at MDR = 90% where BrainCLIP achieves 85.78% accuracy on Epilepsy and 75.03% on ADNI, while the best baseline (SimMLM) reaches 82.68% and 72.18%, respectively.

From MDR = 10% to MDR = 90%, the classification accuracy of BrainCLIP degrades by 10.64 percentage points on Epilepsy, from 96.42% to 85.78%, and by 17.62 percentage points on ADNI, from 92.65% to 75.03%. Although the absolute degradation is slightly larger than that of the strongest baseline SimMLM (9.77 and 17.50 percentage points respectively), the absolute superiority of BrainCLIP is preserved across the entire MDR spectrum. The smallest absolute gap to SimMLM, observed at MDR = 70% on ADNI, still exceeds 2.0 percentage points in ACC. These trends indicate that the proposed dual-contrastive recovery framework remains effective even under severe missingness, which is the regime for which BrainCLIP is designed.

To further validate the diagnostic capability under challenging scenarios, we evaluate the ROC curves at MDR = 90%. As shown in Fig. 3, BrainCLIP consistently achieves the highest AUC across all classification tasks on both datasets, with its curves positioned furthest from the random classifier diagonal. These ROC results confirm that BrainCLIP's dual contrastive learning framework effectively learns discriminative representations even when 90% of samples have missing imaging modalities, demonstrating strong generalization capability for real-world clinical applications.

4.5. Recovery performance

To evaluate the model's capability in restoring missing imaging modalities, we assessed recovery performance under an extremely high MDR of 90%. This setting simulates a worst-case clinical scenario. We employed three widely used quantitative metrics to evaluate the quality of recovered fMRI and DTI modalities:

- **Recovery Modality Similarity (RMS)** (Xia et al., 2015): Measures the cosine similarity between the recovered and ground-truth modalities, reflecting semantic consistency at the feature level.
- **Mean Squared Error (MSE)** (Chicco et al., 2021): Calculates the average squared difference at the feature level between recovered and ground-truth modalities. Lower values indicate better numerical accuracy.
- **Pearson Correlation Coefficient (PCC)**: Measures the linear correlation between the recovered and ground-truth connectome or fiber-count features, with values closer to 1 indicating better preservation of the network topology.

The reported RMS, MSE, and PCC values represent the average performance across both modalities (fMRI and DTI). As shown in Table 3, BrainCLIP consistently outperforms all baseline methods on both the Epilepsy and ADNI datasets, achieving the highest RMS and PCC, along with the lowest MSE, even under MDR = 90%. Specifically, on the Epilepsy dataset, BrainCLIP achieves RMS of 88.93%, MSE of 0.094, and PCC of 0.849, surpassing the best baseline (SimMLM) by 7.25% in

Table 4
Ablation study results with MDR = 90%. (%) All BrainCLIP-vs-variant comparisons yield $p < 0.01$ in the paired Wilcoxon test with Bonferroni correction; the p-values are interpreted as supportive because repeated cross-validation folds overlap across seeds.

Variant	Epilepsy			ADNI		
	ACC	F1	AUC	ACC	F1	AUC
w/o SCL	78.15 ± 3.84	77.47 ± 3.97	78.83 ± 3.71	68.32 ± 3.42	67.28 ± 3.59	69.11 ± 3.36
w/o CRL	58.49 ± 4.21	56.93 ± 4.38	59.04 ± 4.09	54.08 ± 3.91	53.72 ± 4.05	55.36 ± 3.78
w/o Dual-Sim Matrix	81.34 ± 3.05	80.12 ± 3.18	81.25 ± 2.97	70.87 ± 2.94	69.94 ± 3.07	71.42 ± 2.86
w/o Generator	72.61 ± 2.74	70.38 ± 2.86	73.17 ± 2.65	67.52 ± 2.61	66.84 ± 2.74	68.95 ± 2.53
w/o Text Encoder	80.42 ± 2.18	79.16 ± 2.31	81.07 ± 2.09	71.05 ± 2.27	70.68 ± 2.39	72.43 ± 2.21
w/o Two-Stage	79.62 ± 2.41	79.10 ± 2.55	79.41 ± 2.38	68.85 ± 2.05	68.34 ± 2.21	68.62 ± 2.18
BrainCLIP (Ours)	**85.78** ± 2.89	**85.27** ± 3.13	**85.58** ± 2.94	**75.03** ± 2.63	**74.45** ± 2.88	**74.78** ± 2.71

RMS, 36.9% reduction in MSE, and 3.2% improvement in PCC. On the ADNI dataset, BrainCLIP achieves RMS of 85.51%, MSE of 0.110, and PCC of 0.854, exceeding the best baseline (SimMLM) by 7.33% in RMS, 38.2% reduction in MSE, and 5.2% improvement in PCC. These results highlight the robustness and effectiveness of BrainCLIP in recovering high-quality modalities under severe missingness.

The superior recovery performance of BrainCLIP under such extreme conditions can be attributed to its dual contrastive learning strategy. The modality level contrastive objective aligns recovered features with reference features from complete samples sharing the same label, ensuring semantic consistency at the feature level. This alignment ensures that recovered modalities maintain disease-related semantic patterns rather than merely minimizing feature-level reconstruction errors. The sample level contrastive objective enforces alignment between textual embeddings and fused multi-modal representations, preserving cross-modal semantic relationships learned during pretraining. Together, these objectives enable BrainCLIP to minimize feature-level reconstruction errors while preserving structural integrity and semantic consistency, ensuring that recovered modalities are both numerically accurate and semantically aligned with disease-related patterns. Because the reference embedding is always drawn from complete samples with the same diagnostic label as the incomplete subject, the modality-level constraint applies symmetrically to NC and patient subjects. Empirically, the recovered-modality RMS, MSE, and PCC remain comparable between NC and patient subgroups on both datasets, indicating that enforcing disease-related semantic patterns does not systematically bias normal scans.

Importantly, this high-quality recovery directly translates into better downstream classification performance. By integrating the results from Tables 2 and 3, we plotted scatter plots of RMS vs. ACC and MSE vs. ACC under the 90% MDR condition. As shown in Fig. 4, there is a clear positive correlation between RMS and ACC, and a negative correlation between MSE and ACC across all methods. This correlation pattern indicates that models achieving higher recovery fidelity consistently achieve superior classification results, demonstrating that effective modality recovery is crucial for accurate disease diagnosis. BrainCLIP consistently achieves the highest recovery quality (highest RMS, lowest MSE) and the best classification accuracy, positioned at the optimal region of both scatter plots.

We also observe interesting patterns when comparing different baseline methods. SimMLM achieves the second-best recovery performance on both datasets but ranks second in classification accuracy, suggesting that its recovery strategy, while effective, may not fully preserve discriminative semantic patterns for classification. In contrast, MMBNet shows relatively lower recovery quality but maintains competitive classification performance, indicating a trade-off between reconstruction fidelity and discriminative power. BrainCLIP's dual contrastive strategy successfully bridges this gap by ensuring that recovered modalities are both numerically accurate (low MSE) and semantically aligned with disease-related patterns (high RMS), enabling superior performance in both recovery and classification tasks. This comprehensive advantage demonstrates that BrainCLIP's approach of combining semantic alignment with explicit modality recovery creates a synergistic effect, where high-fidelity reconstruction and discriminative representation learning mutually reinforce each other.

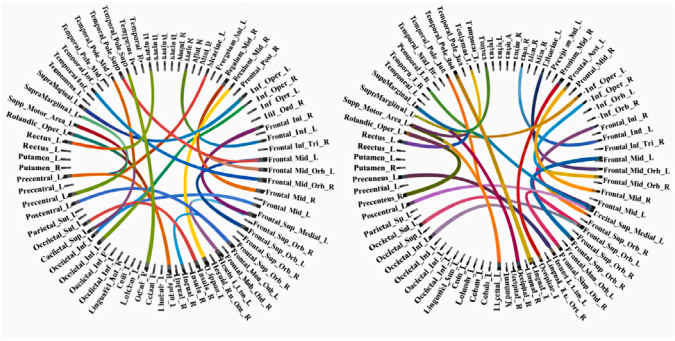
4.6. Hyperparameter sensitivity analysis

The two loss weights λ_1 and λ_2 in Eq. (12) are fixed to $\lambda_1 = \lambda_2 = 0.7$ throughout all experiments. To assess robustness, we performed a 5×5 sensitivity sweep on both datasets at MDR = 90% with $\lambda_1, \lambda_2 \in \{0.3, 0.5, 0.7, 1.0, 1.5\}$. As shown in Fig. 5, the ACC and RMS of BrainCLIP vary within ± 0.8 percentage points across all 25 configurations, and any choice in $[0.5, 1.0]$ stays within ± 0.4 percentage points of the reported best value. This indicates that BrainCLIP is robust to a moderate variation of the loss weights, and the value 0.7 is selected as a balanced operating point that performs consistently well on both classification and recovery metrics.

4.7. Ablation study

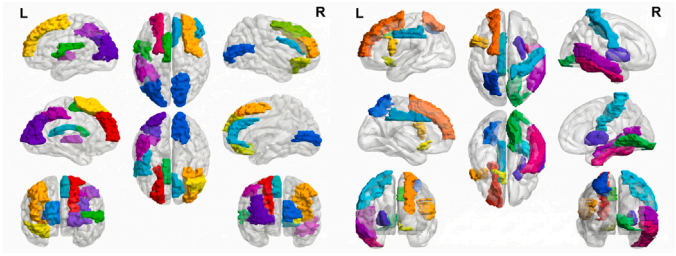
To comprehensively evaluate the impact of key components in BrainCLIP, we performed an ablation study with the MDR fixed at 90%. Six model variants were designed: (1) **w/o SCL.** Removes the similarity contrastive learning stage, directly training encoders without contrastive alignment. (2) **w/o CRL.** Excludes the cross-modal recovery learning stage, handling incomplete data without modality recovery. (3) **w/o Dual-Sim Matrix.** Uses a single similarity matrix instead of dual matrices S_f and S_d, reducing alignment granularity. (4) **w/o Generator.** Replaces the modality generator with attention-based feature completion. (5) **w/o Text Encoder.** Removes diagnostic text information, relying solely on imaging modalities. (6) **w/o Two-Stage.** Replaces the proposed two-stage training pipeline with a single-stage end-to-end joint optimization (E2E joint), in which all encoders, generators, and the fusion module are simultaneously trained from scratch with the combined loss $\mathcal{L}_{con} + \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{feat} + \lambda_2 \mathcal{L}_{text}$ on both complete and incomplete samples. This variant is included to directly compare end-to-end joint training with the two-stage design. The experimental results are presented in Table 4, where the best results are marked in bold, and the second-best results are underlined. These results provide the following insights:

- **w/o SCL:** Removing the similarity contrastive learning stage leads to significant performance degradation. Without contrastive alignment, the encoders fail to learn semantically consistent cross-modal representations, resulting in misaligned embeddings that cannot effectively guide modality recovery. The contrastive learning stage establishes semantic alignment between textual descriptions and imaging modalities, enabling textual information to serve as semantic guidance for recovery tasks.
- **w/o CRL:** Excluding the cross-modal recovery learning stage causes the most severe performance drop, confirming that cross-modal recovery is crucial for handling incomplete imaging modalities. Without explicit modality recovery, the model cannot reconstruct missing features, leading to incomplete multi-modal representations that severely degrade classification performance.
- **w/o Dual-Sim Matrix:** Using a single similarity matrix instead of dual matrices reduces alignment granularity. The dual similarity matrices S_f and S_d separately capture relationships between text



(a) Epilepsy (b) ADNI

Fig. 6. The top 20 most key brain connections in the Epilepsy and ADNI datasets.



(a) Epilepsy (b) ADNI

Fig. 7. Spatial distribution of the top 10 most discriminative brain regions in the Epilepsy and ADNI datasets.

and fMRI, and text and DTI, enabling fine-grained semantic alignment, while a single matrix cannot distinguish between different imaging modalities.

- **w/o Generator:** This variant is intended to isolate the contribution of the explicit MLP generators G_f and G_d rather than to introduce confounding architectural changes. Concretely, we keep *exactly the same input and output interface* as the full BrainCLIP, namely the inputs $[d_i; t_i]$ and $[f_i; t_i]$ and the output dimensionality $d = 512$, but replace the two-layer MLP with a single self-attention layer over the concatenated tokens, while preserving all other modules including Stage 1 contrastive pretraining, the dual-contrastive recovery losses, the fusion module, and the classifier. The only altered component is therefore the generator backbone itself. The generator reconstructs missing features from available modalities and textual embeddings, ensuring semantic consistency with the learned language-informed representation space, while attention-based completion lacks this dual-contrastive supervision-friendly inductive bias.
- **w/o Text Encoder:** Removing diagnostic text information causes moderate performance degradation, suggesting that textual diagnosis information plays an auxiliary but non-negligible role. The text encoder provides semantic anchors for modality recovery, and the sample level contrastive objective enforces alignment between textual embeddings and fused multi-modal representations. Without text guidance, the model loses the semantic supervision that enhances recovery quality.
- **w/o Two-Stage:** The single-stage end-to-end variant attains 79.62% accuracy on Epilepsy and 68.85% on ADNI, which is substantially lower than the proposed two-stage BrainCLIP at 85.78% and 75.03% by 6.16 and 6.18 percentage points respectively. This confirms the rationality of the two-stage design. When the contrastive alignment loss and the recovery and classification losses are simultaneously optimized on incomplete samples from scratch, the text-anchored semantic space is contaminated by recovery noise during training, resulting in a less discriminative shared embedding space. In contrast, the two-stage scheme first establishes a clean semantic anchor on complete samples in Stage 1, and then adapts to the incomplete-modality regime in Stage 2 within this anchored space, avoiding the joint-optimization conflict and yielding consistently better performance.

4.8. Biomarker detection

To investigate the effectiveness of the proposed method in identifying discriminative higher-order brain connections related to Alzheimer's disease and epilepsy, we identified the 20 most key brain connections using significant connectivity changes (SAC) (Zhu et al., 2018), as visualized in Fig. 6. The identification is based on the fused features generated by BrainCLIP, which incorporate both complete and recovered modalities. In the ADNI dataset, discriminative connections are mainly located in the temporal lobe, particularly the hippocampus, middle and superior temporal gyri, and parahippocampal gyrus, which are closely related to memory function and commonly affected in Alzheimer's disease. Additional key connections in the precentral and middle frontal gyri highlight their importance in distinguishing NC, SMC, and EMCI. In the Epilepsy dataset, the top 20 most influential brain connections are mainly in the temporal lobe, including the hippocampus, with additional discriminative connections in frontal and parietal regions, playing a crucial role in distinguishing NC, TLE, and FLE patients.

Some discriminative brain connections are located outside the temporal lobe and hippocampus (Fig. 6), which may be due to abnormal neural electrical activity and compensatory mechanisms. For instance, degenerative changes in Alzheimer's disease lead to communication disruptions between brain regions (Zhang et al., 2024), while abnormal neuronal discharges during epileptic seizures form seizure foci that spread through neural pathways (Chen et al., 2023), triggering compensatory responses in unaffected brain regions.

We also examine the top 10 most significant brain regions identified by BrainCLIP in both datasets, as visualized in Fig. 7. The discriminative power of each brain region is calculated by summing the regression coefficients of its associated connections from the SAC-based non-negative elastic net model. For epilepsy, the key brain regions include Angular, Frontal, and Cingulum, associated with higher cognitive functions and motor control. For Alzheimer's disease, the key brain regions include the Precuneus, Temporal Mid, and Parietal Inf, associated with memory, language, and cognitive functions. Studies have highlighted the importance of these regions in diagnosing epilepsy (Mueller et al., 2009) and Alzheimer's disease (Grady et al., 2001). These results demonstrate that BrainCLIP can effectively provide biomarkers for epilepsy and Alzheimer's disease diagnosis.

The effectiveness of our modality recovery module is further verified through similarity metrics. As shown in Table 3, BrainCLIP achieves the highest recovery quality across all metrics on both datasets, indicating that the recovered modalities are highly consistent with the original data. This high-fidelity recovery ensures that key disease-related patterns are preserved during the imputation process, making the discriminative brain connections identified by BrainCLIP (e.g., hippocampus and temporal lobes for Alzheimer's disease, frontal–temporal circuitry for epilepsy) more likely to reflect true pathophysiological abnormalities rather than noise or artifacts. This highlights BrainCLIP's potential for both accurate diagnosis and biologically grounded interpretation (Zhou et al., 2010; Mueller et al., 2009).

5. Conclusion

In this paper, we propose **BrainCLIP**, a CLIP inspired two-stage framework for incomplete multi-modal brain disease diagnosis. Unlike conventional methods, BrainCLIP jointly performs modality recovery and disease classification through a dual contrastive strategy that combines modality level and sample level contrastive objectives, effectively addressing missing imaging modality challenges. This approach not only enables effective model training with incomplete multi-modal data but also ensures semantically consistent recovered features, leading to more precise predictions. We conduct extensive experiments on both our private Epilepsy dataset and the open-source ADNI dataset. The results demonstrate that our method outperforms existing incomplete multi-modal data fusion approaches in diagnosing epilepsy and Alzheimer's disease across various missing data ratios. Furthermore, our method can identify biomarkers, offering promising prospects for enhancing the diagnosis of brain disorders. Although the proposed framework achieves promising results, its evaluation is currently limited to small-scale datasets with two imaging modalities, and primarily assumes random missing patterns. In future work, we plan to incorporate large-scale pre-trained models to enhance modality recovery, extend the framework to more diverse or longitudinal neuroimaging datasets, and strengthen collaborations with hospitals and clinical centers to expand cohort size, include more disease types and finer-grained labels, and thereby validate BrainCLIP on larger, more heterogeneous multi-center datasets. When sufficiently large multi-site cohorts become available, it would also be promising to replace the lightweight MLP generator with more expressive generative architectures such as variational autoencoders, GANs, or diffusion-based feature generators, where their additional capacity may be better exploited without the over-fitting risk inherent to the current small-sample regime. In addition, we plan to systematically investigate the influence of imaging protocols, including scanner vendor, magnetic field strength, b-values for DTI, and acquisition parameters for fMRI, on cross-protocol generalization, and to introduce protocol-aware adaptation mechanisms such as protocol embeddings or protocol-adversarial training when multi-protocol data become accessible. We also plan to quantitatively evaluate the robustness of BrainCLIP under noisy, heterogeneous, and partially missing clinical text inputs once larger multi-site cohorts with naturally heterogeneous narrative styles are available, and to develop text-imputation or prompt-tuning strategies for such deployment scenarios.

CRediT authorship contribution statement

Jinrong Cui: Writing – original draft. **Weihao Ye:** Writing – review & editing, Software. **Jie Wen:** Supervision. **Qi Zhu:** Resources, Investigation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported in part by Shenzhen Science and Technology Program under Grant No. JCYJ20240813105135047, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant No. 2024A1515030213, in part by the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University under Grant No. VRLAB2025B01, and in part by the National Natural Science Foundation of China under Grants 62576140, 62371234, and 62076129.

Data availability

Data will be made available on request.

References

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jin, D., Naumann, T., McDermott, M., 2019. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. Association for Computational Linguistics, pp. 72–78.

Asadi-Pooya, A.A., Brigo, F., Lattanzi, S., Blumcke, I., 2023. Adult epilepsy. Lancet 402 (10399), 412–424.

Ashburner, J., Barnes, G., Chen, C., Daunizeau, J., Flandin, G., Friston, K., Gitelman, D., Kiebel, S., Kilner, J., Litvak, V., et al., 2012. SPM8 manual. Funct. Imaging Lab. Inst. Neurol..

Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E., 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations. In: Advances in Neural Information Processing Systems. Vol. 33, pp. 12546–12558.

Chen, Z., Brodie, M.J., Ding, D., Kwan, P., 2023. Epidemiology of epilepsy and seizures. Front. Epidemiol. 3, 1273163.

Chen, X., Wang, X., Zhang, K., Fung, K.-M., Thai, T.C., Moore, K., Mannel, R.S., Liu, H., Zheng, B., Qiu, Y., 2022. Recent advances and clinical applications of deep learning in medical image analysis. Med. Image Anal. 79, 102444.

Chicco, D., Warrens, M.J., Jurman, G., 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. Peerj Comput. Sci. 7, e623.

Cui, Z., Zhong, S., Xu, P., He, Y., Gong, G., 2013. PANDA: A pipeline toolbox for analyzing brain diffusion images. Front. Hum. Neurosci. 7, 42.

Dong, Y., Peng, C.-Y.J., 2013. Principled missing data methods for researchers. SpringerPlus 2 (1), 222.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27 (8), 861–874.

Fotiadis, P., Parkes, L., Davis, K.A., Satterthwaite, T.D., Shinohara, R.T., Bassett, D.S., 2024. Structure–function coupling in macroscale human brain networks. Nature Rev. Neurosci. 25 (10), 688–704.

Grady, C.L., Furey, M.L., Pietrini, P., Horwitz, B., Rapoport, S.I., 2001. Altered brain functional connectivity and impaired short-term memory in Alzheimer's disease. Brain 124 (4), 739–756.

Gu, Z., Jamison, K.W., Sabuncu, M.R., Kuceyeski, A., 2021. Heritability and interindividual variability of regional structure-function coupling. Nat. Commun. 12 (1), 4894.

Han, K., Hu, D., Zhao, F., Liu, T., Yang, F., Li, G., 2025. Incomplete multi-modal disentanglement learning with application to Alzheimer's disease diagnosis. IEEE Trans. Med. Imaging.

Horng, Y.-J., Chen, S.-M., Chang, Y.-C., Lee, C.-H., 2005. A new method for fuzzy information retrieval based on fuzzy hierarchical clustering and fuzzy inference techniques. IEEE Trans. Fuzzy Syst. 13 (2), 216–228.

Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., Smith, S.M., 2012. FSL. NeuroImage 62 (2), 782–790.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Li, S., Chen, C., Han, J., 2025. SimMLM: A simple framework for multi-modal learning with missing modality. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ICCV, pp. 24068–24077.

Liu, Y., Fan, L., Zhang, C., Zhou, T., Xiao, Z., Geng, L., Shen, D., 2021. Incomplete multi-modal representation learning for Alzheimer's disease diagnosis. Med. Image Anal. 69, 101953.

Liu, M., Zhang, J., Yap, P.-T., Shen, D., 2016. Diagnosis of Alzheimer's disease using view-aligned hypergraph learning with incomplete multi-modality data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 308–316.

Meng, X., Sun, K., Xu, J., He, X., Shen, D., 2024. Multi-modal modality-masked diffusion network for brain MRI synthesis with random modality missing. IEEE Trans. Med. Imaging 43 (7), 2587–2598.

Mitra, M., Chaudhuri, B., 2000. Information retrieval from documents: A survey. Inf. Retr. 2, 141–163.

Mueller, S.G., Laxer, K.D., Barakos, J., Cheong, I., Garcia, P., Weiner, M.W., 2009. Widespread neocortical abnormalities in temporal lobe epilepsy with and without mesial sclerosis. Neuroimage 46 (2), 353–359.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR, pp. 8748–8763.

Scheltens, P., De Strooper, B., Kivipelto, M., Holstege, H., Chételat, G., Teunissen, C.E., Cummings, J., van der Flier, W.M., 2021. Alzheimer's disease. Lancet 397 (10284), 1577–1590.

Shen, D., Wu, G., Suk, H.-I., 2017. Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. 19 (1), 221–248.

Wang, R., Benner, T., Sorensen, A.G., Wedeen, V.J., 2007. Diffusion toolkit: A software package for diffusion imaging data processing and tractography. In: Proc Intl Soc Mag Reson Med. Vol. 15, Berlin.

Wang, H., Chen, Y., Ma, C., Avery, J., Hull, L., Carneiro, G., 2023. 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.

Wang, Z., Wu, Z., Agarwal, D., Sun, J., 2022. Medclip: Contrastive learning from unpaired medical images and text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing. Vol. 2022, p. 3876.

Warren, S.L., Moustafa, A.A., 2023. Functional magnetic resonance imaging, deep learning, and Alzheimer's disease: A systematic review. J. Neuroimaging 33 (1), 5–18.

Xia, P., Zhang, L., Li, F., 2015. Learning similarity with cosine similarity ensemble. Inform. Sci. 307, 39–52.

Xie, C., Zhou, W., Peng, C., Hoshyar, A.N., Xu, C., Naseem, U., Xia, F., 2024. Multimodal hyperbolic graph learning for Alzheimer's disease detection. In: Australasian Joint Conference on Artificial Intelligence. Springer, pp. 390–403.

Yan, C.-G., Zang, Y.-F., 2010. DPARSF: A MATLAB toolbox for "Pipeline" data analysis of resting-state fMRI. Front. Syst. Neurosci. 4, 13.

Yang, Y., Chen, H., Chang, Z., Xiang, Y., Ye, C., Ma, T., 2024. Incomplete learning of multi-modal connectome for brain disorder diagnosis via modal-mixup and deep supervision. In: Medical Imaging with Deep Learning. PMLR, pp. 1006–1018.

Yuan, J., Ran, X., Liu, K., Yao, C., Yao, Y., Wu, H., Liu, Q., 2022. Machine learning applications on neuroimaging for diagnosis and prognosis of epilepsy: A review. J. Neurosci. Methods 368, 109441.

Zeng, Z., Peng, Z., Yang, X., Shen, W., 2024. Missing as masking: Arbitrary cross-modal feature reconstruction for incomplete multimodal brain tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 424–433.

Zeng, L., Wang, X., Yang, J., Hu, X., Liu, T., 2023. Inter-subject and intra-subject contrastive learning for functional brain network analysis. Med. Image Anal. 87, 102838.

Zhang, J., Zhang, Y., Wang, J., Xia, Y., Zhang, J., Chen, L., 2024. Recent advances in Alzheimer's disease: Mechanisms, clinical trials and new drug development strategies. Signal Transduct. Target. Ther. 9 (1), 211.

Zhou, J., Greicius, M.D., Gennatas, E.D., Growdon, M.E., Jang, J.Y., Rabinovici, G.D., Kramer, J.H., Weiner, M., Miller, B.L., Seeley, W.W., 2010. Divergent network connectivity changes in behavioural variant frontotemporal dementia and Alzheimer's disease. Brain 133 (5), 1352–1367.

Zhu, Q., Huang, J., Xu, X., 2018. Non-negative discriminative brain functional connectivity for identifying schizophrenia on resting-state fMRI. Biomed. Eng. Online 17, 1–15.

Zhu, Q., Li, H., Ye, H., Zhang, Z., Wang, R., Fan, Z., Zhang, D., 2022. Incomplete multi-modal brain image fusion for epilepsy classification. Inform. Sci. 582, 316–333.

Zhu, W., Sun, L., Huang, J., Han, L., Zhang, D., 2021. Dual attention multi-instance deep learning for Alzheimer's disease diagnosis with structural MRI. IEEE Trans. Med. Imaging 40 (9), 2354–2366.

Jinrong Cui received the Ph.D. degree in computer science and technology from the Harbin Institute of Technology, Shenzhen, China, in 2015. She is currently as an Assistant Professor with the College of Mathematics and Informatics, South China Agricultural University, Guangzhou, China. Her current research interests include image and video enhancement, pattern recognition, and machine learning.

Weihao Ye is currently pursuing the master's degree with the College of Mathematics and Informatics, South China Agricultural University, Guangzhou, China. His current research interests include machine learning and multi-modal learning.

Jie Wen (Senior Member, IEEE) received the Ph.D. degree in Computer Science and Technology at Harbin Institute of Technology, Shenzhen in 2019. He is currently an Associate Professor at the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen. His research interests include image and video enhancement, pattern recognition, and machine learning. For more information, please refer to the homepage: https://sites.google.com/view/jerry-wen-hit/home.

Qi Zhu (Senior Member, IEEE) received his B.S. degree, M.S. degree, and Ph.D. degree from Harbin Institute of Technology in 2007, 2010, and 2014, respectively. He is a professor at College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics. He has published over 100 scientific articles in refereed international journal and conference proceedings. His interests include pattern recognition and multi-modal data fusion.