

Multimodal Decoupled Dynamic Graph Learning for Brain Disease Diagnosis

Supplementary Material

1. Modality Quality Score

A central challenge in multimodal fusion lies in the intrinsic imbalance across modalities, whereby their contributions to the final prediction vary markedly. We posit that such imbalance primarily stems from differences in modality quality. Here, quality is defined as a composite notion that encompasses both the discriminative capability of a modality for the target task and the effective density of information it conveys. This disparity is particularly pronounced in medical multimodal datasets. For instance, medical imaging modalities (e.g., MRI and CT) typically provide high-dimensional and information-rich morphological or functional representations, whereas clinical assessments or laboratory measurements are usually low-dimensional and comparatively sparse. When naively fused, features from highly discriminative and information-dense modalities tend to dominate the representation space, thereby overshadowing modalities with lower-quality signals. In attention-based fusion frameworks, this manifests as substantially lower attention weights assigned to low-quality modalities.

Table 3. Modality Quality Scores

Modality	MacroF1(%)	EffDim	$S(\%)$
COGNITIVE TEST	78.01 ± 6.37	14	28.80
CSF	26.18 ± 4.87	2	23.83
ROI AVERAGE	49.35 ± 8.17	7	23.73
PET	52.58 ± 2.98	42	13.98
MRI	48.96 ± 6.28	70	11.49
RISK FACTOR	34.29 ± 2.75	27	10.29

To quantitatively characterize these quality differences and to inform the subsequent disentanglement-then-fusion strategy, we introduce the Modality Quality Score, which integrates unimodal predictive performance with an estimate of information density. We employ the effective feature dimensionality (EffDim) to quantify the amount of non-redundant information encoded in each modality, avoiding the misconception that high raw dimensionality necessarily implies information richness. EffDim is computed using principal component analysis:

$$\text{EffDim} = \min \left\{ k \in \mathbb{Z}^+ : \sum_{i=1}^k \lambda_i \geq \alpha \sum_{j=1}^D \lambda_j \right\} \quad (15)$$

where λ_i denotes the variance of the i -th principal component in descending order, D represents the original feature dimensionality, and α is the cumulative variance threshold (set to 0.95 in this study). Thus, EffDim corresponds to the minimal number of components needed to explain 95% of the total variance. The unimodal predictive ability of each modality is evaluated using K -fold cross-validation:

$$\text{Performance} = \frac{1}{K} \sum_{i=1}^K \text{MacroF1}_i, \quad (16)$$

where K is the number of folds (set to 5), and MacroF1_i denotes the MacroF1 score of the i -th fold. The final quality score S is defined as:

$$S = \frac{\text{Performance}}{\log(1 + \text{EffDim})}, \quad (17)$$

where the logarithmic term $\log(1 + \text{EffDim})$ provides a mild regularization to prevent modalities with excessively large effective dimensionality—often accompanied by noise or redundancy—from receiving disproportionately high scores. As illustrated in Tab. 3 and Fig. 1, we compute the modality quality scores on the TADPOLE three-class classification task. The COGNITIVE TEST modality achieves the highest score, consistent with its strong cross-modal interactions in the attention maps, whereas several lower-quality modalities exhibit markedly weaker interaction patterns.

2. Proof of the Modality-Specific Embedding Constraint $c_{ii} \rightarrow 1$

In this section, we provide a theoretical justification for the constraint $c_{ii} \rightarrow 1$ imposed on modality-specific embeddings. Although minimizing c_{ii} may appear beneficial for reducing dimension-wise correlations, we show that enforcing $c_{ii} \rightarrow 1$ is essential from the perspectives of normalization consistency, information-theoretic optimality, and optimization stability.

2.1. Normalized Feature Contradiction Analysis

Let $\mathbf{p} \in \mathbb{R}^d$ be an ℓ_2 -normalized feature vector, so $\|\mathbf{p}\|_2^2 = 1$. Consider the outer-product matrix $\mathbf{p}\mathbf{p}^\top$, whose diagonal entries satisfy

$$c_{ii} = p_i^2. \quad (18)$$

Summing over all dimensions yields

$$\sum_{i=1}^d c_{ii} = \sum_{i=1}^d p_i^2 = 1. \quad (19)$$

If one forces $c_{ii} \rightarrow 0$ for all i , then

$$\sum_{i=1}^d c_{ii} \rightarrow 0, \quad (20)$$

which contradicts the normalization constraint. Therefore, not all c_{ii} can be simultaneously reduced to zero without violating the unit-norm condition. This implies that each feature dimension must retain a non-negligible contribution to the total variance, and thus maintaining c_{ii} near 1 (meaningfully preserving variance rather than collapsing it) is essential for keeping dimension-wise information and avoiding degenerate solutions.

2.2. Information-Theoretic Perspective

Interpreting each feature dimension as a random variable Z_i , and assuming centered, normalized features, c_{ii} approximates the variance of Z_i . Minimizing $(c_{ii} - 1)^2$ thus corresponds to maximizing the entropy of Z_i :

$$\min(c_{ii} - 1)^2 \iff \max H(Z_i) \iff \max \mathbb{E}_{z_i \sim p_{Z_i}} [-\log p(z_i)]. \quad (21)$$

Under unit variance, the entropy is maximized when Z_i follows a Gaussian distribution. Meanwhile, minimizing c_{ij}^2 for $i \neq j$ promotes statistical independence:

$$\min c_{ij}^2 \iff \max H(Z_i | Z_j) \iff \min I(Z_i; Z_j), \quad (22)$$

where $I(Z_i; Z_j)$ is the mutual information. Combining both principles yields the information-theoretic objective

$$\max_Z \left[\sum_{i=1}^d H(Z_i) - \sum_{i \neq j} I(Z_i; Z_j) \right], \quad (23)$$

which aligns with inter-modal disentanglement loss $\mathcal{L}_{\text{inter}}$. This formulation shows that maintaining $c_{ii} \approx 1$ while pushing $c_{ij} \approx 0$ encourages each feature dimension to carry informative and independent signals.

2.3. Optimization Feasibility and Stability

Let $Z \in \mathbb{R}^{N \times d}$ be the feature matrix and $C = Z^\top Z$ its correlation matrix. Consider the loss

$$\mathcal{L} = \frac{1}{d} \sum_{i=1}^d (c_{ii} - 1)^2 + \frac{1}{d(d-1)} \sum_{i \neq j} c_{ij}^2, \quad (24)$$

and define the Lyapunov function

$$V = \|C - I_d\|_{\mathbb{F}}^2. \quad (25)$$

Under gradient descent, $\dot{Z} = -\frac{\partial \mathcal{L}}{\partial Z}$, and the derivative of C satisfies

$$\dot{C} = \dot{Z}^\top Z + Z^\top \dot{Z} = -2(Z^\top (J - I_d)Z + Z^\top \text{diag}(C - I_d)Z), \quad (26)$$

where J is the all-ones matrix and $\text{diag}(C - I_d)$ collects the diagonal terms. Substituting into \dot{V} yields

$$\dot{V} = -4\|\text{off}(C)\|_F^2 - 4\|\text{diag}(C - I_d)\|_F^2 \leq 0, \quad (27)$$

with equality if and only if $C = I_d$, meaning $c_{ii} = 1$ and $c_{ij} = 0$. Since V is positive definite and radially unbounded, LaSalle's invariance principle guarantees global convergence to $C = I_d$. This confirms that the optimization dynamics naturally drive $c_{ii} \rightarrow 1$ and $c_{ij} \rightarrow 0$, ensuring both feasibility and stability of the constraint.

827

3. Method

Algorithm 1 MDDGL Training Procedure

```

1: Input: Multimodal data  $X = \{X^1, \dots, X^M\}$ , labels  $Y$ , hyperparameters  $\lambda, \beta, \gamma, \alpha$ , number of layers  $L$ , neighbor size  $k$ .
2: Output: Trained model parameters and predictions  $\hat{Y}$ .
3: for epoch = 1 to MaxEpochs do
4:   // Multimodal Disentanglement Phase
5:   for each modality  $m = 1, \dots, M$  do
6:     Project input features:  $\tilde{X}^m \leftarrow W^m X^m$ .
7:     Compute embeddings:  $Z_m^s \leftarrow f_s(\tilde{X}^m)$  (shared),  $Z_m^{spe} \leftarrow f_m^{spe}(\tilde{X}^m)$  (specific).
8:   end for
9:   Compute cross-orthogonality loss:  $\mathcal{L}_{\text{cos}} \leftarrow \frac{1}{M} \sum_{m=1}^M \|(Z_m^s)^\top Z_m^{spe}\|_F^2$ .
10:  Compute inter-modal losses  $\mathcal{L}_{\text{inter}}^s$  for shared and  $\mathcal{L}_{\text{inter}}^{spe}$  for specific embeddings.
11:   $\mathcal{L}_{\text{dis}} \leftarrow \mathcal{L}_{\text{inter}}^s + \mathcal{L}_{\text{inter}}^{spe} + \gamma \mathcal{L}_{\text{cos}}$ .
12:  // Multimodal Fusion Phase
13:  for each subject  $i = 1, \dots, N$  do
14:    Form sequences  $Z_i^s = [z_{1,i}^s, \dots, z_{M,i}^s]$ ,  $Z_i^{spe} = [z_{1,i}^{spe}, \dots, z_{M,i}^{spe}]$ .
15:    Apply masked multi-head attention:  $\tilde{Z}_i^s \leftarrow \text{Attn}(Z_i^s, \hat{M})$ ,  $\tilde{Z}_i^{spe} \leftarrow \text{Attn}(Z_i^{spe}, \hat{M})$ .
16:    Fuse features:  $h_i \leftarrow f_{\text{fu}}(\tilde{Z}_i^s, \tilde{Z}_i^{spe})$ .
17:  end for
18:  // Graph Construction
19:  Compute adjacency  $A$  from  $X$  via cosine similarity (keep top- $k$  neighbors, apply ReLU); normalize to  $\hat{A}$ .
20:  Initialize node features  $H^{(0)} = [h_1, \dots, h_N]$ .
21:  // Dynamic Message Passing Phase
22:  for  $l = 0$  to  $L - 1$  do
23:    Compute queries, keys, values:  $(Q, K, V) \leftarrow \text{Proj}(H^{(l)})$ .
24:     $\tilde{Q} \leftarrow \ell_2(Q)$ ,  $\tilde{K} \leftarrow \ell_2(K)$ .
25:     $S \leftarrow \tilde{Q}(\tilde{K}^\top V)$  (tensor contraction along head/feature dims).
26:     $\text{SA} \leftarrow \frac{S + \sum_{j=1}^N V_j}{\tilde{Q}(\tilde{K}^\top \mathbf{1}) + N}$ .
27:    Aggregate across heads and combine:  $H' \leftarrow \text{MeanHeads}(\text{SA}) + \hat{A}H^{(0)}$ .
28:    Update embeddings:  $H^{(l+1)} \leftarrow \text{LN}(\alpha H' + (1 - \alpha)H^{(l)})$ .
29:  end for
30:  // Prediction and Optimization
31:  Compute logits and predictions:  $\hat{Y} \leftarrow \text{softmax}(W_o H^{(L)})$ .
32:  Compute class-weighted cross-entropy loss  $\mathcal{L}_P(Y, \hat{Y})$ .
33:  Total loss  $\mathcal{L} \leftarrow \mathcal{L}_P + \mathcal{L}_{\text{dis}}$ .
34:  Update model parameters via  $\nabla_\theta \mathcal{L}$ .
35: end for

```

4. Dataset Details

All imaging data, including fMRI, was preprocessed and represented as phenotypic features. A summary of key dataset statistics—including modalities, age and gender distributions, and other relevant attributes—is provided in Tab. 4. Tab. 5 provides descriptive information for each modality. The complete data preprocessing pipeline and implementation code are available in the paper [35]. Subject IDs for the ABIDE and ABIDE-5 datasets are included in the dataset folders.

Table 4. Dataset overview: modalities and demographic characteristics

Dataset	Modalities	Category	Sample	Female/Male	Total	MMSE	Age
TADPOLE	MRI, CSF, RISK FACTOR, COGNITIVE TEST, ROI AVERAGE	sMCI	490	230/305	535	27.28±2.52	72.07±7.41
		pMCI	45				
	MRI, PET, CSF, RISK FACTOR, COGNITIVE TEST, ROI AVERAGE	AD	72	288/310	598	27.84±2.59	71.85±6.96
		CN	209				
		sMCI	317				
ABIDE	PHENO, ANAT, FUNC, fMRI	ASD NC	403 468	144/727	871	-	16.94±7.58
ABIDE-5	PHENO, ANAT, FUNC, MRI, fMRI	ASD NC	397 467	143/721	864	-	16.95±7.6

Table 5. Detailed description of modalities across datasets

Modality	Description
MRI	Magnetic Resonance Imaging
PET	Positron Emission Tomography
CSF	Cerebrospinal Fluid biomarkers
RISK FACTOR	Demographic and Genetic Risk Factors
COGNITIVE TEST	Neuropsychological Tests
ROI AVERAGE	Region of Interest Averaged measurements
PHENO	Demographic Information
ANAT	Automated Anatomical quality assessment metrics
FUNC	Automated Functional quality assessment metrics
fMRI	Functional Magnetic Resonance Imaging data

4.1. Experimental Setup

All comparative and ablation experiments were conducted using standard ten-fold cross-validation, where the dataset was partitioned into ten subsets. Each subset was sequentially used as the test set, while the remaining nine were used for training. For experiments evaluating the impact of varying training set sizes, the data were randomly split into training and testing sets, and the process was repeated ten times to compute the mean and standard deviation. The proposed model was trained for 1000 epochs using the Adam optimizer with a fixed learning rate of 0.001. The full configuration is included in the code. The partial hyperparameter configurations determined using Optuna [2] are summarized in Tab. 6. All experiments were conducted on a server equipped with a quad-core CPU and an NVIDIA 2080Ti GPU. To limit computational overhead, a unified set of hyperparameters λ and β was applied to the disentanglement losses for both modality-shared and modality-specific embeddings. While tuning them separately may improve model performance, it would significantly increase the Optuna search space.

Table 6. The partial hyperparameter settings for different datasets and tasks.

Dataset (Task)	Hidden Size	λ	β	γ	α	K
TADPOLE (sMCI, pMCI)	32	0.1	0.05	0.3	0.3	30
TADPOLE (AD, CN, sMCI)	32	0.1	0.0005	0.3	0.3	30
ABIDE (NC, ASD)	64	0.2	0.05	0.3	0.3	80
ABIDE-5 (NC, ASD)	32	0.3	0.05	0.1	0.5	80

5. Overall Performance

5.1. Baseline

Tab. 7 summarizes the design strategies of the baseline methods. Graph Construction is categorized into three types: KNN-based, Manually Designed, and Graph Structure Learning (data-driven graph learning). For EVGCN [8], we refer to the conference version rather than the extended journal version, since its code is not publicly available. While MMGK [15] constructs only a single adjacency matrix, it employs multiple modality-specific GCNs and is therefore categorized as a multi-graph method. Notably, although both EVGCN and MMGK are categorized as graph structure learning methods, they only learn the weights of existing edges, with edge connections being manually constructed. All baseline methods were retrained using the multimodal datasets. Although some methods do not explicitly incorporate multimodal designs, they can still be interpreted as implicitly performing data-level fusion.

Table 7. Comparison of graph construction and multimodal designs among baseline methods.

Method	Graph Construction	Multimodal Design	#Graphs
GCN	KNN-based	-	Single
GAT	KNN-based	-	Single
POPGCN	Manually Designed	-	Single
EVGCN	Graph Structure Learning	-	Single
DGM	Graph Structure Learning	-	Single
MMGL	Graph Structure Learning	Feature-level Fusion	Single
MMGK	Graph Structure Learning	Decision-level Fusion	Multi
HMFGL	Graph Structure Learning	Feature-level Fusion	Multi
HierSSL	Graph Structure Learning	Feature-level Fusion	Single

5.2. Quantitative Analysis

The comprehensive quantitative comparison is provided in Tabs. 8, 9, and 10. The reported metrics include AUC, ACC, and weighted F1 score, with Sensitivity (SEN) and Specificity (SPE) additionally reported for the ABIDE and ABIDE-5 datasets. For the TADPOLE (sMCI vs. pMCI) task, SEN and SPE are omitted due to severe class imbalance, which leads to metric instability. Moreover, addressing class imbalance lies beyond the scope of this study. Overall, the proposed method demonstrates consistent superiority over existing approaches across multiple tasks.

To assess whether the performance differences between the MDDGL model and baseline methods are statistically significant, we performed statistical analyses using results from 20 independent experimental repetitions (each initialized with a different random seed). For each evaluation metric (ACC, AUC, and F1), we applied a paired t-test to examine the paired performance differences between MDDGL and each competing approach. The test statistic is defined as:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}, \quad (28)$$

where $d_i = \text{MDDGL}_i - \text{Comp}_i$ represents the difference in performance between MDDGL and the competing method Comp in the i -th repetition, $\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i$ denotes the mean difference, $s_d = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2}$ is the sample standard deviation of the paired differences, and $n = 20$ is the number of repetitions. Under the null hypothesis, this statistic follows

Table 8. Performance comparison on the TADPOLE dataset.

Method	TADPOLE (sMCI,pMCI)			TADPOLE (AD,CN,sMCI)		
	ACC(%)	AUC(%)	F1(%)	ACC(%)	AUC(%)	F1(%)
GCN	82.58±12.61	72.11±11.19	85.75±9.60	73.42±4.57	76.54±4.20	73.36±4.63
GAT	87.28±6.58	77.16±11.42	87.17±4.10	74.09±5.41	76.79±4.36	73.87±5.45
POPGCN	89.71±3.53	75.61±10.62	86.95±2.46	81.43±4.20	90.23±2.83	81.36±4.16
EVGCN	94.21±2.26	82.57±13.70	93.21±3.14	85.28±3.10	91.76±3.59	85.25±3.13
DGM	95.33±2.25	83.41±12.94	90.73±3.29	83.29±3.29	85.65±5.62	83.14±3.31
MMGL	94.02±2.87	<u>85.86±9.12</u>	94.32±2.39	91.64±2.89	92.09±3.01	91.59±2.89
MMGK	93.45±4.56	81.67±9.47	93.08±3.57	83.77±6.03	88.71±4.04	83.54±6.09
HMFGL	<u>95.89±1.63</u>	84.33±13.83	<u>95.49±2.48</u>	<u>92.46±3.70</u>	<u>93.83±2.88</u>	<u>92.45±3.73</u>
HierSSL	94.95±1.90	82.67±7.96	94.90±1.80	91.29±2.72	92.32±2.55	91.29±2.73
MDDGL	97.00±2.25	95.83±4.95	95.51±3.71	93.31±2.47	95.91±2.58	93.29±2.49

Table 9. Performance comparison on the ABIDE dataset.

Method	ABIDE (NC,ASD)				
	ACC(%)	AUC(%)	F1(%)	SEN(%)	SPE(%)
GCN	64.76±2.79	64.50±2.81	64.51±2.84	67.50±8.15	61.50±8.87
GAT	64.87±3.45	64.70±3.66	64.53±3.56	66.41±9.37	62.99±11.90
POPGCN	79.43±6.17	82.47±1.44	78.29±9.03	80.20±12.25	77.76±9.83
EVGCN	84.48±3.84	87.07±3.75	84.47±3.84	84.76±5.06	84.13±5.01
DGM	86.00±4.14	88.65±5.38	83.68±5.65	83.94±5.53	83.41±7.64
MMGL	88.74±3.85	88.66±3.89	88.72±3.88	89.95±4.58	87.37±6.13
MMGK	80.82±5.43	81.08±8.47	80.45±5.99	84.21±9.75	76.87±14.17
HMFGL	<u>90.01±2.96</u>	<u>89.85±3.00</u>	<u>89.98±2.97</u>	<u>92.10±3.93</u>	87.60±5.05
HierSSL	89.89±3.81	89.80±3.85	89.88±3.83	91.23±4.00	<u>88.38±5.62</u>
MDDGL	90.70±2.73	92.63±2.43	90.67±2.76	92.32±4.16	88.83±5.84

Table 10. Performance comparison on the ABIDE-5 datasets.

Method	ABIDE-5 (NC,ASD)				
	ACC(%)	AUC(%)	F1(%)	SEN(%)	SPE(%)
GCN	66.68±3.51	66.24±3.69	66.45±3.62	71.97±5.23	60.51±8.12
GAT	68.29±3.39	67.88±3.50	68.15±3.47	73.02±3.96	62.74±6.23
POPGCN	77.89±1.90	81.21±3.60	77.77±1.91	82.43±4.41	72.54±4.82
EVGCN	84.71±3.06	88.41±2.72	84.63±3.10	88.84±4.11	79.84±5.63
DGM	86.12±3.27	87.99±3.84	84.13±4.93	85.42±7.39	82.65±5.13
MMGL	89.82±3.71	89.71±3.72	89.80±3.71	90.79±5.07	88.64±5.49
MMGK	81.60±3.47	84.74±6.32	81.39±3.55	88.65±4.26	73.31±6.98
HMFGL	<u>90.17±1.99</u>	<u>90.12±2.11</u>	<u>90.16±2.01</u>	<u>90.80±2.68</u>	89.44±4.56
HierSSL	<u>90.05±3.12</u>	<u>90.05±3.10</u>	<u>90.05±3.12</u>	<u>90.15±3.72</u>	<u>89.94±3.68</u>
MDDGL	91.79±2.43	91.82±3.00	91.78±2.43	92.51±3.62	90.94±4.51

p-values:

$$\chi^2 = -2 \sum_{i=1}^k \ln(p_i), \quad (29)$$

where p_i denotes the p-value of the i -th metric and $k = 3$ is the total number of metrics. The resulting statistic χ^2 follows a chi-square distribution with $2k$ degrees of freedom, and the aggregated p-value is computed as:

$$p_{\text{combined}} = P(\chi_{2k}^2 \geq -2 \sum_{i=1}^k \ln(p_i)). \quad (30)$$

As shown in Tab. 11, MDDGL achieves significantly superior overall performance compared with existing multimodal graph learning methods, including MMGL, HMFGL, and HierSSL. The consistently favorable results across all evaluated tasks, together with the combined significance level ($p_{\text{combined}} < 0.05$), further demonstrate the effectiveness of MDDGL.

Table 11. Statistical comparison of p_{acc} , p_{auc} , p_{f1} , and p_{combined} .

Dataset(Task)	Method	p_{acc}	p_{auc}	p_{f1}	p_{combined}
TADPOLE (sMCI, pMCI)	MMGL	$< 10^{-6}$	$< 10^{-6}$	0.0132	4.26×10^{-22}
	HMFGL	2.00×10^{-6}	$< 10^{-6}$	1.34×10^{-4}	4.45×10^{-19}
	HierSSL	$< 10^{-6}$	$< 10^{-6}$	9.00×10^{-5}	1.41×10^{-22}
TADPOLE (AD, CN, sMCI)	MMGL	0.0051	$< 10^{-6}$	0.0075	1.40×10^{-17}
	HMFGL	0.4660	$< 10^{-6}$	0.4700	6.37×10^{-13}
	HierSSL	3.58×10^{-4}	$< 10^{-6}$	6.50×10^{-4}	1.51×10^{-18}
ABIDE (NC, ASD)	MMGL	7.41×10^{-4}	$< 10^{-6}$	0.1890	1.25×10^{-9}
	HMFGL	0.0014	$< 10^{-6}$	0.0180	4.24×10^{-10}
	HierSSL	1.46×10^{-4}	2.00×10^{-6}	0.1580	1.46×10^{-8}
ABIDE-5 (NC, ASD)	MMGL	0.0126	5.00×10^{-5}	0.4510	3.70×10^{-5}
	HMFGL	1.03×10^{-4}	2.00×10^{-6}	0.4370	2.54×10^{-8}
	HierSSL	0.0100	7.10×10^{-5}	0.6960	5.99×10^{-5}

5.3. Qualitative Analysis

An effective graph learning method should preserve the structural integrity of original patient representations in the embedding space while ensuring clear separability between different classes. To provide an intuitive evaluation of the learned node embeddings, we visualize the embeddings generated by the penultimate layer of the dynamic message passing network on the TADPOLE and ABIDE-5 datasets. We employ t-SNE to project the high-dimensional embeddings onto a 2D space. As illustrated in Figure 7, the node embeddings form well-defined clusters corresponding to distinct classes. In the TADPOLE dataset, patients at different stages of Alzheimer’s disease are distributed into clearly separated clusters. Similarly, in the ABIDE-5 dataset, the class boundaries are sharply defined, and samples within each class exhibit high intra-class compactness. These results indicate that MDDGL effectively captures intra-class similarities while maintaining discriminative representations across classes.

6. Complexity Analysis

We compare the proposed MDDGL with representative multimodal graph learning methods, including MMGL [35], HMFGL [31], and HierSSL [25], in terms of model complexity and inference efficiency. The inclusion of the modality-shared encoder $f_s(\cdot)$, modality-specific encoders $f_m^{\text{spe}}(\cdot)$, and the fusion network $f_{\text{fu}}(\cdot)$ introduces additional parameter overhead compared to existing methods. Despite the increased parameter count, our method achieves significantly lower inference time, as shown in Table 12. We attribute this efficiency to the high computational complexity of the VL-Transformer [5] module employed in prior methods, which tends to dominate their inference runtime.

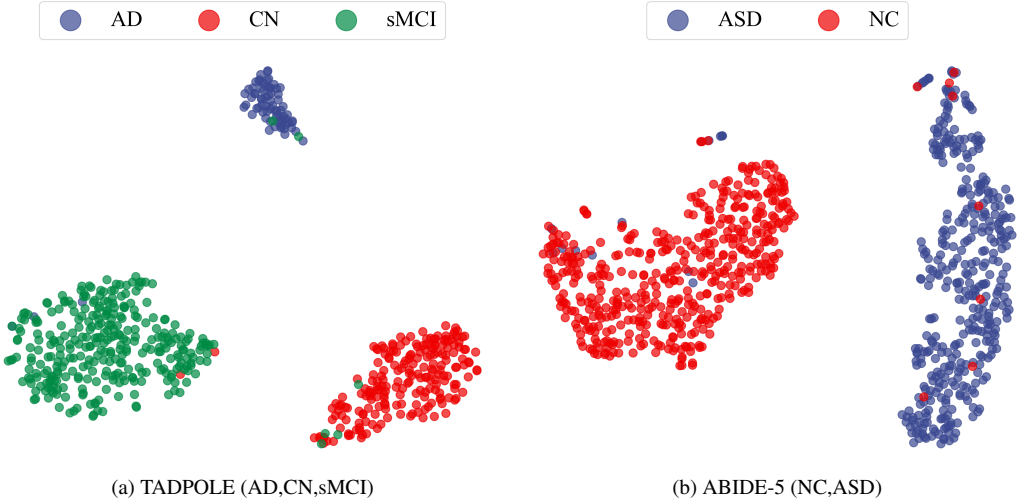


Figure 7. The 2D visualization of the node embeddings.

Table 12. Model inference time comparison across different methods and datasets (in milliseconds).

Method	TADPOLE (sMCI, pMCI)	TADPOLE (AD, CN, sMCI)	ABIDE (NC, ASD)	ABIDE-5 (NC, ASD)
MMGL	58.42	49.93	48.68	48.59
HMFG	51.40	50.10	40.14	41.47
HierSSL	69.66	70.21	81.49	80.71
MDDGL	7.32	6.72	6.37	6.76

895 **7. Case Study**

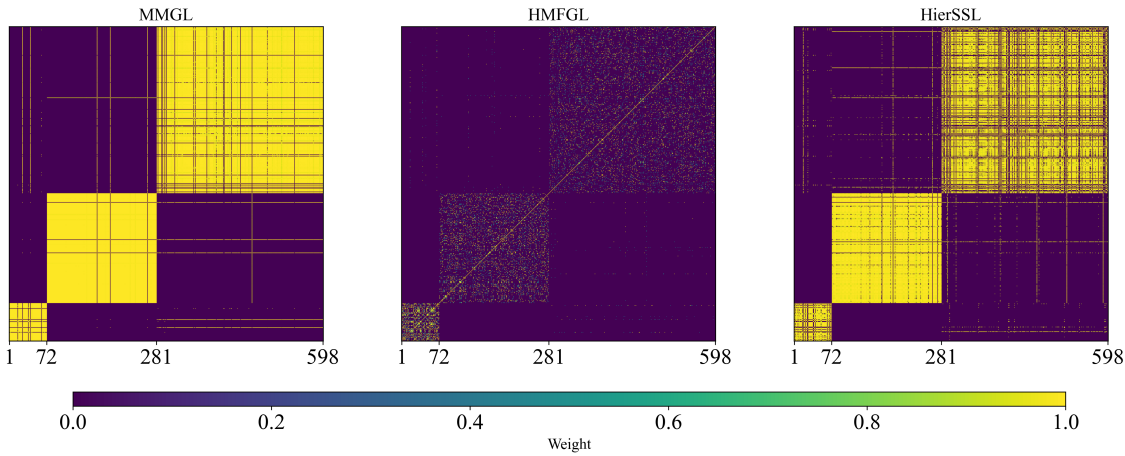


Figure 8. Visualization of inter-subject correlations in a three-class classification task on the TADPOLE dataset.

896 In Fig. 8, we present a visualization of inter-subject associations as learned by the baseline methods MMGL [35], HM-
897 FGL [31], and HierSSL [25]. These methods tend to establish connections predominantly within the same class, while largely
898 neglecting inter-class similarities. As illustrated in Fig. 5, the proposed MDDGL method, unlike the compared approaches,
effectively captures semantically meaningful cross-class similarities.

899

8. Ablation Study

Table 13. Performance comparison of ablation study on the TADPOLE dataset

Model	TADPOLE (sMCI,pMCI)			TADPOLE (AD,CN,sMCI)		
	ACC(%)	AUC(%)	F1(%)	ACC(%)	AUC(%)	F1(%)
w/o Disentanglement	95.89 \pm 2.01	92.32 \pm 7.01	90.97 \pm 5.49	91.13 \pm 2.74	94.33 \pm 3.51	91.07 \pm 2.80
w/o Fusion	95.90 \pm 2.16	94.45 \pm 5.15	93.31 \pm 2.59	91.14 \pm 3.21	95.76 \pm 2.53	90.94 \pm 3.30
w/o Mask	96.82 \pm 1.21	93.91 \pm 6.04	94.97 \pm 3.44	92.48 \pm 1.52	96.44\pm1.63	92.51 \pm 1.57
w/o Multimodal	95.71 \pm 2.62	92.76 \pm 10.72	93.33 \pm 5.36	84.13 \pm 4.06	91.95 \pm 2.96	83.90 \pm 4.20
w/o Multimodal and Adj	95.14 \pm 2.53	89.95 \pm 9.77	90.82 \pm 5.05	81.46 \pm 4.86	90.60 \pm 2.89	81.23 \pm 4.70
Full Model	97.00\pm1.25	95.83\pm2.95	96.51\pm1.71	93.31\pm2.47	95.91 \pm 2.58	93.29\pm2.49

Table 14. Performance comparison of ablation study on the ABIDE and ABIDE-5 datasets.

Model	ABIDE (NC,ASD)			ABIDE-5 (NC,ASD)		
	ACC(%)	AUC(%)	F1(%)	ACC(%)	AUC(%)	F1(%)
w/o Disentanglement	90.01 \pm 2.90	90.60 \pm 3.48	87.61 \pm 5.72	90.39 \pm 3.20	91.51 \pm 3.43	89.57 \pm 3.73
w/o Fusion	89.55 \pm 3.62	90.74 \pm 4.05	88.49 \pm 4.10	90.28 \pm 3.53	91.32 \pm 4.30	88.49 \pm 5.22
w/o Mask	90.81\pm2.13	92.10 \pm 2.63	90.10 \pm 2.01	91.55 \pm 2.39	91.73 \pm 3.12	90.50 \pm 4.42
w/o Multimodal	87.72 \pm 4.12	89.85 \pm 4.63	86.53 \pm 5.11	88.54 \pm 3.02	91.30 \pm 3.21	86.79 \pm 3.45
w/o Multimodal and Adj	87.03 \pm 3.35	89.64 \pm 3.93	85.86 \pm 3.62	87.51 \pm 2.68	90.55 \pm 3.72	86.34 \pm 3.87
Full Model	90.70 \pm 2.73	92.63\pm2.43	90.67\pm2.76	91.79\pm2.43	91.82\pm3.00	91.78\pm2.43

A comprehensive summary of the ablation results is provided in Tab. 13 and Tab. 14. In addition, we conducted an ablation analysis of the Masking strategy. While introducing the Mask leads to minor decreases in several evaluation metrics, the extent of this degradation remains marginal. Thus, we deem the module’s design to be well-justified and consistent with our intuition. For modalities with relatively low information density, highlighting their cross-modal associations may be more beneficial than focusing solely on reconstructing modality-specific information.

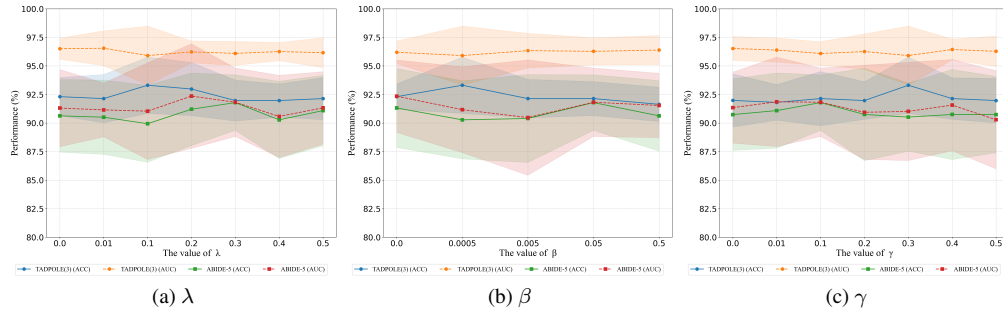


Figure 9. Impact of disentanglement loss hyperparameters on model performance.

9. Hyperparameter Study

To assess the contribution of each component of the disentanglement loss, we performed a sensitivity analysis on its hyperparameters λ , β , and γ . The experiments were conducted within the Optuna-determined search range $[0, 0.5]$, varying one hyperparameter at a time while holding the others constant. As shown in Fig. 9, adjustments to each disentanglement loss component consistently lead to performance gains. Moreover, the relatively coarse search interval suggests that a finer-grained hyperparameter exploration may yield additional improvements in model effectiveness.