HiBrain: Structure-Calibrated Higher-Order Hypergraph Learning for Brain Network Analysis

Anonymous authors
Paper under double-blind review

Abstract

Pairwise connectivity does not fully characterize joint multi-region dependence, while higher-order network construction alone does not establish subject-specific structural support. We propose HiBrain, a structure-calibrated multi-order hypergraph framework for brain disorder classification. HiBrain assesses third- and fourth-order cumulant increments relative to fitted lower-order models and calibrates candidate hyperedge support using individual gray-matter evidence. It integrates these relations through joint-evidence sparse attention and adjacent-order propagation, while conditional innovation admission controls additional higher-order contributions to a multi-order base prediction. On ABIDE-II, REST-MDD, and SRPBS-SCZ for autism spectrum disorder (ASD), major depressive disorder (MDD), and schizophrenia (SCZ) classification, respectively, HiBrain achieved the highest mean accuracy among 12 evaluated methods, exceeding the strongest baseline by 4.54, 6.94, and 6.47 percentage points. Within the same framework, it exceeded the strongest attention alternative in accuracy by 3.04, 3.04, and 2.99 points, respectively. External evaluation on ABIDE-I showed a 4.01-point accuracy gain over the strongest evaluated baseline. Model-derived regional importance showed greater cerebellar emphasis at higher orders in ASD and MDD, and posterior cortical and cerebellar emphasis at higher orders in SCZ. Code is available at `https://anonymous.4open.science/r/HiBrain-47DF`.

1 Introduction

Brain network analysis commonly uses pairwise correlations between regions of interest (ROIs) in resting-state functional magnetic resonance imaging (rs-fMRI) to characterize connectivity and classify brain disorders Li et al. (2021); Xu et al. (2024). However, pairwise connectivity does not fully specify multi-region joint dependence, which has been studied through genuine higher-order interactions, partial entropy decomposition, and multivariate cumulants Herzog et al. (2022); Varley et al. (2023); Hindriks et al. (2024). Higher-order connectomics has explored individual identification, task decoding, and behavior Santoro et al. (2024; 2026), alongside autism spectrum disorder (ASD) classification and network alterations in major depressive disorder (MDD) and schizophrenia (SCZ) Zhao et al. (2021); Kurkin et al. (2024); Li et al. (2025). These studies motivate extending pairwise analysis to third- and fourth-order relations.

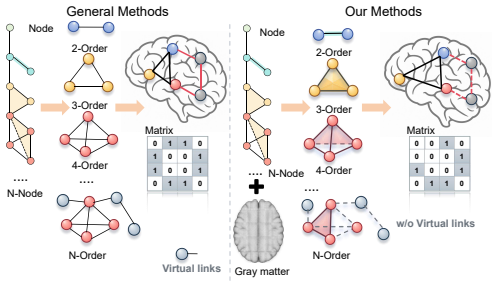


Figure 1: General construction versus gray-matter-calibrated hyperedge support in HiBrain. Nodes represent regions of interest; order indicates the number of participating ROIs.

The usefulness of higher-order representations depends on the evidence supporting their constituent relations. Existing hypergraph approaches use general construction rules to generate individual brain networks Ji et al. (2022); Guo et al. (2024); Han et al. (2026). However, such rules may introduce

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

spurious hyperedges that don't reflect genuine multi-region dependencies in an individual (Fig. 1). Adaptive directed hypergraph learning addresses construction reliability through joint estimation of robust sparse representations and relational structure Yang et al. (2025), while topological signal processing infers higher-order brain structures through multivariate statistics and joint topology–signal learning Bispo et al. (2025). Beyond identifying candidate relations, it is necessary to assess the statistical information they add relative to lower-order models and their individual structural support. These considerations motivate combining lower-order-conditioned statistical modeling with individual gray-matter calibration of hyperedge support.

Beyond construction, higher-order relations require appropriate encoding, weighting, and integration. Graph and spatio-temporal Transformers extend brain-network modeling Yu et al. (2024); Pei et al. (2025); Zhang et al. (2025a), while hypergraph learning and cross expansion enrich multi-region relational representations Yan et al. (2024); Ma et al. (2026b). However, preserving hyperedge structure does not by itself ensure that joint multi-region statistics inform computation. For relation weighting, analysis under contextual stochastic block models shows that graph attention benefits depend on structure and feature noise Ma et al. (2024). Attention weights constrained to sum to one over candidate relations still allocate mass when all candidates are weakly supported. At the prediction stage, statistical increments need not improve classification. These considerations motivate joint-evidence attention with an abstention option and selective higher-order integration conditioned on lower-order representations.

To address these limitations, we propose HiBrain, a structure-calibrated framework for higher-order brain network learning. HiBrain constructs third- and fourth-order relations using lower-order-conditioned cumulant statistics and calibrates candidate hyperedge support with individual gray-matter measurements. Its ROI–hyperedge attention combines joint relation statistics, temporal reliability, and structural support, while competition with a reference state reduces reliance on weakly supported candidates. Adjacent-order propagation exchanges information across orders, and conditional innovation admission controls additional higher-order contributions to the multi-order base prediction. The main contributions are summarized as follows:

- We construct multi-order functional relations using lower-order-conditioned cumulants and individual gray-matter calibration. This construction characterizes statistical increments relative to fitted lower-order models and calibrates candidate hyperedge support for each subject.
- We develop evidence-calibrated ROI–hyperedge attention that incorporates joint multi-region statistics, temporal reliability, and structural support into information aggregation. A reference state enables abstention from weakly supported candidates, while conditional innovation admission controls additional higher-order contributions to the multi-order base prediction.
- We evaluate HiBrain on ASD, MDD, and SCZ classification, achieving the highest mean accuracy among the evaluated methods on all three datasets. External ASD evaluation supports cross-dataset generalization, and interpretability analyses reveal model-derived regional importance patterns across disorders and interaction orders.

2 Related Work

2.1 Higher-Order Brain Network Representations

Higher-order brain network representations differ in what constitutes a multi-region relation. Dynamic approaches characterize relationships among time-varying pairwise connections Chen et al. (2016); Teng et al. (2024), whereas multivariate statistical approaches quantify joint dependence using cumulants Hindriks et al. (2024), total correlation Li et al. (2022), dual total correlation Herzog et al. (2022), and partial entropy decomposition Varley et al. (2023). Topological and hypergraph-based approaches represent multi-region organization through simplicial complexes Billings et al. (2021) and hyperedges constructed using similarity or sparse representation Guo et al. (2024); Han et al. (2026). Other approaches learn or refine relational structures through global constraints Zhan et al. (2026) and adaptive hyperedge masks Zuo et al. (2025). Beyond brain-network applications, differentiable lifting provides a general mechanism for learning higher-order structures Franco et al.

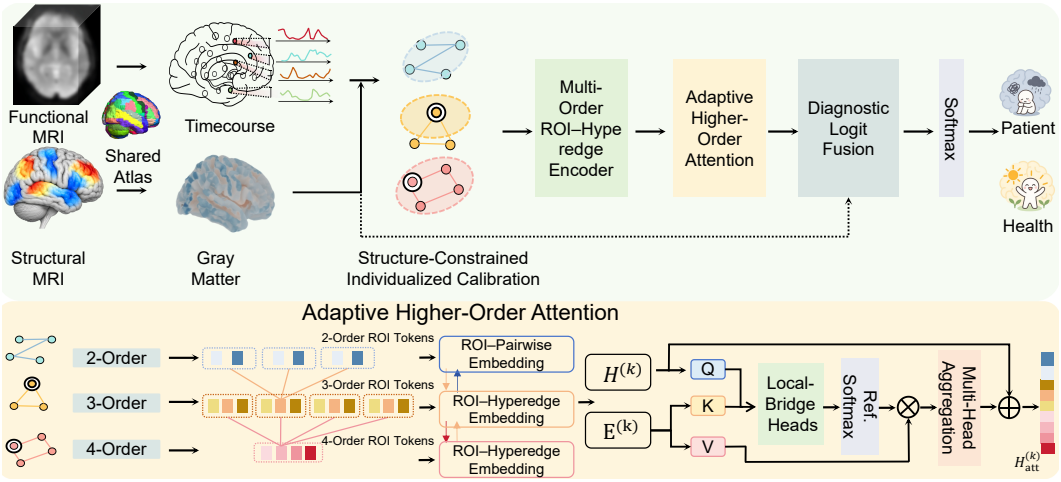


Figure 2: Workflow of HiBrain, where $\mathbf{H}^{(k)}$ and $\mathbf{E}^{(k)}$ denote the order-k ROI and candidate-relation embeddings used to generate queries (Q), keys (K), and values (V); Local/Bridge Heads capture within- and cross-network interactions, while Ref. Softmax jointly normalizes supported candidates and a zero-valued reference state. $\mathbf{H}_{\text{att}}^{(k)}$ denotes the attention-updated order-k ROI representation obtained by adding the aggregated hyperedge message to the original ROI representation $\mathbf{H}^{(k)}$.

(2026). Constructing a hyperedge does not by itself establish statistical information beyond lower-order dependencies or subject-specific anatomical support. HiBrain assesses statistical increments relative to fitted lower-order models and calibrates candidate hyperedge support using individual gray matter evidence.

2.2 Learning with Higher-Order Brain Networks

Learning methods for higher-order brain networks differ in how they encode, propagate, and select relational information for prediction. Dynamic weighted hypergraph convolution supports functional connectome learning Wang et al. (2023), while multiview hyperedge embedding integrates complementary views of functional connectivity Wang et al. (2024). Other frameworks use prompted hypergraph models for diagnosis Liu & Rossi (2025) or temporal–regional hypergraphs Ma et al. (2026b), while spatio-temporal hypergraph attention combines temporal priors, hierarchical structure generation, and spatial propagation Li et al. (2026). Hypergraph-based learning has also been investigated for anomalous co-activation detection Sadeghian et al. (2024). For information selection, higher-order information bottlenecks retain task-relevant features Zhang et al. (2025b). Hyperedge structure alone does not ensure that joint statistics guide aggregation, while task relevance does not establish added predictive value beyond lower-order representations. HiBrain incorporates joint statistical evidence into ROI–hyperedge attention and uses conditional innovation admission to control additional higher-order contributions to a multi-order base prediction.

3 Methods

3.1 Overview of the Proposed Method

As illustrated in Fig. 2, HiBrain constructs functional relations, calibrates their subject-specific support using gray-matter measurements, and aggregates them through ROI–hyperedge attention and adjacent-order propagation. A conditional admission branch then adds a bounded correction to the multi-order base prediction. Population transformations and candidate selection use training subjects only and remain fixed during evaluation.

For subject u, let T_u be scan length and N the number of ROIs, indexed by i. From rs-fMRI signals $\mathbf{X}_u \in \mathbb{R}^{T_u \times N}$ and a gray matter (GM) image $\mathbf{V}_u$, HiBrain predicts control/patient probabilities $\widehat{\mathbf{p}}_u \in [0,1]^2$ for label $y_u \in \{0,1\}$.

3.2 Multi-Order Functional Hypergraph Construction

Orders $k \in \mathcal{K} = \{2, 3, 4\}$ count distinct member ROIs; $\mathcal{K}_+ = \{3, 4\}$ denotes the higher orders. Let $\mathcal{E}_{\mathrm{tr}}^{(k)}$ be the training-derived candidate bank, e a k-ROI candidate, and $m_{u,e}^{\mathrm{tar},(k)} \in \{0, 1\}$ its target mask. The dense functional-connectivity (FC) anchor uses positive original blood-oxygen-level-dependent (BOLD) correlations; the statistical dictionary uses difference-based Fisher correlations at 2-Order and connected cumulants $c_{u,e}^{(k)}$ of standardized differences at higher orders. Let $r_{u,e}^{(2)}$ and $\widetilde{c}_{u,e}^{(k)}$ denote their training-standardized values, $\mathbf{b}_{u,e}^{(k)}$ the lower-face attributes, Φ_{quad} a clipped quadratic expansion, and $\widehat{g}^{(k)}$ a cross-fitted ridge predictor. The higher-order increment $r_{u,e}^{(k)}$ is:

$$r_{u,e}^{(k)} = \widetilde{c}_{u,e}^{(k)} - \widehat{g}^{(k)}\left(\Phi_{\mathrm{quad}}(\mathbf{b}_{u,e}^{(k)})\right), \tag{1}$$

where $\widehat{g}^{(k)}$ predicts higher-order values from lower-face evidence. Candidate banks combine pair-guided and independently sampled higher-order proposals through training-only reliability and stability screening, retaining required lower-order faces for conditioning. For width D, ROI tokens $\mathbf{h}_{u,i}^{(k)} \in \mathbb{R}^D$ encode FC profiles; edge tokens $\mathbf{e}_{u,e}^{(k)} \in \mathbb{R}^D$ combine member pooling, multiplicative interactions, residual attributes, and temporal agreement $\eta_{u,e}^{(k)} \in [0, 1]$ (Appendices A.2.1 and A.2.3).

3.3 Structure-Constrained Hyperedge Calibration

Adjusted gray-matter measurements yield candidate embeddings $\mathbf{g}_{u,e}^{(k)}$ through member encoding and mean/dispersion pooling. Using $\|$ for concatenation and σ for the logistic sigmoid, a learned scalar map $f_\rho^{(k)}$ gives support $a_{u,e}^{(k)}$ for $k \in \mathcal{K}_+$:

$$a_{u,e}^{(k)} = m_{u,e}^{\mathrm{tar},(k)} \sigma\left(f_\rho^{(k)}\left(\mathbf{e}_{u,e}^{(k)} \| \mathbf{g}_{u,e}^{(k)} \| r_{u,e}^{(k)}\right)\right), \tag{2}$$

where $m_{u,e}^{\mathrm{tar},(k)}$ excludes ineligible targets; pairwise support is $a_{u,e}^{(2)} = m_{u,e}^{\mathrm{tar},(2)}$. Support-weighted propagation yields structural order embeddings $\mathbf{s}_u^{(k)}$, complemented by a regional/spatial encoder (Appendices A.3.1 and A.3.2).

3.4 Evidence-Calibrated ROI–Hyperedge Attention

For an even head count H that divides D, let h index a head and $d_h = D/H$. Projected ROI queries $\mathbf{q}_{u,i}^{(k,h)}$ and edge keys/values $\mathbf{k}_{u,e}^{(k,h)}, \mathbf{v}_{u,e}^{(k,h)}$ lie in $\mathbb{R}^{d_h}$. Heads are divided equally into local and bridge families, with eligible incidences partitioned by shared ROI-network membership. Their network compatibility is $\nu_{i,e}^{(k,h)}$. For scalar x, define its positive part as $[x]_+ = \max(x, 0)$. With learned $\kappa_h^{\mathrm{net}}, \kappa_h^{\mathrm{rel}} \geq 0$, unrestricted δ_h^+, δ_h^-, and floor $\varepsilon_{\mathrm{rel}} > 0$, the score $\ell_{u,i,e}^{(k,h)}$ is:

$$\begin{aligned}\ell_{u,i,e}^{(k,h)} = &\frac{(\mathbf{q}_{u,i}^{(k,h)})^\top \mathbf{k}_{u,e}^{(k,h)}}{\sqrt{d_h}} + \kappa_h^{\mathrm{net}} \nu_{i,e}^{(k,h)} \\ &+ \kappa_h^{\mathrm{rel}} \log \max(\eta_{u,e}^{(k)}, \varepsilon_{\mathrm{rel}}) + \delta_h^+ [r_{u,e}^{(k)}]_+ + \delta_h^- [-r_{u,e}^{(k)}]_+,\end{aligned} \tag{3}$$

where $\kappa_h^{\mathrm{net}}, \kappa_h^{\mathrm{rel}}$ weight network and temporal evidence, and δ_h^+, δ_h^- weight signed relations.

For each ROI and head, $\mathcal{S}_{u,i}^{(k,h)}$ contains at most K_{attn} candidates selected by score within its eligible incidence domain, where K_{attn} is the selection budget. Let $d_{u,i,h}^{(k)}$ be the preselection eligible degree floored at one, $\tau_a > 0$ the score temperature, and $\ell_{u,i,\varnothing}^{(k,h)}$ the score of the reference state $\varnothing$, whose value vector is zero. Weights $\alpha_{u,i,e}^{(k,h)}$ and messages $\mathbf{m}_{u,i}^{(k,h)}$ are:

$$\begin{aligned}\alpha_{u,i,e}^{(k,h)} &= \frac{a_{u,e}^{(k)} \exp(\ell_{u,i,e}^{(k,h)}/\tau_a)}{d_{u,i,h}^{(k)} \exp(\ell_{u,i,\varnothing}^{(k,h)}) + \sum_{e' \in \mathcal{S}_{u,i}^{(k,h)}} a_{u,e'}^{(k)} \exp(\ell_{u,i,e'}^{(k,h)}/\tau_a)}, \\ \mathbf{m}_{u,i}^{(k,h)} &= \sum_{e \in \mathcal{S}_{u,i}^{(k,h)}} \alpha_{u,i,e}^{(k,h)} \mathbf{v}_{u,e}^{(k,h)},\end{aligned} \tag{4}$$

where $\alpha_{u,i,e}^{(k,h)}$ weights supported incidences and the reference receives the remaining probability, enabling abstention. Projected head messages update ROI tokens residually; empty neighborhoods give zero feedback (Appendix A.4.2).

3.5 Cross-Order Propagation and Conditional Innovation Admission

Let $M_k = |\mathcal{E}_{\mathrm{tr}}^{(k)}|$ and $\mathbf{E}_{u,\ell}^{(k)} \in \mathbb{R}^{M_k \times D}$ stack layer-ℓ edge states. Face/coface messages $\mathbf{M}_{\mathrm{up},u,\ell}^{(k)}$ and $\mathbf{M}_{\mathrm{down},u,\ell}^{(k)}$ have the same shape; learned $\lambda_{\mathrm{up},\ell}^{(k)}, \lambda_{\mathrm{down},\ell}^{(k+1)}$ scale their updates, with $\mathbb{I}$ denoting an indicator:

$$\begin{aligned}\mathbf{E}_{u,\ell+1}^{(k)} = \mathbf{E}_{u,\ell}^{(k)} &+ \mathbb{I}(k>2)\lambda_{\mathrm{up},\ell}^{(k)}\mathbf{M}_{\mathrm{up},u,\ell}^{(k)} \\ &+ \mathbb{I}(k<4)\lambda_{\mathrm{down},\ell}^{(k+1)}\mathbf{M}_{\mathrm{down},u,\ell}^{(k)},\end{aligned} \tag{5}$$

where the indicators restrict exchange to adjacent orders. Updated edges recompute incidence attention and hierarchical readout, yielding functional order embeddings $\mathbf{f}_u^{(k)}$. Pairwise-preserving fusion routes norm-adjusted higher orders into a feature $\mathbf{z}_u^{\mathrm{base}}$ of width D_z and functional logits $\boldsymbol{\ell}_u^{\mathrm{func}}$. Let $\Delta_u^{\mathrm{GM}}, \Delta_u^{\mathrm{profile}}$ be structural-hypergraph and regional/spatial log-odds corrections. The base logits $\boldsymbol{\ell}_u^{\mathrm{base}} \in \mathbb{R}^2$ are:

$$\boldsymbol{\ell}_u^{\mathrm{base}} = \boldsymbol{\ell}_u^{\mathrm{func}} + \frac{1}{2}\begin{bmatrix} -(\Delta_u^{\mathrm{GM}} + \Delta_u^{\mathrm{profile}}) \\ \Delta_u^{\mathrm{GM}} + \Delta_u^{\mathrm{profile}} \end{bmatrix}, \tag{6}$$

where the two corrections add structural evidence to the functional logits (Appendix A.5.1).

For $k \in \mathcal{K}_+$, let $\mathbf{j}_u^{(k)}$ combine functional, structural, and lower-order context. Admission scores $A_u^{(3)}, A_u^{(4)} \in [0,1]$ use this context, support, reliability, and reference probability. The allocation $\boldsymbol{\pi}_u$ has entries corresponding to 2-Order, 3-Order, and 4-Order:

$$\boldsymbol{\pi}_u = \begin{bmatrix} 1 - A_u^{(3)} \\ A_u^{(3)}(1 - A_u^{(4)}) \\ A_u^{(3)} A_u^{(4)} \end{bmatrix}, \tag{7}$$

where $A_u^{(3)}$ controls higher-order mass and $A_u^{(4)}$ its fourth-order share. Let $\pi_u^{(k)}$ be the order-k entry. HA denotes hierarchical admission within the conditional innovation admission branch. Fusing weighted innovations $\pi_u^{(k)}\mathbf{j}_u^{(k)}$ yields $\mathbf{h}_u$ of width D_{HA}. For bound $g_{\max} > 0$, base-conditioned gate $g_u \in [0, g_{\max}]$, projection $\mathbf{w}_\Delta \in \mathbb{R}^{D_{\mathrm{HA}}}$, and saturation scale $c_\Delta > 0$, the correction Δ_u^{HA} is:

$$\Delta_u^{\mathrm{HA}} = g_u c_\Delta \tanh\left(\frac{\mathbf{w}_\Delta^\top \mathbf{h}_u}{c_\Delta}\right), \qquad |\Delta_u^{\mathrm{HA}}| \le g_{\max} c_\Delta, \tag{8}$$

where g_u controls admission and $g_{\max} c_\Delta$ bounds the log-odds change. The posterior is:

$$\widehat{\mathbf{p}}_u = \mathrm{softmax}\left(\boldsymbol{\ell}_u^{\mathrm{base}} + \frac{1}{2}\begin{bmatrix} -\Delta_u^{\mathrm{HA}} \\ \Delta_u^{\mathrm{HA}} \end{bmatrix}\right), \tag{9}$$

where Δ_u^{HA} supplements the multi-order base prediction (Appendix A.6).

3.6 Loss Function

The classification loss $\mathcal{L}_{\mathrm{cls}}$ is binary cross-entropy. Let $\mathcal{L}_{\mathrm{attn}}$, $\mathcal{L}_{\mathrm{route}}$, and $\mathcal{L}_{\mathrm{link}}$ denote the attention-entropy band, order-routing, and responsibility-alignment penalties, respectively (Appendix A.7). The contribution-budget and candidate-generation penalties are zero under our configuration, so the structural regularizer $\mathcal{L}_{\mathrm{str}}$ retains the five-component scaling:

$$\mathcal{L}_{\mathrm{str}} = \frac{\mathcal{L}_{\mathrm{attn}} + \mathcal{L}_{\mathrm{route}} + \mathcal{L}_{\mathrm{link}}}{5}, \tag{10}$$

where the three terms regularize attention distributions, order routing, and responsibility alignment, respectively (Appendix A.7).

Let A, B denote alternating-block views, $\mathcal{U}_{AB}$ their paired subject set, and $\varepsilon_z > 0$ a numerical floor. For $u \in \mathcal{U}_{AB}$ and $v \in \{A, B\}$, define the normalized base feature

$\overline{\mathbf{z}}_u^v = \mathbf{z}_u^{\text{base},v} / \max(\|\mathbf{z}_u^{\text{base},v}\|_2, \varepsilon_z)$, where $\mathbf{z}_u^{\text{base},v}$ is the base feature. Let $\mathcal{T}_{AB}$ index shared representation-family/order combinations and, for $t \in \mathcal{T}_{AB}$, let $\mathcal{J}_t^{AB}$ denote the mean Jensen–Shannon (JS) divergence. The temporal regularizer $\mathcal{L}_{\text{tmp}}$ is:

$$\mathcal{L}_{\text{tmp}} = \frac{\sum_{u \in \mathcal{U}_{AB}} \|\overline{\mathbf{z}}_u^A - \overline{\mathbf{z}}_u^B\|_2^2}{\max(1, |\mathcal{U}_{AB}|) D_z} + \sum_{t \in \mathcal{T}_{AB}} \mathcal{J}_t^{AB} \Big/ (1 + |\mathcal{T}_{AB}|), \tag{11}$$

where D_z is feature width and $\mathcal{J}_t^{AB}$ compares attention and ROI distributions; empty comparisons contribute zero. Let $\lambda_{\text{str}}, \lambda_{\text{tmp}} \geq 0$ control the strengths of structural and temporal regularization, respectively, relative to the classification loss. The total training loss is:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{str}} \mathcal{L}_{\text{str}} + \lambda_{\text{tmp}} \mathcal{L}_{\text{tmp}}. \tag{12}$$

4 Experiments

4.1 Datasets

This study used four publicly available multicenter neuroimaging cohorts: ABIDE-II (19 sites; 1,090 participants), REST-MDD (24 sites; 2,380 participants) Yan et al. (2019), SRPBS-SCZ (9 sites; 294 participants) Tanaka et al. (2021), and an external cohort selected from ABIDE-I Di Martino et al. (2017) using prespecified site-level criteria (11 sites; 499 participants) (Table 1). ABIDE-II, REST-MDD, and SRPBS-SCZ were used for internal ASD, MDD, and SCZ classification, respectively, with healthy controls (HC) as the comparison group in each dataset. ABIDE-I was reserved exclusively for external testing and was excluded from model training, validation, hyperparameter selection, and threshold calibration.

Table 1: Demographic summary of the evaluated cohorts. HC denotes healthy controls; M and F denote male and female, respectively.

Dataset	Subjects	Patients	HC	Sex (M/F)
ABIDE-II	1090	518	572	847/243
REST-MDD	2380	1276	1104	925/1455
SRPBS-SCZ	294	147	147	177/117
ABIDE-I	499	251	248	426/73

4.2 Preprocessing

The fMRI data were preprocessed using DPABI (`http://www.rfmri.org/dpabi`). After removing the first 10 volumes, preprocessing included slice-timing correction, six-parameter rigid-body realignment, nuisance-covariate regression, T1 coregistration, tissue segmentation, and normalization to MNI-152 space using Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra (DARTEL). Linear trends were regressed out, and a 0.01–0.1 Hz band-pass filter was applied. Structural features were derived from the T1-weighted images. Gray-matter maps obtained by tissue segmentation were spatially normalized using DARTEL. The brain was parcellated into 160 regions using the Dosenbach-160 atlas.

4.3 Baseline Models

We compared HiBrain with 11 baseline models: BrainNetCNN Kawahara et al. (2017), GCM Zhan et al. (2026), BrainIB Zheng et al. (2024), BrainHGT Ma et al. (2026a), CrossGNN Huang et al. (2023), XG-GNN Qiu et al. (2024), HiGCN Huang et al. (2024), PHGNN Liu & Rossi (2025), MvHo-IB Zhang et al. (2025b), HoloGraph Dan et al. (2025), and SMGD Shi et al.. Attention-replacement experiments included Graph Attention (GAT) Veličković et al. (2017), Self-Attention Vaswani et al. (2017), Cross Attention Gheini et al. (2021), GATv2 Brody et al. (2021), and TeCh-CoTAR Guoqi et al. (2026).

4.4 Experimental Settings

Implementation Details. HiBrain uses 4 ROI–hyperedge attention heads with top-8 sparse candidate selection and two cross-order attention layers with 4 heads each. We trained HiBrain with

Table 2: Classification results on the three internal datasets (%, mean ± standard deviation across 10 folds). **Bold** and underlined means denote the best and second-best results, respectively.

	ABIDE-II				REST-MDD				SRPBS-SCZ			
Method	ACC	F1	AUC	SPE	ACC	F1	AUC	SPE	ACC	F1	AUC	SPE
BrainNetCNN	66.89 ± 2.86	65.88 ± 2.66	70.29 ± 3.07	65.39 ± 2.88	61.01 ± 2.52	61.60 ± 2.97	66.47 ± 3.15	63.74 ± 2.83	83.37 ± 2.66	83.03 ± 2.90	86.71 ± 3.13	84.38 ± 2.53
BrainIB	58.99 ± 3.01	57.73 ± 3.43	64.13 ± 3.18	58.57 ± 3.24	59.07 ± 3.06	60.62 ± 3.13	62.64 ± 3.23	58.14 ± 3.07	65.67 ± 3.42	65.12 ± 3.23	70.83 ± 3.12	65.14 ± 3.20
BrainHGT	63.12 ± 2.72	61.03 ± 2.93	70.13 ± 2.72	63.95 ± 2.39	64.92 ± 2.95	66.89 ± 3.07	68.38 ± 2.46	62.51 ± 2.52	78.93 ± 3.07	78.80 ± 2.89	80.55 ± 2.46	78.81 ± 2.30
CrossGNN	57.79 ± 3.09	55.14 ± 3.37	62.52 ± 3.45	59.95 ± 3.09	58.03 ± 3.13	59.81 ± 3.29	63.67 ± 3.31	56.98 ± 3.30	58.83 ± 3.14	59.36 ± 3.34	67.65 ± 3.49	57.05 ± 3.39
GCM	65.22 ± 3.12	64.42 ± 3.04	69.38 ± 3.10	63.47 ± 3.48	60.55 ± 3.44	61.94 ± 3.17	64.49 ± 3.10	60.42 ± 3.21	81.29 ± 3.03	80.77 ± 3.16	84.37 ± 3.22	81.62 ± 3.46
HiGCN	59.91 ± 3.36	58.07 ± 3.22	65.74 ± 3.14	60.47 ± 3.37	59.45 ± 3.23	60.49 ± 3.46	65.62 ± 3.32	60.22 ± 3.14	68.71 ± 3.49	68.44 ± 3.10	74.62 ± 3.33	68.57 ± 3.08
HoloGraph	62.01 ± 2.45	61.47 ± 2.94	67.42 ± 2.46	59.94 ± 2.55	60.84 ± 2.65	61.31 ± 2.76	65.63 ± 2.46	63.11 ± 3.04	72.79 ± 2.39	73.21 ± 3.17	78.55 ± 2.44	70.62 ± 3.03
MvHo-IB	61.10 ± 2.71	59.22 ± 2.91	66.12 ± 2.89	62.07 ± 2.92	59.97 ± 2.80	62.05 ± 2.47	64.52 ± 2.71	58.22 ± 2.66	75.86 ± 2.92	76.29 ± 2.76	78.85 ± 2.42	74.09 ± 2.38
PHGNN	59.64 ± 3.46	57.59 ± 3.22	63.73 ± 3.50	61.02 ± 3.13	58.45 ± 3.40	60.29 ± 3.27	61.99 ± 3.43	57.06 ± 3.07	64.66 ± 3.33	64.59 ± 3.30	68.99 ± 3.04	63.05 ± 3.31
XG-GNN	69.17 ± 2.75	67.06 ± 3.03	72.85 ± 3.19	70.96 ± 2.59	62.86 ± 2.94	65.02 ± 2.90	67.03 ± 2.45	60.29 ± 2.66	87.44 ± 2.51	86.95 ± 2.95	90.96 ± 3.05	89.24 ± 2.59
SMGD	72.10 ± 2.92	70.86 ± 2.89	76.25 ± 2.37	72.03 ± 2.32	66.93 ± 2.46	67.70 ± 3.03	70.55 ± 2.58	67.85 ± 2.61	89.79 ± 2.52	89.55 ± 3.17	89.03 ± 2.86	91.19 ± 2.38
HiBrain	**76.64** ± 2.43	**75.21** ± 2.48	**87.16** ± 2.50	**76.54** ± 2.10	**73.87** ± 2.24	**76.27** ± 2.11	**78.01** ± 2.16	**68.59** ± 2.42	**96.26** ± 2.25	**96.47** ± 2.33	**99.49** ± 1.04	**92.66** ± 2.25

Table 3: Performance of different attention mechanisms across the three datasets (%, mean ± standard deviation across 10 folds). **Bold** and underlined means denote the best and second-best results, respectively.

	ABIDE-II				REST-MDD				SRPBS-SCZ			
Method	ACC	F1	AUC	SPE	ACC	F1	AUC	SPE	ACC	F1	AUC	SPE
Graph Attention	66.28 ± 2.96	65.17 ± 3.04	76.81 ± 2.40	66.19 ± 3.71	63.58 ± 3.49	66.73 ± 3.18	67.74 ± 3.12	58.32 ± 3.75	86.44 ± 3.29	86.92 ± 2.02	89.61 ± 3.77	82.75 ± 3.25
Self-Attention	69.95 ± 2.65	68.90 ± 3.83	80.47 ± 3.58	69.86 ± 2.67	67.21 ± 3.32	70.12 ± 2.53	71.36 ± 2.79	61.94 ± 3.28	88.15 ± 2.82	88.56 ± 3.12	91.35 ± 2.39	84.48 ± 3.16
Cross Attention	71.77 ± 2.40	70.76 ± 2.63	82.30 ± 2.51	71.68 ± 2.01	69.02 ± 2.24	71.81 ± 2.05	73.17 ± 2.20	63.75 ± 2.72	89.89 ± 2.77	90.25 ± 2.25	93.02 ± 3.86	86.22 ± 2.44
GATv2	73.60 ± 3.26	72.63 ± 2.41	84.13 ± 3.39	73.51 ± 3.41	70.83 ± 2.60	73.48 ± 3.50	74.98 ± 3.76	65.56 ± 3.82	91.58 ± 2.83	91.88 ± 3.14	94.74 ± 2.10	87.91 ± 2.84
TeCh-CoTAR	68.12 ± 2.86	67.04 ± 3.85	78.64 ± 2.99	68.02 ± 2.83	65.40 ± 2.26	68.44 ± 3.98	69.55 ± 2.53	60.13 ± 2.46	93.27 ± 3.12	93.51 ± 2.23	96.43 ± 2.38	89.60 ± 3.28
HiBrain	**76.64** ± 2.43	**75.21** ± 2.48	**87.16** ± 2.50	**76.54** ± 2.10	**73.87** ± 2.24	**76.27** ± 2.11	**78.01** ± 2.16	**68.59** ± 2.42	**96.26** ± 2.25	**96.47** ± 2.33	**99.49** ± 1.04	**92.66** ± 2.25

AdamW, a batch size of 32, and an initial learning rate of 10^{-4}. Internal evaluation used 10-fold cross-validation stratified by diagnosis within acquisition sites, with an 8/1/1 training/validation/test split. External evaluation comprised 10 runs, each using a different ABIDE-II fold for validation and the remaining nine for training. Each model was tested on the same ABIDE-I cohort, with metrics reported as mean ± standard deviation across runs. Further settings are provided in Appendix C.

Evaluation Metrics. Performance was evaluated using accuracy (ACC), F1 score, area under the receiver operating characteristic curve (AUC), and specificity (SPE).

4.5 Classification and Attention Performance

Classification Performance. Table 2 shows that HiBrain achieved the highest mean performance across all datasets and metrics. Compared with the second-best results, HiBrain improved ACC, F1, AUC, and SPE by 4.54, 4.35, 10.91, and 4.51 percentage points on ABIDE-II; 6.94, 8.57, 7.46, and 0.74 points on REST-MDD; and 6.47, 6.92, 8.53, and 1.47 points on SRPBS-SCZ, respectively. SMGD was the runner-up for all comparisons except AUC on SRPBS-SCZ, where XG-GNN ranked second. The gains were particularly pronounced in AUC across all three datasets, whereas SPE improvements were smaller on REST-MDD and SRPBS-SCZ.

Attention Performance. Table 3 shows that HiBrain achieved the highest mean ACC, F1, AUC, and SPE across all three datasets. Compared with GATv2, HiBrain improved these metrics by 3.04, 2.58, 3.03, and 3.03 percentage points on ABIDE-II and by 3.04, 2.79, 3.03, and 3.03 percentage points on REST-MDD, respectively. On SRPBS-SCZ, the corresponding gains over TeCh-CoTAR were 2.99, 2.96, 3.06, and 3.06 percentage points. These improvements over the strongest alternative on each dataset support the effectiveness of HiBrain's evidence-calibrated ROI–hyperedge attention within the evaluated framework.

4.6 Ablation Experiment

Ablation settings. We compared HiBrain with six ablated variants to assess functional interaction orders, structural components, and conditional innovation admission. **(1) w/o 3-Order, 4-Order, and structural modules:** retains only the pairwise 2-Order representation. **(2) w/o 4-Order and structural modules:** retains and fuses the 2-Order and 3-Order representations. **(3) w/o structural modules:** retains all three functional orders but removes all structural components. **(4) w/o structural propagation and direct fusion:** retains all three functional orders with subject-specific

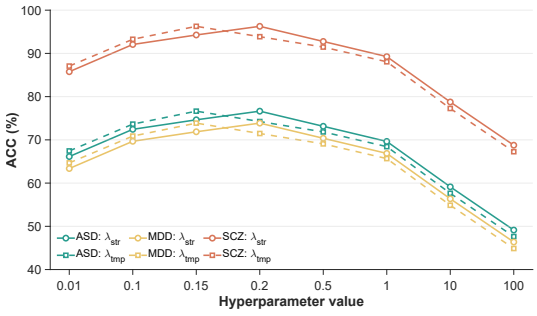


Figure 3: Sensitivity of classification accuracy to λ_{str} and λ_{tmp} on 3 datasets. Each coefficient was varied independently while the other coefficient was held fixed.

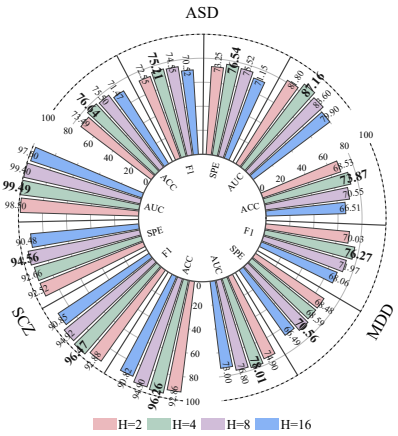


Figure 4: Sensitivity of ACC, F1, SPE, and AUC to the number of attention heads on ABIDE-II, REST-MDD, and SRPBS-SCZ.

structural support calibration but removes structural message passing and direct structural fusion. **(5) w/o direct structural fusion:** retains structural support calibration and message passing but removes direct structural fusion. **(6) w/o conditional admission:** replaces hierarchical conditional admission with an unconstrained softmax over orders, while retaining the functional and structural components, higher-order innovation representations, and bounded residual correction. All configurations followed the same site- and diagnosis-stratified 10-fold evaluation protocol.

Ablation results. As shown in Table 5, HiBrain achieved the highest mean ACC, F1, AUC, and SPE across all three datasets. Removing all structural modules caused the largest decline, followed by replacing conditional admission with unconstrained order weighting. Progressively incorporating structural calibration, message passing, and direct fusion improved all four metrics. Among functional-only variants, adding 3-Order representations improved the pairwise model, whereas further adding 4-Order representations reduced performance. These results support structural integration and conditional innovation admission for multi-order classification.

4.7 Generalization Performance on The External Dataset

Models trained on ABIDE-II were evaluated on ABIDE-I. HiBrain achieved the highest mean ACC, F1, AUC, and SPE among the evaluated methods (Table 4). Compared with SMGD, HiBrain improved these metrics by 4.01, 3.73, 4.96, and 4.84 percentage points, respectively, supporting its cross-dataset generalization.

Table 4: External validation results (%, mean ± standard deviation across 10 runs). **Bold** and underlined means denote the best and second-best results, respectively.

Method	ACC	F1	AUC	SPE
BrainNetCNN	64.33 ± 2.93	64.68 ± 3.76	69.15 ± 2.01	63.71 ± 2.79
XG-GNN	63.53 ± 3.49	63.75 ± 3.52	70.01 ± 2.45	63.31 ± 3.81
SMGD	66.93 ± 2.91	67.33 ± 2.76	73.18 ± 2.06	66.13 ± 3.42
HiBrain	**70.94** ± 2.78	**71.06** ± 2.19	**78.14** ± 3.68	**70.97** ± 3.72

4.8 Hyperparameter Analysis

Loss-Weight Coefficients. We independently varied λ_{str} and λ_{tmp} over $\{0.01, 0.10, 0.15, 0.20, 0.50, 1.00, 10.00, 100.00\}$ while holding the other fixed. As shown in Fig. 3, ACC peaked at $\lambda_{\text{str}} = 0.20$ and $\lambda_{\text{tmp}} = 0.15$ across all datasets and declined with larger weights. These values were selected using validation accuracy within the evaluated range.

Number of Attention Heads. We varied $H \in \{2, 4, 8, 16\}$ in both ROI–hyperedge and cross-order attention. As shown in Fig. 4, the validation-selected default, $H = 4$, achieved the highest ACC, F1, and AUC across all three datasets.

Table 5: Performance of HiBrain and ablated variants across the three datasets (%, mean ± standard deviation across 10 folds). **Bold** and underlined means denote the best and second-best results, respectively.

	ABIDE-II				REST-MDD				SRPBS-SCZ			
Configuration	ACC	F1	AUC	SPE	ACC	F1	AUC	SPE	ACC	F1	AUC	SPE
w/o 3-Order, 4-Order, and structural modules	69.27 ± 3.14	67.93 ± 2.01	75.16 ± 2.17	70.11 ± 2.20	67.19 ± 2.55	67.63 ± 3.28	72.68 ± 3.21	66.21 ± 3.18	86.17 ± 3.27	86.19 ± 3.01	91.24 ± 3.27	85.18 ± 2.45
w/o 4-Order and structural modules	71.36 ± 2.06	70.12 ± 2.74	77.48 ± 2.51	72.26 ± 2.17	69.36 ± 3.39	69.82 ± 2.62	74.55 ± 3.37	67.42 ± 2.86	89.43 ± 2.70	89.61 ± 2.79	94.18 ± 2.54	88.47 ± 2.31
w/o structural modules	64.57 ± 2.98	62.91 ± 2.05	69.18 ± 3.47	65.46 ± 2.56	62.81 ± 2.41	63.69 ± 3.20	68.17 ± 2.57	61.57 ± 3.04	79.07 ± 2.15	79.18 ± 2.59	84.69 ± 3.26	78.47 ± 2.64
w/o structural propagation and direct fusion	70.61 ± 2.89	69.16 ± 3.24	77.18 ± 3.42	71.53 ± 2.12	68.11 ± 2.09	68.63 ± 2.05	74.18 ± 2.31	66.59 ± 2.52	87.16 ± 3.12	87.34 ± 3.10	92.21 ± 3.24	86.13 ± 2.53
w/o direct structural fusion	74.69 ± 2.07	73.41 ± 2.93	81.37 ± 2.51	75.48 ± 3.45	72.17 ± 2.07	72.56 ± 3.05	76.86 ± 3.13	67.82 ± 3.19	92.17 ± 3.12	92.27 ± 2.95	96.14 ± 3.33	91.13 ± 3.07
w/o conditional admission	67.84 ± 3.10	66.17 ± 2.96	73.28 ± 2.09	68.79 ± 2.25	65.87 ± 2.51	66.62 ± 3.21	71.13 ± 3.48	64.73 ± 2.92	83.91 ± 2.86	83.94 ± 3.43	89.17 ± 2.12	83.67 ± 2.82
HiBrain	**76.64** ± 2.43	**75.21** ± 2.48	**87.16** ± 2.50	**76.54** ± 2.10	**73.87** ± 2.24	**76.27** ± 2.11	**78.01** ± 2.16	**68.59** ± 2.42	**96.26** ± 2.25	**96.47** ± 2.33	**99.49** ± 1.04	**92.66** ± 2.25

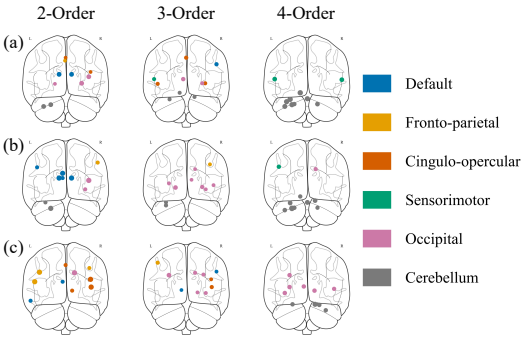


Figure 5: Top-10 ROIs ranked by HiBrain-derived regional importance for (a) ASD, (b) MDD, and (c) SCZ across 2-Order, 3-Order, and 4-Order representations.

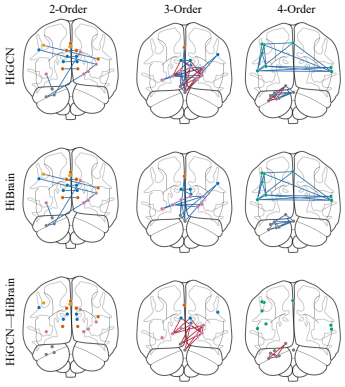


Figure 6: ASD connections in HiGCN and HiBrain: shared in blue and unique to HiGCN in red.

4.9 Interpretability Experiment

Attention-Derived Regional Importance. HiBrain showed disorder- and order-specific regional importance (Fig. 5). In ASD and MDD, importance shifted from posterior cortical and cerebellar regions at 2-Order toward predominantly cerebellar regions at 4-Order. 3-Order patterns were more distributed in ASD and mainly occipital–cerebellar in MDD. SCZ showed prefrontal, temporal, and occipital contributions at 2-Order. Its 3-Order top-ranked regions were mainly posterior cortical, whereas its 4-Order profile was dominated by the occipital and cerebellar networks.

Structural Calibration and Hyperedge Organization. We examined hyperedge importance and compared higher-order relations between HiBrain and HiGCN in ASD (Fig. 6). Shared and HiGCN-specific connections qualitatively illustrate similarities and differences in the displayed relational support. In HiBrain, gray-matter measurements calibrate functional candidate support, influencing attention and propagation. Corresponding MDD and SCZ visualizations appear in Fig. 7 (Appendix D).

5 Conclusion

We presented HiBrain, which combines lower-order-conditioned functional relations and individual gray-matter evidence through joint-evidence attention and conditional innovation admission. HiBrain achieved the highest mean accuracy among the evaluated methods across three disorders. External evaluation supported cross-dataset generalization in ASD, while generalizability to independent MDD and SCZ cohorts remains to be established. Interpretability analyses revealed model-derived regional importance patterns across disorders and interaction orders.

AI USE STATEMENT

Generative AI tools assisted with background research, language polishing, software implementation, and manuscript editing. The authors retained responsibility for the scientific framing, experimental design and controls, and all final scientific and editorial decisions. Portions of the code for data analysis, experimental evaluation, and visualization were developed with AI assistance. The authors reviewed all content and the code used in the reported experiments, and take full responsibility for the manuscript and accompanying research artifacts.

ETHICS STATEMENT

This study used existing, publicly available neuroimaging datasets distributed under their respective access and data-use policies. All analyses were conducted on de-identified data, with no new participant recruitment or data collection. We complied with the applicable data-use agreements and did not attempt to identify individual participants. Differences in demographic composition and acquisition settings across datasets may affect model performance and limit generalizability to underrepresented populations. HiBrain is intended for research purposes; its predictions should not be interpreted as clinical diagnoses or used for individual clinical decisions without further validation.

REPRODUCIBILITY STATEMENT

The model architecture and learning objectives are described in the Methods. The Experiments section specifies the datasets, evaluation protocol, and performance metrics. The source code is available via the link provided in the abstract. Additional details on preprocessing, data partitioning, hyperparameter settings, implementation, and metric computation are provided in the supplementary materials to facilitate reproduction of the experiments.

REFERENCES

Jacob Billings, Manish Saggar, Jaroslav Hlinka, Shella Keilholz, and Giovanni Petri. Simplicial and topological descriptions of human brain dynamics. *Network Neuroscience*, 5(2):549–568, 2021.

Breno C Bispo, Stefania Sardellitti, Fernando AN Santos, and Juliano B Lima. Learning higher-order interactions in brain networks via topological signal processing. In *2025 33rd European Signal Processing Conference (EUSIPCO)*, pp. 2652–2656. IEEE, 2025.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*, 2021.

Xiaobo Chen, Han Zhang, Yue Gao, Chong-Yaw Wee, Gang Li, Dinggang Shen, and Alzheimer's Disease Neuroimaging Initiative. High-order resting-state functional connectivity network for mci classification. *Human brain mapping*, 37(9):3282–3296, 2016.

Tingting Dan, Jiaqi Ding, and Guorong Wu. Explore brain-inspired machine intelligence for connecting dots on graphs through holographic blueprint of oscillatory synchronization. *Nature Communications*, 16(1):9425, 2025.

Adriana Di Martino, David O'connor, Bosi Chen, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Joshua H Balsters, Leslie Baxter, Anita Beggiato, Sylvie Bernaerts, et al. Enhancing studies of the connectome in autism using the autism brain imaging data exchange ii. *Scientific data*, 4(1): 170010, 2017.

Jorge Franco, Gabriel Duarte, Alexander Nikitin, Moacir Ponti, Diego Mesquita, and Amauri Souza. Differentiable lifting for topological neural networks. In *International Conference on Learning Representations*, volume 2026, pp. 44265–44284, 2026.

Mozhdeh Gheini, Xiang Ren, and Jonathan May. Cross-attention is all you need: Adapting pretrained transformers for machine translation. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 1754–1765, 2021.

Jiujiang Guo, Monan Wang, and Xiaojing Guo. Hierarchical hyperedge graph transformer: Toward dynamic interactions of brain networks for neurodevelopmental disease diagnosis. *IEEE Access*, 12:162145–162156, 2024.

Yu Guoqi, Juncheng Wang, Chen Yang, Jing Qin, Angelica Aviles-Rivero, and Shujun Wang. Decentralized attention fails centralized signals: Rethinking transformers for medical time series. In *International Conference on Learning Representations*, volume 2026, pp. 108935–108955, 2026.

Xiangmin Han, Mengqi Lei, and Junchang Li. Hypergraph-based semantic and topological self-supervised learning for brain disease diagnosis. *Pattern Recognition*, 169:111921, 2026.

Rubén Herzog, Fernando E Rosas, Robert Whelan, Sol Fittipaldi, Hernando Santamaria-Garcia, Josephine Cruzat, Agustina Birba, Sebastian Moguilner, Enzo Tagliazucchi, Pavel Prado, et al. Genuine high-order interactions in brain networks and neurodegeneration. *Neurobiology of Disease*, 175:105918, 2022.

Rikkert Hindriks, Tommy AA Broeders, Menno M Schoonheim, Linda Douw, Fernando Santos, Wessel van Wieringen, and Prejaas KB Tewarie. Higher-order functional connectivity analysis of resting-state functional magnetic resonance imaging data using multivariate cumulants. *Human brain mapping*, 45(5):e26663, 2024.

Qihe Huang, Lei Shen, Ruixin Zhang, Shouhong Ding, Binwu Wang, Zhengyang Zhou, and Yang Wang. Crossgnn: Confronting noisy multivariate time series via cross interaction refinement. *Advances in Neural Information Processing Systems*, 36:46885–46902, 2023.

Yiming Huang, Yujie Zeng, Qiang Wu, and Linyuan Lü. Higher-order graph convolutional network with flower-petals laplacians on simplicial complexes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 12653–12661, 2024.

Junzhong Ji, Yating Ren, and Minglong Lei. Fc–hat: Hypergraph attention network for functional brain network classification. *Information Sciences*, 608:1301–1316, 2022.

Jeremy Kawahara, Colin J Brown, Steven P Miller, Brian G Booth, Vann Chau, Ruth E Grunau, Jill G Zwicker, and Ghassan Hamarneh. Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038–1049, 2017.

Semen A Kurkin, Nikita M Smirnov, Rositsa Paunova, Sevdalina Kandilarova, Drozdstoy Stoyanov, Larisa Mayorova, and Alexander E Hramov. Beyond pairwise interactions: higher-order q-analysis of fmri-based brain functional networks in patients with major depressive disorder. *IEEE Access*, 12:197168–197186, 2024.

Chaojun Li, Peiliang Gong, Shengrong Li, Chunwei Tian, Yinbo Yu, Ran Wang, Daoqiang Zhang, and Qi Zhu. Spatio-temporal hypergraph attention networks for brain disease analysis. *IEEE Transactions on Image Processing*, 35:2727–2739, 2026. doi: 10.1109/TIP.2026.3671657.

Qiang Li, Greg Ver Steeg, Shujian Yu, and Jesus Malo. Functional connectome of the human brain with total correlation. *Entropy*, 24(12):1725, 2022.

Qiang Li, Shujian Yu, Jesus Malo, Godfrey D Pearlson, Yu-Ping Wang, and Vince D Calhoun. Higher-order triadic interactions: Insights into the multiscale network organization in schizophrenia. *Human Brain Mapping*, 46(16):e70399, 2025.

Xiaoxiao Li, Yuan Zhou, Nicha Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H Staib, Pamela Ventola, and James S Duncan. Braingnn: Interpretable brain graph neural network for fmri analysis. *Medical image analysis*, 74:102233, 2021.

Chenyu Liu and Luca Rossi. Phgnn: a novel prompted hypergraph neural network to diagnose alzheimer's disease. *arXiv preprint arXiv:2503.14577*, 2025.

Jiajun Ma, Yongchao Zhang, Chao Zhang, Zhao Lv, and Shengbing Pei. Brainhgt: A hierarchical graph transformer for interpretable brain network analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17617–17625, 2026a.

Yulan Ma, Fangkun Li, Wenchao Yang, Qian Si, Chenglong Yu, and Yang Li. Hyperdiag: Temporal–regional hypergraph learning via topology-enhanced state propagation for brain disease diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7874–7882, 2026b.

Zhongtian Ma, Qiaosheng Zhang, Bocheng Zhou, Yexin Zhang, Shuyue Hu, and Zhen Wang. Graph attention is not always beneficial: A theoretical analysis of graph attention mechanisms via contextual stochastic block models. *arXiv preprint arXiv:2412.15496*, 2024.

Shengbing Pei, Jiajun Ma, Zhao Lv, Chao Zhang, and Jihong Guan. Community-aware graph transformer for brain disorder identification. In *Ijcai*, pp. 4191–4199, 2025.

Xinmei Qiu, Fan Wang, Yongheng Sun, Chunfeng Lian, and Jianhua Ma. Towards graph neural networks with domain-generalizable explainability for fmri-based brain disorder diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 454–464. Springer, 2024.

Sadaf Sadeghian, Xiaoxiao Li, and Margo Seltzer. Hyperbrain: Anomaly detection for temporal hypergraph brain networks, 2024. URL `https://arxiv.org/abs/2410.02087`.

Andrea Santoro, Federico Battiston, Maxime Lucas, Giovanni Petri, and Enrico Amico. Higher-order connectomics of human brain function reveals local topological signatures of task decoding, individual identification, and behavior. *Nature Communications*, 15(1):10244, 2024.

Andrea Santoro, Matteo Neri, Simone Poetto, Davide Orsenigo, Matteo Diano, Marilyn Gatica, and Giovanni Petri. Charting higher-order models of brain function beyond pairwise interactions. *Nature Communications*, 2026.

Hongyu Shi, Kaizhong Zheng, Wensheng Zhai, Shuai Jiang, Badong Chen, and Liangjun Chen. Structured multi-modal graph disentanglement for psychiatric diagnosis. In *Forty-third International Conference on Machine Learning*.

Saori C Tanaka, Ayumu Yamashita, Noriaki Yahata, Takashi Itahashi, Giuseppe Lisi, Takashi Yamada, Naho Ichikawa, Masahiro Takamura, Yujiro Yoshihara, Akira Kunimatsu, et al. A multi-site, multi-disorder resting-state magnetic resonance image database. *Scientific data*, 8(1):227, 2021.

Yingzhi Teng, Kai Wu, Jing Liu, Yifan Li, and Xiangyi Teng. Constructing high-order functional connectivity networks with temporal information from fmri data. *IEEE Transactions on Medical Imaging*, 43(12):4133–4145, 2024.

Thomas F Varley, Maria Pope, Maria Grazia, Joshua, and Olaf Sporns. Partial entropy decomposition reveals higher-order information structures in human brain activity. *Proceedings of the National Academy of Sciences*, 120(30):e2300888120, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Junqi Wang, Hailong Li, Gang Qu, Kim M Cecil, Jonathan R Dillman, Nehal A Parikh, and Lili He. Dynamic weighted hypergraph convolutional network for brain functional connectome analysis. *Medical image analysis*, 87:102828, 2023.

Wei Wang, Li Xiao, Gang Qu, Vince D Calhoun, Yu-Ping Wang, and Xiaoyan Sun. Multiview hyperedge-aware hypergraph embedding learning for multisite, multiatlas fmri based functional connectivity network analysis. *Medical Image Analysis*, 94:103144, 2024.

Jiaxing Xu, Qingtian Bian, Xinhang Li, Aihu Zhang, Yiping Ke, Miao Qiao, Wei Zhang, Wei Khang Jeremy Sim, and Balázs Gulyás. Contrastive graph pooling for explainable classification of brain networks. *IEEE Transactions on Medical Imaging*, 43(9):3292–3305, 2024.

Chao-Gan Yan, Xiao Chen, Le Li, Francisco Xavier Castellanos, Tong-Jian Bai, Qi-Jing Bo, Jun Cao, Guan-Mao Chen, Ning-Xuan Chen, Wei Chen, et al. Reduced default mode network functional connectivity in patients with recurrent major depressive disorder. *Proceedings of the National Academy of Sciences*, 116(18):9078–9083, 2019.

Yuguang Yan, Yuanlin Chen, Shibo Wang, Hanrui Wu, and Ruichu Cai. Hypergraph joint representation learning for hypervertices and hyperedges via cross expansion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 9232–9240, 2024.

Lan Yang, Yao Li, and Chen Qiao. Constructing effective hyper-connectivity networks through adaptive directed hypergraph embedded dictionary learning: Application to early mild cognitive impairment detection. *IEEE Transactions on Medical Imaging*, 2025.

Shuo Yu, Shan Jin, Ming Li, Tabinda Sarwar, and Feng Xia. Long-range brain graph transformer. *Advances in Neural Information Processing Systems*, 37:24472–24495, 2024.

Ling Zhan, Junjie Huang, Xiaoyao Yu, Wenyu Chen, and Tao Jia. Beyond pairwise connections: Extracting high-order functional brain network structures under global constraints. *Advances in Neural Information Processing Systems*, 38:171568–171600, 2026.

Jing Zhang, Yanjun Lyu, Xiaowei Yu, Lu Zhang, Chao Cao, Tong Chen, Minheng Chen, Yan Zhuang, Tianming Liu, and Dajiang Zhu. Classiffication of mild cognitive impairment based on dynamic functional connectivity using spatio-temporal transformer. In *2025 IEEE 22nd international symposium on biomedical imaging (ISBI)*, pp. 1–5. IEEE, 2025a.

Kunyu Zhang, Qiang Li, and Shujian Yu. Mvho-ib: Multi-view higher-order information bottleneck for brain disorder diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 407–417. Springer, 2025b.

Feng Zhao, Xiangfei Zhang, Kim-Han Thung, Ning Mao, Seong-Whan Lee, and Dinggang Shen. Constructing multi-view high-order functional connectivity networks for diagnosis of autism spectrum disorder. *IEEE Transactions on Biomedical Engineering*, 69(3):1237–1250, 2021.

Kaizhong Zheng, Shujian Yu, Baojuan Li, Robert Jenssen, and Badong Chen. Brainib: Interpretable brain network-based psychiatric diagnosis with graph information bottleneck. *IEEE Transactions on Neural Networks and Learning Systems*, 36(7):13066–13079, 2024.

Shijia Zuo, Yu Li, Jie Wen, Xun Chen, and Aiping Liu. Adaptive hypergraph contrastive learning for asd classification using fmri connectome. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–7. IEEE, 2025.

A METHODOLOGICAL DERIVATIONS

This appendix follows the main-text computation sequence: functional relation estimation, structural calibration, evidence-calibrated attention, cross-order propagation and base readout, conditional innovation admission, and the learning objective. It expands the estimators, normalization domains, and boundary cases while retaining the notation and objective of the main text.

A.1 NOTATION AND TRAINING PROTOCOL

Let $\mathcal{U}_{\mathrm{tr}}$ denote the training subjects in an outer evaluation fold, excluding the designated validation and test subjects. Candidate selection and population preprocessing are fitted on $\mathcal{U}_{\mathrm{tr}}$ and held fixed thereafter. For boundary regression, $\mathcal{U}_{\mathrm{tr}}$ is divided into Q_{cv} subject-level partitions, with all observations from each subject assigned to the same partition. The boundary predictor and its feature standardization are fitted on the remaining partitions and applied to the held-out partition to obtain cross-fitted training residuals. For validation and testing, both are refitted using all subjects in $\mathcal{U}_{\mathrm{tr}}$. Hyperparameters are selected based on validation accuracy, while the test set is reserved for final evaluation. The separate scan-length-based block-count rule consults all loaded subjects, including held-out scans; its domain and timing are specified in Appendix A.2.2.

We use the main-text inputs $\mathbf{X}_u \in \mathbb{R}^{T_u \times N}$ and $\mathbf{V}_u$ for subject u, with ROI index i, scan length T_u, and atlas size N. Classes 0 and 1 denote control and patient. Interaction orders are $\mathcal{K} = \{2, 3, 4\}$ and $\mathcal{K}_+ = \{3, 4\}$; the order is the number of distinct member ROIs. For order k, the training-derived bank is $\mathcal{E}_{\mathrm{tr}}^{(k)}$, its size is M_k, and e denotes a candidate. Target eligibility $m_{u,e}^{\mathrm{tar},(k)} \in \{0, 1\}$ controls attention and readout, whereas $m_{u,f}^{\mathrm{bdry}} \in \{0, 1\}$ indicates whether a lower-order face f is available for conditioning. A conditioning-only face can have boundary availability one and target eligibility zero.

The common token width is D. ROI–hyperedge attention uses an even number H of heads, indexed by h, with width $d_h = D/H$. Cross-order propagation uses L_{cross} layers, each with H_{cross} attention heads; both head counts divide D. We use $\|$ for concatenation, $\odot$ for elementwise multiplication, $\mathbb{I}$ for an indicator, sg for stop-gradient, and $\sigma(x) = (1 + e^{-x})^{-1}$. For scalar x, $[x]_+ = \max(x, 0)$. Feature vectors are columns; matrices stack their transposes as rows. Let $\mathbf{M} \in \{0, 1\}^{N \times G}$ be the fixed atlas ROI–network membership matrix, where G is the number of networks; overlapping memberships are allowed. Let LN denote layer normalization (LN) and GELU the Gaussian error linear unit (GELU) activation. Positive constants ε with distinct subscripts stabilize the corresponding numerical operations.

A.2 MULTI-ORDER FUNCTIONAL RELATIONS

A.2.1 POSITIVE PAIRWISE SUPPORT AND CONNECTED CUMULANTS

Let t index consecutive observations and $\varepsilon_{\mathrm{corr}} \in (0, 1)$ be the correlation clipping margin. For distinct ROIs i and j, define original connectivity by $F_{u,ij} = \operatorname{arctanh}(\operatorname{clip}(\operatorname{corr}_t(X_{u,:,i}, X_{u,:,j}), -1 + \varepsilon_{\mathrm{corr}}, 1 - \varepsilon_{\mathrm{corr}}))$. Pairwise support is restricted to $F_{u,ij} > 0$ before training-derived transformation. The dense FC anchor transforms these retained original-BOLD correlations using training-fitted parameters; excluded pairs remain zero. Its positive-support rule is separate from the difference-based statistical dictionary. In that dictionary, the raw second-order attribute is the Fisher-transformed correlation of temporal differences, and $r_{u,e}^{(2)}$ denotes its order-wise standardized value. Dictionary target eligibility follows the candidate-selection and subject-support rules below. Define temporal differences $d_{u,t,i} = X_{u,t+1,i} - X_{u,t,i}$. Let $W_{u,\mathrm{full}} = \{1, \ldots, T_u - 1\}$ be the full difference-series domain and $W_{u,b}$ its temporal block b. Denote full-domain means and population standard deviations by $\mu_{u,i}^{\Delta}$ and $\sigma_{u,i}^{\Delta}$. With threshold $\varepsilon_{\Delta} = 10^{-8}$, let $s_{u,i}^{\Delta} = \sigma_{u,i}^{\Delta}$ when $\sigma_{u,i}^{\Delta} > \varepsilon_{\Delta}$ and $s_{u,i}^{\Delta} = 1$ otherwise. The standardized differences are

$$z_{u,t,i} = \frac{d_{u,t,i} - \mu_{u,i}^{\Delta}}{s_{u,i}^{\Delta}}, \tag{A.1}$$

where nonfinite inputs are replaced by zero before centering. For an estimation domain W, let $z_{u,t,i}^{W}$ denote these observations recentered and rescaled within W, using the same standard-deviation

threshold and unit-denominator fallback. Thus block estimates use block-specific centering and scaling.

For $e = \{i_1, \ldots, i_k\}$, let Π_k contain partitions of $\{1, \ldots, k\}$, π denote a partition, and A one of its member-index subsets. Write $\mathbb{E}_W[g] = |W|^{-1} \sum_{t \in W} g_t$ for empirical averaging and $c^{(k)}_{u,e}[W]$ for the corresponding connected-cumulant estimate. Define the block estimate $c^{(k)}_{u,e,b} = c^{(k)}_{u,e}[W_{u,b}]$ and the primary subject-level estimate $c^{(k)}_{u,e} = c^{(k)}_{u,e}[W_{u,\mathrm{full}}]$:

$$c^{(k)}_{u,e}[W] = \sum_{\pi \in \Pi_k} (|\pi| - 1)!(-1)^{|\pi|-1} \prod_{A \in \pi} \mathbb{E}_W \left[\prod_{j \in A} z^W_{u,t,i_j} \right], \tag{A.2}$$

where the partition sum removes products of lower-order moments. The current estimator uses the full-domain value as its primary statistic; the median of block estimates is a distinct quantity used for the reliability scale below. These plug-in empirical cumulants are not assumed to be finite-sample unbiased. Let $z^{(b)}_i$ abbreviate $z^{W_{u,b}}_{u,t,i}$. With all variables centered and scaled within $W_{u,b}$, the 3- and 4-Order expansions are

$$c^{(3)}_{u,\{i,j,l\},b} = \mathbb{E}_{W_{u,b}}[z^{(b)}_i z^{(b)}_j z^{(b)}_l], \tag{A.3}$$

where $\mathbb{E}_{W_{u,b}}$ averages over block b. For 4-Order,

$$\begin{aligned} c^{(4)}_{u,\{i,j,l,m\},b} = \mathbb{E}_{W_{u,b}}[z^{(b)}_i z^{(b)}_j z^{(b)}_l z^{(b)}_m] \\ - \mathbb{E}_{W_{u,b}}[z^{(b)}_i z^{(b)}_j]\mathbb{E}_{W_{u,b}}[z^{(b)}_l z^{(b)}_m] \\ - \mathbb{E}_{W_{u,b}}[z^{(b)}_i z^{(b)}_l]\mathbb{E}_{W_{u,b}}[z^{(b)}_j z^{(b)}_m] \\ - \mathbb{E}_{W_{u,b}}[z^{(b)}_i z^{(b)}_m]\mathbb{E}_{W_{u,b}}[z^{(b)}_j z^{(b)}_l], \end{aligned} \tag{A.4}$$

where centering removes singleton-partition terms, leaving the three pair partitions. The full-scan expansions use the same expressions with $W_{u,b}$ replaced by $W_{u,\mathrm{full}}$.

Training-fitted statistical coordinates. Let μ^{tr}_k and s^{tr}_k be the mean and population standard deviation of finite primary statistics pooled over selected order-k candidates and training subjects; the scale is floored at 10^{-5}. For $k \in \mathcal{K}_+$, define $\widetilde{c}^{(k)}_{u,e} = (c^{(k)}_{u,e} - \mu^{\mathrm{tr}}_k)/s^{\mathrm{tr}}_k$. The same fitted transformation is applied to temporal views and evaluation subjects. The boundary predictor targets $\widetilde{c}^{(k)}_{u,e}$, whereas the original magnitude $\zeta^{(k)}_{u,e} = |c^{(k)}_{u,e}|$ remains in the raw statistical scale. The second-order dictionary uses the analogous transformation of its difference-based Fisher correlation.

Quadratic lower-face predictor. For $k \in \mathcal{K}_+$, the lower-order boundary is $\mathcal{F}_e = \{f \subset e : 2 \le |f| < k\}$. Face attributes are concatenated in a fixed atlas-based combinatorial order, and unavailable faces contribute zero through m^{bdry}. With d_{attr} attributes per face, the boundary dimension is $d_k = d_{\mathrm{attr}} \sum_{j=2}^{k-1} \binom{k}{j}$. For face f, let $v_{u,f}$ be its training-standardized difference-based Fisher correlation at order two or cumulant at order three, $\eta_{u,f}$ its temporal reliability, and $\zeta_{u,f}$ the absolute value of its original statistic before training standardization. With clipping threshold $c_b > 0$, the boundary vector $\mathbf{b}^{(k)}_{u,e}$ and bounded vector $\mathbf{x}_{u,e}$ are:

$$\mathbf{b}^{(k)}_{u,e} = \operatorname*{Concat}_{f \in \mathcal{F}_e} m^{\mathrm{bdry}}_{u,f}[v_{u,f}, |v_{u,f}|, \eta_{u,f}, \zeta_{u,f}].$$

$$\mathbf{x}_{u,e} = \mathrm{clip}(\mathbf{b}^{(k)}_{u,e}, -c_b, c_b).$$

Unavailable face attributes are finite zeros. For a generic boundary vector $\mathbf{b}$, write $\mathbf{x} = \mathrm{clip}(\mathbf{b}, -c_b, c_b)$ and let p, q index its coordinates. The fixed quadratic map is:

$$\Phi_{\mathrm{quad}}(\mathbf{b}) = [\mathbf{x} \| \mathbf{x}^{\odot 2} \| (x_p x_q)_{p<q}]. \tag{A.5}$$

The expanded dimension is $d^{\mathrm{quad}}_k = 2d_k + \binom{d_k}{2}$, giving 90 features at 3-Order and 860 at 4-Order for four attributes per face, excluding the intercept. 4-Order conditioning uses third-order attributes before residualization. Atlas ordering fixes the face coordinates, and ridge regularization stabilizes correlated boundary features.

Order-shared, subject-level cross-fitting. A shared predictor is estimated for each target order by pooling valid subject–hyperedge observations. Let $q(u) \in \{1, \dots, Q_{\mathrm{cv}}\}$ index the internal partition of subject u. The predictor excluding partition q uses the standardized feature vector

$$\mathbf{t}_{u,e,-q}^{(k)} = T_{-q}^{(k)}\left(\Phi_{\mathrm{quad}}(\mathbf{b}_{u,e}^{(k)})\right).$$

The transformation $T_{-q}^{(k)}$ uses featurewise means and standard deviations estimated from fitting observations with $q(u) \neq q$; constant-feature scales are set to one.

Let $\mathcal{E}_{u,\mathrm{valid}}^{(k)}$ contain valid order-k candidates for subject u, and let $\lambda_{\mathrm{ridge}} > 0$ control shrinkage. The fitted coefficient vector $\widehat{\boldsymbol{\beta}}_{-q}^{(k)}$ and unpenalized intercept $\widehat{b}_{-q}^{(k)}$ minimize the objective over trial coefficients $\boldsymbol{\beta}$ and intercept b:

$$(\widehat{\boldsymbol{\beta}}_{-q}^{(k)}, \widehat{b}_{-q}^{(k)}) = \arg\min_{\boldsymbol{\beta},b} \sum_{\substack{u\in\mathcal{U}_{\mathrm{tr}}:\, q(u)\neq q \\ e\in\mathcal{E}_{u,\mathrm{valid}}^{(k)}}} \left[\widetilde{c}_{u,e}^{(k)} - b - (\mathbf{t}_{u,e,-q}^{(k)})^{\top}\boldsymbol{\beta}\right]^2 + \lambda_{\mathrm{ridge}}\|\boldsymbol{\beta}\|_2^2. \tag{A.6}$$

Valid subject–hyperedge observations receive equal weight. Let $\widehat{\widetilde{c}}_{u,e}^{(k)}$ denote the prediction excluding the subject's internal partition. The cross-fitted prediction and residual are:

$$\widehat{\widetilde{c}}_{u,e}^{(k)} = \widehat{b}_{-q(u)}^{(k)} + (\mathbf{t}_{u,e,-q(u)}^{(k)})^{\top}\widehat{\boldsymbol{\beta}}_{-q(u)}^{(k)}. \tag{A.7}$$

$$r_{u,e}^{(k)} = \widetilde{c}_{u,e}^{(k)} - \widehat{\widetilde{c}}_{u,e}^{(k)}. \tag{A.8}$$

For validation and testing, the predictor and its feature standardization are estimated once using all valid observations in $\mathcal{U}_{\mathrm{tr}}$. The predictor $\widehat{g}^{(k)}$ in Eq. 1 includes the corresponding feature transformation.

The residual $r_{u,e}^{(k)}$ in Eq. 1 measures the component unexplained by the fitted quadratic boundary model. Subsequent high-order encoding uses

$$[r_{u,e}^{(k)}, |r_{u,e}^{(k)}|, \eta_{u,e}^{(k)}, \zeta_{u,e}^{(k)}].$$

For higher-order candidates, $\zeta_{u,e}^{(k)} = |c_{u,e}^{(k)}|$ denotes the original cumulant magnitude before residualization. The predictor is fixed during classifier optimization. Signed and absolute residuals are encoded together with reliability and original magnitude. The residual is a model-relative conditional increment; ridge regularization and cross-fitting do not impose exact orthogonality to the boundary features.

Training-derived candidate generation. Let B_k be the pre-closure proposal budget for order k, and let $C_{u,ij}^{\mathrm{prop}}$ be the precomputed second-order difference statistic used for proposals. Population pair scores are $P_{ij} = \mathrm{median}_{u\in\mathcal{U}_{\mathrm{tr}}} |C_{u,ij}^{\mathrm{prop}}|$, symmetrized with a zero diagonal. Nonfinite population scores are replaced by zero. The top 1600 unordered pairs form the pair pool. For each order, candidate tuples are sorted by ROI index, repeated members are rejected, and duplicate tuples are removed.

An independent higher-order pool does not use pairwise scores. For $k = 3, 4$, draw up to $\lceil \rho_k B_k \rceil$ unique tuples, with $\rho_3 = 8$ and $\rho_4 = 12$. Sampling cycles through within-network, two-network bridge, and unrestricted proposals; infeasible network draws revert to unrestricted draws. Bridge proposals allocate $\lfloor k/2 \rfloor$ and $k - \lfloor k/2 \rfloor$ members to two randomly chosen networks. Sampling stops at the requested pool size or $\max(2000, 200\lceil \rho_k B_k \rceil)$ attempts. A fixed seed $17 + 1009k$ controls this pool. Select at most 64 training subjects without replacement using seed 17, independent of labels. For subject u and temporal block b, let $c_{u,e,b}^{(k)}$ be its block cumulant, and let $\mathcal{U}_{\mathrm{scr}}$ denote this training subset. Let B_{time} denote the shared number of temporal blocks, specified in Appendix A.2.2. Independent proposals are ranked by

$$\mathrm{Score}_{\mathrm{ind}}(e) = \mathrm{median}_{u\in\mathcal{U}_{\mathrm{scr}}}\, \mathrm{median}_b\, |c_{u,e,b}^{(k)}| \cdot \frac{1}{|\mathcal{U}_{\mathrm{scr}}|} \sum_{u\in\mathcal{U}_{\mathrm{scr}}} \left| \frac{1}{B_{\mathrm{time}}} \sum_b \mathrm{sgn}(c_{u,e,b}^{(k)}) \right|. \tag{A.9}$$

The sign-consistency exponent is one. Retain up to B_k ranked independent proposals for dictionary assembly.

The dictionary first admits up to $\mathrm{round}(0.6B_k)$ independent proposals at orders three and four. Pair-guided extension fills the nonrandom allocation up to $\mathrm{round}(0.8B_k)$: extend a lower-order tuple f by nodes ranked by $|f|^{-1}\sum_{i\in f} P_{ij}$. Third-order extensions start from the pair pool; fourth-order extensions start from assembled third-order candidates. At order two, take the highest-scoring pairs for the nonrandom allocation. Random draws fill the remainder up to B_k, alternating within-network and unrestricted draws, with seed 17 and at most $\max(2000, 100B_k)$ attempts. Deduplication applies across all sources. Then append all missing proper faces of sizes $2, \ldots, k-1$. Appended faces are marked conditioning-only unless already proposed as targets; consequently the final coordinate-bank size can exceed B_k.

Training-only stability selection. For each candidate, let $\mathcal{U}_e$ contain training subjects with a finite value of the quantity being summarized, $n_e = |\mathcal{U}_e|$, and $\bar{x}_e$ its finite-sample mean. Define the uncertainty-penalized ranking statistic using $z_{\mathrm{LCB}} = 1.645$:

$$\mathrm{LCB}_e(x) = \bar{x}_e - z_{\mathrm{LCB}}\sqrt{\frac{\sum_{u\in\mathcal{U}_e}(x_{u,e}-\bar{x}_e)^2}{\max(n_e-1,1)\max(n_e,1)}}. \tag{A.10}$$

For $n_e = 1$, the variance term is zero; for $n_e = 0$, mean and ranking score are set to zero. Thus the implementation does not automatically reject a singleton finite sample. The coefficient penalizes uncertainty for ranking; no calibrated confidence-coverage claim is made.

Let p_e^+ be the finite-subject fraction with $\eta_{u,e} > 0$, and let $t_{u,e} = \eta_{u,e}\tanh(\zeta_{u,e}/3)$ be magnitude-weighted reliability. First require original target status and $p_e^+ \geq 0.1$. For finite sigmoid reliabilities, this prevalence condition is automatically satisfied, including the neutral value 0.5; it does not distinguish weak from strong reliability. Reliability discrimination is provided by the subsequent LCB and stability filters. Among these candidates, retain those with $\mathrm{LCB}_e(\eta)$ at or above its training-fold median (quantile 0.5). Let $n_{\mathrm{elig}}^{(k)}$ count survivors, $K_{\min}^{(k)} = 3$, $K_{\max}^{(k)}$ the order-specific cap, and q_{sel} the selection quantile. Rank by $\mathrm{LCB}_e(t)$ and retain

$$K_{\mathrm{keep}}^{(k)} = \min\Big(n_{\mathrm{elig}}^{(k)}, K_{\max}^{(k)}, \max\left(3, \left\lceil(1-q_{\mathrm{sel}})n_{\mathrm{elig}}^{(k)}\right\rceil\right)\Big). \tag{A.11}$$

An order with fewer than three candidates at either eligibility stage is inactive. Repeat the selector in 20 subject bootstraps and in leave-one-training-site-out subsets when at least three sites are available; each site-excluded subset must contain at least $\max(K_{\min}^{(k)}, 2)$ subjects. Retain only candidates meeting both selection-frequency thresholds. If a stability family has no valid resamples, its frequency is one. No label-effect ranking or network-coverage repair is enabled. If fewer than three targets survive the final stability filters, deactivate the order without backfilling unreliable candidates.

The proposal budgets are $(B_2, B_3, B_4) = (256, 384, 128)$, the selection quantile is $q_{\mathrm{sel}} = 0.75$, and the target caps are $(K_{\max}^{(2)}, K_{\max}^{(3)}, K_{\max}^{(4)}) = (64, 48, 24)$. Both bootstrap and site-selection frequencies must be at least 0.60. These bounds specify proposal and selection capacities.

Individual target eligibility and coordinate banks. Let $b_e^{\mathrm{sel},(k)}$ indicate survival of training-only target selection. For higher orders, the subject-specific target mask is

$$m_{u,e}^{\mathrm{tar},(k)} = b_e^{\mathrm{sel},(k)}\mathbb{I}\Big(\eta_{u,e}^{(k)}, \zeta_{u,e}^{(k)} \text{ are finite}, \quad \eta_{u,e}^{(k)} > 0, \quad \zeta_{u,e}^{(k)} \geq 0\Big). \tag{A.12}$$

The subject reliability and magnitude thresholds are zero; higher-order signs are retained. Positive-support restrictions for the original BOLD pairwise branch remain as specified above. Conditioning-only faces have target mask zero but can remain available for boundary regression. Let $\mathcal{V} = \{1, \ldots, N\}$ be the ROI node set, $B_{i,e}^{(k)}$ the incidence indicator, $\mathbf{B}^{(k)}$ the matrix collecting these indicators, and let $\mathbf{r}_u^{(k)}, \boldsymbol{\eta}_u^{(k)}, \boldsymbol{\zeta}_u^{(k)}, \mathbf{m}_u^{\mathrm{tar},(k)}$ stack the candidate attributes and masks. The shared coordinate bank and individual hypergraph are

$$B_{i,e}^{(k)} = \mathbb{I}(i \in e). \tag{A.13}$$

$$\mathcal{H}_u^{(k)} = (\mathcal{V}, \mathcal{E}_{\mathrm{tr}}^{(k)}, \mathbf{B}^{(k)}, \mathbf{r}_u^{(k)}, \boldsymbol{\eta}_u^{(k)}, \boldsymbol{\zeta}_u^{(k)}, \mathbf{m}_u^{\mathrm{tar},(k)}). \tag{A.14}$$

Candidate identities are frozen within an evaluation fold. Validation/test subjects do not contribute to candidate rankings, selector thresholds, or fitted population transformations. Boundary availability alone does not grant attention, readout, or cross-order propagation support.

A.2.2 Temporal Views and Reliability

The continuous-reliability path uses no surrogate-null calibration. Let $\mathcal{U}_{\mathrm{load}}$ be the subjects whose ROI files are supplied together to fold-feature preparation, and define $n_{\min} = \min_{u \in \mathcal{U}_{\mathrm{load}}}(T_u - 1)$. For internal evaluation, $\mathcal{U}_{\mathrm{load}}$ is the entire retained internal cohort, including its training, validation, and test partitions. The external-evaluation runner combines the retained ABIDE-II source cohort and ABIDE-I external cohort before feature preparation; its $\mathcal{U}_{\mathrm{load}}$ is this union. The common block count is resolved before candidate screening and feature estimation for each fold-preparation call, and is then fixed for all subjects and views in that call. With an unchanged loaded set it is identical across folds; it is not selected separately for each external subject.

A shared block count $B_{\mathrm{time}} = 4$ is used when $n_{\min} \geq 120$, and $B_{\mathrm{time}} = 2$ when $60 \leq n_{\min} < 120$. For $n_{\min} < 60$, feature construction raises an error. For a subject with difference-series length n, block boundaries are $\lfloor bn/B_{\mathrm{time}} \rfloor$ for $b = 0, \ldots, B_{\mathrm{time}}$, with at least 30 observations per block. Differences are formed between consecutive acquired observations before partitioning. The block-count rule uses scan lengths, including held-out lengths, but not diagnostic labels. Adding a shorter held-out scan can therefore change the shared view partition and the resulting reliability and candidate-selection statistics. This rule is distinct from the training-only fitting of candidate rankings and population transformations; it is not a source-training-only choice frozen before external data are loaded.

For a fixed subject and candidate, let v_b be the statistic estimated in block b, with $b = 1, \ldots, B_{\mathrm{time}}$; for higher orders, $v_b = c_{u,e,b}^{(k)}$. Odd and even blocks form views A and B. Let v^A, v^B denote their estimates, m_{blk} the median of all block statistics, MAD their median absolute deviation, d_{rob} their shared scale, and z_A, z_B the standardized view magnitudes:

$$\begin{aligned} m_{\mathrm{blk}} &= \mathrm{median}_b\, v_b, \\ v^A &= \mathrm{median}_{b\ \mathrm{odd}}\, v_b, \\ v^B &= \mathrm{median}_{b\ \mathrm{even}}\, v_b, \\ \mathrm{MAD} &= \mathrm{median}_b\, |v_b - m_{\mathrm{blk}}|, \\ d_{\mathrm{rob}} &= \max(1.4826\,\mathrm{MAD}, 10^{-8}), \\ z_A &= |v^A|/d_{\mathrm{rob}}, \\ z_B &= |v^B|/d_{\mathrm{rob}}, \end{aligned} \tag{A.15}$$

where m_{blk} centers the dispersion estimate only and does not replace the full-scan primary statistic $c_{u,e}^{(k)}$. Neither view numerator is centered. Finite-value medians are used; nonfinite aggregated views are replaced by zero, a nonfinite MAD by zero, and standardized magnitudes by zero for NaN or 20 for positive infinity. Let $\tau_\eta = 2$ be the reliability sensitivity, $c_\eta = 8$ the logit clip, and η the temporal agreement:

$$\eta = \sigma\big(\mathrm{clip}\left[\mathrm{sgn}(v^A)\,\mathrm{sgn}(v^B) z_A z_B/\tau_\eta, -c_\eta, c_\eta\right]\big), \tag{A.16}$$

where nonfinite view inputs receive the neutral value $\eta_{\mathrm{neutral}} = 0.5$; either zero view also gives 0.5. This numerical convention does not replace the insufficient-length error above. The views share candidate coordinates, population transformations, and model parameters, with separately estimated functional attributes.

A.2.3 ROI and Hyperedge Encoding

Let $\mathbf{F}_u^{\mathrm{tr}}$ be the transformed FC input and let $\mathcal{H}_u^{(k)}$ collect the precomputed candidate indices, relation attributes, and masks. The encoder Enc_k returns matrices $\mathbf{H}_u^{(k)} \in \mathbb{R}^{N \times D}$ and $\mathbf{E}_u^{(k)} \in \mathbb{R}^{M_k \times D}$, whose rows are ROI and hyperedge tokens:

$$(\mathbf{H}_u^{(k)}, \mathbf{E}_u^{(k)}) = \mathrm{Enc}_k(\mathbf{F}_u^{\mathrm{tr}}, \mathcal{H}_u^{(k)}). \tag{A.17}$$

The time series $\mathbf{X}_u$ supplies FC and statistical attributes through preprocessing; there is no additional neural encoder acting directly on raw time points in this interface.

FC row encoder and independent higher-order projections. Let LN denote LayerNorm, GELU the activation, and Drop_p dropout with probability p during training and the identity at evaluation. We use token width $D = 32$, FC dropout probability $p_{\mathrm{FC}} = 0.40$, and hidden width $d_{\mathrm{FC}} = 64$. The affine maps (W_1, b_1) and (W_2, b_2) have dimensions $N \to d_{\mathrm{FC}}$ and $d_{\mathrm{FC}} \to D$. For $k \in \{3, 4\}$, let W_k, b_k be order-specific $D \to D$ projection parameters. The common FC encoder and the higher-order ROI projections are

$$\begin{aligned} \mathbf{h}_{u,i}^{(2)} &= \mathrm{LN}_2\big(W_2\,\mathrm{Drop}_{p_{\mathrm{FC}}}\big(\mathrm{GELU}(\mathrm{LN}_1(W_1(\mathbf{F}_{u,i,:}^{\mathrm{tr}})^\top + b_1))\big) + b_2\big)\,, \\ \mathbf{h}_{u,i}^{(k)} &= \mathrm{LN}_k\Big(\mathrm{GELU}(W_k\,\mathrm{sg}(\mathbf{h}_{u,i}^{(2)}) + b_k)\Big)\,, \qquad k \in \{3, 4\}. \end{aligned} \tag{A.18}$$

The row encoder shares parameters across ROIs. The higher-order projections have independent parameters, and their source tokens are detached; higher-order losses therefore do not update the FC row encoder through these projections. The encoder omits ROI-identity embeddings and a separate ROI self-attention module.

Shared set and canonical polyadic content branches. For member $i \in e$, define $\overline{\mathbf{h}}_{u,i}^{(k)} = \mathrm{LN}_{\mathrm{mem}}(\mathbf{h}_{u,i}^{(k)})$. The shared member map ψ_{set} applies $D \to D$ affine, GELU, dropout, and $D \to D$ affine operations. The shared readout ϕ_{set} applies LayerNorm followed by the same affine–GELU–dropout–affine sequence. The set token $\mathbf{t}_{u,e}^{\mathrm{set},(k)} \in \mathbb{R}^D$ is

$$\mathbf{t}_{u,e}^{\mathrm{set},(k)} = \phi_{\mathrm{set}}\left(\frac{1}{\sqrt{k}} \sum_{i \in e} \psi_{\mathrm{set}}(\overline{\mathbf{h}}_{u,i}^{(k)})\right). \tag{A.19}$$

Let $R = 6$ be the canonical polyadic (CP) rank, $\mathbf{o}_k \in \mathbb{R}^{16}$ a learned order embedding, and $U, V \in \mathbb{R}^{R \times D}$ shared factor matrices. Shared affine maps $\mathcal{A}_U, \mathcal{A}_V : \mathbb{R}^{16} \to \mathbb{R}^R$ generate row modulation; diag forms a diagonal matrix. The order-specific factors U_k, V_k are

$$\begin{aligned} U_k &= \mathrm{diag}(\mathbf{1} + 0.1\tanh(\mathcal{A}_U(\mathbf{o}_k)))U, \\ V_k &= \mathrm{diag}(\mathbf{1} + 0.1\tanh(\mathcal{A}_V(\mathbf{o}_k)))V. \end{aligned} \tag{A.20}$$

For rank index r, let $U_{k,r,:}$ be row r of U_k and let $q_{u,i,r}^{(k)}$ denote its bounded member factor. Let $\mathbf{p}_{u,e}^{(k)}$ stack the R stabilized products $p_{u,e,r}^{(k)}$. Define these products and CP token $\mathbf{t}_{u,e}^{\mathrm{CP},(k)}$ by

$$\begin{aligned} q_{u,i,r}^{(k)} &= \tanh(U_{k,r,:}\overline{\mathbf{h}}_{u,i}^{(k)}), \\ p_{u,e,r}^{(k)} &= \left(\prod_{i \in e} \mathrm{sgn}(q_{u,i,r}^{(k)})\right) \exp\left(\mathrm{clip}\left[\sum_{i \in e} \log(|q_{u,i,r}^{(k)}| + 10^{-6}), -20, 10\right]\right), \\ \mathbf{t}_{u,e}^{\mathrm{CP},(k)} &= V_k^\top \mathbf{p}_{u,e}^{(k)}. \end{aligned} \tag{A.21}$$

The product uses the sum of log-magnitudes, without division by k; the sign product preserves zero factors. Logarithms and exponentiation are evaluated in FP32. The optional learned member gate is disabled.

Attribute injection and parameter sharing. For $k \in \mathcal{K}_+$, let $\mathbf{a}_{u,e}^{\mathrm{attr},(k)} = [r_{u,e}^{(k)}, |r_{u,e}^{(k)}|, \eta_{u,e}^{(k)}, \zeta_{u,e}^{(k)}]^\top$. Pairwise tokens use the corresponding unresidualized relation attributes. Let ψ_{attr} be a shared $4 \to D \to D$ affine–GELU–dropout–affine encoder, γ_{CP} a shared unconstrained scalar initialized at zero, and $m_{u,e}^{\mathrm{enc},(k)}$ the computation mask, which may include conditioning-only faces. The hybrid token is

$$\mathbf{e}_{u,e}^{(k)} = m_{u,e}^{\mathrm{enc},(k)}\,\mathrm{LN}_{\mathrm{out}}\Big(\mathbf{t}_{u,e}^{\mathrm{set},(k)} + \gamma_{\mathrm{CP}}\mathbf{t}_{u,e}^{\mathrm{CP},(k)} + \psi_{\mathrm{attr}}(\mathbf{a}_{u,e}^{\mathrm{attr},(k)})\Big). \tag{A.22}$$

The FC encoder, set/CP/attribute encoder, and order hypernetwork are shared across orders. The learned order embeddings and higher-order ROI projections distinguish orders. ROI–hyperedge attention and its feedback projections have separate parameters for each order. Encoder computation masks do not override target eligibility in later attention or readout.

A.3 Structure-Constrained Calibration

A.3.1 Structural Covariate Adjustment and Hyperedge Support

Negative gray-matter voxel values are clipped to zero before ROI averaging. Let $\bar{V}_{u,i}$ be the ROI mean, $p_{u,i}$ its centered log-intensity, and $r^s_{u,i}$ its covariate-adjusted residual. A training-fitted nuisance ridge predictor $\widehat{f}_i$ uses subject age age_u, its square, and sex sex_u:

$$p_{u,i} = \log(1 + \bar{V}_{u,i}) - \frac{1}{N}\sum_j \log(1 + \bar{V}_{u,j}). \tag{A.23}$$

$$r^s_{u,i} = p_{u,i} - \widehat{f}_i(\mathrm{age}_u, \mathrm{age}_u^2, \mathrm{sex}_u). \tag{A.24}$$

The nuisance ridge is fitted only on the designated outer-training partition. For site s, let n_s be its training-subject count, $\mu_{s,i}$ its training residual mean at ROI i, μ_i the corresponding global training mean, and $\kappa = 40$ the shrinkage parameter. Let $s(u)$ identify the site of subject u, σ_i be the global training residual scale, $\varepsilon_s > 0$ its numerical floor, and $c_s = 8$ the clipping bound. With shrinkage weight $w_s = n_s/(n_s + \kappa)$ and center $\mu^*_{s,i} = w_s\mu_{s,i} + (1 - w_s)\mu_i$, the adjusted ROI scalar $s_{u,i}$ is:

$$s_{u,i} = \mathrm{clip}\left((r^s_{u,i} - \mu^*_{s(u),i})/\max(\sigma_i, \varepsilon_s), -c_s, c_s\right).$$

The training-population center is used for unseen sites, and all adjustment parameters are frozen for evaluation. The direct structural representation also includes mean log-intensity and spatial summaries over fixed MNI blocks.

For each higher-order candidate, a shared member multilayer perceptron (MLP) ψ_s encodes the adjusted ROI scalars. Mean and centered root-mean-square (RMS) pooling are invariant to member ordering; the square root below acts elementwise. Write $\bar{\boldsymbol{\psi}}_{u,e} = k^{-1}\sum_{i\in e}\psi_s(s_{u,i})$. Let $\phi^{(k)}_s$ be the order-specific pooling map and $\varepsilon_{\mathrm{pool}} > 0$ the numerical floor. The member mean and RMS dispersion form

$$\mathbf{g}^{(k)}_{u,e} = \phi^{(k)}_s\left(\frac{1}{k}\sum_{i\in e}\psi_s(s_{u,i}) \,\middle\|\, \sqrt{\frac{1}{k}\sum_{i\in e}(\psi_s(s_{u,i}) - \bar{\boldsymbol{\psi}}_{u,e})^{\odot 2} + \varepsilon_{\mathrm{pool}}}\right). \tag{A.25}$$

The sigmoid factor in Eq. 2 defines $\rho^{(k)}_{u,e}$, and target validity gives $a^{(k)}_{u,e} = m^{\mathrm{tar},(k)}_{u,e}\rho^{(k)}_{u,e}$. This quantity expresses multimodal support for a functional candidate. Valid pairwise relations have $a^{(2)}_{u,e} = 1$.

A.3.2 Structural Hypergraph and Spatial Contributions

For $k \in \mathcal{K}_+$, let $\psi_{\mathrm{scalar}} : \mathbb{R} \to \mathbb{R}^D$ encode adjusted ROI intensity, $\mathbf{e}^{\mathrm{ROI}}_i, \mathbf{e}^{\mathrm{order}}_k \in \mathbb{R}^D$ be learned ROI and order embeddings, and $W_{\mathrm{net}} \in \mathbb{R}^{D\times G}$ map the already defined network membership. Let $\mathbf{h}^{s,(k,0)}_{u,i}$ denote the initial structural ROI state. The GM HGNN initializes it by

$$\mathbf{h}^{s,(k,0)}_{u,i} = \mathrm{LN}\left(\psi_{\mathrm{scalar}}(s_{u,i}) + \mathbf{e}^{\mathrm{ROI}}_i + W_{\mathrm{net}}\mathbf{M}^\top_{i,:} + \mathbf{e}^{\mathrm{order}}_k\right).$$

L_{str} order-specific layers perform ROI-to-hyperedge and support-weighted hyperedge-to-ROI updates. Final member mean/RMS summaries give structural edge embeddings. These are fused with functional edge embeddings through their concatenation, product, and absolute difference. The subject embedding combines support-weighted edge and node pooled representations and is multiplied by the mean valid-candidate support strength $S^{(k)}_u \in [0, 1]$.

Structural message updates and readout. Within a layer, write $\mathbf{h}_i$ for its structural ROI input and a_e for candidate support; subject and order indices are suppressed. Let ϕ_E and ϕ_V be the edge and ROI update maps, and b_{layer} a learned scalar gate logit. The ROI-to-hyperedge embedding is $\mathbf{u}_e = \phi_E(k^{-1}\sum_{i\in e}\mathbf{h}_i)$, masked to zero for ineligible targets. Returning messages $\mathbf{v}_i$ and updated ROI states $\mathbf{h}^{\mathrm{next}}_i$ are

$$\mathbf{v}_i = \frac{\sum_{e\ni i} a_e\mathbf{u}_e}{\max(1, \sum_{e\ni i} a_e)}. \tag{A.26}$$

$$\mathbf{h}_i^{\text{next}} = \mathbf{h}_i + \sigma(b_{\text{layer}})\mathbb{I}(\sum_{e \ni i} a_e > 0)\phi_V(\mathbf{v}_i). \tag{A.27}$$

Both maps apply LayerNorm, a bias-free D-to-D linear map, and GELU; ϕ_V additionally applies dropout. Each order has its own layer parameters. The denominator's unit floor retains attenuation when total support is below one. An unsupported ROI receives no update.

Final structural edge embeddings use member mean and centered RMS pooling followed by a $2D$-to-D linear map, GELU, and LayerNorm. Let $\mathbf{t}_e$ be the functional–structural fused edge token and $w_i = \sum_{e \ni i} a_e$. Edge and ROI summaries are $\sum_e a_e \mathbf{t}_e / \max(10^{-6}, \sum_e a_e)$ and $\sum_i w_i \mathbf{h}_i / \max(10^{-6}, \sum_i w_i)$, respectively. Their concatenation passes through a $2D$-to-D linear map, GELU, and LayerNorm, then is multiplied by $S_u^{(k)}$. Relative pooling and final absolute support modulation therefore occur at separate steps. Let $\mathcal{E}_{u,\text{tar}}^{(k)} = \{e \in \mathcal{E}_{\text{tr}}^{(k)} : m_{u,e}^{\text{tar},(k)} = 1\}$. The order-level structural support is

$$S_u^{(k)} = \frac{\sum_{e \in \mathcal{E}_{u,\text{tar}}^{(k)}} a_{u,e}^{(k)}}{\max(1, |\mathcal{E}_{u,\text{tar}}^{(k)}|)}. \tag{A.28}$$

This mean is zero when no valid target relation is present.

The first GM residual uses an order selector over $\mathcal{K}$ with the pairwise state representing abstention from structural higher-order evidence. The selected higher-order structural embeddings produce a clipped scalar log-odds direction multiplied by a bounded functional-conditioned gate and the structural-availability factor. The second path encodes the full N-ROI profile with an MLP and the fixed-MNI block profile with a global MLP and a spatial convolutional network. Their fused profile embedding produces another bounded log-odds residual, whose effective gate includes the combined support S_u^{union}, defined as

$$S_u^{\text{union}} = 1 - (1 - S_u^{(3)})(1 - S_u^{(4)}).$$

The combined support lies in $[0, 1]$ and vanishes when both order-specific supports are zero. The hypergraph and profile corrections, Δ_u^{GM} and $\Delta_u^{\text{profile}}$, enter the base logits through Eq. 6. Each path also accounts for structural availability.

A.4 Evidence-Calibrated ROI–Hyperedge Attention

A.4.1 Sparse Incidence and Head Families

Let $\mathbf{h}_{u,i}^{(k)}$ and $\mathbf{e}_{u,e}^{(k)}$ be the backbone ROI and joint functional candidate embeddings. Learned projections produce $\mathbf{q} = W_Q\mathbf{h}$, $\mathbf{k} = W_K\mathbf{e}$, and $\mathbf{v} = W_V\mathbf{e}$ with $Hd_h = D$. Candidate content includes explicit relation statistics; the projection compares only observed member incidences, not every ROI with every candidate. Using the fixed membership matrix $\mathbf{M}$ defined in Appendix A.1, define local network coherence $\gamma_{i,e}^{(k)}$ by

$$\gamma_{i,e}^{(k)} = \frac{1}{k-1} \sum_{j \in e \setminus \{i\}} \mathbb{I}(\mathbf{M}_{i,:}\mathbf{M}_{j,:}^\top > 0). \tag{A.29}$$

Let the equally sized sets $\mathcal{H}_{\text{loc}}$ and $\mathcal{H}_{\text{br}}$ partition the head-index set, with cardinalities H_{loc} and H_{br} satisfying $H_{\text{loc}} + H_{\text{br}} = H$. Local heads admit $\gamma_{i,e}^{(k)} \geq \gamma_{\text{c}}$ and use $\nu = \gamma$; bridge heads admit the complementary incidences and use $\nu = 1 - \gamma$. The threshold γ_{c} determines the within-network fraction required for local membership. Each domain additionally requires target eligibility and $\eta > 0$; the latter is automatically satisfied by finite sigmoid outputs.

For each subject and query ROI, define the eligible incidence neighborhood

$$\mathcal{I}_{u,i}^{(k)} = \{e \in \mathcal{E}_{u,\text{tar}}^{(k)} : i \in e,\ \eta_{u,e}^{(k)} > 0\}. \tag{A.30}$$

For $h \in \mathcal{H}_{\text{loc}}$, the neighborhood $\mathcal{I}_{u,i,h}^{(k)}$ contains incidences with $\gamma_{i,e}^{(k)} \geq \gamma_{\text{c}}$; for $h \in \mathcal{H}_{\text{br}}$, it contains the complementary subset. The degree used in reference normalization is

$$d_{u,i,h}^{(k)} = \max(1, |\mathcal{I}_{u,i,h}^{(k)}|). \tag{A.31}$$

Let $K_{\text{attn}} = 8$ be the per-ROI, per-head selection budget. The selected set $\mathcal{S}_{u,i}^{(k,h)} \subseteq \mathcal{I}_{u,i,h}^{(k)}$ contains the highest-ranked candidates, with unselected real incidences assigned zero attention, and satisfies $|\mathcal{S}_{u,i}^{(k,h)}| = \min(K_{\text{attn}}, |\mathcal{I}_{u,i,h}^{(k)}|)$.

A.4.2 Evidence-Modulated Attention and Soft Relation Support

Real-candidate scores follow Eq. 3, with nonnegative network and reliability coefficients and unrestricted signed-relation coefficients. During training, ranking scores receive Gumbel perturbations with scale σ_{G}; evaluation uses unperturbed scores. Structural support weights the retained exponential masses, and the reference state does not consume a top-K slot.

Before attention dropout, the head-specific aggregated representation is

$$\mathbf{m}_{u,i}^{(k,h)} = \sum_{e \in \mathcal{S}_{u,i}^{(k,h)}} \alpha_{u,i,e}^{(k,h)} \mathbf{v}_{u,e}^{(k,h)}. \tag{A.32}$$

The reference probability is

$$\alpha_{u,i,\varnothing}^{(k,h)} = 1 - \sum_{e \in \mathcal{S}_{u,i}^{(k,h)}} \alpha_{u,i,e}^{(k,h)}. \tag{A.33}$$

Its value vector is zero, so it contributes no feature vector to the aggregate.

Query-conditioned reference. Within each order k, let $d_{u,i}^{\mathrm{ctx},(k)}$ count valid, positive-reliability incidences of ROI i before the local/bridge split. The context vector comprises reliability, unsupported mass, reliability dispersion, and inverse-degree scaling:

$$\boldsymbol{\xi}_{u,i}^{(k)} = \left[\overline{\eta}_{u,i},\ 1 - \overline{a}_{u,i},\ \sqrt{\max(\overline{\eta^2}_{u,i} - \overline{\eta}_{u,i}^2, 0) + \varepsilon_{\mathrm{ctx}}},\ \max(d_{u,i}^{\mathrm{ctx},(k)}, 1)^{-1/2} \right].$$

Overbars denote means over $\mathcal{I}_{u,i}^{(k)}$ before the local/bridge split, with $d_{u,i}^{\mathrm{ctx},(k)} = |\mathcal{I}_{u,i}^{(k)}|$. An empty neighborhood uses $\boldsymbol{\xi}_{u,i}^{(k)} = \mathbf{1}$.

Let $b_h^{\varnothing} \in \mathbb{R}$ be a learned head-specific reference bias, $\mathbf{k}_h^{\varnothing} \in \mathbb{R}^{d_h}$ a learned reference key, and $f_{\varnothing} : \mathbb{R}^4 \to \mathbb{R}^H$ an affine–GELU–affine context map with hidden width 16. The vector $\boldsymbol{\xi}_{u,i}^{(k)}$ above has four coordinates, and $c_{\varnothing} = 4$ bounds the context correction. The reference score is

$$\ell_{u,i,\varnothing}^{(k,h)} = b_h^{\varnothing} + \frac{(\mathbf{q}_{u,i}^{(k,h)})^{\top} \mathbf{k}_h^{\varnothing}}{\sqrt{d_h}} + \left[c_{\varnothing} \tanh(f_{\varnothing}(\mathrm{sg}(\boldsymbol{\xi}_{u,i}^{(k)}))/c_{\varnothing}) \right]_h. \tag{A.34}$$

Here sg is the stop-gradient operator. The reference mass $d_{u,i,h}^{(k)} \exp(\ell_{u,i,\varnothing}^{(k,h)})$ and real masses $a_{u,e}^{(k)} \exp(\ell_{u,i,e}^{(k,h)}/\tau_a)$ are jointly normalized by Eq. 4. Empty neighborhoods assign unit probability to the reference state, whose value vector is zero. With the selected set and all scores fixed, attention increases with candidate support. During training, attention dropout is applied before value aggregation. Concatenated head messages are projected and added to the ROI representation, with zero feedback for unsupported ROIs. Order-level norm adjustment is applied separately before base fusion (Eq. A.48).

ROI importance. Let $\widehat{\mathbf{m}}_{u,i}^{(k,h)}$ denote the aggregated head message used to form ROI feedback, including attention dropout during training and before output projection. The nonnegative ROI-importance mass $q_{u,i}^{(k)}$ is

$$q_{u,i}^{(k)} = \frac{1}{H} \sum_{h=1}^{H} \left\| \widehat{\mathbf{m}}_{u,i}^{(k,h)} \right\|_2. \tag{A.35}$$

Let $\iota_{u,i}^{(k)}$ denote the normalized importance over ROIs within subject u and order k. With a numerical floor of 10^{-12},

$$\iota_{u,i}^{(k)} = \frac{q_{u,i}^{(k)}}{\max(10^{-12}, \sum_{j=1}^{N} q_{u,j}^{(k)})}. \tag{A.36}$$

An all-zero mass vector yields all-zero importance. This measure depends on both attention weights and value vectors through the aggregated messages; it is not a sum of attention probabilities. It is distinct from the within-network hub distribution derived from member responsibilities. Temporal comparisons apply the normalization and valid-mass rule in Appendix A.7.

Computational complexity. Unless stated otherwise, operation counts are given per subject and layer. For given embeddings, scoring and aggregation over actual incidences require $O(ID)$ operations, where $I = \sum_k kM_k$. Linear feature projections cost $O((|\mathcal{K}|N+M)D^2)$, with $M = \sum_k M_k$; for a fixed number of orders, this reduces to $O((N+M)D^2)$. For fixed head count and selection budget, top-K selection can be performed in time linear in the number of candidate scores inspected.

Padding variable-degree neighborhoods to the maximum degree introduces $P = \sum_k N d_{\max}^{(k)}$ slots, where $d_{\max}^{(k)}$ is the maximum order-k neighborhood size. If padded slots are evaluated before masking, interaction computation includes an $O(PD)$ term. For a batch of B_{batch} subjects, attention-score storage is bounded by $O(B_{\text{batch}}H(I+P))$.

Adjacent-order propagation contains at most $J_{\max} = \sum_{k\in\mathcal{K}_+} kM_k$ face–coface pairs and requires $O(J_{\max}D)$ interaction operations plus $O(MD^2)$ projections per layer. Across L_{cross} layers, these costs scale by L_{cross}. Candidate estimation, boundary regression, and image encoding are accounted for separately. For comparison, dense attention among S representations requires $O(S^2D)$ interaction operations.

A.5 Cross-Order Propagation and Base Readout

Only available direct faces participate in cross-order propagation: $C_{f,e}^{(k)} = \mathbb{I}(f \subset e)$ for $|f| = k-1$, $k \in \mathcal{K}_+$. Direction-specific Q/K/V projections produce separately normalized upward and downward attention. Source and receiver support weight real face–coface masses, with zero messages for receivers without a positive-mass neighborhood. For a receiving relation e, let $\mathcal{N}(e)$ contain available adjacent-order sources and let $s_{e,f}$ denote a direction-specific, head-specific compatibility score. The normalized source weight is

$$\omega_{e\leftarrow f} = \frac{a_e a_f \exp(s_{e,f})}{\sum_{f'\in\mathcal{N}(e)} a_e a_{f'} \exp(s_{e,f'})}. \tag{A.37}$$

This formula applies only when the denominator is positive; otherwise every incoming message is zero. For $a_e > 0$, its common factor cancels, so positive receiver support alone does not attenuate the normalized message magnitude. Absolute support attenuation is governed separately by incidence attention and structural weighting. Source embeddings are aggregated using these weights and direction-specific value projections; upward and downward neighborhoods are normalized separately. For layer ℓ, let $\mathbf{E}_{u,\ell}^{(k)}$ be the edge-state matrix and let $\mathbf{M}_{\text{up},u,\ell}^{(k)}$ and $\mathbf{M}_{\text{down},u,\ell}^{(k)}$ be face and coface messages, each of shape $M_k \times D$. Learned coefficients $\lambda_{\text{up},\ell}^{(k)}$ and $\lambda_{\text{down},\ell}^{(k+1)}$ scale the updates. The main-text update in Eq. 5 is

$$\mathbf{E}_{u,\ell+1}^{(k)} = \mathbf{E}_{u,\ell}^{(k)} + \mathbb{I}(k>2)\lambda_{\text{up},\ell}^{(k)}\mathbf{M}_{\text{up},u,\ell}^{(k)} + \mathbb{I}(k<4)\lambda_{\text{down},\ell}^{(k+1)}\mathbf{M}_{\text{down},u,\ell}^{(k)}. \tag{A.38}$$

Each of the H_{cross} heads has width D/H_{cross}. Propagation uses consecutive orders, whereas statistical boundary regression can condition on all lower-order faces.

A.5.1 Hierarchical Readout and Base Prediction

Let $\widetilde{\mathbf{E}}_u^{(k)}$ contain final edge states $\widetilde{\mathbf{e}}_{u,e}^{(k)}$, and let $\mathbf{H}_{u,\text{att}}^{(k)}$ contain attention-updated ROI states. Vectors $\mathbf{m}_u^{\text{tar},(k)}$ and $\mathbf{a}_u^{(k)}$ collect target masks and support weights. The functional order embedding $\mathbf{f}_u^{(k)} \in \mathbb{R}^D$ is

$$\mathbf{f}_u^{(k)} = \text{Read}_k(\mathbf{H}_{u,\text{att}}^{(k)}, \widetilde{\mathbf{E}}_u^{(k)}, \mathbf{m}_u^{\text{tar},(k)}, \mathbf{a}_u^{(k)}). \tag{A.39}$$

This is a summary interface: it suppresses the original ROI states, relation attributes, membership indices, and retained member-effect scores used by the detailed operations below. Following cross-order exchange, the affected order readouts and incidence messages are recomputed from the updated edge tokens.

Masked normalizations. For score vector $\mathbf{x}$, eligible coordinate set $\mathcal{I}$, temperature $\tau > 0$, and normalization choice m, write $\mathscr{P}_{\mathcal{I}}^m(\mathbf{x};\tau)$ for normalization over $\mathcal{I}$, with zero at other coordinates and an all-zero result for an empty set. Softmax uses $\exp(x_j/\tau)$ divided by its sum. Entmax with parameter 1.5 uses $[x_j/(2\tau) - \theta]_+^2$, with threshold θ chosen so the eligible probabilities sum to one. The hierarchy uses entmax-1.5, denoted below by $m = \text{ent}$.

Hyperedge-to-network-to-subject aggregation. Let $g \in \{1, \dots, G\}$ index atlas networks and let M_{ig} be the binary ROI membership indicator defined in Appendix A.1. Define member fraction $\mu_{g,e} = k^{-1} \sum_{i \in e} M_{ig}$ and eligible edge set $\mathcal{I}_{u,g}^{(k)} = \{e : m_{u,e}^{\mathrm{tar},(k)} = 1, \mu_{g,e} > 0\}$. Shared learned maps produce edge and order queries $\mathbf{q}_k^E, \mathbf{q}_k^O \in \mathbb{R}^D$ from $\mathbf{o}_k$; $\mathbf{p}_g \in \mathbb{R}^D$ is a learned network prototype. Let B_E be a $(D+6) \to D \to 1$ affine–GELU–dropout–affine scorer, with two zero topology coordinates because topology features are disabled. Its base edge score $b_{u,e}^{(k)}$ and network score $s_{u,g,e}^{(k)}$ are

$$\begin{aligned} b_{u,e}^{(k)} &= B_E(\widetilde{\mathbf{e}}_{u,e}^{(k)} \| \mathbf{a}_{u,e}^{\mathrm{attr},(k)} \| \mathbf{0}_2) + (\widetilde{\mathbf{e}}_{u,e}^{(k)})^\top \mathbf{q}_k^E / \sqrt{D}, \\ s_{u,g,e}^{(k)} &= b_{u,e}^{(k)} + (\widetilde{\mathbf{e}}_{u,e}^{(k)})^\top \mathbf{p}_g / \sqrt{D} + \log \max(\mu_{g,e}, 10^{-8}). \end{aligned} \tag{A.40}$$

Let $\mathbf{s}_{u,g}^{(k)}$ stack the edge scores $s_{u,g,e}^{(k)}$, and let τ_E be the hierarchy temperature, $\bar{p}_{u,g,e}^{(k)}$ the edge-normalized probability before structural reweighting, $p_{u,g,e}^{(k)}$ its support-weighted probability, and $\mathbf{n}_{u,g}^{(k)}$ the network embedding:

$$\begin{aligned} \bar{\mathbf{p}}_{u,g}^{(k)} &= \mathscr{P}^{\mathrm{ent}}_{\mathcal{I}_{u,g}^{(k)}}(\mathbf{s}_{u,g}^{(k)}; \tau_E), \\ p_{u,g,e}^{(k)} &= \frac{\bar{p}_{u,g,e}^{(k)} a_{u,e}^{(k)}}{\sum_{e'} \bar{p}_{u,g,e'}^{(k)} a_{u,e'}^{(k)}}, \\ \mathbf{n}_{u,g}^{(k)} &= \sum_e p_{u,g,e}^{(k)} \mu_{g,e} \widetilde{\mathbf{e}}_{u,e}^{(k)}. \end{aligned} \tag{A.41}$$

When the reweighting denominator is zero, all probabilities in that row are zero. The membership fraction enters both the score and the pooled vector; it is not removed by a second normalization of the pooled vector. Let B_G be a shared $2D \to D \to 1$ affine–GELU–affine scorer, and let $\mathcal{G}_u^{(k)}$ contain networks with positive real-edge attention. The network weights $\vartheta_{u,g}^{(k)}$ and edge-based subject representation $\mathbf{f}_{u,E}^{(k)}$ are

$$\begin{aligned} \boldsymbol{\vartheta}_u^{(k)} &= \mathscr{P}^{\mathrm{ent}}_{\mathcal{G}_u^{(k)}}\left((B_G(\mathbf{n}_{u,g}^{(k)} \| \mathbf{q}_k^O))_{g=1}^G; \tau_E \right), \\ \mathbf{f}_{u,E}^{(k)} &= \sum_g \vartheta_{u,g}^{(k)} \mathbf{n}_{u,g}^{(k)}. \end{aligned} \tag{A.42}$$

The hierarchy temperature is $\tau_E = 1.25$. The scoring maps and network prototypes are shared across orders; order-dependent queries come from the shared order hypernetwork. The optional second-pass hub bias is disabled.

Member responsibility and ROI feedback. At the encoder stage, replace member i by the hypernetwork's learned neutral token and re-encode the edge. Let $d_{u,e,i}^{\mathrm{mem},(k)}$ be the absolute difference between its original base edge score and this neutralized score; these member effects are retained when the hierarchy is recomputed after cross-order propagation. Normalize them over the k members to obtain responsibility $\beta_{u,e,i}^{(k)} = \mathscr{P}_e^{\mathrm{ent}}((d_{u,e,j}^{\mathrm{mem},(k)})_{j \in e}; 1)_i$. Let $w_{u,g,i}^{(k)}$ be unnormalized hub mass and $b_{u,g,i}^{(k)}$ the within-network hub weight:

$$\begin{aligned} w_{u,g,i}^{(k)} &= M_{ig} \sum_{e \ni i} p_{u,g,e}^{(k)} \beta_{u,e,i}^{(k)} \eta_{u,e}^{(k)} m_{u,e}^{\mathrm{tar},(k)}, \\ b_{u,g,i}^{(k)} &= \frac{w_{u,g,i}^{(k)}}{\max(10^{-12}, \sum_j w_{u,g,j}^{(k)})}. \end{aligned} \tag{A.43}$$

Define ROI update $\Delta \mathbf{h}_{u,i}^{(k)} = \mathbf{h}_{u,i,\mathrm{att}}^{(k)} - \mathbf{h}_{u,i}^{(k)}$. Let P_k be an order-specific bias-free $D \to D$ linear map, GELU, and bias-free $D \to D$ linear map. The sparse order readout $\mathbf{f}_{u,\mathrm{sparse}}^{(k)}$ is

$$\mathbf{f}_{u,\mathrm{sparse}}^{(k)} = \mathbf{f}_{u,E}^{(k)} + P_k \left(\sum_g \vartheta_{u,g}^{(k)} \sum_i b_{u,g,i}^{(k)} \Delta \mathbf{h}_{u,i}^{(k)} \right). \tag{A.44}$$

Thus feedback uses the attention-induced change, rather than pooling the entire ROI state a second time. Higher orders use $\mathbf{f}_u^{(k)} = \mathbf{f}_{u,\mathrm{sparse}}^{(k)}$.

Dense and low-rank pairwise anchor. Two learned affine scalar scorers B_R, B_2 pool the original pairwise ROI tokens. With $\mathcal{V}_g = \{i : M_{ig} = 1\}$, define dense ROI weights $b^{\text{dense}}_{u,g,i}$, network vectors $\mathbf{n}^{\text{dense}}_{u,g}$, network weights $\vartheta^{\text{dense}}_{u,g}$, and dense feature $\mathbf{f}^{\text{dense}}_u$ by

$$\begin{aligned}
\mathbf{b}^{\text{dense}}_{u,g} &= \mathscr{P}^{\text{ent}}_{\mathcal{V}_g}((B_R(\mathbf{h}^{(2)}_{u,i}))_i; 1),\\
\mathbf{n}^{\text{dense}}_{u,g} &= \sum_i b^{\text{dense}}_{u,g,i}\mathbf{h}^{(2)}_{u,i},\\
\boldsymbol{\vartheta}^{\text{dense}}_u &= \mathscr{P}^{\text{ent}}_{\{1,\dots,G\}}((B_2(\mathbf{n}^{\text{dense}}_{u,g}))_g; 1),\\
\mathbf{f}^{\text{dense}}_u &= \sum_g \vartheta^{\text{dense}}_{u,g}\mathbf{n}^{\text{dense}}_{u,g}.
\end{aligned} \tag{A.45}$$

For the low-rank branch, let $\mathbf{x}^{\text{FC}}_u$ be the strict upper triangle of the harmonized FC matrix, with negative entries clipped to zero. Let $\boldsymbol{\mu}_F, \boldsymbol{\sigma}_F$ be training-edge means and population standard deviations; replace nonfinite or at-most-10^{-6} scales by one. Fit principal component analysis (PCA) on these standardized training vectors only. Let $\boldsymbol{\mu}_P$, W_P, and $\boldsymbol{\lambda}_P$ denote the fitted PCA center, component rows, and explained variances. Division and square root below are elementwise. The whitened score vector $\mathbf{v}^{\text{PCA}}_u$ is

$$\mathbf{v}^{\text{PCA}}_u = \frac{W_P((\mathbf{x}^{\text{FC}}_u - \boldsymbol{\mu}_F)/\boldsymbol{\sigma}_F - \boldsymbol{\mu}_P)}{\sqrt{\max(\boldsymbol{\lambda}_P, 10^{-6})}}. \tag{A.46}$$

The component count is $\max(2, \min(d_P, |\mathcal{U}_{\text{tr}}| - 1, N(N-1)/2))$, with requested dimension $d_P = 32$. All transformations remain frozen for validation/testing; no additional score scaling is applied. Let E_P be the low-rank token encoder, b_P a learned scalar logit, and $g_P = \sigma(b_P)$ its sigmoid gate initialized at 0.1. Let $d_{P,\text{hid}}$ and p_P be its hidden width and dropout probability: affine–LayerNorm–GELU–dropout maps the PCA input to this hidden width, followed by affine–LayerNorm to output width D. The MLP uses hidden width $d_{P,\text{hid}} = 32$, output width $D = 32$, and dropout $p_P = 0.35$. The pairwise feature is

$$\mathbf{f}^{(2)}_u = \mathbf{f}^{\text{dense}}_u + g_P E_P(\mathbf{v}^{\text{PCA}}_u) + \mathbf{f}^{(2)}_{u,\text{sparse}}. \tag{A.47}$$

Order-level norm adjustment and base routing. For $k \in \{3, 4\}$, define detached scale ratio $s^{(k)}_u$ and adjusted feature $\widetilde{\mathbf{f}}^{(k)}_u$; this scalar $s^{(k)}_u$ is distinct from structural vector $\mathbf{s}^{(k)}_u$. With lower bound $s_{\min} = 0.25$, upper bound $s_{\max} = 3$, and $\varepsilon_{\text{norm}} = 10^{-6}$,

$$\begin{aligned}
s^{(k)}_u &= \text{clip}\left(\frac{\text{sg}(\|\mathbf{f}^{(2)}_u\|_2)}{\max(\text{sg}(\|\mathbf{f}^{(k)}_u\|_2), \varepsilon_{\text{norm}})}, s_{\min}, s_{\max}\right),\\
\widetilde{\mathbf{f}}^{(k)}_u &= s^{(k)}_u \mathbf{f}^{(k)}_u.
\end{aligned} \tag{A.48}$$

Set $\widetilde{\mathbf{f}}^{(2)}_u = \mathbf{f}^{(2)}_u$. Let $Q^{(k)}_u$ be mean temporal reliability over target-eligible candidates for higher orders (Eq. A.52), and set $Q^{(2)}_u = 1$. The effective router domain is $\mathcal{K}^{\text{valid}}_u = \{k \in \mathcal{K} : Q^{(k)}_u > 0\}$. Let $\mathbf{v}_u$ stack the order scores $v^{(k)}_u$. Let B_O be the shared $2D \to D \to 1$ affine–GELU–dropout–affine scorer and $\chi_Q \geq 0$ the quality-logit coefficient, with $\chi_Q = 0$. The router uses $m_{\text{route}} = \text{ent}$ (entmax-1.5). Router scores $v^{(k)}_u$ and probabilities $\omega^{(k)}_u$ are

$$\begin{aligned}
v^{(k)}_u &= B_O(\widetilde{\mathbf{f}}^{(k)}_u \| \mathbf{q}^O_k) + \chi_Q \log\max(Q^{(k)}_u, 10^{-8}),\\
\boldsymbol{\omega}_u &= \mathscr{P}^{m_{\text{route}}}_{\mathcal{K}^{\text{valid}}_u}(\mathbf{v}_u; 1).
\end{aligned} \tag{A.49}$$

For zero-based training epoch t and planned epoch count T_{train}, let $p_t = \text{clip}(t/\max(T_{\text{train}} - 1, 1), 0, 1)$. The deterministic curriculum multiplier is $\lambda_{\text{curr}} = 0.05 + 0.95\,\text{clip}((p_t - 0.10)/0.30, 0, 1)$; evaluation uses its saved checkpoint value. The identity-preserving feature and actual branch coefficients $c^{(k)}_u$ are

$$\begin{aligned}
\mathbf{z}^{\text{base}}_u &= \mathbf{f}^{(2)}_u + \lambda_{\text{curr}} \sum_{k=3,4} \omega^{(k)}_u \widetilde{\mathbf{f}}^{(k)}_u,\\
c^{(2)}_u &= 1,\\
c^{(k)}_u &= \lambda_{\text{curr}}\omega^{(k)}_u \ (k = 3, 4).
\end{aligned} \tag{A.50}$$

Additional reliability multipliers and nested order gates are disabled in this path. The realized contribution vectors used by the budget loss are $\mathbf{d}_u^{(2)} = \mathbf{f}_u^{(2)}$ and $\mathbf{d}_u^{(k)} = c_u^{(k)}\widetilde{\mathbf{f}}_u^{(k)}$ for $k = 3, 4$. The router loss acts on $\omega_u^{(k)}$, not on $c_u^{(k)}$.

Let $d_{\mathrm{cls}} = 32$ be the classifier hidden width and $p_{\mathrm{cls}} = 0.45$ its dropout probability. Let C_{cls} apply $D \to d_{\mathrm{cls}}$ affine, LayerNorm, GELU, dropout with probability p_{cls}, and $d_{\mathrm{cls}} \to 2$ affine operations; the functional logits are $\boldsymbol{\ell}_u^{\mathrm{func}} = C_{\mathrm{cls}}(\mathbf{z}_u^{\mathrm{base}})$. With auxiliary quantities $\mathcal{A}_u$ collecting the probabilities, realized contributions, and diagnostics above, the summarized interface is

$$(\mathbf{z}_u^{\mathrm{base}}, \boldsymbol{\ell}_u^{\mathrm{func}}, \mathcal{A}_u) = \mathrm{Fuse}_{\mathrm{base}}(\{\mathbf{f}_u^{(k)}\}_{k\in\mathcal{K}}, \mathbf{F}_u^{\mathrm{tr}}). \tag{A.51}$$

Here $D_z = D$. Structural corrections then form the base logits in Eq. 6; conditional admission adds the separate bounded correction in Eq. 9.

A.6 Conditional Innovation Admission

A.6.1 Lower-Order-Conditioned Cross-Modal Representations

Let $\mathbf{f}_u^{(k)}$ be the functional order embedding and $\mathbf{s}_u^{(k)}$ the GM HGNN subject embedding. For equal-width vectors $\mathbf{x}, \mathbf{y}$, define $\mathcal{C}(\mathbf{x}, \mathbf{y}) = \mathrm{Concat}(\mathbf{x}, \mathbf{y}, \mathbf{x} \odot \mathbf{y}, |\mathbf{x} - \mathbf{y}|)$. Let F_k, L_k, J_k be learned maps producing the cross-modal embedding $\mathbf{c}_u^{(k)}$, lower-order context $\mathbf{l}_u^{(k)}$, and contextual innovation $\mathbf{j}_u^{(k)}$, respectively. Then

$$\begin{aligned}
\mathbf{c}_u^{(k)} &= F_k(\mathcal{C}(\mathbf{f}_u^{(k)}, \mathbf{s}_u^{(k)})).\\
\mathbf{l}_u^{(3)} &= L_3(\mathbf{f}_u^{(2)}).\\
\mathbf{l}_u^{(4)} &= L_4(\mathbf{f}_u^{(2)} \| \mathbf{f}_u^{(3)}).\\
\mathbf{j}_u^{(k)} &= J_k(\mathcal{C}(\mathbf{c}_u^{(k)}, \mathbf{l}_u^{(k)})).
\end{aligned}$$

The learned contextual representation $\mathbf{j}_u^{(k)}$ is distinct from the statistical residual $r_{u,e}^{(k)}$.

A.6.2 Admission Evidence and Averaging Domains

For the fixed candidate bank, define order quality $Q_u^{(k)}$ as mean temporal reliability over target-eligible relations:

$$Q_u^{(k)} = \frac{\sum_{e\in\mathcal{E}_{\mathrm{tr}}^{(k)}} m_{u,e}^{\mathrm{tar},(k)} \eta_{u,e}^{(k)}}{\max(1, \sum_{e\in\mathcal{E}_{\mathrm{tr}}^{(k)}} m_{u,e}^{\mathrm{tar},(k)})}. \tag{A.52}$$

Thus $Q_u^{(k)} \in [0, 1]$, with zero for an empty target set. It summarizes temporal evidence, whereas $S_u^{(k)}$ summarizes learned multimodal support. The mean reference probability $n_u^{(k)}$ averages equally over all ROI–head coordinates within a subject:

$$n_u^{(k)} = \frac{1}{NH} \sum_{i=1}^{N} \sum_{h=1}^{H} \alpha_{u,i,\varnothing}^{(k,h)}. \tag{A.53}$$

Empty incidence neighborhoods are included with reference probability one. Neither average pools across subjects or weights an ROI by its incidence degree. For order-specific support $S_u^{(k)}$, quality $Q_u^{(k)}$, and mean reference probability $n_u^{(k)}$, let G_k be a learned scalar map, $t_u^{(k)}$ the admission logit, and $A_u^{(k)}$ its sigmoid score. Nonnegative coefficients $\beta_S, \beta_Q, \beta_n$ weight support, quality, and reference evidence. For scalar x and $0 < \varepsilon < 1/2$, define $\widetilde{x} = \mathrm{clip}(x, \varepsilon, 1 - \varepsilon)$ and $\mathrm{logit}_\varepsilon(x) = \log(\widetilde{x}/(1 - \widetilde{x}))$. Using floor $\varepsilon_{\mathrm{logit}}$, admission is

$$t_u^{(k)} = G_k(\mathbf{j}_u^{(k)} \| \,\mathrm{sg}(\mathbf{l}_u^{(k)})) + \beta_S \,\mathrm{logit}_{\varepsilon_{\mathrm{logit}}} S_u^{(k)} + \beta_Q \,\mathrm{logit}_{\varepsilon_{\mathrm{logit}}} Q_u^{(k)} - \beta_n \,\mathrm{logit}_{\varepsilon_{\mathrm{logit}}} n_u^{(k)}. \tag{A.54}$$

$$A_u^{(k)} = \sigma(t_u^{(k)}). \tag{A.55}$$

The allocation in Eq. 7 sums to one and assigns total higher-order mass $A_u^{(3)}$, with the maximal-order share determined conditionally by $A_u^{(4)}$. These soft weights control the additional innovation branch.

Gradient paths. Support and reference evidence remain differentiable in the admission score; quality is computed from fixed relation attributes. The stopped lower-order input blocks only its direct path into G_k: lower-order context still participates in $\mathbf{j}_u^{(k)}$. Likewise, stopping the base representation in the residual gate does not freeze the base classifier. The boundary predictor and population transformations remain fixed during classifier training.

A.6.3 Bounded Diagnostic Correction and Base Recovery

Let F_{HA} be the innovation fusion map and G_{res} the scalar residual-gate map. Define $\mathbf{w}_3 = \pi_u^{(3)}\mathbf{j}_u^{(3)}$ and $\mathbf{w}_4 = \pi_u^{(4)}\mathbf{j}_u^{(4)}$. Their joint representation is $\mathbf{h}_u = F_{\mathrm{HA}}(\mathcal{C}(\mathbf{w}_3, \mathbf{w}_4))$. Let D_{HA} be the width of $\mathbf{h}_u$, $\mathbf{w}_\Delta \in \mathbb{R}^{D_{\mathrm{HA}}}$ the learned direction, $g_{\max} > 0$ the gate bound, and $c_\Delta > 0$ the saturation scale. Conditioned on $\mathbf{z}_u^{\mathrm{base}}$, the gate g_u and correction Δ_u^{HA} in Eq. 8 are

$$g_u = g_{\max}\sigma(G_{\mathrm{res}}(\mathrm{sg}(\mathbf{z}_u^{\mathrm{base}})\|\mathbf{h}_u)).$$

$$\Delta_u^{\mathrm{HA}} = g_u c_\Delta \tanh(\mathbf{w}_\Delta^\top \mathbf{h}_u / c_\Delta).$$

Since $0 \le g_u \le g_{\max}$ and $|\tanh x| \le 1$, the correction satisfies $|\Delta_u^{\mathrm{HA}}| \le g_{\max} c_\Delta$. Substitution into Eq. 9 yields the diagnostic posterior.

For class ordering control then patient, subtraction of the two logits gives

$$\log \frac{\widehat{p}_{u,1}}{\widehat{p}_{u,0}} = (\ell_{u,1}^{\mathrm{base}} - \ell_{u,0}^{\mathrm{base}}) + \Delta_u^{\mathrm{HA}}. \tag{A.56}$$

Zero initialization of the residual direction preserves the base prediction initially. After training, zero weighted innovations give exact fallback when the fusion is zero-preserving, $F_{\mathrm{HA}}(\mathbf{0}) = \mathbf{0}$; zero initialization alone does not ensure this property. The correction bound applies to the additional admission branch, and the base representation already contains multi-order information.

A.7 Learning Objective

The main text describes binary cross-entropy and displays the structural regularizer (Eq. 10), temporal regularizer (Eq. 11), and total objective (Eq. 12). Below, we specify their component definitions and averaging domains.

A.7.1 Classification Loss

The classification loss $\mathcal{L}_{\mathrm{cls}}$ is binary cross-entropy.

Distribution normalization. Let $\mathcal{U}_B$ denote a nonempty subject minibatch for the regularizers below. Let $\varepsilon_{\mathrm{mass}} > 0$ be a numerical threshold. For a nonnegative-mass instance $\mathbf{x}$, write $s(\mathbf{x}) = \sum_j [x_j]_+$ and $\mathcal{N}(\mathbf{x})_j = [x_j]_+ / s(\mathbf{x})$ whenever $s(\mathbf{x}) > \varepsilon_{\mathrm{mass}}$. Comparisons use only observations whose total mass exceeds $\varepsilon_{\mathrm{mass}}$ in both arguments. For normalized distributions P, Q on the same frozen coordinate domain, let $\overline{P} = (P + Q)/2$ and let KL denote Kullback–Leibler divergence. Their Jensen–Shannon divergence is

$$\mathrm{JS}(P\|Q) = \tfrac{1}{2}\,\mathrm{KL}(P\|\overline{P}) + \tfrac{1}{2}\,\mathrm{KL}(Q\|\overline{P}). \tag{A.57}$$

Missing coordinates carry zero mass, with $0 \log 0 = 0$. Comparisons without valid observations return zero. Natural logarithms are used.

A.7.2 Relational Structural Regularization

For scalar x and prescribed interval $[b_-, b_+]$, define $\mathcal{B}(x; b_-, b_+) = [b_- - x]_+^2 + [x - b_+]_+^2$. Edge and incidence entropies use different support counts and validity conditions.

For a subject–network instance, let $P_e = p_{u,g,e}^{(k)}$ be the support-reweighted hyperedge distribution and let $n_+ = |\{e : P_e > 0\}|$ count its positive probabilities. Its normalized edge entropy is

$$H_{u,g}^{E,(k)} = -\frac{\sum_{e:P_e>0} P_e \log(P_e + 10^{-8})}{\log \max(n_+, 2)}. \tag{A.58}$$

Instances with $n_+ > 0$ enter the band penalty. A singleton distribution therefore participates; with a positive entropy lower bound it incurs a penalty. Empty instances are excluded.

For an ROI–head instance, let P_j^I be the joint normalized probabilities of selected real incidences and the reference state, and let n_{real} count selected, eligible real incidences with positive support weight. The reference contributes one entropy coordinate but is not counted in n_{real}. The normalized incidence entropy is

$$H_{u,i,h}^{I,(k)} = -\frac{\sum_{j:P_j^I>0} P_j^I \log\max(P_j^I, 10^{-12})}{\log\max(n_{\text{real}}+1, 2)}. \tag{A.59}$$

Only instances with $n_{\text{real}} > 1$ enter its band penalty. This support count is determined from selected eligible incidences and positive gates, not from the number of floating-point probabilities remaining positive after normalization.

For each order, average edge penalties over valid subject–network instances and incidence penalties over valid subject–ROI–head instances. Let $\langle\cdot\rangle_E$ and $\langle\cdot\rangle_I$ then denote equal averages of these order-level means over orders with at least one valid instance in the respective group; an empty group returns zero. With entropy intervals $[b_-^E, b_+^E]$ and $[b_-^I, b_+^I]$, the attention penalty is

$$\mathcal{L}_{\text{attn}} = \frac{1}{2}\left(\left\langle \mathcal{B}(H^E; b_-^E, b_+^E)\right\rangle_E + \left\langle \mathcal{B}(H^I; b_-^I, b_+^I)\right\rangle_I\right). \tag{A.60}$$

The entropy interval is $[0.15, 0.75]$ for both groups. The reference participates in incidence entropy; it is excluded from the real-incidence distributions used by the temporal and responsibility JS losses.

Let $\mathbf{d}_u^{(k)}$ denote the actual vector contributed by order k to base fusion, including its effective coefficient and any norm adjustment. With numerical floor $\varepsilon_R = 10^{-8}$, define realized contribution ratio $R_u^{(k)}$ by

$$R_u^{(k)} = \frac{\|\mathbf{d}_u^{(k)}\|_2}{\max(\varepsilon_R, \sum_{j\in\mathcal{K}} \|\mathbf{d}_u^{(j)}\|_2)}. \tag{A.61}$$

The denominator sums separate branch norms, so cancellation between branch vectors does not reduce it. Each ratio lies in $[0, 1]$; all ratios are zero when every contribution vanishes. These quantities remain differentiable through the realized contributions. For prescribed higher-order contribution cap $R_{\max}$, the contribution-budget penalty $\mathcal{L}_{\text{budget}}$ is

$$\mathcal{L}_{\text{budget}} = \frac{1}{|\mathcal{U}_B|}\sum_{u\in\mathcal{U}_B}\left[\sum_{k\in\mathcal{K}_+} R_u^{(k)} - R_{\max}\right]_+^2. \tag{A.62}$$

Here $R_u^{(k)}$ measures the realized contribution of order k to base fusion after order-level norm adjustment, distinct from the additional admission weights. With $R_{\max} = 1$, the higher-order ratios sum to at most one, giving $\mathcal{L}_{\text{budget}} = 0$.

For the identity-preserving fusion in Eq. A.50, let $\omega_u^{(k)}$ be the router probability and $c_u^{(k)}$ the actual fusion coefficient. These differ: $c_u^{(2)} = 1$ and $c_u^{(k)} = \lambda_{\text{curr}}\omega_u^{(k)}$ for $k = 3, 4$. The router loss is evaluated on probabilities rather than actual fusion coefficients. Define batch-mean router mass $\bar{\omega}^{(k)} = |\mathcal{U}_B|^{-1}\sum_{u\in\mathcal{U}_B}\omega_u^{(k)}$ and normalized entropy $H_u^{\text{route}} = -\sum_{k\in\mathcal{K}}\omega_u^{(k)}\log\max(\omega_u^{(k)}, 10^{-12})/\log|\mathcal{K}|$. For prescribed router-mass interval $[\omega_{\min}, \omega_{\max}]$ and entropy interval $[h_{\min}, h_{\max}]$, the post-curriculum penalty is

$$\begin{aligned}\mathcal{L}_{\text{route}} = {} & \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}[\omega_{\min} - \bar{\omega}^{(k)}]_+^2 + [\max_{k\in\mathcal{K}}\bar{\omega}^{(k)} - \omega_{\max}]_+^2 \\ & + \frac{1}{|\mathcal{U}_B|}\sum_{u\in\mathcal{U}_B}\mathcal{B}(H_u^{\text{route}}; h_{\min}, h_{\max}).\end{aligned} \tag{A.63}$$

The three terms have unit weights. Mass constraints act on batch-mean router probabilities; the contribution-budget penalty separately constrains realized branch contributions. The router penalty is zero while $\lambda_{\text{curr}} < 1 - 10^{-8}$ or when only one order is represented. Configured mass bounds are $[0, 0.9]$ and entropy bounds are $[0.15, 0.85]$.

Responsibility target construction. Let $p^{(k)}_{u,g,e}$ be the base hierarchy's hyperedge attention for network g, and let $\beta^{(k)}_{u,e,i}$ be the member responsibility obtained by masked normalization of the encoder's member-effect scores within e. The nonnegative incidence score is

$$t^{(k)}_{u,i,e} = m^{\mathrm{tar},(k)}_{u,e}\eta^{(k)}_{u,e}\beta^{(k)}_{u,e,i}\sum_{g} p^{(k)}_{u,g,e}. \tag{A.64}$$

For each head, positive-score incidences are passed to the segmented ROI-incidence selection and normalization routine, using $\log\max(t^{(k)}_{u,i,e}, 10^{-12})$ as the input score. Target eligibility is determined by positive incidence scores; the local/bridge partition and structural gate are omitted. The resulting real-incidence masses define $\mathbf{\Gamma}^{(k)}_{u,h}$. The target is recomputed on training forward passes that request alignment and detached for the corresponding differentiation step.

For responsibility alignment, let $\mathbf{x}^{(k)}_{u,h}$ and $\mathbf{\Gamma}^{(k)}_{u,h}$ be predicted attention and detached target masses on the same flattened real-incidence coordinates. Normalization $\mathcal{N}$ excludes the reference state. Let $\mathcal{K}_\Gamma$ contain orders with a defined target, and let $\mathcal{V}_k$ contain subject–head pairs whose predicted and target masses both exceed $\varepsilon_{\mathrm{mass}}$. Define the order-level discrepancy J^{link}_k and the complete penalty by

$$\begin{aligned} J^{\mathrm{link}}_k &= \frac{1}{\max(1, |\mathcal{V}_k|)} \sum_{(u,h)\in\mathcal{V}_k} \mathrm{JS}\Big(\mathcal{N}(\mathbf{x}^{(k)}_{u,h}) \Big\| \mathcal{N}(\mathrm{sg}(\mathbf{\Gamma}^{(k)}_{u,h}))\Big), \\ \mathcal{L}_{\mathrm{link}} &= \frac{1}{\max(1, |\mathcal{K}_\Gamma|)} \sum_{k\in\mathcal{K}_\Gamma} J^{\mathrm{link}}_k. \end{aligned} \tag{A.65}$$

For each order, divergence is averaged over valid subject–head instances; the resulting losses are averaged equally over orders with a defined target. An order with no valid comparisons contributes zero and remains in the outer denominator. This differs from pooling all valid subject–order–head instances into one mean.

The prediction and target can have different supports. Local/bridge masks exclude some incidences from each prediction head, while top-K selection introduces additional zeros that can change across forward passes. If the normalized target assigns total mass $q > 0$ to coordinates where the prediction is zero, those coordinates contribute $q\log(2)/2$ to the Jensen–Shannon divergence. Thus, target mass outside the prediction support introduces a positive contribution to the alignment loss.

The implementation averages five penalty slots: attention entropy, contribution budget, routing, candidate generation, and responsibility alignment. The contribution-budget term is zero for $R_{\max} = 1$, and the candidate-generation term is zero for the fixed candidate bank. Retaining the original division by five gives the three-term regularizer in Eq. 10.

A.7.3 Temporal Regularization

Let $\mathcal{U}_{AB}$ contain minibatch subjects with both temporal views. The base representation in Appendix A.5.1 has dimension D_z. For view $v \in \{A, B\}$, let $\mathbf{z}^{\mathrm{base},v}_u$ be its base feature and $\varepsilon_z > 0$ a numerical floor. Define $\overline{\mathbf{z}}^v_u = \mathbf{z}^{\mathrm{base},v}_u / \max(\|\mathbf{z}^{\mathrm{base},v}_u\|_2, \varepsilon_z)$. The elementwise mean squared discrepancy is

$$\mathcal{D}^{AB}_z = \frac{\sum_{u\in\mathcal{U}_{AB}} \|\overline{\mathbf{z}}^A_u - \overline{\mathbf{z}}^B_u\|_2^2}{\max(1, |\mathcal{U}_{AB}|)D_z}. \tag{A.66}$$

The feature term is zero when no paired subjects are available.

Let t index a representation-family and interaction-order combination. The families comprise hyperedge attention, within-network ROI hub distributions, ROI importance, and real-incidence attention. The set $\mathcal{T}_{AB}$ contains the combinations defined for both temporal views.

For comparison instance r, $\mathbf{x}^v_{t,r}$ is a nonnegative measure on a shared coordinate domain. Hyperedge distributions are compared over candidate edges within each subject and functional network. Hub distributions are compared over ROIs within each subject and functional network, whereas ROI importance from Eq. A.36 is compared over ROIs within each subject and order. Real-incidence distributions are compared over flattened ROI–hyperedge coordinates within each subject and attention head, excluding the reference state.

Using the mass function s and normalization $\mathcal{N}$ defined above, the valid comparison set is

$$\mathcal{R}_t^{AB} = \left\{ r : s(\mathbf{x}_{t,r}^A) > \varepsilon_{\text{mass}}, \quad s(\mathbf{x}_{t,r}^B) > \varepsilon_{\text{mass}} \right\}.$$

Its mean discrepancy is

$$\mathcal{J}_t^{AB} = \begin{cases} \dfrac{1}{|\mathcal{R}_t^{AB}|} \displaystyle\sum_{r \in \mathcal{R}_t^{AB}} \mathrm{JS}\big(\mathcal{N}(\mathbf{x}_{t,r}^A) \big\| \mathcal{N}(\mathbf{x}_{t,r}^B)\big), & |\mathcal{R}_t^{AB}| > 0, \\ 0, & |\mathcal{R}_t^{AB}| = 0. \end{cases} \tag{A.67}$$

Combining these terms gives the main-text temporal loss in Eq. 11: $\mathcal{L}_{\text{tmp}} = (\mathcal{D}_z^{AB} + \sum_{t \in \mathcal{T}_{AB}} \mathcal{J}_t^{AB})/(1 + |\mathcal{T}_{AB}|)$. The feature term and each representation-family and order discrepancy receive equal scalar weights. Combinations defined for both views remain in the denominator even when they have no valid comparisons, in which case their discrepancy is zero. Batches without paired views have zero temporal loss. Temporal regularization acts on normalized base features and the specified relation distributions.

B Datasets Cohort Selection Criteria

B.1 SRPBS-SCZ Cohort

The SRPBS-SCZ analysis cohort was assembled according to an age- and sex-matching principle between participants with schizophrenia and healthy controls (HC). Eligibility therefore required valid imaging data together with the age and sex information needed for matching. The resulting cohort contained 147 participants with schizophrenia and 147 healthy controls, with the matching intended to limit diagnostic-group imbalance attributable to these two demographic variables.

B.2 External Cohort

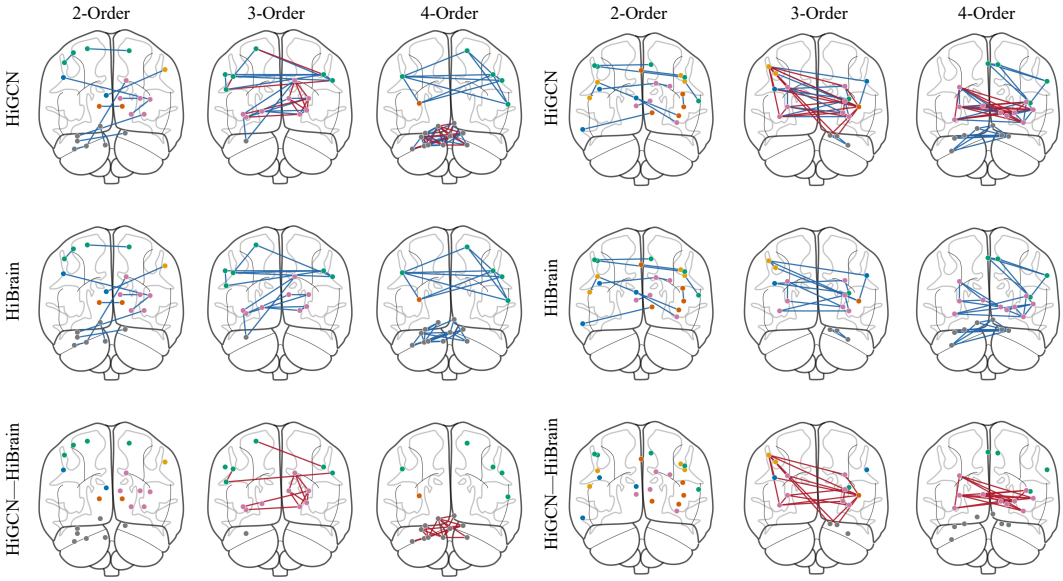
ABIDE-I was screened at the acquisition-site level after preprocessing and quality control. For site s, let n_s^{valid}, n_s^{ASD}, and n_s^{HC} denote the numbers of valid participants, participants with autism spectrum disorder, and healthy controls, respectively. A site was retained only when all three criteria were satisfied simultaneously: $n_s^{\text{valid}} > 20$, $n_s^{\text{ASD}} \geq 10$, and $n_s^{\text{HC}} \geq 10$. The strict total-sample threshold excluded sites with 20 or fewer valid participants, while the two class-specific thresholds prevented inclusion of sites with insufficient representation of either diagnostic group. Applying these rules yielded the locked external cohort of 11 sites and 499 participants reported in Table 1; this cohort was used only for external testing.

C Implementation and Training Details

HiBrain was implemented in PyTorch 2.4.1 and trained on an NVIDIA GeForce RTX 4090 GPU. The latent dimension was $D = 32$. The FC encoder and classifier used widths $160 \to 64 \to 32$ and $32 \to 32 \to 2$, respectively. The structural scalar encoder used $1 \to 32 \to 32$, and the direct regional GM encoder used $160 \to 32 \to 32$. Each 3-Order and 4-Order structural branch contained two hypergraph propagation layers. The spatial GM encoder combined a global projection with hidden width 64 and two 3-D convolutional layers with channels $(8, 16)$. ROI–hyperedge attention used 4 heads and top-8 selection; cross-order propagation used two attention layers with 4 heads each.

Dropout rates were 0.40 for the core FC encoder, 0.45 for the classifier, 0.35 for ROI–hyperedge attention, and 0.25 for cross-order attention. Structural hypergraph, direct GM-profile, spatial GM, and order-admission dropout rates were 0.15, 0.20, 0.25, and 0.15, respectively. The structure-conditioned support encoder used dropout 0.10, and GM modality dropout was 0.20.

We used AdamW with batch size 32, learning rate 10^{-4}, and weight decay 0.01. Structural and admission parameters formed a separate parameter group with the same learning rate and weight decay. Gradient norms were clipped at 2.0. An exponential moving average (EMA) of model parameters was maintained with decay 0.99 from the first epoch.



(a) MDD (b) SCZ

Figure 7: Comparison of connections represented by HiGCN and HiBrain in (a) MDD and (b) SCZ. Blue indicates connections shared by both models, whereas red indicates connections unique to HiGCN.

Models were trained for up to 100 epochs with an early-stopping patience of 15. ReduceLROnPlateau monitored validation accuracy with a reduction factor of 0.5, patience of 3, and a minimum learning rate of 10^{-6}. Internal evaluation used a single model seed across ten site- and diagnosis-stratified outer folds, with eight folds for training, one for validation, and one for testing. External evaluation used 10 ABIDE-II training/validation splits, with each fold used once for validation and the remaining nine for training. Each model was evaluated separately on the same ABIDE-I cohort, and each metric was summarized across the 10 runs.

D Structural Interpretability in MDD and SCZ

We extended the comparison between HiGCN and HiBrain to MDD and SCZ to examine differences in their higher-order relational representations. Blue connections are shared by both models, whereas red connections are unique to HiGCN. These visualizations complement the ASD analysis by showing the correspondence and differences between the two models across disorders (Fig. 7). They provide qualitative context for structural calibration.