DECOUPLED GRAPH EXPERTS FOR ROBUST MULTIMODAL LEARNING WITH MISSING MODALITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal clinical data routinely arrive with *missing* modalities. Bipartite graph formulations address this missingness by encoding patients and modality-derived units as a bipartite graph, but modality-level graph construction forces heterogeneous sub-signals within one modality to share an operator (*false grouping*) and separates related sub-signals across modalities (*false separation*). Both restrictions are costly for patients with missing modalities, who rely on evidence propagated from other patients. False separation places related evidence from a missing modality in a view the patient cannot reach, and false grouping mixes the evidence the patient does receive with unrelated sub-signals. Which signals share an operator should therefore be learned from data: signals that co-vary across modalities through a shared process should share a view, while co-variation within one modality, which acquisition effects confound, should not decide the grouping. We propose DECENT, which encodes each observed modality into index-aligned proto-tokens and learns the signal-node partition by self-supervised sparse routing, trained to route signals that co-vary across modalities to the same view. With this partition fixed, DECENT runs message passing on the resulting view-specific sample–signal graphs and fuses the views observed for each patient. On ADNI and MIMIC-IV, DECENT achieves the highest average macro-F1 and macro-AUC among the evaluated methods across four missingness settings. Partition controls with the tokenizer and graph encoder fixed attribute these gains to the learned partition. Post hoc analyses on ADNI further show that the learned views align more closely with disease-informed axes than with source modalities. Code is available at `https://anonymous.4open.science/r/DECENT-7202/`.

1 INTRODUCTION

Clinical prediction draws on complementary evidence from imaging, laboratory assays, genetic profiles, and structured records (Kline et al., 2022; Liang et al., 2024; Ramachandram & Taylor, 2017), yet patients often lack one or more of these modalities (Wu et al., 2026; 2024). Sample-independent methods, such as imputation (Yoon et al., 2018; Ma et al., 2021), robust architectures (Ma et al., 2022; Peng et al., 2025), and modality experts (Yun et al., 2024; Li et al., 2025), share knowledge across patients only through fitted parameters. Graph formulations instead propagate evidence across patients at inference, for example through bipartite graphs that connect patients to feature- or modality-derived signal nodes (You et al., 2020; Wu et al., 2024). They either process all signals with one shared message function (Gilmer et al., 2017) or fix the views in advance (Chen & Zhang, 2020; Zhang et al., 2024). These design choices raise a structural question: which signals should share a message-passing operator?

Acquisition modality need not be the appropriate boundary. A modality may carry signals from several disease processes, and one process may span modalities (Jack et al., 2018) (Figure 1a). Predefined modality-level partitions then impose two restrictions: *false grouping*, in which heterogeneous sub-signals within one modality share a subgraph and operator, and *false separation*, in which related sub-signals across modalities cannot interact within the same view. Both are costly for patients with missing modalities, who rely on evidence propagated from other patients. False separation leaves related evidence from a missing modality in a view the patient cannot reach, and false grouping mixes the evidence the patient does receive with unrelated sub-signals.

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.

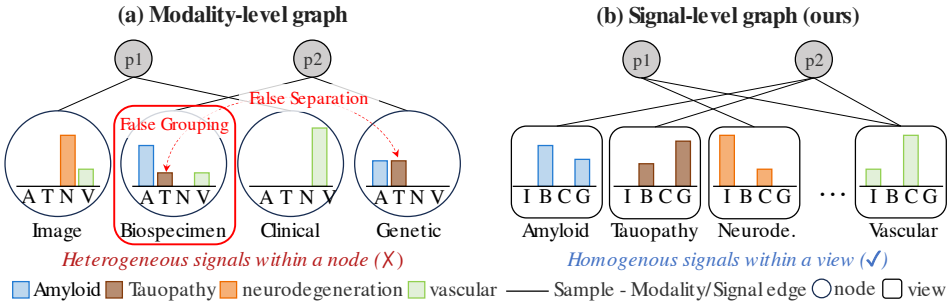


Figure 1: **Modality-level versus signal-level graph construction.** A/T/N/V denote disease-relevant processes in ADNI, following the AT(N) research framework (Jack et al., 2018) with a vascular axis motivated by white matter hyperintensities (Ng et al., 2023): amyloid deposition, tau pathology, neurodegeneration, and vascular burden. I/B/C/G denote acquisition modalities: image, biospecimen, clinical, and genetic data. **(a)** A modality-level graph groups signals by acquisition source, which can mix distinct disease axes within one modality and separate related axes across modalities. **(b)** DECENT learns signal-level views (schematic) before message passing, allowing cross-modal grouping and intra-modal splitting when disease processes and acquisition sources do not align.

Learning the partition requires a criterion for when two signals reflect the same process. Co-variation across patients within a modality is confounded: signals acquired together share scanner or assay-batch effects and co-vary even when they measure different processes, so grouping by co-variation over all signal pairs drifts back toward modality-level views. When these effects are independent across modalities, co-variation between signals of different modalities reflects only processes that the signals share or that depend on each other.

We propose DECENT (**Dec**oupled Graph **E**xperts for Multimodal Lear**n**ing wi**t**h Missing Modalities), which learns which sub-signals share a graph operator (Figure 1b). DECENT encodes each observed modality into index-aligned proto-tokens, for which co-variation across patients is well defined, and assigns their signal nodes to views with a sparse router trained without outcome labels to route signals that co-vary across modalities to the same view. It then fixes the partition and trains prediction on the resulting view-specific sample–signal graphs, fusing the views available to each patient. Neither the view count nor the assignments are fixed by acquisition source, allowing cross-modal grouping and intra-modal splitting while preserving connections across patients. Across four missingness settings on ADNI and MIMIC-IV, DECENT achieves the highest average macro-F1 and macro-AUC among the compared methods. With the tokenizer and graph encoder fixed, the learned partition improves over modality-based, single-view, and random partitions under induced missingness.

Our contributions are as follows.

- We identify false grouping and false separation as two restrictions of modality-level graph construction that limit the evidence reaching patients with missing modalities, and formalize them as properties of the signal-node partition (Section 2.1).
- We develop DECENT, which addresses both restrictions by training a sparse router without outcome labels to partition signal nodes by the cross-modal co-variation of index-aligned proto-tokens, and then trains shared graph encoding and availability-aware fusion on the fixed partition (Sections 2.2–2.5).
- We evaluate on two clinical benchmarks over four missingness settings and five seeds, with partition controls, routing-target comparisons, and ADNI analyses of the learned views (Section 3).

2 METHOD

DECENT maps each observed modality to index-aligned proto-tokens, routes their signal nodes into views, and predicts through view-specific graph encoding and availability-aware fusion (Figure 2). For N samples and M modalities, sample i has observation $\boldsymbol{x}^{(i)} = (\boldsymbol{x}_1^{(i)}, \ldots, \boldsymbol{x}_M^{(i)})$ and availability

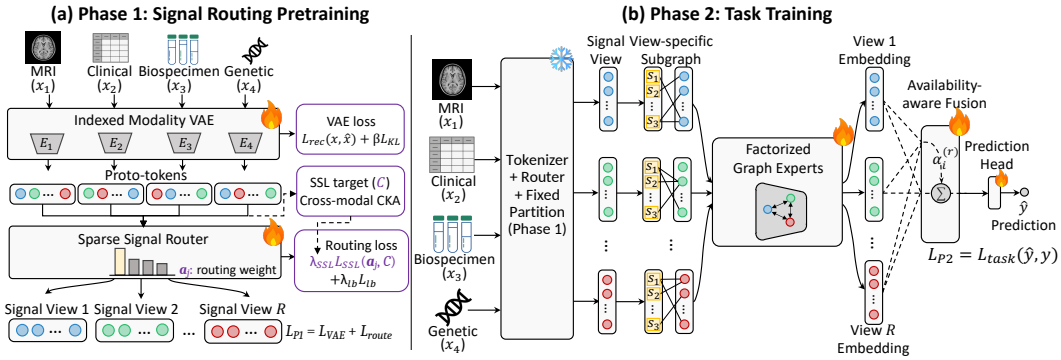


Figure 2: **Overview of DECENT.** Phase 1, without outcome labels, trains an indexed modality VAE and a sparse router whose routing similarities follow the cross-modal co-variation of the proto-tokens (posterior means), and then fixes the partition of signal nodes into views. Phase 2 encodes each view on its sample–signal bipartite graph, fuses the views available to each patient, and trains the graph encoder, fusion module, and classifier.

mask $\boldsymbol{M} \in \{0,1\}^{N\times M}$, where $\boldsymbol{M}[i,m] = 1$ iff modality m is observed and $\sum_m \boldsymbol{M}[i,m] \geq 1$ (Section 3.1). Each modality contributes k signal nodes: p_j, $j = (m,t)$, corresponds to token t of modality $m(j) = m$, and $J = Mk$.

2.1 PROBLEM FORMULATION

We view each modality m as expressing a subset $\mathcal{Q}_m \subseteq \{1,\ldots,Q\}$ of latent disease processes $z_1,\ldots,z_Q$, which may be correlated, together with modality-local variation u_m such as scanner or assay-batch effects and measurement noise:

$$\boldsymbol{x}_m = \sum_{q\in\mathcal{Q}_m} f_{m,q}(z_q) + f_{m,0}(u_m). \tag{1}$$

Process q is shared if its footprint $F_q = \{m : q \in \mathcal{Q}_m\}$ contains at least two modalities. Let $v^\star(j)$ be the process that signal j reflects, with $v^\star(j) = 0$ for modality-local variation. A partition π of the signal nodes into views is *process-consistent*, with no false grouping, if $\pi(j) = \pi(j')$ implies $v^\star(j) = v^\star(j')$, and *process-complete*, with no false separation, if $v^\star(j) = v^\star(j')$ implies $\pi(j) = \pi(j')$. On the shared signals $\mathcal{J}_{\text{sh}} = \{j : v^\star(j) \geq 1,\ |F_{v^\star(j)}| \geq 2\}$, the target $\pi^\star = v^\star$ has both properties, whereas a modality partition violates consistency when a modality expresses two shared processes and completeness when a process spans two modalities.

The partition criterion must not be confounded by acquisition effects. Signals of one modality share u_m and can co-vary across patients even when they reflect different processes. When $u_1,\ldots,u_M$ are independent across modalities and of the processes, signals of different modalities can co-vary only through processes that they share or that depend on each other, for example through overall disease severity (Proposition 1 in Appendix B). DECENT therefore uses cross-modal co-variation across patients as the supervisory signal for learning which signals share a view.

2.2 INDEX-ALIGNED PROTO-TOKENIZATION

Measuring how two signals co-vary across patients requires each token index to denote the same unit for every patient, which exchangeable slots (Locatello et al., 2020) do not provide. We use an indexed modality VAE (Kingma & Welling, 2014), whose reconstruction term keeps input information in the tokens. For modality m, the encoder E_m maps $\boldsymbol{x}_m$ to k diagonal Gaussian posteriors, $\{(\boldsymbol{\mu}_{m,t}, \boldsymbol{\sigma}_{m,t})\}_{t=1}^k = E_m(\boldsymbol{x}_m)$ with $\boldsymbol{\mu}_{m,t}, \boldsymbol{\sigma}_{m,t} \in \mathbb{R}^{d_z}$, and the decoder D_m reconstructs $\boldsymbol{x}_m$ from reparameterized samples of all k groups. Group t is the output at a fixed encoder index, so its computational correspondence across patients is fixed by construction. Sample i contributes the proto-token $\boldsymbol{p}_{i,j} = \boldsymbol{\mu}_{m,t}^{(i)}$ for signal $j = (m,t)$ and the edge evidence

$$\boldsymbol{e}_{i,j}^{(0)} = \phi_m(\boldsymbol{p}_{i,j}) \in \mathbb{R}^d, \tag{2}$$

where the projection ϕ_m is shared by the groups of modality m and does not encode the group index itself. Posterior means are used in both phases to prevent sampling noise from affecting routing and graph construction.

2.3 Signal Routing by Cross-Modal Co-Variation

A sparse router assigns the signal nodes to R views. For signal j, let $\mathcal{I}_j$ be the samples in the current mini-batch that observe $m(j)$ and $\bar{\boldsymbol{e}}_j$ the mean of $\boldsymbol{e}_{i,j}^{(0)}$ over $\mathcal{I}_j$. Sparse mixture-of-experts gating (Shazeer et al., 2017) gives

$$\boldsymbol{a}_j = \mathrm{TopK}\big(\mathrm{softmax}(H(\bar{\boldsymbol{e}}_j + \boldsymbol{\eta}_j)/\tau),\, K\big) \;\in\; \Delta^{R-1}, \tag{3}$$

where $\boldsymbol{\eta}_j \in \mathbb{R}^d$ is a learned embedding of signal j initialized at zero, H gives R noisy linear logits (Eq. 12), TopK keeps the K largest probabilities and renormalizes them, and the temperature τ is annealed to $\tau_{\min}$. The embedding lets the router place groups of one modality in different views even when their summaries are close.

Routing target. For signals j and j' of different modalities, let $\mathcal{T}_{jj'}$ be the training patients who observe both, with $n_{jj'} = |\mathcal{T}_{jj'}|$, and let $\boldsymbol{K}_j = \boldsymbol{X}_j \boldsymbol{X}_j^\top$ and $\boldsymbol{K}_{j'} = \boldsymbol{X}_{j'} \boldsymbol{X}_{j'}^\top$ be the linear Gram matrices of their column-centered proto-tokens $\boldsymbol{X}_j, \boldsymbol{X}_{j'} \in \mathbb{R}^{n_{jj'} \times d_z}$ over $\mathcal{T}_{jj'}$. We measure their co-variation by linear centered kernel alignment (CKA) (Kornblith et al., 2019) computed with the unbiased HSIC estimator (Song et al., 2012), known as debiased CKA (Nguyen et al., 2021),

$$C_{jj'} = \frac{\max\big\{0, \widehat{\mathrm{HSIC}}_u(\boldsymbol{K}_j, \boldsymbol{K}_{j'})\big\}}{\sqrt{\widehat{\mathrm{HSIC}}_u(\boldsymbol{K}_j, \boldsymbol{K}_j)\, \widehat{\mathrm{HSIC}}_u(\boldsymbol{K}_{j'}, \boldsymbol{K}_{j'})}}, \tag{4}$$

where $\widehat{\mathrm{HSIC}}_u$ denotes this estimator. Plain linear CKA is biased upward by about $d_z/n_{jj'}$ for independent tokens (Appendix A.1), which would favor partners with few co-observed patients. C is computed without gradient at the start of every Phase-1 epoch after the warm-up (Section 2.5) over all co-observed training patients, as a mini-batch contains few co-observed patients per pair. Unlike similarity between signal summaries, which compares the tokens of an average patient, Eq. 4 measures whether two signals vary together across patients. Let Ω_j be the signals j' of other modalities with $n_{jj'} \geq n_{\min}$, and $\mathcal{J}_c$ the signals with $\Omega_j \neq \emptyset$, both restricted to signals observed in the current mini-batch. For each anchor $j \in \mathcal{J}_c$, the distributions $P_j^{\mathrm{co}} = \mathrm{softmax}\big((C_{jj'})_{j' \in \Omega_j}/\tau_c\big)$ and $P_j^{\mathrm{route}} = \mathrm{softmax}\big((\cos(\boldsymbol{a}_j, \boldsymbol{a}_{j'}))_{j' \in \Omega_j}/\tau_r\big)$ are aligned by

$$\mathcal{L}_{\mathrm{SSL}} = \frac{1}{2|\mathcal{J}_c|} \sum_{j \in \mathcal{J}_c} \Big[\mathrm{KL}\big(P_j^{\mathrm{co}} \,\|\, P_j^{\mathrm{route}}\big) + \mathrm{KL}\big(P_j^{\mathrm{route}} \,\|\, P_j^{\mathrm{co}}\big)\Big]. \tag{5}$$

Two signals of one modality are thus routed together when they co-vary with the same signals of other modalities and apart when they co-vary with different ones, while co-variation within a modality, which u_m confounds, does not enter the target. A coefficient-of-variation penalty $\mathcal{L}_{\mathrm{lb}}$ on $\sum_j \boldsymbol{a}_j$ (Shazeer et al., 2017) spreads the signals across views (Eq. 13).

Fixed partition. After Phase 1, the router is applied once to the summaries $\bar{\boldsymbol{e}}_j$ over all training patients, the population counterpart of the mini-batch summaries used in Phase 1, and each signal is assigned to $g(j) = \arg\max_r H(\bar{\boldsymbol{e}}_j + \boldsymbol{\eta}_j)_r$ without routing noise. This partition is fixed in Phase 2 and at test time, so it does not depend on the test batch.

2.4 View Graphs and Prediction

Samples and signals form an edge-attributed bipartite graph with sample nodes $\mathcal{S} = \{s_i\}_{i=1}^N$ and signal nodes $\mathcal{P} = \{p_j\}_{j=1}^J$ shared by all samples. The edge set $\mathcal{E}$ contains (s_i, p_j) iff $\boldsymbol{M}[i, m(j)] = 1$ and carries $\boldsymbol{e}_{i,j}^{(0)}$ (Eq. 2). With views $\mathcal{P}^{(r)} = \{p_j : g(j) = r\}$, view r keeps all sample nodes and the edges incident to $\mathcal{P}^{(r)}$:

$$\mathcal{G}^{(r)} = (\mathcal{S} \cup \mathcal{P}^{(r)},\, \mathcal{E}^{(r)}), \quad \mathcal{E}^{(r)} = \{(s_i, p_j) \in \mathcal{E} \mid p_j \in \mathcal{P}^{(r)}\}. \tag{6}$$

Samples observing the same signal exchange information through $s_i \to p_j \to s_{i'}$, so a view can group signals across modalities and split them within one. At inference, the graphs contain only the current test batch, and a patient's prediction depends on other patients only through message passing.

Factorized encoding and fusion. Sample nodes start from all-ones vectors and signal nodes from learned token-position embeddings, so patient-specific evidence enters only through edges. Each $\mathcal{G}^{(r)}$ is processed by L layers of edge-attributed message passing in the GraphSAGE family (Hamilton et al., 2017) with edge updates as in GRAPE (You et al., 2020) (Eq. 15), whose message weights use one basis per view (Schlichtkrull et al., 2018),

$$\boldsymbol{W}_j^{(\ell)} = \sum_{r=1}^{R} \hat{a}_{j,r}\, \boldsymbol{B}_r^{(\ell)} = \boldsymbol{B}_{g(j)}^{(\ell)}, \qquad \hat{a}_{j,r} = \mathbb{1}[r = g(j)], \tag{7}$$

while aggregation and edge updates are shared. A view-specific readout gives $\boldsymbol{h}_i^{(r)}$, attention over the views $\mathcal{K}_i$ with observed signals of sample i fuses them into $\boldsymbol{z}_i$ (Eq. 17), and $\hat{y}_i = \mathrm{MLP}(\boldsymbol{z}_i)$.

2.5 Two-Phase Training

Phase 1 uses no outcome labels and minimizes, over the parameters marked below each term,

$$\mathcal{L}_{\mathrm{P1}} = \underbrace{\mathcal{L}_{\mathrm{rec}} + \beta\, \mathcal{L}_{\mathrm{KL}}}_{E_m,\, D_m} + \underbrace{\lambda_{\mathrm{SSL}}\, \mathcal{L}_{\mathrm{SSL}} + \lambda_{\mathrm{lb}}\, \mathcal{L}_{\mathrm{lb}}}_{\phi_m,\, \boldsymbol{W}_g,\, \boldsymbol{W}_{\mathrm{noise}},\, \boldsymbol{\eta}}, \tag{8}$$

where $\mathcal{L}_{\mathrm{rec}}$ is the dimension-normalized squared error over observed modalities and $\mathcal{L}_{\mathrm{KL}}$ the divergence of the posteriors from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ (Appendix A.1). $\mathcal{L}_{\mathrm{SSL}}$ is switched on after E_w warm-up epochs of the VAE, when C is first computed. The proto-tokens enter the routing losses with stopped gradients, so the VAE is optimized only by its variational objective and receives no routing gradients. Clinical-axis annotations are not used in training. Phase 2 freezes the VAE, ϕ_m, $\boldsymbol{\eta}_j$, the router, and the partition g, and trains the bases $\{\boldsymbol{B}_r^{(\ell)}\}$, the remaining graph-encoder parameters, the fusion module, and the classifier with the task loss

$$\mathcal{L}_{\mathrm{P2}} = \mathcal{L}_{\mathrm{task}}(\hat{y}, y). \tag{9}$$

Appendix A.1 gives the training and inference algorithms.

3 Experiments

We ask whether signal-level partitioning improves prediction under missing modalities (Section 3.2), whether the gain comes from the learned partition and the components of Section 2 that produce it (Section 3.3), and what the label-free Phase 1 learns (Section 3.4).

3.1 Experimental Setup

Datasets. We evaluate on two multimodal clinical benchmarks. **ADNI** (Jack et al., 2008) contains four modalities (Image, Genetic, Clinical, Biospecimen) for diagnosis stage prediction (CN/MCI/AD). **MIMIC-IV** (Johnson et al., 2023b), accessed through PhysioNet (Pollard et al., 2026), contains three modalities (lab and vital values, clinical notes (Johnson et al., 2023a), ICD-9 codes) for one-year mortality prediction. We use preprocessed modality features following Flex-MoE (Yun et al., 2024).

Missing-modality protocol. We independently mask originally observed modalities with probability $\rho \in \{0.2, 0.5, 0.7\}$ using fixed uniform scores. If all are masked, we restore the modality with the largest score. This retain-one rule applies to training, validation, and testing; ρ is the removal probability before restoration, and $\rho = 0$ preserves the original observation pattern, which already contains missing modalities. All methods and component comparisons use five paired random seeds with the same data splits and missingness manifests.

Baselines. We compare against five graph-based methods (GRAPE (You et al., 2020), M3Care (Zhang et al., 2022), MUSE (Wu et al., 2024), HGMF (Chen & Zhang, 2020), GTP-4o (Li et al., 2024)) and three non-graph methods (Flex-MoE (Yun et al., 2024), AMC (Peng et al., 2025), SimMLM (Li et al., 2025)). Hyperparameters follow published settings; otherwise we tune the learning rate, hidden dimension, and batch size on the validation set.

Table 1: **Main results on ADNI and MIMIC-IV.** Macro-F1 and macro-AUC (%) without additional masking (ρ=0) and with induced missingness ($\rho \in \{0.2, 0.5, 0.7\}$). Means over five seeds on fixed data splits; standard deviations are reported in Appendix C.1. Best values are bold. Ours is shaded blue. Values lower than ours are gray.

	ADNI								MIMIC-IV							
	ρ=0		ρ=0.2		ρ=0.5		ρ=0.7		ρ=0		ρ=0.2		ρ=0.5		ρ=0.7	
Method	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
Graph-based																
GRAPE	63.3	80.4	**62.1**	78.0	53.6	74.6	52.0	71.5	57.0	74.7	55.7	69.9	54.6	67.8	55.5	67.1
GTP-4o	55.9	74.5	54.8	71.7	49.9	69.0	49.6	68.1	55.8	71.4	54.4	68.4	54.0	63.6	54.9	60.3
HGMF	51.8	69.4	49.3	67.6	47.1	64.5	49.1	66.3	55.9	72.8	53.4	65.1	52.0	58.8	53.0	61.7
M3Care	61.3	77.2	55.2	74.8	52.9	72.4	49.2	70.6	54.1	66.7	51.2	50.1	50.9	47.8	51.9	48.3
MUSE	**63.7**	**81.3**	60.2	77.8	56.1	75.4	53.6	72.4	59.6	**76.0**	56.3	71.1	55.8	69.0	55.6	65.9
Non-graph																
AMC	49.5	68.4	45.5	66.2	40.4	61.1	34.9	57.5	55.2	63.5	54.6	59.2	53.0	55.3	53.5	56.4
Flex-MoE	59.9	78.2	50.8	71.9	48.7	69.1	47.8	65.9	55.8	71.3	54.8	67.4	54.5	65.1	54.0	63.3
SimMLM	62.3	79.1	56.9	75.2	52.7	71.4	51.3	70.6	55.8	72.7	54.6	70.0	55.0	66.3	54.1	64.6
Ours																
DECENT	63.0	80.6	61.9	**79.6**	**57.9**	**77.8**	**55.8**	**74.3**	**59.8**	75.9	**58.7**	**73.3**	**57.4**	**71.1**	**56.5**	**69.2**

Evaluation and implementation. We report macro-F1 (primary) and macro-AUC as means over five seeds, with standard deviations in Appendix C.1. DECENT is implemented in PyTorch (Paszke et al., 2019) with GraphSAGE (Hamilton et al., 2017) as the view encoder, $L = 3$ layers on ADNI and $L = 2$ on MIMIC-IV, $k = 4$ groups per modality, and $R = 4$ and $R = 6$ views. Its test graphs hold batches of 128 (ADNI) and 256 (MIMIC-IV) patients; as the partition is fixed, the test batch affects a prediction only through message passing (Appendix C.3). One configuration per dataset is used across all missingness settings; other hyperparameters and schedules are in Appendix A.2.

3.2 PRIMARY RESULTS

Prediction from incomplete observations. Section 1 argues that false grouping and false separation cost most for patients who must rely on evidence propagated from other patients. Table 1 tests this against graph methods that propagate evidence between samples and features (GRAPE) or between samples and whole modalities (MUSE), and against non-graph methods that predict from each patient's own observations. DECENT attains the highest average macro-F1 and macro-AUC on both datasets. The comparison with MUSE, which uses the modality-level construction, locates the gain. With the natural observation pattern ($\rho = 0$) the two are within 0.7 pp, and MUSE is ahead on ADNI and in MIMIC-IV macro-AUC. Once modalities are removed, DECENT leads macro-AUC in every setting and macro-F1 in all but one, and its macro-AUC declines less from $\rho = 0$ to 0.7, by 6.3 and 6.7 pp against 8.9 and 10.1 pp for MUSE. The gain therefore appears where patients depend most on propagated evidence, and Section 3.3 tests whether the learned partition produces it. Appendix C.2 extends the comparison to a non-clinical dataset.

3.3 COMPONENT ANALYSIS

For Table 2, we replace one component at a time and keep the others fixed according to the protocols in Appendix A.2. Blocks (a) and (b) test what learning the partition requires, and blocks (c) and (d) test how the partition is used. Appendix D further examines the routing target, the routing losses, and the partition granularity.

Proto-tokenization. Section 2.2 requires several signals per modality, so that a view can split a modality, and a fixed index per signal, so that co-variation across patients is defined. Removing either requirement lowers macro-F1. A single token per modality keeps the capacity but can group only whole modalities, and it is lower in every setting. Slot Attention (Locatello et al., 2020) keeps several slots but no index shared across patients and is the weakest tokenizer throughout. Classical

Table 2: **Component comparisons.** Macro-F1 (%) under induced missingness ($\rho \in \{0.2, 0.5, 0.7\}$), averaged over five seeds. Each block changes one component of DECENT and keeps the others fixed. Best values within each block are bold. Ours is shaded blue. Values lower than ours are gray.

Configuration	ADNI			MIMIC-IV		
Missing rate ρ	0.2	0.5	0.7	0.2	0.5	0.7
(a) Proto-tokenization						
Indexed VAE	**61.9**	**57.9**	55.8	**58.7**	**57.4**	**56.5**
Single token (k=1)	59.7	55.4	53.8	56.6	55.7	54.8
Slot Attention	57.5	52.4	50.4	54.7	52.3	51.6
PCA	59.0	**57.9**	56.2	56.3	56.3	55.4
FA-inspired	59.4	57.5	**56.4**	58.0	57.2	56.0
(c) Training strategy						
Frozen (two-phase)	**61.9**	**57.9**	55.8	**58.7**	**57.4**	**56.5**
Phase-2 fine-tuning	59.8	55.5	**56.3**	57.1	56.2	55.6
Task-only training	59.2	55.0	53.1	56.8	56.2	55.7

Configuration	ADNI			MIMIC-IV		
Missing rate ρ	0.2	0.5	0.7	0.2	0.5	0.7
(b) Signal-node partition						
Learned (fixed)	**61.9**	**57.9**	**55.8**	58.7	**57.4**	**56.5**
Learned (dynamic)	61.2	55.9	54.7	**58.9**	56.8	55.6
Modality-based	60.5	55.9	53.8	57.2	56.2	55.8
Balanced random	57.0	53.7	50.3	54.7	52.1	51.4
Single view (R=1)	58.4	54.5	50.1	56.0	52.7	50.5
(d) Encoder sharing						
Factorized experts	**61.9**	57.9	**55.8**	**58.7**	**57.4**	**56.5**
Shared GNN	61.0	57.4	54.5	57.1	55.8	54.8
Independent GNNs	61.4	**58.6**	55.0	58.0	56.9	55.7

index-aligned bases, PCA (Jolliffe & Cadima, 2016) and FA-inspired decomposition, come within 1.1 pp of the indexed VAE at $\rho \in \{0.5, 0.7\}$ and exceed it on ADNI at $\rho = 0.7$, but trail it at $\rho = 0.2$ by up to 2.9 pp. At higher missing rates, index alignment thus matters more than the choice of the VAE, which adds most at $\rho = 0.2$.

Signal-node partition. The central claim is that which signals share a view matters beyond having several views. Block (b) keeps the tokenizer, signal nodes, and graph encoder, and R except in the single-view control, and changes only the assignment rule. The learned partition outperforms the modality-based, balanced random, and single-view partitions in all six settings. The modality-based partition, which imposes false grouping and false separation with the same number of views, trails by up to 2.0 pp, and a balanced random partition falls below it, so the gain requires the learned assignment rather than more operators. Re-running the router on every mini-batch and test batch (*learned (dynamic)*) trails the fixed partition in five of the six settings, so a population-level partition fixed after Phase 1 suffices and keeps the assignment independent of the test batch.

Training strategy. Section 2.5 learns the partition without labels and freezes it before task fitting. Block (c) instead lets the task loss update the tokenizer and router, after Phase 1 (*Phase-2 fine-tuning*) or from the start (*task-only training*). Fine-tuning is lower than the frozen schedule in five of the six settings, all but ADNI at $\rho = 0.7$, and task-only training in all six, by up to 2.9 pp. Starting from the learned partition is thus not enough when task gradients can still move it, and Section 3.4 shows the direction in which task supervision moves it: toward modality-level views.

Encoder sharing. The partition acts through view-specific operators: views use separate message bases and share aggregation and edge updates (Eq. 7). A fully shared GNN keeps the view graphs but applies one operator to all of them, and it is lower in every setting, so routing signals into separate graphs helps only when the graphs are processed differently. Independent GNNs, with all message-passing parameters separate, exceed the factorized experts in only one setting despite more parameters, so the separate message bases carry the benefit and the rest can be shared.

3.4 ANALYSIS OF THE LEARNED PARTITION

We test what the label-free Phase 1 learns: whether the partition captures cross-modal co-variation that holds beyond the training patients and matters for prediction, and how it relates to disease-informed axes on ADNI.

Held-out co-variation and prediction. The routing target is formed on training patients, so we measure on validation patients whether the partition groups signals that co-vary: Δ_C is the mean debiased CKA between cross-modal signals in the same view minus that between cross-modal signals in different views. It needs no annotations and applies to both datasets. Figure 3 places the learned

partitions, the balanced random partitions of Table 2b, and the routing-target controls of Appendix D.3 by their held-out Δ_C and macro-F1. The learned partitions attain the highest mean Δ_C on both datasets, whereas balanced random partitions stay near zero, so the router captures co-variation that holds beyond the patients used to form the target. Across partitions, macro-F1 increases with Δ_C (Spearman correlation 0.81 on ADNI and 0.64 on MIMIC-IV): targets that ignore co-variation or add within-modality pairs group signals with weaker held-out co-variation and predict worse, and spectral clustering of C reaches a similar Δ_C with lower macro-F1 than the router. The quantity that Phase 1 optimizes without labels thus tracks prediction.

Semantic reference. We annotate input feature groups using amyloid (A), tau (T), and neurodegeneration (N) axes from ATN (Jack et al., 2018) and a vascular (V) axis motivated by white matter hyperintensities (Ng et al., 2023), including related risk factors. These groups cross acquisition boundaries: for example, amyloid-related evidence includes biospecimen Aβ and genetic APOE $\epsilon 4$, while biospecimen measurements span A, T, and N. The annotations are used only to interpret the learned partition. Agreement with the axis and modality labels is measured by adjusted Rand index (Hubert & Arabie, 1985), denoted A_{axis} and A_{mod}. Each group takes the annotated axis whose features contribute most to its posterior mean on training patients, and groups dominated by unannotated features are excluded (Appendix A.2).

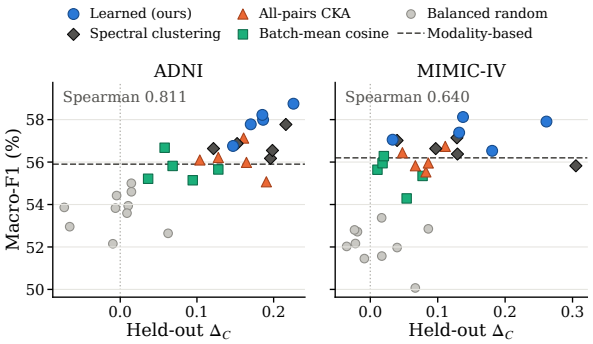


Figure 3: **Held-out co-variation tracks prediction.** Each point is one partition and seed at $\rho = 0.5$: Δ_C on validation patients and test macro-F1 of the model trained on that partition. Dashed line: modality-based partition, for which Δ_C is undefined.

Partition alignment during training. On ADNI, A_{axis} increases over Phase 1 and ends at 0.62, against 0.21 for A_{mod} (Figure 4), while random partitions have an expected ARI of zero; A_{axis} exceeds A_{mod} at every checkpoint after initialization. Task-only training instead moves the partition toward acquisition source within its first 20 epochs, and over its first 100 epochs A_{mod} ends at 0.55 and A_{axis} at 0.12. Without axis or modality labels, the partition thus departs from acquisition-based grouping toward disease-informed axes, whereas task supervision pulls the partition back to modality-level views. These opposite trajectories are the reason DECENT freezes the partition before task fitting. Appendix E.2 shows the ADNI view composition, and Appendix E.3 tests whether cross-modal co-variation recovers planted views on synthetic data.

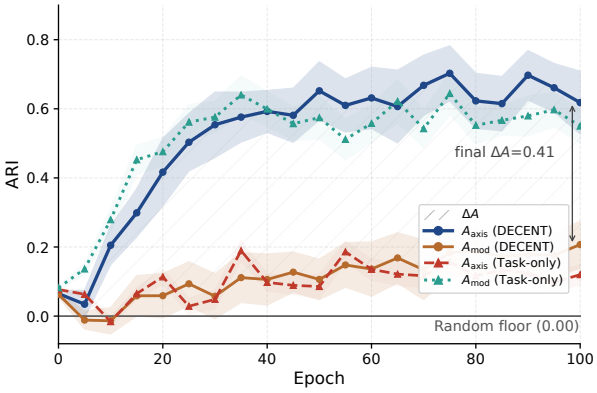


Figure 4: **Label-free pretraining favors axis over modality alignment.** A_{axis} and A_{mod} of the ADNI partition during Phase 1 of DECENT and the first 100 epochs of task-only training ($\rho = 0.5$; five-seed mean $\pm$ SD). Hatching marks $\Delta A = A_{\text{axis}} - A_{\text{mod}}$ for DECENT.

4 Related Work

From imputation to graphs for missing modalities. Imputation reconstructs missing inputs or latent features with generative models (Tran et al., 2017; Yoon et al., 2018; Ngiam et al., 2011) or Bayesian meta-learning (Ma et al., 2021), which fewer observed modalities can limit (Wu et al.,

2026). Direct prediction instead uses robust fusion (Havaei et al., 2016; Ma et al., 2022), knowledge distillation (Wang et al., 2023b; Chen et al., 2024), prompt adaptation (Lee et al., 2023; Jang et al., 2024), adaptive token re-balancing (Peng et al., 2025), token replacement (Wang et al., 2022), or modality experts (Yun et al., 2024; Li et al., 2025). These sample-independent methods capture cross-patient regularities through learned parameters. Graph-based methods add explicit inference-time exchange: sample-feature bipartite graphs (You et al., 2020), sample-modality bipartite graphs (Wu et al., 2024), and task-guided patient-similarity aggregation (Zhang et al., 2022) provide relational evidence under missingness (Wu et al., 2024; Zhang et al., 2024). See Ektefaie et al. (2023) for a broader survey.

From predefined to learned graph partitions. HGMF uses a heterogeneous hypernode graph to relate modality-availability patterns (Chen & Zhang, 2020); MPI separates learned biological factors from patient–phenotype relations (Zhang et al., 2024); GTP-4o uses modality-specific node and edge types (Li et al., 2024). Heterogeneous GNNs organize messages by predefined relations (Schlichtkrull et al., 2018) or meta-paths (Wang et al., 2019). These designs specify view roles or relation types before message passing. Disentangled GNNs learn latent factors for edges from the task, by neighborhood routing (Ma et al., 2019) or by factorizing a graph into subgraphs (Yang et al., 2020). DECENT instead learns, without labels, which sub-modal signals share an operator, and fixes this partition while preserving cross-sample propagation.

Shared and modality-specific structure. Multi-view factor models such as group factor analysis (Klami et al., 2015), oi-VAE (Ainsworth et al., 2018), and MOFA (Argelaguet et al., 2018) separate variation shared across views from view-specific variation, and multimodal VAEs infer joint posteriors from any subset of modalities (Sutter et al., 2021). These models learn representations or impute data but do not define how patients exchange evidence at inference. DECENT uses cross-modal co-variation only to decide which signals share a graph operator, and predicts through message passing across patients; debiased CKA, usually used to compare trained representations (Kornblith et al., 2019; Nguyen et al., 2021), serves here as a label-free routing target.

Mixture-of-experts and graph routing. Mixture-of-experts models learn assignments (Jacobs et al., 1991; Ma et al., 2018), often with sparse gates and balancing losses (Shazeer et al., 2017; Fedus et al., 2022). Multimodal experts route tokens or predictions (Mustafa et al., 2022; Han et al., 2024; Li et al., 2025); Flex-MoE specializes by observed modality combination (Yun et al., 2024). Other mechanisms include soft dispatch, factorized weights, and token re-balancing (Puigcerver et al., 2024; Oldfield et al., 2024; Peng et al., 2025). GMoE selects aggregation experts and hop ranges per node (Wang et al., 2023a). These routers are trained with the task loss. The router of DECENT is trained without labels to follow cross-modal co-variation, and its fixed partition defines view-specific subgraphs before sample–signal propagation.

5 CONCLUSION

DECENT makes signal-node partitioning a learned part of multimodal graph construction under missing modalities. It encodes each modality into index-aligned proto-tokens and trains a sparse router without labels to follow their cross-modal co-variation, so that related signals share operators across modalities and heterogeneous signals of one modality can use different operators. On ADNI and MIMIC-IV, it achieves the highest average macro-F1 and macro-AUC among the evaluated methods; under induced missingness, the learned partition outperforms modality-based and random partitions and a single view, and the learned views align more closely with disease-informed axes than with acquisition source. Cross-modal co-variation can also arise from factors that reach several modalities, such as overall disease severity, so it serves as a supervisory signal for the partition rather than an identification of disease processes. Appendix F.1 discusses further limitations, including the dependence of predictions on the test batch and the scope of the semantic analyses.

AI USE STATEMENT

We used generative AI tools as assistants in this work. For tasks with required disclosure, they gave feedback on the research methodology and experimental design, helped refine the problem formulation and the statement and proof of Proposition 1, assisted in implementing parts of the method, and helped

describe and interpret the reported results. For tasks with recommended disclosure, they assisted with editing the manuscript for clarity, drafting and revising text, identifying related literature, creating figures from our results, and formatting references. Generative AI tools were not used to generate, clean, or modify data sets, to translate, to perform qualitative data analysis, or to obtain experimental results. We reviewed all AI-assisted work: reported numbers were checked against our experiment logs, citations were checked against the original publications, the proof was checked by the authors, and AI-assisted code was reviewed and tested before use. The authors take full responsibility for the final content of this work, including its text, claims, results, and conclusions.

Ethics Statement

This work studies retrospective prediction using de-identified clinical data. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). The models have not been validated for clinical decision support. Cohort selection and modality availability may introduce biases that limit generalization across patient populations and care settings. Appendix F.2 discusses these considerations and the intended research use of the framework. MIMIC-IV was accessed through PhysioNet under its credentialed data use agreement. The authors have made every effort to conduct this study in accordance with relevant ethical standards and responsible research practices.

Reproducibility Statement

Section 2 and Appendix A.1 specify the model, objectives, and training and inference algorithms. Appendix A.2 documents preprocessing, patient splits, missingness protocols, and hyperparameters. Appendix C.1 reports full predictive results with standard deviations across five seeds; the diagnostic experiments include their protocols in the corresponding appendix sections. Code is available at `https://anonymous.4open.science/r/DECENT-7202/`.

References

Samuel K. Ainsworth, Nicholas J. Foti, Adrian K. C. Lee, and Emily B. Fox. oi-VAE: Output interpretable VAEs for nonlinear group factor analysis. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 119–128, 2018.

Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C. Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6):MSB178124, 2018.

Jiayi Chen and Aidong Zhang. HGMF: Heterogeneous graph-based fusion for multimodal data with incompleteness. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1295–1305, 2020.

Mengxi Chen, Fei Zhang, Zihua Zhao, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Probabilistic conformal distillation for enhancing missing modality robustness. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pp. 36218–36242, 2024.

Biplab Deka, Zifeng Huang, Chad Franzen, Joshua Hibschman, Daniel Afergan, Yang Li, Jeffrey Nichols, and Ranjitha Kumar. Rico: A mobile app dataset for building data-driven design applications. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*, pp. 845–854, 2017.

Yasha Ektefaie, George Dasoulas, Ayush Noori, Maha Farhat, and Marinka Zitnik. Multimodal learning with graphs. *Nature Machine Intelligence*, 5(4):340–350, 2023.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1263–1272, 2017.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

Xing Han, Huy Nguyen, Carl Harris, Nhat Ho, and Suchi Saria. FuseMoE: Mixture-of-experts transformers for fleximodal fusion. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pp. 67850–67900, 2024.

Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. HeMIS: Hetero-modal image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 469–477, 2016.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

Clifford R. Jack, Matt A. Bernstein, Nick C. Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J. Britson, Jennifer L. Whitwell, Chadwick Ward, Anders M. Dale, Joel P. Felmlee, Jeffrey L. Gunter, Derek L. G. Hill, Ron Killiany, Norbert Schuff, Sabrina Fox-Bosetti, Chen Lin, Colin Studholme, Charles S. DeCarli, Gunnar Krueger, Heidi A. Ward, Gregory J. Metzger, Katherine T. Scott, Richard Mallozzi, Daniel Blezek, Joshua Levy, Josef P. Debbins, Adam S. Fleisher, Marilyn Albert, Robert Green, George Bartzokis, Gary Glover, John Mugler, and Michael W. Weiner. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008.

Clifford R. Jack, David A. Bennett, Kaj Blennow, Maria C. Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M. Holtzman, William Jagust, Frank Jessen, Jason Karlawish, Enchi Liu, Jose Luis Molinuevo, Thomas Montine, Creighton Phelps, Katherine P. Rankin, Christopher C. Rowe, Philip Scheltens, Eric Siemers, Heather M. Snyder, Reisa Sperling, Cerise Elliott, Eliezer Masliah, Laurie Ryan, and Nina Silverberg. NIA-AA research framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia*, 14(4):535–562, 2018.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Jaehyuk Jang, Yooseung Wang, and Changick Kim. Towards robust multimodal prompting with missing modalities. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8070–8074, 2024.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV-Note: Deidentified free-text clinical notes. *PhysioNet*, 2023a. Version 2.2.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023b.

Ian T. Jolliffe and Jorge Cadima. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

Arto Klami, Seppo Virtanen, Eemeli Leppaaho, and Samuel Kaski. Group factor analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 26(9):2136–2147, 2015.

Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1):171, 2022.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 3519–3529, 2019.

H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14943–14952, 2023.

Luis A. Leiva, Asutosh Hota, and Antti Oulasvirta. Enrico: A dataset for topic modeling of mobile UI designs. In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services, Adjunct (MobileHCI)*, pp. 1–4, 2020.

Chenxin Li, Xinyu Liu, Cheng Wang, Yifan Liu, Weihao Yu, Jing Shao, and Yixuan Yuan. GTP-4o: Modality-prompted heterogeneous graph learning for omni-modal biomedical representation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 168–187, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 19730–19742, 2023.

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

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10): 1–42, 2024.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pp. 11525–11538, 2020.

Jianxin Ma, Peng Cui, Kun Kuang, Xin Wang, and Wenwu Zhu. Disentangled graph convolutional networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 4212–4221, 2019.

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1930–1939, 2018.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. SMIL: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 2302–2310, 2021.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18156–18165, 2022.

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with LIMoE: The language-image mixture of experts. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pp. 9564–9576, 2022.

Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2001.

Kok Pin Ng, J. Y. Shen, H. J. Chiew, A. S. L. Ng, N. Kandiah, P. Rosa-Neto, and S. Gauthier. White matter hyperintensity as a vascular contribution to the AT(N) framework. *The Journal of Prevention of Alzheimer's Disease*, 10(3):387–400, 2023.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 689–696, 2011.

Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? Uncovering how neural network representations vary with width and depth. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

James Oldfield, Markos Georgopoulos, Grigorios G. Chrysos, Christos Tzelepis, Yannis Panagakis, Mihalis Nicolaou, Jiankang Deng, and Ioannis Patras. Multilinear mixture of experts: Scalable expert specialization through factorization. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pp. 53022–53063, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Jie Peng, Jenna L. Ballard, Mohan Zhang, Sukwon Yun, Jiayi Xin, Qi Long, Yanyong Zhang, and Tianlong Chen. Modalities contribute unequally: Enhancing medical multi-modal learning through adaptive modality token re-balancing. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 48789–48807, 2025.

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 1(8):792–795, 2026.

Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 28435–28445, 2024.

Dhanesh Ramachandram and Graham W. Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *Proceedings of the European Semantic Web Conference (ESWC)*, pp. 593–607, 2018.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 3145–3153, 2017.

Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(47):1393–1434, 2012.

Thomas M. Sutter, Imant Daunhawer, and Julia E. Vogt. Generalized multimodal ELBO. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4971–4980, 2017.

Haotao Wang, Ziyu Jiang, Yuning You, Yan Han, Gaowen Liu, Jayanth Srinivasa, Ramana Kompella, and Zhangyang "Atlas" Wang. Graph mixture of experts: Learning on large-scale graphs with explicit diversity modeling. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pp. 50825–50837, 2023a.

Hu Wang, Congbo Ma, Jianpeng Zhang, Yuan Zhang, Jodie Avery, Louise Hull, and Gustavo Carneiro. Learnable cross-modal knowledge distillation for multi-modal learning with missing modality. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 216–226, 2023b.

Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S. Yu. Heterogeneous graph attention network. In *Proceedings of The Web Conference (WWW)*, pp. 2022–2032, 2019.

Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12176–12185, 2022.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *Transactions on Machine Learning Research*, 2026.

Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Mike Nalls, Jimeng Sun, and Faraz Faghri. Multimodal patient representation learning with missing modalities and labels. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 55538–55557, 2024.

Yiding Yang, Zunlei Feng, Mingli Song, and Xinchao Wang. Factorizable graph convolutional networks. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pp. 20286–20296, 2020.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GAIN: Missing data imputation using generative adversarial nets. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 5689–5698, 2018.

Jiaxuan You, Xiaobai Ma, Yi Ding, Mykel J. Kochenderfer, and Jure Leskovec. Handling missing data with graph representation learning. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pp. 19075–19087, 2020.

Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. Flex-MoE: Modeling arbitrary modality combination via the flexible mixture-of-experts. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pp. 98782–98805, 2024.

Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. M3Care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 2418–2428, 2022.

Qiannan Zhang, Weishen Pan, Zilong Bai, Chang Su, and Fei Wang. Unified insights: Harnessing multi-modal data for phenotype imputation via view decoupling. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pp. 3332–3353, 2024.

A Method and Implementation Details

A.1 Training and Inference Algorithms

Algorithm 1 specifies label-free pretraining and supervised prediction, and Algorithm 2 inductive inference. Phase 1 trains the tokenizer and the router without outcome labels and fixes the partition; Phase 2 freezes them and trains the factorized graph encoder, fusion module, and classifier. At inference, graph propagation is computed within the current test batch, without training samples or test labels, and the partition is not recomputed.

Tokenizer. The tokenizer receives preprocessed modality vectors rather than raw scans or text; tokens of unobserved modalities are not computed. For an observed modality m of sample i, the decoder reconstructs $\hat{\boldsymbol{x}}_m^{(i)} = D_m\big([\boldsymbol{\mu}_{m,1} + \boldsymbol{\sigma}_{m,1} \odot \boldsymbol{\epsilon}_{m,1}; \ldots; \boldsymbol{\mu}_{m,k} + \boldsymbol{\sigma}_{m,k} \odot \boldsymbol{\epsilon}_{m,k}]\big)$ with $\boldsymbol{\epsilon}_{m,t} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where the posterior parameters are those of sample i. With $\mathcal{O}$ the observed sample–modality pairs of a mini-batch and d_m the input dimension of modality m,

$$\begin{aligned} \mathcal{L}_{\text{rec}} &= \frac{1}{|\mathcal{O}|} \sum_{(i,m)\in\mathcal{O}} \frac{1}{d_m} \big\| \boldsymbol{x}_m^{(i)} - \hat{\boldsymbol{x}}_m^{(i)} \big\|_2^2, \\ \mathcal{L}_{\text{KL}} &= \frac{1}{|\mathcal{O}|} \sum_{(i,m)\in\mathcal{O}} \frac{1}{k d_z} \sum_{t=1}^{k} \sum_{l=1}^{d_z} \frac{1}{2} \big(\mu_{m,t,l}^2 + \sigma_{m,t,l}^2 - \log \sigma_{m,t,l}^2 - 1 \big), \end{aligned} \tag{10}$$

where $\mu_{m,t,l}$ and $\sigma_{m,t,l}$ are the l-th entries of the posterior parameters of sample i.

Tokenizer alternatives. *Single-node-per-modality* representations, as in MUSE (Wu et al., 2024), provide one node per modality and cannot assign its sub-signals to separate views. *Domain-specific decomposition*, such as genomic groups and image patches (Li et al., 2024), yields fine-grained units but relies on modality-specific construction rules. *Slot Attention* (Locatello et al., 2020) lets slots compete for input features, but its slots are exchangeable, so a slot index need not denote the same content across patients. *Bottleneck queries with key-axis softmax attention* (Li et al., 2023) are indexed and differentiable but require a training objective of their own; the VAE supplies one through reconstruction, which requires the tokens to retain the information needed to reconstruct the input.

Routing. *Signal summary.* The router input is the masked mean of the edge evidence on edges incident to p_j,

$$\bar{\boldsymbol{e}}_j = \frac{1}{\max\{1, |\mathcal{I}_j|\}} \sum_{i\in\mathcal{I}_j} \boldsymbol{e}_{i,j}^{(0)}, \qquad \mathcal{I}_j = \{i : \boldsymbol{M}[i, m(j)] = 1\}, \tag{11}$$

computed over the current mini-batch in Phase 1 and over all training patients for the fixed partition. Signals whose modality no patient in the mini-batch observes are excluded from $\mathcal{L}_{\text{SSL}}$, as anchors and as partners, and from $\mathcal{L}_{\text{lb}}$.

Gate logit. With the gate input $\boldsymbol{\zeta}_j = \bar{\boldsymbol{e}}_j + \boldsymbol{\eta}_j$, the noisy linear logit is

$$H(\boldsymbol{\zeta}_j)_r = (\boldsymbol{\zeta}_j^\top \boldsymbol{W}_g)_r + \epsilon_{j,r} \cdot \text{softplus}\big((\boldsymbol{\zeta}_j^\top \boldsymbol{W}_{\text{noise}})_r\big), \qquad \epsilon_{j,r} \sim \mathcal{N}(0,1), \tag{12}$$

with $\boldsymbol{W}_g, \boldsymbol{W}_{\text{noise}} \in \mathbb{R}^{d\times R}$. The temperature τ in Eq. 3 is annealed from $\tau_{\max}$ to $\tau_{\min}$ during Phase 1.

Load-balancing penalty. With the per-view importance vector $\sum_j \boldsymbol{a}_j \in \mathbb{R}^R$ aggregated over the signal nodes present in the mini-batch (Shazeer et al., 2017),

$$\mathcal{L}_{\text{lb}} = \text{CV}\Big(\textstyle\sum_j \boldsymbol{a}_j\Big)^2, \qquad \text{CV}(\boldsymbol{v}) = \frac{\text{std}(\boldsymbol{v})}{\text{mean}(\boldsymbol{v}) + \varepsilon}. \tag{13}$$

Co-variation target. For a cross-modal pair with $n = n_{jj'}$ co-observed training patients, let $\tilde{\boldsymbol{K}}_j$ and $\tilde{\boldsymbol{K}}_{j'}$ be the Gram matrices of Eq. 4 with their diagonals set to zero. The unbiased estimator (Song et al., 2012) is

$$\widehat{\text{HSIC}}_u(\boldsymbol{K}_j, \boldsymbol{K}_{j'}) = \frac{1}{n(n-3)} \left[\text{tr}(\tilde{\boldsymbol{K}}_j \tilde{\boldsymbol{K}}_{j'}) + \frac{\mathbf{1}^\top \tilde{\boldsymbol{K}}_j \mathbf{1}\, \mathbf{1}^\top \tilde{\boldsymbol{K}}_{j'} \mathbf{1}}{(n-1)(n-2)} - \frac{2}{n-2} \mathbf{1}^\top \tilde{\boldsymbol{K}}_j \tilde{\boldsymbol{K}}_{j'} \mathbf{1} \right], \tag{14}$$

which requires $n \geq 4$, so $n_{\min} \geq 4$. For linear kernels, every term follows from $\boldsymbol{X}_j^\top \boldsymbol{X}_{j'}$, the row norms, and the column sums in $O(n\, d_z^2)$ time, without forming the Gram matrices. $C_{jj'} = 0$ if either self term is not positive. The plain estimator is biased for small n: for independent tokens with isotropic unit-variance coordinates, $\mathbb{E}\|\boldsymbol{X}_{j'}^\top \boldsymbol{X}_j\|_F^2 \approx n d_z^2$ while $\|\boldsymbol{X}_j^\top \boldsymbol{X}_j\|_F \approx n\sqrt{d_z}$ for $n \gg d_z$, so plain linear CKA is about d_z/n instead of zero. For the first E_w epochs of Phase 1, $\lambda_{\mathrm{SSL}} = 0$ and C is not computed, so the router is trained by $\mathcal{L}_{\mathrm{lb}}$ alone. From then on, at the start of each epoch, the proto-tokens of all training patients are computed with the current encoders in evaluation mode, and Eq. 4 is evaluated for every cross-modal pair with $n_{jj'} \geq n_{\min}$. The matrix C is held constant during the epoch and carries no gradient. Signals outside $\mathcal{J}_{\mathrm{c}}$ do not enter $\mathcal{L}_{\mathrm{SSL}}$ and are placed by $\mathcal{L}_{\mathrm{lb}}$ alone. In Phase 1, the edge evidence that enters the router is computed as $\phi_m(\mathrm{sg}(\boldsymbol{p}_{i,j}))$, where sg stops the gradient, so $\mathcal{L}_{\mathrm{SSL}}$ and $\mathcal{L}_{\mathrm{lb}}$ update ϕ_m, $\boldsymbol{\eta}_j$, $\boldsymbol{W}_g$, and $\boldsymbol{W}_{\mathrm{noise}}$, and E_m and D_m are optimized only by $\mathcal{L}_{\mathrm{rec}} + \beta\, \mathcal{L}_{\mathrm{KL}}$.

Factorized graph encoding and fusion. *Factorized View-GNN.* For each view r and layer ℓ,

$$\begin{aligned} \boldsymbol{h}_v^{(\ell+1,r)} &= \boldsymbol{U}^{(\ell)} \left[\boldsymbol{h}_v^{(\ell,r)} \,\middle\|\, \frac{1}{\max\{1, |\mathcal{N}_r(v)|\}} \sum_{u \in \mathcal{N}_r(v)} \sigma\Big(\boldsymbol{W}_{j(u,v)}^{(\ell)} \boldsymbol{h}_u^{(\ell,r)} + \boldsymbol{O}^{(\ell)} \boldsymbol{e}_{u,v}^{(\ell,r)} \Big) \right], \\ \boldsymbol{e}_{u,v}^{(\ell+1,r)} &= \boldsymbol{P}^{(\ell)} \left[\boldsymbol{h}_u^{(\ell,r)} \,\middle\|\, \boldsymbol{h}_v^{(\ell,r)} \,\middle\|\, \boldsymbol{e}_{u,v}^{(\ell,r)} \right], \end{aligned} \tag{15}$$

where σ is an element-wise nonlinearity, $\|$ denotes concatenation, and $j(u,v)$ is the signal-node index of edge (u,v) (each edge in $\mathcal{E}^{(r)}$ has exactly one signal endpoint), $\mathcal{N}_r(v)$ is v's neighborhood in $\mathcal{G}^{(r)}$, $\boldsymbol{W}_j^{(\ell)}$ is the basis combination from Eq. 7, and $\boldsymbol{U}^{(\ell)}, \boldsymbol{O}^{(\ell)}, \boldsymbol{P}^{(\ell)}$ are aggregator, edge-transformation, and edge-update matrices shared across views. Empty-neighborhood message aggregates are defined as zero.

View-specific readout and fusion. Message passing starts from the sample-node and signal-node initializations in Section 2.4, with proto-token evidence as the initial observed edge attributes. The final sample-node state is mapped to a per-view representation:

$$\boldsymbol{h}_i^{(r)} = \mathrm{Readout}_r\Big(\boldsymbol{h}_i^{(L,r)}\Big) = \mathrm{Lin}_{r,2}\Big(\mathrm{Dropout}\Big(\mathrm{GELU}\Big(\mathrm{Lin}_{r,1}\Big(\mathrm{LN}_r\Big(\boldsymbol{h}_i^{(L,r)}\Big)\Big)\Big)\Big)\Big), \tag{16}$$

where LN_r and the affine maps $\mathrm{Lin}_{r,1}$ and $\mathrm{Lin}_{r,2}$ have view-specific parameters. The factorized experts thus share aggregation and edge-processing components while retaining view-specific message bases and readouts. For each sample, $\mathcal{K}_i \subseteq [R]$ contains the views with at least one observed signal; the retain-one protocol ensures $\mathcal{K}_i \neq \emptyset$. Attention scores are normalized over $\mathcal{K}_i$ only:

$$\begin{aligned} \omega_i^{(r)} &= \boldsymbol{w}^\top \tanh\Big(\boldsymbol{W}_{\mathrm{attn}}\, \mathrm{LN}(\boldsymbol{h}_i^{(r)}) + \mathbf{b}_{\mathrm{attn}}\Big), \qquad \alpha_i^{(r)} = \frac{\exp(\omega_i^{(r)})}{\sum_{r' \in \mathcal{K}_i} \exp(\omega_i^{(r')})}, \quad r \in \mathcal{K}_i, \\ \boldsymbol{z}_i &= \phi_{\mathrm{fuse}}\left(\sum_{r \in \mathcal{K}_i} \alpha_i^{(r)}\, \boldsymbol{h}_i^{(r)} \right), \end{aligned} \tag{17}$$

where LN is layer normalization and ϕ_{fuse} is a shared post-fusion block of LayerNorm, GELU, and dropout. Views outside $\mathcal{K}_i$ receive zero weight. A view with no assigned signal has no edges, is not encoded, and is never in $\mathcal{K}_i$. Views without assigned signals occur on MIMIC-IV in most seeds (Appendix D.2) and whenever $R > J$ (Appendix D.5).

Training and inference procedures. Let Θ_{VAE} denote the encoders E_m and decoders D_m, and Θ_{route} the projections ϕ_m, the signal embeddings $\boldsymbol{\eta}_j$, and the gate parameters $\{\boldsymbol{W}_g, \boldsymbol{W}_{\mathrm{noise}}\}$. We use Θ_{enc} for the graph encoder parameters, including learned signal-node initial embeddings, view-specific message bases, shared aggregation and edge-processing parameters, and view-specific readout parameters. The fusion parameters Θ_{fus} include attention scoring and its input normalization, together with the shared post-fusion block ϕ_{fuse} in Eq. 17; Θ_{cls} denotes the classifier parameters. Algorithm 2 sets all modules to evaluation mode so that readout and fusion dropout are disabled.

Algorithm 1 DECENT: two-phase training

Require: Training data $\{(\boldsymbol{x}^{(i)}, y^{(i)}, \boldsymbol{M}[i,:])\}$; views R; groups per modality k; top-K; minimum count $n_{\min}$; warm-up epochs E_w

Phase 1: label-free pretraining
repeat
  if the warm-up of E_w epochs is completed **then**
    Compute $\{\boldsymbol{p}_{i,j}\}$ for all training patients; compute $C_{jj'}$ for cross-modal pairs with $n_{jj'} \geq n_{\min}$. ▷ Eqs. 4, 14
  end if
  for each training mini-batch $(\boldsymbol{x}, \boldsymbol{M})$ **do**
    Encode observed modalities; reconstruct from reparameterized samples. ▷ Eq. 10
    $\boldsymbol{e}_{i,j}^{(0)} \leftarrow \phi_m(\mathrm{sg}(\boldsymbol{p}_{i,j}))$; $\{\bar{\boldsymbol{e}}_j\} \leftarrow \textsc{SignalSummary}(\{\boldsymbol{e}_{i,j}^{(0)}\}, \boldsymbol{M})$. ▷ Eq. 11
    $\{\boldsymbol{a}_j\} \leftarrow \textsc{Gate}(\{\bar{\boldsymbol{e}}_j + \boldsymbol{\eta}_j\})$ with routing noise and the Phase-1 temperature schedule. ▷ Eq. 3
    Update $\Theta_{\mathrm{VAE}}, \Theta_{\mathrm{route}}$ using $\nabla\mathcal{L}_{\mathrm{P1}}$, with $\lambda_{\mathrm{SSL}} = 0$ during warm-up. ▷ Eqs. 5, 8
  end for
until the Phase-1 epoch budget is completed
Phase transition
Freeze $\Theta_{\mathrm{VAE}}, \Theta_{\mathrm{route}}$; compute $\{\bar{\boldsymbol{e}}_j\}$ over all training patients.
Set $\epsilon_{j,r} = 0$; fix $g(j) = \arg\max_r H(\bar{\boldsymbol{e}}_j + \boldsymbol{\eta}_j)_r$ and $\hat{a}_{j,r}$.
Phase 2: task training
Initialize $\Theta_{\mathrm{enc}}, \Theta_{\mathrm{fus}}, \Theta_{\mathrm{cls}}$.
repeat
  Fetch a training mini-batch $(\boldsymbol{x}, \mathbf{y}, \boldsymbol{M})$.
  Compute $\{\boldsymbol{e}_{i,j}^{(0)}\}$ with the frozen tokenizer. ▷ Eq. 2
  Construct the view graphs $\{\mathcal{G}^{(r)}\}_{r=1}^{R}$ under g and available-view sets $\{\mathcal{K}_i\}$. ▷ Eq. 6
  For each view, compute $\boldsymbol{h}_i^{(L,r)}$ by factorized message passing. ▷ Eqs. 7, 15
  $\boldsymbol{h}_i^{(r)} \leftarrow \mathrm{Readout}_r(\boldsymbol{h}_i^{(L,r)})$ for each sample and view.
  Compute $\boldsymbol{z}_i$ by availability-aware fusion; $\hat{y}_i \leftarrow \mathrm{MLP}(\boldsymbol{z}_i)$. ▷ Eq. 17
  Update $\Theta_{\mathrm{enc}}, \Theta_{\mathrm{fus}}, \Theta_{\mathrm{cls}}$ using $\nabla\mathcal{L}_{\mathrm{P2}}$. ▷ Eq. 9
until the Phase-2 epoch budget is completed

Algorithm 2 DECENT: inductive inference per test batch

Require: Test batch $\{(\boldsymbol{x}^{(i)}, \boldsymbol{M}[i,:])\}_{i=1}^{B}$, with $\sum_m \boldsymbol{M}[i,m] \geq 1$ for every sample; trained parameters $\Theta_{\mathrm{VAE}}, \Theta_{\mathrm{route}}, \Theta_{\mathrm{enc}}, \Theta_{\mathrm{fus}}, \Theta_{\mathrm{cls}}$; fixed partition g

Set all modules to evaluation mode; keep their parameters fixed.
Compute $\{\boldsymbol{e}_{i,j}^{(0)}\}$ from the posterior means. ▷ Eq. 2
Construct $\{\mathcal{G}^{(r)}\}_{r=1}^{R}$ under g using only the test batch and determine $\{\mathcal{K}_i\}$.
for $r = 1, \ldots, R$ **do**
  Compute $\boldsymbol{h}_i^{(L,r)}$ by factorized message passing on $\mathcal{G}^{(r)}$.
  $\boldsymbol{h}_i^{(r)} \leftarrow \mathrm{Readout}_r(\boldsymbol{h}_i^{(L,r)})$ for each sample.
end for
Compute $\boldsymbol{z}_i$ by availability-aware fusion over $\mathcal{K}_i$; $\hat{y}_i \leftarrow \mathrm{MLP}(\boldsymbol{z}_i)$. ▷ Eq. 17
return $\{\hat{y}_i\}_{i=1}^{B}$

A.2 IMPLEMENTATION DETAILS

Environment. We implement DECENT in Python 3.11 with PyTorch 2.5.1, CUDA 12.1, cuDNN 9.1, and PyTorch Geometric 2.7 (Fey & Lenssen, 2019) with mean aggregation. Experiments run on NVIDIA RTX 3090 GPUs with 24 GB memory. Each training run uses a single GPU; multiple GPUs are used only to parallelize hyperparameter search.

Data splits and missingness. We follow the feature-preprocessing pipeline of Flex-MoE (Yun et al., 2024). ADNI uses patient-aligned clinical, biospecimen, MRI-derived, and genomic feature

vectors. The supplied patient split contains 1,667 / 352 / 361 subjects for train / validation / test (2,380 total). MIMIC-IV uses pre-extracted lab/vital, note, and code feature matrices aligned to subject IDs and binary one-year mortality labels, with 6,303 / 1,351 / 1,349 subjects (9,003 total). These subject splits and the original modality-availability masks are inputs to the evaluation pipeline; following the preprocessing does not make the reported cohorts identical to the cohorts in the Flex-MoE paper. The evaluation concerns retrospective prediction from benchmark features, not prospective risk estimation at a validated clinical index time. All primary comparisons and component ablations use the same fixed splits and missingness manifests across five paired seeds. Means and SDs describe training-run variability on these fixed cohorts.

For each originally observed modality, we preassign an independent score $u_{im} \sim \mathrm{Uniform}(0, 1)$ and mask it when $u_{im} < \rho$. If the masking removes all observed modalities of sample i, we restore the one with the largest u_{im}; we neither resample scores nor exclude the sample. The same scores are reused across missing rates, so the restoration candidate is fixed and the sets of removed modalities are nested as ρ increases. This rule is used in every split. The final masks are not independent Bernoulli masks, and ρ denotes the pre-restoration removal probability rather than the realized missing fraction.

Hyperparameter selection. For each dataset, we screen a default configuration, three targeted variants, and 20 Sobol-sampled configurations at $\rho = 0.5$ using seed 0, with 100 epochs per phase, and rank them by validation macro-F1. The two highest-ranked configurations are also evaluated with 200 Phase 1 epochs. We then evaluate the default and the three highest-ranked non-default configuration–budget pairs using seeds 0, 1, and 2. The final configuration is selected by mean validation macro-F1, with mean validation macro-AUC and then configuration identifier used to break ties. The selected configuration for each dataset is held fixed across all four values of ρ. Test results are not used for this selection. The number of views R controls the granularity of the learned signal partition and is not tied to the number of input modalities. Table 3 lists the selected configuration.

Training protocol. We use AdamW in both training phases. Phase 1 trains the VAE, the projections ϕ_m, the signal embeddings $\boldsymbol{\eta}_j$, and the router for 100 epochs. Separate learning rates are used for the VAE and the routing parameters, where the latter comprise ϕ_m, $\boldsymbol{\eta}_j$, and the gate and noise matrices $\boldsymbol{W}_g$ and $\boldsymbol{W}_{\mathrm{noise}}$. The VAE is updated only by reconstruction and KL losses; posterior means are detached before projection, so SSL and load-balancing gradients update only the routing parameters. The first E_w epochs optimize the VAE losses and $\mathcal{L}_{\mathrm{lb}}$ only; afterwards, C is recomputed without gradients at the start of every epoch using co-observed training patients, and $\mathcal{L}_{\mathrm{SSL}}$ is added (Eqs. 5, 13, 8). No outcome labels are used in Phase 1. The fixed partition g is then computed from training-set summaries with routing noise disabled and stored as one-hot assignments. Phase 2 freezes the VAE, ϕ_m, $\boldsymbol{\eta}_j$, and the router, constructs view-specific graphs under g, and trains the View-GNN, fusion module, and classifier for 100 epochs using the supervised task loss (Eq. 9). The partition is not recomputed during Phase 2 or inference. In both phases, training applies modality-edge dropout with probability 0.1, retaining at least one modality whenever any is observed; CKA refreshes and final partition fitting do not apply this augmentation. We select the Phase 2 checkpoint with the best validation macro-F1 without early stopping. For MIMIC-IV, the binary classification threshold is selected on the validation set at each epoch, and the threshold associated with the selected checkpoint is reused at test time.

Component-comparison protocol. Table 2(a) compares the indexed VAE with a single token per modality ($k = 1$, one group of dimension 32, the total kd_z of the main configuration), Slot Attention (Locatello et al., 2020) with stochastic slot initialization in Phase 1 and fixed draws afterwards, PCA (Jolliffe & Cadima, 2016), and FA-inspired decomposition. Classical bases are fitted on the training split; groups of their factor scores replace the posterior means as proto-tokens, and ϕ_m, $\boldsymbol{\eta}_j$, and the router are trained with the same Phase-1 objective. Block (b) compares learned (fixed), learned (dynamic), modality-based, balanced random, and single-view partitions. Multi-view controls share the pretrained tokenizer, signal-node and view counts, and downstream model; only the assignment rule changes. Modality-based partitioning uses one view per modality on ADNI ($R = 4$) and two per modality on MIMIC-IV ($R = 6$), splitting each modality's four groups into pairs. Random partitions are fixed within a run and balanced across views. Two partition draws are averaged within each seed before computing five-seed means and standard deviations. Learned (dynamic) applies the frozen router to the summaries of each mini-batch in Phase 2 and of each test batch at inference. Block (c) compares the frozen two-phase schedule with Phase-2 fine-tuning of

Table 3: **Hyperparameter configuration** for the full DECENT model, selected by a prespecified search using validation macro-F1.

Hyperparameter	ADNI	MIMIC-IV
Optimization		
Batch size	128	256
Learning rate (Phase 1: VAE)	3×10^{-4}	10^{-3}
Learning rate (Phase 1: routing)	3×10^{-4}	3×10^{-4}
Learning rate (Phase 2)	5×10^{-5}	4×10^{-5}
Weight decay	10^{-4}	10^{-6}
Epochs (Phase 1 / Phase 2)	100 / 100	100 / 100
Warm-up epochs E_w	10	40
Early stopping	Disabled	Disabled
Modality-edge dropout (training)	0.10	0.10
Tokenizer		
Groups per modality k	4	4
Group dimension d_z	8	8
Encoder / decoder hidden widths	128 / 128	128 / 128
Encoder / decoder hidden layers	1 / 1	1 / 1
Edge-evidence dimension d	384	384
KL weight β	10^{-4}	10^{-2}
Routing		
Number of views R	4	6
Routing top-K	3	2
Routing temperature $\tau_{\max}/\tau_{\min}$	2.0 / 0.5	0.5 / 0.1
SSL weight λ_{SSL}	10.0	10.0
Co-variation temperature τ_c	0.20	0.05
Routing similarity temperature τ_r	2.0	1.0
Load-balancing weight λ_{lb}	0.20	0.10
Minimum co-observed count $n_{\min}$	32	32
Graph encoder and fusion		
View GNN type (edge-aware)	GraphSAGE	GraphSAGE
View GNN layers L	3	2
View GNN dropout	0.10	0.25
Fusion dropout	0.10	0.10
Classifier hidden width	256	256
Classifier dropout	0.30	0.30
Task loss		
Loss type	Cross-entropy	BCE (logits)
Positive-class weight	—	2.0

the VAE, ϕ_m, and the router after the same Phase 1, and with task-only training of the whole model for the combined epoch budget. Block (d) varies encoder sharing while retaining the view graphs and per-view readouts. Shared GNNs share message and other GNN weights; factorized experts retain view-specific message bases; independent GNNs use separate message-passing parameters. The routing-target controls (Appendix D.3) change only the routing target, except for the spectral clustering control, which replaces the router. The reruns use the same initial parameters for shared components and the same batch order.

Task gradients through routing. Task-only training and Phase-2 fine-tuning retain hard graph membership and one-hot message-basis selection. Their straight-through estimator passes task gradients through soft assignments in the basis combination; the graph-membership selection itself is not differentiated.

Partition analysis protocol. *Axis labels.* For group t of modality m and input feature f, the attribution $c_{m,t,f}$ is the mean over training patients observing m of $\|x_{i,m,f}\, \partial \boldsymbol{\mu}^{(i)}_{m,t} / \partial x_{i,m,f}\|_2$, a gradient-times-input attribution (Shrikumar et al., 2017) with the norm taken over the d_z coordinates, computed with the final Phase-1 encoder. The score of axis a is the sum of $c_{m,t,f}$ over the features

annotated with a (Appendix E.1), and the group takes the axis with the largest score. Groups whose largest attribution falls on unannotated features are excluded from A_{axis} and A_{mod}. The labels are computed once, from training patients only, and held fixed across checkpoints, so the curves in Figure 4 track the partition rather than the labels.

Alignment trajectory. Every 5 epochs of Phase 1 and of the first 100 epochs of task-only training, the current router is applied to the summaries over all training patients without routing noise, and the resulting partition is scored by A_{axis} and A_{mod}. The reference line marks the expected ARI of random partitions, which is zero.

Held-out co-variation contrast. For each partition used in Phase 2 (learned, the balanced random draws, and the routing-target controls), debiased CKA (Eq. 4) is computed between every cross-modal pair of signals on the validation patients who observe both modalities, using the final Phase-1 proto-tokens of the corresponding run and the same $n_{\min}$. Δ_C is the mean over same-view pairs minus the mean over different-view pairs. Δ_C is undefined for the modality-based partition, which has no cross-modal pair within a view.

Within-modality group similarity. For each modality, debiased CKA (Eq. 4) is computed between the posterior means of every pair of its groups on the test patients who observe it, using the final Phase-1 encoder, and each pair is marked as same-view or different-view under the fixed partition.

Spectral clustering control. The affinity matrix is the co-variation matrix C of the last Phase-1 epoch, with zero entries within modalities and for ineligible pairs. The rows of the eigenvectors of the normalized Laplacian for the R smallest eigenvalues are normalized to unit length and clustered into R views by k-means (Ng et al., 2001); signals without an eligible partner are assigned to the smallest view. Phase 2 is then trained with this fixed partition.

Cross-seed agreement. For each dataset and routing target, the cross-seed ARI is the mean ARI between the fixed partitions of all pairs of the five seeds. Group indices are comparable across seeds only through the input features they encode, so each seed's groups are matched to the groups of a reference seed by maximizing the summed cosine similarity of their attribution vectors $(c_{m,t,f})_f$ within each modality (Hungarian matching (Kuhn, 1955)) before computing the ARI.

B Formal Statement and Proof

B.1 What Cross-Modal Dependence Reflects

Proposition 1 (What cross-modal dependence reflects)**.** *Assume Eq. 1, with $u_1, \dots, u_M$ mutually independent and independent of $(z_1, \dots, z_Q)$. Let $a = \psi(\boldsymbol{x}_m)$ and $b = \psi'(\boldsymbol{x}_{m'})$ be features of single modalities, and let $S \subseteq \mathcal{Q}_m$ and $S' \subseteq \mathcal{Q}_{m'}$ be process sets such that a is a function of (z_S, u_m) and b a function of $(z_{S'}, u_{m'})$, where $z_S = (z_q)_{q \in S}$.*

(a) If $m \neq m'$ and z_S is independent of $z_{S'}$, then a and b are independent. Given $(z_1, \dots, z_Q)$, features of different modalities are always conditionally independent.

(b) If $m = m'$, a and b can be dependent even when $S \cap S' = \emptyset$ and z_S is independent of $z_{S'}$, and even given $(z_1, \dots, z_Q)$.

Proof. All random variables are defined on one probability space, and each modality follows Eq. 1, so $\boldsymbol{x}_m$ is a function of $(z_{\mathcal{Q}_m}, u_m)$. By assumption, $u_1, \dots, u_M$ are mutually independent and the vector $(u_1, \dots, u_M)$ is independent of $z = (z_1, \dots, z_Q)$. We write $X \perp Y$ for independence.

Part (a). Let $a = \tilde{\psi}(z_S, u_m)$ and $b = \tilde{\psi}'(z_{S'}, u_{m'})$ with $m \neq m'$ and $z_S \perp z_{S'}$. The pair $(z_S, z_{S'})$ is independent of $(u_m, u_{m'})$, and $u_m \perp u_{m'}$, so the joint law of $(z_S, z_{S'}, u_m, u_{m'})$ factorizes into the laws of z_S, $z_{S'}$, u_m, and $u_{m'}$. Hence $(z_S, u_m) \perp (z_{S'}, u_{m'})$, and $a \perp b$ as functions of independent variables. Given z, the feature a is a function of u_m and b a function of $u_{m'}$. As $(u_m, u_{m'})$ is independent of z and $u_m \perp u_{m'}$, the two variables remain independent given z, so a and b are conditionally independent given z.

Part (b). Let modality m express two processes with scalar, independent z_1 and z_2 of finite variance, and let $\boldsymbol{x}_m = (z_1 + u_m,\ z_2 + u_m)$ with scalar u_m of positive, finite variance, independent of

(z_1, z_2). This construction is an instance of Eq. 1 with $f_{m,1}(z) = (z, 0)$, $f_{m,2}(z) = (0, z)$, and $f_{m,0}(u) = (u, u)$. The features $a = z_1 + u_m$ and $b = z_2 + u_m$ satisfy $S = \{1\}$, $S' = \{2\}$, and $z_S \perp z_{S'}$, yet $\mathrm{Cov}(a, b) = \mathrm{Var}(u_m) > 0$, so they are dependent. Given (z_1, z_2), both are shifts of u_m and remain dependent. □

C Additional Predictive Evaluation

C.1 Full Predictive Results

Tables 4 and 5 report the full metric breakdown (accuracy, recall, macro-F1, macro-AUC) across all four missing rates $\rho \in \{0, 0.2, 0.5, 0.7\}$ for the methods in Table 1. The macro-F1 and macro-AUC means match the main table; accuracy and recall appear here only.

Table 4: **Full results on ADNI** (3-class AD diagnosis, %). ρ=0 denotes no additional masking. Each entry shows the mean with the sample standard deviation in parentheses below. All results are averaged over five seeds. Best values are bold. Ours is shaded blue. Values lower than ours are gray.

	ρ=0				ρ=0.2				ρ=0.5				ρ=0.7			
Method	Acc	Rec	F1	AUC	Acc	Rec	F1	AUC	Acc	Rec	F1	AUC	Acc	Rec	F1	AUC
Graph-based																
GRAPE	63.9 (1.61)	**63.7** (1.16)	63.3 (1.40)	80.4 (0.67)	63.3 (1.56)	61.3 (1.75)	**62.1** (1.77)	78.0 (0.89)	55.5 (1.34)	53.0 (1.19)	53.6 (1.27)	74.6 (0.60)	53.3 (1.47)	50.7 (1.28)	52.0 (1.20)	71.5 (0.68)
GTP-4o	60.3 (1.94)	59.0 (2.65)	55.9 (3.58)	74.5 (2.71)	58.1 (2.00)	55.5 (2.51)	54.8 (2.45)	71.7 (2.75)	53.6 (2.70)	50.2 (3.61)	49.9 (2.75)	69.0 (2.17)	52.3 (2.34)	49.1 (2.63)	49.6 (2.43)	68.1 (2.18)
HGMF	53.5 (6.64)	53.2 (5.90)	51.8 (8.21)	69.4 (4.20)	51.7 (2.37)	50.7 (4.07)	49.3 (3.12)	67.6 (1.94)	50.1 (2.31)	47.1 (2.32)	47.1 (2.38)	64.5 (1.67)	52.6 (2.45)	48.5 (2.55)	49.1 (2.55)	66.3 (2.28)
M3Care	62.2 (1.67)	61.3 (1.70)	61.3 (1.52)	77.2 (1.05)	57.3 (1.52)	54.7 (1.55)	55.2 (1.64)	74.8 (0.51)	55.8 (1.90)	51.8 (1.73)	52.9 (1.70)	72.4 (0.81)	55.0 (1.67)	48.6 (1.43)	49.2 (1.41)	70.6 (0.54)
MUSE	64.1 (1.30)	62.7 (1.12)	**63.7** (0.91)	**81.3** (0.60)	61.1 (1.73)	60.0 (1.44)	60.2 (1.90)	77.8 (1.34)	58.9 (1.46)	55.4 (1.50)	56.1 (1.27)	75.4 (0.66)	57.0 (1.93)	52.8 (1.84)	53.6 (2.19)	72.4 (1.00)
Non-graph																
AMC	54.9 (2.80)	50.9 (2.90)	49.5 (3.46)	68.4 (2.61)	52.1 (2.21)	47.0 (2.91)	45.5 (3.24)	66.2 (2.12)	47.5 (2.36)	40.9 (3.11)	40.4 (4.73)	61.1 (3.57)	44.9 (1.67)	37.1 (1.86)	34.9 (3.48)	57.5 (2.54)
Flex-MoE	61.1 (2.27)	60.1 (2.24)	59.9 (2.19)	78.2 (1.32)	54.1 (1.68)	51.6 (1.95)	50.8 (1.55)	71.9 (0.94)	52.3 (2.66)	48.5 (1.73)	48.7 (1.85)	69.1 (2.02)	51.5 (1.36)	47.5 (1.86)	47.8 (1.53)	65.9 (1.25)
SimMLM	62.9 (1.46)	62.1 (1.71)	62.3 (1.47)	79.1 (1.53)	58.3 (2.47)	56.8 (1.94)	56.9 (1.86)	75.2 (1.03)	54.4 (3.00)	52.0 (2.62)	52.7 (2.88)	71.4 (2.49)	53.8 (1.83)	50.4 (1.42)	51.3 (1.35)	70.6 (1.52)
Ours																
DECENT	**65.5** (1.47)	63.1 (1.53)	63.0 (1.31)	80.6 (1.51)	**64.3** (1.74)	**62.0** (1.43)	61.9 (2.14)	**79.6** (1.30)	**61.3** (1.82)	**57.6** (1.27)	**57.9** (1.01)	**77.8** (1.54)	**60.2** (1.69)	**54.9** (1.04)	**55.8** (0.98)	**74.3** (2.09)

DECENT attains the highest macro-F1 and macro-AUC when averaged across the four settings on each dataset, and the full tables locate the exceptions. MUSE has higher macro-F1 and macro-AUC on ADNI and higher macro-AUC on MIMIC-IV without additional masking, and GRAPE has higher ADNI macro-F1 at $\rho = 0$ and 0.2. Under induced missingness, DECENT leads macro-AUC on both datasets and macro-F1 on MIMIC-IV at all three rates. Accuracy does not follow the macro metrics on MIMIC-IV: AMC has the highest accuracy in three of the four settings but the lowest macro-AUC without additional masking, so accuracy alone would misrank the methods.

AMC evaluation. AMC results are five-seed evaluations at each missingness setting, and prediction-level checks found no single-class runs among them on either dataset.

C.2 Non-Medical Domain Extension

We examine a non-medical application on ENRICO (Leiva et al., 2020), a curated subset of the Rico mobile UI dataset (Deka et al., 2017). We retain 1,458 screens from 20 design classes with all four modalities present and readable images. A class-stratified, screen-level 70/15/15 split gives 1,020 training, 219 validation, and 219 test screens, shared across five training seeds. The split does not

Table 5: **Full results on MIMIC-IV** (binary mortality prediction, %). ρ=0 denotes no additional masking. Each entry shows the mean with the sample standard deviation in parentheses below. All results are averaged over five seeds. Best values are bold. Ours is shaded blue. Values lower than ours are gray.

	ρ=0				ρ=0.2				ρ=0.5				ρ=0.7			
Method	Acc	Rec	F1	AUC	Acc	Rec	F1	AUC	Acc	Rec	F1	AUC	Acc	Rec	F1	AUC
Graph-based																
GRAPE	80.7 (2.25)	**63.6** (3.68)	57.0 (1.65)	74.7 (4.81)	80.4 (1.22)	**59.2** (0.82)	55.7 (1.02)	69.9 (0.45)	83.2 (1.01)	55.3 (1.25)	54.6 (0.90)	67.8 (0.98)	81.6 (1.59)	**57.4** (1.12)	55.5 (1.18)	67.1 (0.92)
GTP-4o	85.7 (2.63)	56.8 (2.66)	55.8 (1.14)	71.4 (1.45)	85.3 (1.54)	55.1 (1.43)	54.4 (1.10)	68.4 (1.60)	85.8 (2.06)	54.4 (1.90)	54.0 (1.58)	63.6 (2.65)	85.8 (3.46)	55.5 (1.48)	54.9 (1.49)	60.3 (4.20)
HGMF	86.2 (1.20)	56.5 (1.31)	55.9 (0.83)	72.8 (1.05)	82.5 (1.74)	55.1 (3.55)	53.4 (2.42)	65.1 (2.90)	86.4 (1.36)	52.2 (1.78)	52.0 (1.39)	58.8 (1.06)	86.3 (1.88)	53.1 (1.32)	53.0 (1.04)	61.7 (3.09)
M3Care	85.2 (5.24)	56.7 (6.25)	54.1 (3.22)	66.7 (14.63)	87.4 (3.90)	51.4 (1.49)	51.2 (1.82)	50.1 (10.84)	85.0 (4.27)	51.4 (2.55)	50.9 (2.89)	47.8 (6.24)	86.9 (2.81)	52.2 (2.66)	51.9 (3.06)	48.3 (4.30)
MUSE	84.6 (2.89)	60.6 (2.57)	59.6 (1.99)	**76.0** (0.83)	84.0 (2.34)	58.6 (3.23)	56.3 (1.97)	71.1 (1.48)	86.0 (2.14)	56.3 (2.56)	55.8 (1.75)	69.0 (1.68)	85.2 (3.73)	57.1 (3.00)	55.6 (1.44)	65.9 (7.97)
Non-graph																
AMC	**91.0** (1.70)	54.8 (1.43)	55.2 (1.44)	63.5 (2.48)	**89.9** (1.46)	54.2 (1.46)	54.6 (1.15)	59.2 (5.20)	87.3 (1.68)	52.7 (1.52)	53.0 (1.76)	55.3 (3.70)	**88.0** (1.55)	52.8 (1.17)	53.5 (1.41)	56.4 (5.47)
Flex-MoE	86.2 (2.25)	56.6 (3.34)	55.8 (1.92)	71.3 (2.52)	84.2 (2.57)	56.4 (2.62)	54.8 (1.48)	67.4 (2.18)	86.6 (2.05)	54.8 (1.45)	54.5 (1.07)	65.1 (1.96)	86.2 (1.10)	54.2 (1.02)	54.0 (0.78)	63.3 (1.27)
SimMLM	86.4 (1.49)	56.3 (1.93)	55.8 (1.44)	72.7 (0.79)	87.9 (1.31)	54.3 (1.26)	54.6 (1.45)	70.0 (0.85)	83.9 (1.28)	56.5 (1.07)	55.0 (0.81)	66.3 (1.41)	85.9 (2.62)	54.5 (0.83)	54.1 (0.56)	64.6 (1.57)
Ours																
DECENT	90.1 (1.41)	60.3 (0.98)	**59.8** (1.15)	75.9 (1.26)	89.7 (1.22)	58.6 (1.59)	**58.7** (1.48)	**73.3** (1.17)	**88.1** (1.84)	**57.8** (1.47)	**57.4** (1.04)	**71.1** (1.02)	86.8 (0.63)	**57.4** (0.75)	**56.5** (0.38)	**69.2** (1.03)

group screens by app, so this experiment evaluates held-out screens rather than generalization to unseen apps.

Screenshot and wireframe inputs are RGB images resized to 48×48, scaled by $1/255$, and flattened into 6,912-dimensional vectors each. UI hierarchy features combine 16 tree statistics and 128 widget-class frequencies (144 dimensions); app metadata combines 512 text-hashing dimensions and five numeric features (517 dimensions). Hierarchy and metadata features are standardized using training-set means and standard deviations. No pretrained CNN or ViT features are used. We use the same retain-one missingness protocol across the training, validation, and test splits (Section 3.1).

Table 6 tests whether the framework transfers to a non-clinical task whose modalities differ in form and dimension. The pattern of Table 1 recurs: with complete inputs DECENT ranks second in macro-F1, and under induced missingness ($\rho \in \{0.2, 0.5, 0.7\}$) it attains the best or tied-best macro-F1 and the highest macro-recall, with the highest macro-AUC in all four settings. Accuracy is highest for SimMLM in three of the four settings.

C.3 SENSITIVITY TO TEST-BATCH COMPOSITION

The tokenizer, router, and partition are fixed at inference (Algorithm 2), so batch composition affects a prediction only through message passing among the patients of the batch. We evaluate prediction sensitivity at $\rho = 0.5$ over five paired seeds, with 20 neighbor re-draws per intermediate batch size. The operating sizes are 128 for ADNI and 256 for MIMIC-IV, as in Section 3.1. Prediction flip counts patients whose predicted class differs from the full-batch reference. Solo evaluation has no neighbor re-draw; full-batch composition is fixed.

Table 7 tests a property stated in Section 2.4: because the partition is fixed, the test batch affects a prediction only through message passing. At the operating batch sizes, fewer than 0.5% of predictions differ from full-batch evaluation, macro-F1 and macro-AUC stay within 0.11 pp of their full-batch values, and variation decreases monotonically with batch size. Evaluating each patient alone removes all propagated evidence and lowers macro-F1 relative to the operating size, by 0.2 pp on ADNI and

Table 6: **Full results on ENRICO** (20-class Android screen design classification, %). ρ=0 denotes complete inputs. Recall, F1, and AUC use macro averaging. Each entry shows the mean over five seeds with the sample standard deviation in parentheses below. Best values are bold. Ours is shaded blue. Values lower than ours are gray.

	ρ=0				ρ=0.2				ρ=0.5				ρ=0.7			
Method	Acc	Rec	F1	AUC	Acc	Rec	F1	AUC	Acc	Rec	F1	AUC	Acc	Rec	F1	AUC
Graph-based																
GRAPE	12.1 (3.27)	11.4 (2.20)	6.1 (1.69)	55.3 (2.34)	13.7 (2.41)	10.0 (2.33)	5.1 (1.07)	56.6 (2.98)	9.1 (3.11)	7.2 (1.78)	4.7 (1.08)	57.8 (3.38)	12.5 (3.01)	10.8 (1.84)	5.6 (1.66)	56.2 (2.06)
GTP-4o	46.4 (3.57)	36.4 (2.93)	**35.5** (3.02)	81.9 (1.70)	42.4 (2.41)	32.7 (1.09)	32.4 (1.37)	80.6 (2.61)	33.6 (2.11)	28.1 (2.24)	**28.5** (2.32)	72.9 (2.06)	27.4 (1.38)	22.6 (1.32)	22.1 (1.45)	67.7 (2.11)
HGMF	47.5 (1.04)	35.0 (1.77)	33.9 (2.06)	76.9 (2.89)	32.0 (2.12)	23.6 (1.27)	23.2 (1.41)	65.1 (3.34)	26.2 (2.48)	13.4 (1.26)	13.0 (1.44)	58.9 (3.11)	24.8 (1.88)	17.9 (1.12)	17.3 (1.37)	59.9 (3.13)
M3Care	34.1 (11.20)	25.4 (8.22)	24.1 (7.93)	71.9 (7.23)	33.9 (7.57)	25.1 (6.51)	24.5 (6.58)	72.0 (7.57)	28.1 (4.00)	19.4 (3.29)	19.6 (3.29)	69.1 (2.40)	26.1 (1.46)	18.8 (4.66)	18.0 (4.91)	64.3 (1.66)
MUSE	34.8 (2.35)	30.3 (1.85)	27.7 (2.48)	76.9 (1.78)	30.2 (2.28)	27.0 (2.71)	24.5 (2.06)	76.4 (1.17)	24.4 (2.95)	20.7 (3.85)	18.1 (3.06)	68.2 (1.74)	17.2 (2.09)	15.4 (2.25)	13.3 (2.37)	64.0 (1.61)
Non-graph																
AMC	43.2 (2.36)	25.0 (3.35)	22.7 (3.34)	74.2 (1.54)	37.9 (1.41)	21.7 (2.98)	19.7 (3.14)	71.2 (1.79)	28.0 (1.54)	15.0 (1.63)	14.1 (2.03)	67.5 (2.16)	21.0 (1.52)	9.2 (1.90)	7.8 (1.77)	64.1 (2.03)
Flex-MoE	42.1 (2.66)	32.1 (2.61)	29.8 (2.13)	80.1 (1.72)	32.7 (2.57)	25.5 (1.83)	23.8 (2.25)	73.6 (1.95)	26.8 (2.93)	20.7 (2.63)	19.4 (2.65)	69.5 (1.11)	24.9 (2.98)	19.3 (2.13)	17.0 (2.36)	65.6 (2.41)
SimMLM	**48.0** (2.25)	**36.7** (1.64)	35.1 (1.60)	79.1 (1.68)	**43.5** (1.93)	33.9 (1.16)	33.0 (1.63)	76.0 (1.61)	33.7 (1.51)	26.0 (1.11)	26.7 (1.33)	70.5 (2.43)	**31.5** (2.52)	22.4 (1.46)	22.3 (1.75)	66.6 (2.42)
Ours																
DECENT	46.3 (2.81)	36.6 (1.65)	35.3 (2.14)	**83.1** (1.43)	41.3 (2.20)	**36.1** (1.77)	**35.0** (2.20)	**82.9** (2.01)	**35.7** (2.10)	**29.4** (2.26)	**28.5** (2.37)	**76.4** (2.69)	29.8 (1.24)	**24.2** (2.76)	**23.9** (3.37)	**72.5** (2.43)

Table 7: **Prediction sensitivity to test-batch composition.** Five-seed means at $\rho = 0.5$. Prediction flip is disagreement with the full-batch predicted class; probability SD ($\times 10^{-3}$) measures variation over 20 neighbor re-draws. The dagger marks the operating batch size.

Dataset	Batch size	Prediction flip (%)	Macro-F1 (%)	Macro-AUC (%)	Probability SD
ADNI	1	1.88	57.69	77.68	—
	32	0.47	57.86	77.81	1.55
	128†	0.18	57.89	77.82	0.67
	256	0.09	57.88	77.83	0.30
	361 (full)	0.00	57.88	77.81	0.00
MIMIC-IV	1	4.14	54.34	69.79	—
	64	1.07	57.38	71.02	3.82
	256†	0.40	57.42	71.06	1.68
	512	0.23	57.40	71.07	0.98
	1349 (full)	0.00	57.31	71.08	0.00

3.1 pp on MIMIC-IV, so message passing across patients contributes to accuracy. Batch size and batch construction remain part of the evaluation protocol.

D TOKENIZATION AND ROUTING

This appendix examines the two Phase-1 components in the order of Section 2. The tokenizer must give each modality several distinct signals, so that a view can split it (Appendix D.1), and neither the tokenizer nor the learned partition may degenerate (Appendix D.2). The remaining subsections isolate the parts of routing that Table 2 holds fixed: the routing target (Appendix D.3), the two routing losses (Appendix D.4), and the hyperparameters that set the granularity of the partition and the strength of the target (Appendix D.5).

D.1 Group Specialization across Modalities

Splitting a modality into views requires its k groups to carry different information. Figure 5 reports, for one run per dataset at $\rho = 0.5$, the debiased linear CKA (Eq. 4) between the proto-tokens of every pair of groups of a modality on test patients; a value of one means that one group is a linear transform of the other. Unlike distances between embedding distributions, which the KL term pulls toward the same prior for every group, CKA compares how two groups vary across patients and does not depend on their coordinates.

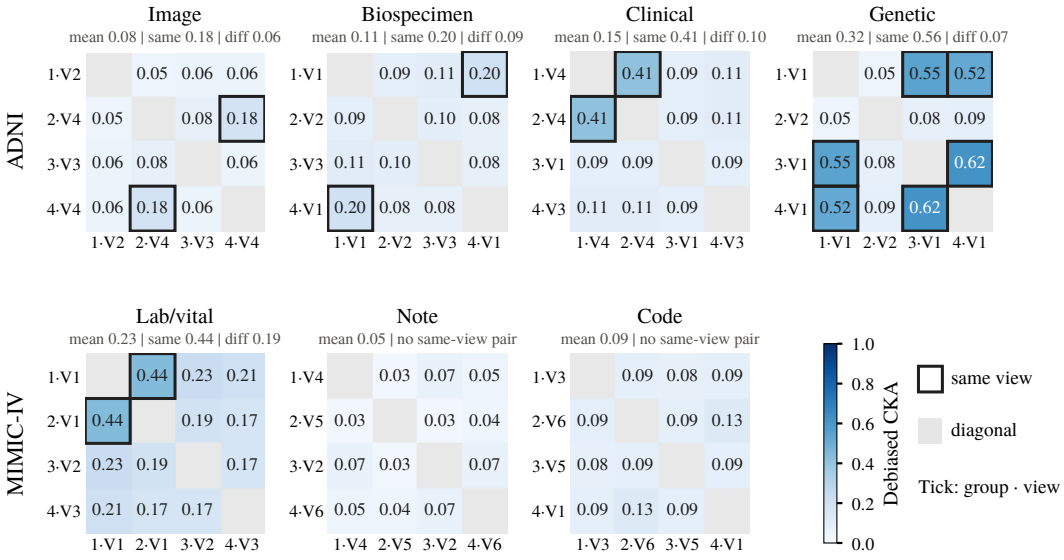


Figure 5: **Proto-token groups of one modality are not redundant.** Debiased linear CKA between the proto-tokens of every pair of groups of each modality on test patients ($\rho = 0.5$; one run per dataset). Tick labels present the group index and its view in the fixed partition; outlined cells mark groups in the same view. Subtitles report the mean over all pairs and over same-view and different-view pairs.

Every pair stays well below one. The largest value is 0.62, and the mean over the pairs of a modality ranges from 0.08 to 0.32 on ADNI and from 0.05 to 0.23 on MIMIC-IV, so the groups of a modality are not redundant copies. In every modality with a same-view pair, the same-view pairs have the largest values of their modality: the three genetic groups that share one view reach 0.52 to 0.62 with each other, whereas the fourth, placed in another view, stays at or below 0.09 with each of them. In note and code, whose four groups occupy four different views, no pair exceeds 0.13. The fixed partition thus keeps groups of one modality together when they carry overlapping variation and separates them otherwise, although the routing target never compares groups of one modality. Groups of one modality share acquisition effects, so these values may overstate their semantic overlap.

D.2 Tokenizer and Partition Diagnostics

Table 8 reports, at $\rho = 0.5$ as medians with ranges over the five seeds, statistics of the tokenizer, the routing target, and the fixed partition. Reconstruction R^2 is computed on test patients from the posterior means and averaged over modalities; 0 corresponds to predicting the feature means. A group is active if its mean KL on training patients exceeds 0.01 nats. $\mathcal{J}_c$ is the set of signals with at least one cross-modal partner with $n_{jj'} \geq n_{\min}$ in the last Phase-1 epoch. Non-empty views and modalities per view describe the fixed partition. The assignment-change rate is the fraction of (test batch, signal) pairs whose view under learned (dynamic) routing differs from the fixed partition.

Table 8 rules out degenerate solutions that could produce the results of Section 3 without learning which signals share a view. The tokenizer retains input information and uses every group, so each signal node carries patient-specific content. Every signal has a cross-modal partner that meets $n_{\min}$,

Table 8: **Tokenizer and partition diagnostics** at $\rho = 0.5$ (median and range over five seeds).

Statistic	ADNI	MIMIC-IV
Reconstruction R^2, mean over modalities	0.58 (0.57–0.60)	0.88 (0.86–0.90)
Active groups (of Mk)	16 (16–16)	12 (12–12)
Signals with an eligible partner, $\lvert\mathcal{J}_c\rvert/J$	1.00	1.00
Non-empty views (of R)	4 (4–4)	5 (5–6)
Modalities per view, mean	2.50 (2.25–3.00)	1.80 (1.50–2.00)
Assignment-change rate (%)	15.5 (11.2–22.3)	10.3 (8.1–12.9)

so the routing target reaches all signals rather than leaving some to load balancing. The fixed partition uses all views on ADNI and at least five of six on MIMIC-IV and mixes modalities within views, so it is neither collapsed nor the modality partition. Learned (dynamic) routing changes 15.5% and 10.3% of the assignments, far below the 75% and 83% of uniform reassignment, so its comparison in Table 2b isolates batch-wise reassignment of an otherwise similar partition.

D.3 Routing Target

Table 9 varies the routing target at $\rho = 0.5$ with the tokenizer, R, and graph encoder fixed. *All-pairs* adds within-modality pairs to Ω_j; *batch-mean cosine* matches routing similarity to the cosine similarity of mini-batch signal summaries, which carries no co-variation across patients; *spectral clustering of* C forms R views directly from the final Phase-1 co-variation matrix (Ng et al., 2001), without a router (Appendix A.2). A_{mod} and A_{axis} are defined in Section 3.4. Cross-seed ARI compares the fixed partitions of different seeds after matching groups by their input attributions, so it measures stability rather than agreement with labels.

Table 9: **Routing target.** Macro-F1 (%) at $\rho = 0.5$ over five seeds, ARI of the ADNI partition with modality and axis labels (A_{mod}, A_{axis}), and cross-seed ARI between the fixed partitions of different seeds, which uses no labels. Ours is shaded blue.

	Macro-F1		ADNI labels		Cross-seed ARI	
Partition from	ADNI	MIMIC-IV	A_{mod}	A_{axis}	ADNI	MIMIC-IV
Router, cross-modal debiased CKA	57.9	57.4	0.21	0.62	0.72	0.55
Router, all-pairs debiased CKA	56.1	56.1	0.44	0.38	0.80	0.69
Router, batch-mean cosine	55.7	55.5	0.64	0.17	0.71	0.69
Spectral clustering of C	56.8	56.6	0.32	0.56	0.59	0.47

Table 9 supports the design of Section 2.3 at three points. Restricting the target to cross-modal pairs matters: adding within-modality pairs moves the partition toward the modality partition (A_{mod} 0.44 vs 0.21) and away from the axes, and lowers macro-F1 on both datasets, as expected when acquisition effects confound co-variation within a modality (Section 2.1). Measuring co-variation across patients matters: matching the similarity of mini-batch summaries recovers the acquisition source (A_{mod} 0.64) and falls below the modality-based partition in macro-F1 (Table 2b), so similar summaries reproduce false grouping. The router matters beyond the target: spectral clustering of the same C keeps most of the axis alignment but predicts worse and is less reproducible across seeds. Reproducibility alone does not indicate a better partition, since the two confounded targets approach the modality structure, which does not depend on the seed, and are more reproducible than ours on MIMIC-IV.

D.4 Routing Losses

We separately remove $\mathcal{L}_{\mathrm{SSL}}$ or $\mathcal{L}_{\mathrm{lb}}$ while keeping the two-phase schedule. Without $\mathcal{L}_{\mathrm{SSL}}$, the router is trained by $\mathcal{L}_{\mathrm{lb}}$ alone and receives no co-variation target. These variants retain Phase 1 and therefore differ from the task-only condition in Table 2(c). Figure 6 reports macro-F1 and macro-AUC, and Table 10 the corresponding structural quantities at the end of Phase 1.

Figure 6 and Table 10 show that the two routing losses have separate roles and that both are needed. Removing either lowers macro-F1 and macro-AUC at every missing rate. Without $\mathcal{L}_{\mathrm{SSL}}$, the router

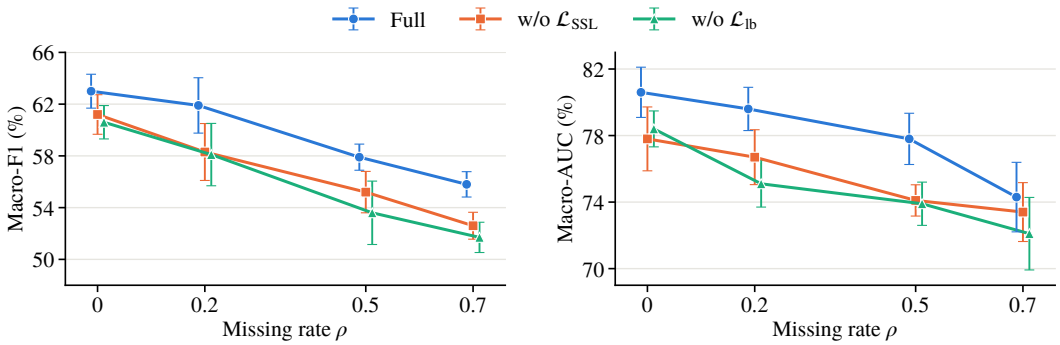


Figure 6: **Effect of the routing losses.** ADNI macro-F1 and macro-AUC (five-seed mean ± SD) with $\mathcal{L}_{\text{SSL}}$ or $\mathcal{L}_{\text{lb}}$ removed and the two-phase schedule held fixed. Lines connect the points corresponding to the measured missing rates.

Table 10: **Phase-1 structural diagnostics on ADNI.** Target fit is the symmetric KL of Eq. 5; load balance is the normalized variance of view importance. Deterioration counts seed–missing-rate pairs with deterioration after removing the controlling term. Non-empty views is the mean number of the $R = 4$ views that receive at least one signal after removing the controlling term; with both terms, all four are non-empty at $\rho = 0.5$ (Table 8). Ours is shaded blue.

Quantity	Full	Controlling term removed	Deterioration	Non-empty views
Target fit (symmetric KL)	0.0833	0.1754	20/20	4.0
Load balance (normalized variance)	0.0069	0.3888	19/20	3.2

no longer fits the co-variation target, whose symmetric KL deteriorates in all 20 seed–missing-rate pairs and roughly doubles on average, although all four views stay in use, so the loss in accuracy comes from which signals share a view. Without $\mathcal{L}_{\text{lb}}$, view importance becomes unbalanced in 19 of 20 pairs and only 3.2 of the four views receive a signal on average, so co-varying signals concentrate in fewer views. The 20 seed–missing-rate pairs are repeated conditions over five seeds; we report directional consistency without treating the pairs as independent observations.

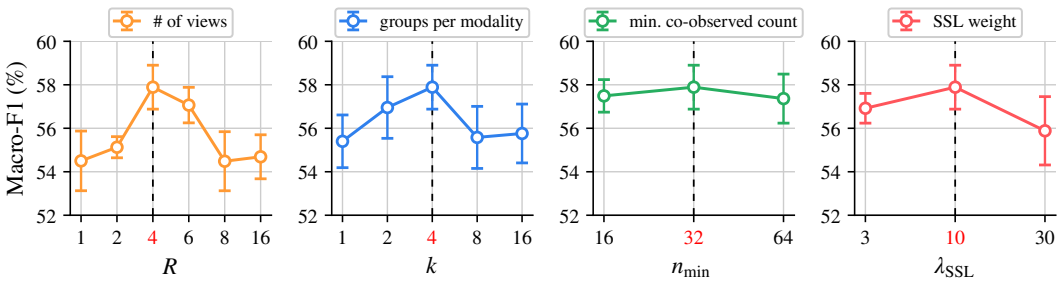
D.5 Hyperparameter Sensitivity

Figure 7 changes one hyperparameter at a time from the configuration in Table 3: the number of views R, the groups per modality k with $kd_z = 32$ fixed, the minimum co-observed count $n_{\min}$, and the SSL weight λ_{SSL}. All other settings are fixed; when R is smaller than the selected top-K, the router keeps all R views. $R = 1$ and $k = 1$ are the single-view and single-token controls of Table 2.

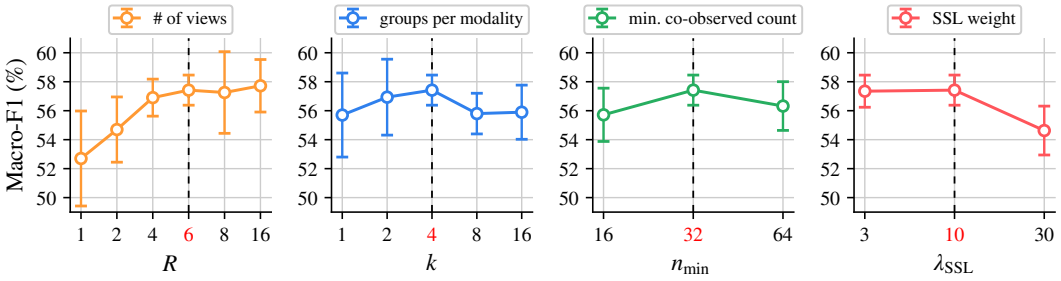
Figure 7 asks whether the gains depend on a narrow choice of granularity or of the routing target's settings. Granularity matters in both directions. The coarsest partitions, a single view or a single group per modality, give the lowest or near-lowest macro-F1 on both datasets, consistent with Table 2. Finer partitions also lose accuracy: groups finer than $k = 4$, which keep $d_z \leq 4$ dimensions each under the fixed capacity, fall to near the single-group level, and on ADNI more than six views return macro-F1 to the single-view level, whereas on MIMIC-IV macro-F1 is flat for $R \geq 6$. The routing target tolerates a moderate range but not every setting. Macro-F1 stays within 1.0 pp for λ_{SSL} between 3 and 10, whereas an overly strong target ($\lambda_{\text{SSL}} = 30$) degrades the partition. Macro-F1 changes little with $n_{\min}$ on ADNI, but on MIMIC-IV lowering $n_{\min}$ to 16 costs 1.7 pp, consistent with noisier estimates of C from pairs with fewer co-observed patients.

E Semantic Structure of the Learned Partition

This appendix relates the learned partition to semantic structure. Appendix E.1 defines the ADNI annotation used in Section 3.4, Appendix E.2 shows how the ADNI views combine modalities, and



(a) ADNI



(b) MIMIC-IV

Figure 7: **Sensitivity.** Macro-F1 at $\rho = 0.5$ (five-seed mean $\pm$ SD) when changing the number of views R, the groups per modality k ($kd_z = 32$), the minimum co-observed count $n_{\min}$, or the SSL weight λ_{SSL}. Tested settings are equally spaced; dashed lines and red ticks mark the selected values.

Appendices E.3 and E.4 test whether cross-modal co-variation recovers planted views on synthetic data, where the ground truth is known.

E.1 ATN(V) Axis Annotation on ADNI

Table 11 summarizes the ADNI axis annotation used in the partition analysis. The *Axis* column denotes the disease-related process, the *Modality* column denotes the ADNI acquisition source, and the *Key variables* column lists representative feature groups assigned to each axis. The annotation is defined at the feature-group level because several preprocessed variables represent one biological or clinical construct. ATN(V) labels are used only for post hoc analysis; the tokenizer, router, View-GNN, fusion module, and classifier do not receive axis labels during training.

Table 11: **ATN(V) annotation for ADNI.** Rows show how disease-related axes are represented across ADNI modalities. The annotation is used only to interpret learned signal partitions.

Axis	Modality	Key variables
A (amyloid)	Biospecimen Genetic	Aβ42/Aβ40 ratio, Aβ42 APOE ϵ4, amyloid-associated loci
T (tau pathology)	Biospecimen Genetic	phosphorylated tau (p-tau) MAPT region, tau-associated loci
N (neurodegeneration)	Biospecimen Image	total tau (t-tau) hippocampal volume, cortical thickness, ventricular enlargement
V (vascular burden)	Image Clinical	white-matter hyperintensity (WMH) volume vascular risk factors: blood pressure, BMI, cardiovascular history

The annotation follows the AT(N) research framework (Jack et al., 2018): CSF Aβ42 and the Aβ42/Aβ40 ratio index amyloid, CSF p-tau indexes tau, and CSF t-tau and anatomic MRI index the non-specific neurodegeneration group, which we represent by ADNI hippocampal volume, cortical thickness, and ventricular enlargement. Following the argument that white matter hyperintensities can be added to the framework as a cerebrovascular marker (Ng et al., 2023), WMH volume represents

cerebrovascular burden, although it is not specific to vascular injury. For our post hoc analysis, we additionally assign genetic and clinical risk-factor groups to these axes on the basis of reported associations, as listed in Table 11. These feature-group assignments are our reference annotations, not additional diagnostic biomarkers specified by the AT(N) framework.

Each proto-token group takes the axis with the largest attribution over annotated feature groups, computed on training patients with the final Phase-1 encoder (Appendix A.2); groups dominated by *Other* are excluded. We hold these labels and the retained groups fixed across epochs to track partition alignment relative to the final semantic reference. Both A_{axis} and A_{mod} use the same retained groups, with ATN(V) and source-modality labels, respectively.

E.2 View Composition

We examine ADNI at $\rho = 0.5$ over five seeds with the partition analysis protocol (Appendix A.2). The reference configuration is $R = k = 4$. View composition includes all sixteen signal nodes, with views matched across seeds by their dominant post hoc ATN(V) axis. Groups excluded from A_{axis} and A_{mod} remain in their learned views for the composition counts, which are averaged across seeds.

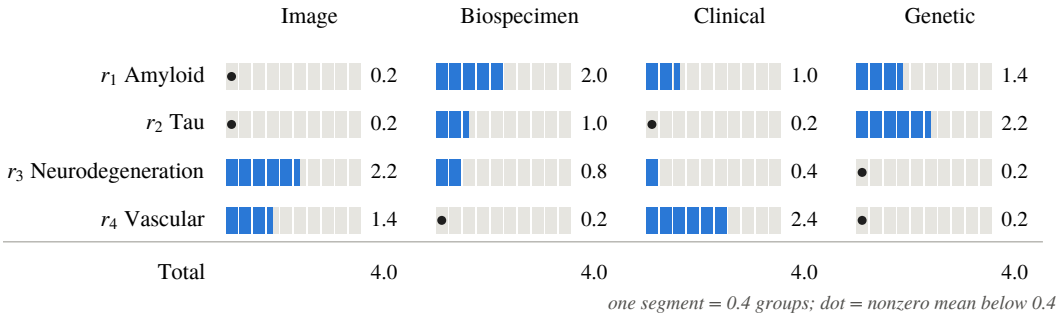


Figure 8: **Views cross modality boundaries.** Mean number of proto-token groups of each modality (columns) assigned to each view (rows) on ADNI ($\rho = 0.5$, $R = k = 4$; five seeds). Each modality has four groups; a bar spans the four groups of its modality in segments of 0.4, and a dot marks a nonzero mean below 0.4. Views are matched across seeds by their dominant post hoc ATN(V) axis.

Figure 8 shows the two departures from modality-level construction that the partition is meant to make. Views draw groups from several modalities, which counters false separation: no view takes more than 61.1% of its groups from one modality, and the amyloid-aligned view combines biospecimen, genetic, and clinical groups. Modalities are split across views, which counters false grouping: no modality places more than 60.0% of its groups in one view, image groups divide between the neurodegeneration- and vascular-aligned views, and genetic groups between the tau- and amyloid-aligned views. Each modality can therefore reach different operators according to the content of its groups (Appendix D.1), instead of one modality-specific operator. Shares are computed from the mean counts.

E.3 Synthetic Benchmark with Planted Views

Clinical cohorts do not provide ground-truth latent partitions. We therefore construct a synthetic benchmark that separates semantic content from source modality and makes false grouping and false separation measurable.

Data generation. We generate 1,500 samples with balanced three-class labels and split them into 1,000/250/250 training/validation/test samples. Each of four modalities contains four blocks. Block dimensions are 8, 32, 128, and 512 for the four modalities, giving modality dimensions 32, 128, 512, and 2,048. The placement in Figure 9(a) mixes three planted views within each modality and places each view in three modalities. For sample i, modality m, and block b, features are

$$x_{i,m,b} = A_{m,b} z_{i,v(m,b)} + \lambda G_m u_{i,m} + 0.5\epsilon_{i,m,b}, \tag{18}$$

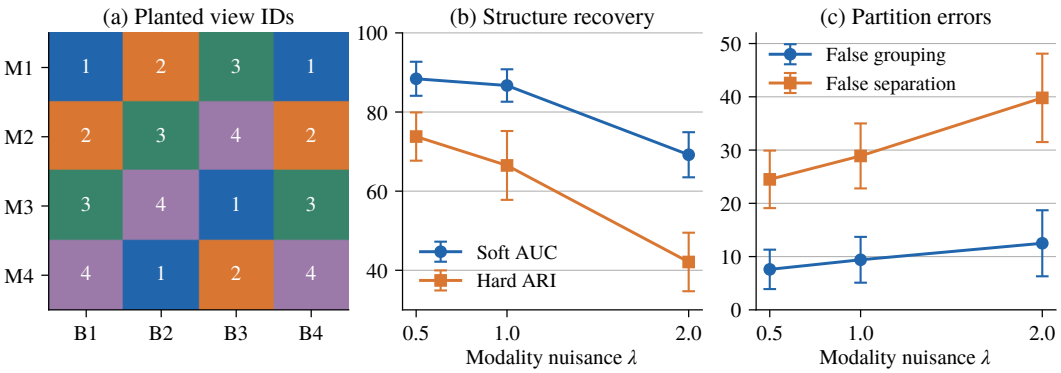


Figure 9: **Planted structure and recovery under modality nuisance.** (a) Each cell denotes the planted semantic view of a feature block. (b,c) Recovery and partition errors of DECENT as modality nuisance λ increases under MCAR at $\rho = 0.5$, multiplied by 100. Error bars show one SD over five held-out seeds. For reference, a balanced random partition gives false grouping 20 and false separation 80 in expectation, and the modality partition gives 100 for both.

where $v(m, b)$ is the planted view, $z_{i,r} \in \mathbb{R}^8$ is its shared latent content, $u_{i,m} \in \mathbb{R}^8$ is modality-specific nuisance, $A_{m,b}$ and G_m are random projections, and ϵ is independent Gaussian noise. Labels are terciles of the severity $s_i = \frac{1}{2}\sum_{r=1}^{4} \beta_r^\top z_{i,r}$, with nonzero random β_r. Every view contributes to the label.

Missingness. Following the missing-data taxonomy of Rubin (1976), we use MCAR dropout independent of sample attributes, MAR probabilities depending on an observed three-level site variable c_i, and MNAR probabilities depending on latent severity. With $\boldsymbol{M}[i, m] = 1$ for an observed modality, the pre-adjustment dropout probabilities are

$$\Pr(\boldsymbol{M}[i,m] = 0) = \begin{cases} \rho & \text{MCAR}, \\ \text{sigmoid}(b_m + \gamma_{m,c_i}) & \text{MAR}, \\ \text{sigmoid}(b_m + 1.5\delta_m \widetilde{s}_i) & \text{MNAR}, \end{cases} \tag{19}$$

where $\widetilde{s}_i$ is standardized severity, $\boldsymbol{\delta} = (-1, -1, +1, +1)$, and γ has rows $(-1, .2, 1)$, $(.8, -1.1, .3)$, $(.2, 1, -1)$, and $(-.9, .9, .1)$. Intercepts b_m calibrate each modality to the target missing rate. Sampled masks are adjusted to the exact marginal counts while preventing all modalities from being missing for a sample. These constraints induce dependence among the final mask entries.

Training and recovery metrics. The Phase-1 objective and frozen Phase 2 follow the main Method, with $R = 4$ views and $k = 4$ groups per modality, matching the numbers of planted views and of blocks per modality, and the ADNI settings of Table 3 otherwise. Development seeds are separate from the five evaluation seeds. Each learned group takes as its semantic label the planted view whose blocks receive the largest attribution (Appendix A.2), computed on training samples. Soft pairwise AUC measures whether $\langle \boldsymbol{a}_j, \boldsymbol{a}_{j'} \rangle$ distinguishes pairs with equal semantic labels; hard ARI compares these labels with the fixed partition. Mixed groups can make the mapping ambiguous. FG is the fraction of same-modality, different-label pairs placed in one view; FS is the fraction of cross-modality, same-label pairs placed in different views. A zero denominator makes the corresponding rate undefined. The modality-fixed reference preserves the original blocks and groups them by acquisition source: its planted ARI is $-1/24$, and FG and FS are both one. The modality-fixed reference is structural, not a tuned competitor with an identical tokenizer. A balanced random partition of the 16 groups into four views places two groups together with probability $3/15$, so its expected FG and FS are 0.2 and 0.8, respectively.

E.4 RECOVERY OF PLANTED VIEWS UNDER MISSINGNESS

The benchmark tests the mechanism of Section 2 where the ground truth is known: cross-modal co-variation should recover which signals share a latent view even when modality-specific variation

Table 12: **Recovery and prediction under distinct missingness mechanisms.** Results at $\rho = 0.5$ and $\lambda = 1.0$ over five held-out seeds. Each entry is mean (standard deviation), with all metrics multiplied by 100. Soft AUC evaluates recovery, not classification. A balanced random partition gives FG 20 and FS 80 in expectation, and the modality partition gives 100 for both. Mask-only macro-F1 uses availability bits without content. Ours is shaded blue.

Mechanism	Soft AUC	Hard ARI	FG	FS	Macro-F1	Mask-only F1
MCAR	86.7 (4.1)	66.5 (8.7)	9.4 (4.3)	28.9 (6.1)	63.0 (2.2)	34.4 (2.1)
MAR	87.8 (4.6)	67.1 (7.3)	8.0 (3.9)	26.8 (5.7)	62.3 (3.4)	33.8 (2.7)
MNAR	83.8 (7.2)	62.7 (9.6)	12.6 (3.6)	31.1 (7.9)	66.5 (2.4)	42.4 (1.7)

confounds co-variation within a modality. At $\rho = 0.5$ and $\lambda = 1$, the learned partition recovers the planted views under all three missingness mechanisms (Table 12), and its false grouping and false separation stay far below both the modality partition and a balanced random partition. Recovery is lowest under MNAR, where missingness depends on latent severity, whereas macro-F1 is highest there because the mask itself carries label information, as the mask-only control shows; recovery thus measures partition quality separately from prediction. Stronger nuisance lowers recovery (Figure 9b,c), and false separation rises more than false grouping, but both remain below the random level at $\lambda = 2$. Strong modality-specific variation thus weakens the grouping of same-view signals across modalities more than the splitting within a modality.

F Limitations and Broader Impacts

F.1 Limitations

Dependence on test-batch composition. The partition is fixed, but message passing uses the current batch, so predictions can vary with its composition. Appendix C.3 reports disagreement with full-batch evaluation at the operating batch sizes and larger changes in solo evaluation. Deployment in new batch settings requires further evaluation.

Co-variation as a supervisory signal. Cross-modal co-variation can also arise from factors that reach several modalities, such as overall disease severity, site, or care intensity. The routing target therefore supervises which signals share an operator but does not identify disease processes, and linear CKA captures only linear co-variation between groups.

Training-strategy comparison. The two-phase and task-only variants use the same total number of epochs, allocated to different objectives. Their comparison evaluates the complete training strategies, including tokenizer adaptation; equal epoch counts do not imply equal compute.

Uniform group counts across modalities. A uniform k assigns the same candidate signal budget to every modality, which may limit allocation to modalities with more heterogeneous sub-signals. Modality-specific group counts can be helpful to change each modality's contribution to view-level load balancing.

Domain annotations for semantic validation. The ADNI analysis (Section 3.4) uses annotations informed by the AT(N) framework (Jack et al., 2018) and a vascular marker (Ng et al., 2023). MIMIC-IV verification is limited to label-free diagnostics (Figure 3 and Appendix D.1) because comparable clinical-axis annotations are not provided. Index alignment fixes the computational correspondence of groups across patients but not their semantic content, which can differ across training runs. Verifying semantic-axis recovery requires an external target structure informed by domain knowledge.

Scope of the mechanism evidence. The partition controls use one tokenizer per run and compare the learned fixed partition with batch-wise reassignment and static partitions. The single-view control changes the number of message bases, and encoder-sharing variants have different parameter counts. The semantic analyses measure post hoc agreement; establishing a causal link between axis alignment and predictive gains requires independent semantic validation. Matched patient splits and missingness masks also leave differences in the relational context used by each graph architecture.

F.2 BROADER IMPACTS

Intended use. DECENT is a research classifier evaluated on retrospective, de-identified data, not validated clinical decision support. Prediction from available modalities could broaden applicability when acquisition is limited (Wu et al., 2026), but benefits to care access remain untested. Partition analyses (Section 3.4) describe model organization rather than clinical causation; agreement with ATN(V) annotations does not establish diagnostic validity.

Cohort representativeness. ADNI is a selected research cohort (Jack et al., 2008), and MIMIC-IV records care at one academic medical center (Johnson et al., 2023b). Missingness can depend on acquisition practice (Wu et al., 2026); its mechanisms in these cohorts remain unidentified. Observation patterns may carry cohort-specific predictive signals. Induced modality removal does not eliminate cohort selection effects, so broader generalization requires external validation.