PROGMF: PROGRESSIVE MULTIMODAL REPRESENTATION LEARNING FOR INCOMPLETE BIOMEDICAL DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal learning in biomedical research is often limited by unequal modality coverage, particularly in Alzheimer's disease (AD) neuroimaging, where acquisition costs and limited access to specialized imaging reduce the number of subjects observed across all modalities. Conventional complete-view fusion methods exclude partially observed subjects, while imputation methods reconstruct missing observations and may introduce reconstruction errors. We introduce Progressive Multimodal Fusion (PROGMF), a framework that learns from larger cohorts with fewer modalities and transfers the learned representations as additional modalities are incorporated. Using independent vector analysis (IVA) as a testbed, PROGMF independently learns a joint representation at each stage and uses the preceding stage's representation as an anchor to guide subsequent refinement. This enables knowledge transfer without imputing missing observations while allowing adaptation to dependencies introduced by new modalities. Controlled simulations involving four modalities show that PROGMF achieves lower latent source recovery error than complete-view fusion and imputation-based baselines. A controlled ablation further demonstrates the contribution of subjects available only at earlier stages. Evaluation on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset also shows higher accuracy and F1 Score than the evaluated baselines for cognitively normal versus dementia classification, providing preliminary evidence of improved generalization to held-out subjects. The implementation code will be made publicly available upon acceptance of the paper.

1 INTRODUCTION

Multimodal learning (MML) has become increasingly prevalent in biomedical research due to its ability to learn complementary information about biological processes from different modalities(Acosta et al., 2022). However, measurements for these modalities are often collected for partially overlapping groups of subjects, creating a fundamental tradeoff between modality coverage and cohort size (Hayat et al., 2022). This challenge is particularly evident in neuroimaging studies of Alzheimer's disease (AD) due factors such as acquisition costs and varying study protocols leading to unequal coverage across modalities. Large studies such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) collect diverse measurements of brain structure, function, and pathology, yet not every subject has every measurement (Veitch et al., 2024). Consequently, incorporating additional modalities can substantially reduce the number of subjects available for MML, potentially excluding observations that contain useful information about the the latent variables otherwise known as latent sources responsible for these biological processes.

Existing approaches to incomplete MML addresses this challenge from several perspectives. Conventional complete-view fusion restricts learning to subjects having all modalities, while imputation methods reconstruct missing observations or latent representations (Tao et al., 2021). Other methods such as MedFuse integrates clinical observations that are not necessarily paired (Hayat et al., 2022), while ShaSpec learns shared and modality specific features from available modalities (Wang et al., 2023). Knowledge distillation has also been used to transfer information from individual modality models trained on larger datasets to a multimodal model (Wang et al., 2020). In neuroimaging, MSR-IVA uses subject masking and shared structural representations to integrate structural MRI

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.

with dynamic functional connectivity states (Solomon et al., 2026). Despite these advances, an important question remains. How can joint representations learned from larger cohorts with fewer modalities be transferred to subsequent fusion stages when additional modalities are available for progressively smaller groups of subjects?

We hypothesize that representations learned from larger cohorts with fewer modalities can improve subsequent fusion as additional modalities become available. However, effective transfer must preserve previously learned information while allowing the representations to adapt to new multimodal observations. Motivated by this, we introduce Progressive Multimodal Fusion (PROGMF), a framework that combines learning across progressively smaller cohorts with anchor guided representation transfer.

Learning Across Nested Cohorts. Rather than requiring all subjects to have all modalities at the beginning, PROGMF initially learns a joint representation from two modalities and progressively incorporates additional modalities. Each stage uses the intersection of subjects observed across its active modalities, resulting in nested training cohorts that become smaller as the number of modalities increases. This formulation allows the model to exploit observations that would otherwise be excluded by complete-view fusion. Importantly, subjects available only at earlier stages contribute through the representations learned from their observed modalities, without requiring the reconstruction of unavailable observations. The resulting challenge is to preserve this information when learning from smaller cohorts with additional modalities.

Anchor Guided Representation Transfer. To transfer information across stages, we introduce an anchor mechanism that uses representations learned from larger earlier cohorts to guide subsequent fusion. Each stage independently estimates a joint representation, which is then refined using the preceding stage as an anchor. In our IVA instantiation, we implement this mechanism using a weighted squared Frobenius norm penalty between normalized unmixing matrices and their corresponding anchors. The penalty encourages the retention of previously learned information while allowing all active modalities to adapt jointly. In this study, we use IVA as testbed for this framework, following the convention in Silva et al. (2021).

Contributions. Our main contributions are summarized as follows.

- We formulate progressive MML over nested subject cohorts, enabling observations outside the final complete cohort to contribute to subsequent fusion.
- We introduce PROGMF, a framework that combines independent representation learning at successive stages with an anchor mechanism that transfers previously learned information while allowing adaptation to newly introduced modalities.
- We conduct controlled four-modality simulations demonstrating improved latent source recovery under shrinking subject overlap and examine the contribution of earlier observations, fusion order, and observation noise. We further evaluate the downstream generalization of PROGMF on ADNI neuroimaging data, where it achieves higher accuracy and F1 Score than the evaluated baselines on an independent held-out test set following five-fold cross-validation for model selection.

2 METHOD

We propose Progressive Multimodal Fusion (PROGMF), a framework for learning particularly from biomedical datasets in which different modalities are available for varying number of subjects. As illustrated in Figure 1, the framework initially learns from a larger cohort sharing two modalities and progressively incorporates additional modalities as the number of shared subjects decreases.

At each subsequent stage, a new joint representation is estimated from the current cohort and all active modalities contributing in that stage. Information learned at the preceding stage is then transferred through an anchor mechanism that guides refinement while allowing the current representation to adapt to newly introduced modalities. This enables subjects available only at earlier stages to contribute without requiring the reconstruction of unavailable observations.

We instantiate the framework using independent vector analysis (IVA). The complete procedure is summarized in Algorithm 1.

2.1 Problem Setting

Consider a dataset containing $M \geq 2$ modalities. For modality m, let $X_m \in \mathbb{R}^{D_m \times N_m}$ denote its observations, where D_m is the number of features, N_m is the number of observed subjects, and $\mathcal{I}_m$ contains their identifiers. Subjects may not have observations from every modality and therefore these subject sets can differ.

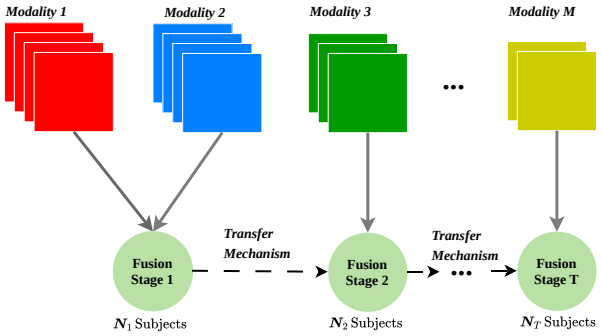

Figure 1: **Overview of ProgMF.** Modalities are progressively incorporated over nested subject cohorts. Dashed arrows indicate anchor-guided transfer between fusion stages. Each stage learns from all its active modalities.

Given a specified modality order $\omega = (m_1, \ldots, m_M)$, ProgMF begins with the first two modalities and introduces one additional modality at each subsequent stage, producing $T = M - 1$ stages. At stage t, the active modalities are $\mathcal{M}_t = \{m_1, \ldots, m_{t+1}\}$, and the corresponding cohort is $\mathcal{C}_t = \{s \mid s \in \mathcal{I}_m \text{ for every } m \in \mathcal{M}_t\}$. Let $N_t = |\mathcal{C}_t|$ and $X_{m,t} \in \mathbb{R}^{D_m \times N_t}$ denote the observations of active modality m, with columns aligned by subject identifier. The cohorts are nested, so $N_1 \geq \cdots \geq N_T$.

Complete-view IVA is limited to learning only from the final cohort $\mathcal{C}_T$, whose subjects have all modalities. In contrast, ProgMF learns from subjects available in larger earlier cohorts and transfers the resulting representations to later stages. For each modality, the feature transformation is kept fixed across stages to ensure that previously learned and current model parameters are directly comparable. In held-out evaluations, these transformations are estimated using training observations only.

2.2 Independent Vector Analysis

We use IVA as the underlying representation model, following the multidataset formulation of MISA (Silva et al., 2021; 2024). At stage t, each active modality has an unmixing matrix $W_{m,t} \in \mathbb{R}^{K \times D_m}$, where $K \leq D_m$ for every active modality and K is the common number of source component vectors (SCVs). The estimated latent sources are $Y_{m,t} = W_{m,t} X_{m,t} \in \mathbb{R}^{K \times N_t}$. We denote the collection of unmixing matrices by $\Theta_t = \{W_{m,t}\}_{m \in \mathcal{M}_t}$.

IVA groups corresponding latent sources across modalities into SCVs. For subject n in the current cohort, the k-th SCV observation contains the k-th estimated source from each active modality: $\mathbf{y}_{k,t}^{(n)} = ([Y_{m_1,t}]_{kn}, \ldots, [Y_{m_{t+1},t}]_{kn})^\top$. Let $\mathbf{y}_{k,t} \in \mathbb{R}^{d_t}$ denote the corresponding random vector, where $d_t = |\mathcal{M}_t|$. IVA seeks independence between different SCVs while allowing statistical dependence among modalities within each SCV. This objective can be expressed using the mutual information among the K SCVs,

$$\mathcal{J}_{\mathrm{MI},t}(\Theta_t) = \sum_{k=1}^{K} h(\mathbf{y}_{k,t}) - h(\mathbf{y}_{1,t}, \ldots, \mathbf{y}_{K,t}), \tag{1}$$

where $h(\cdot)$ denotes differential entropy. Minimizing this quantity encourages independence between distinct SCVs.

Source separation objective. We model each SCV similarly to (Silva et al., 2021) using a multivariate Kotz distribution, with $\beta = 0.5$, $\lambda_{\mathrm{K}} = 1$, and $\eta = 1$. Here, λ_{K} denotes the Kotz distribution parameter and is distinct from the progressive transfer coefficient introduced below. We use the scale-controlled formulation, in which the distribution of each SCV depends on its estimated cross-modal correlation matrix.

For square, invertible unmixing matrices, conventional IVA combines the SCV density terms with the log determinants of the unmixing matrices. To accommodate rectangular matrices, we replace the determinant term with a corresponding measure of the volume of the learned source space. The resulting objective is

$$\begin{aligned}\mathcal{L}_{\mathrm{IVA},t}(\Theta_t) = &\sum_{k=1}^{K} \widehat{\mathbb{E}}_t \left[-\log p_{k,t}(\mathbf{y}_{k,t})\right] \\ &- \frac{1}{2} \sum_{m \in \mathcal{M}_t} \log\det\left(W_{m,t} W_{m,t}^{\top} + \delta I_K\right),\end{aligned} \tag{2}$$

where $p_{k,t}$ is the Kotz density of SCV k, $\widehat{\mathbb{E}}_t$ denotes the empirical mean over the current subjects, and $\delta > 0$ provides numerical stability. The first term models the estimated source distributions, while the second discourages the unmixing matrices from collapsing the source space.

For square, nonsingular matrices, the volume term approaches the conventional IVA log determinant as δ approaches zero. For rectangular matrices, Eq. equation 2 is a source separation objective rather than an exact likelihood obtained by transforming the original observations. The Kotz normalization constants are retained when evaluating the objective. The density and numerical implementation are detailed in Appendix A.1.

2.3 Progressive Learning

Independent stage estimation. The first stage fits IVA using the two initial modalities and their shared cohort, producing $\widehat{\Theta}_1$. At each subsequent stage $t > 1$, we first fit a new IVA model using the current cohort and all active modalities. This fit starts from a fresh initialization and produces an independent estimate $\bar{\Theta}_t$.

The preceding progressive estimate $\widehat{\Theta}_{t-1}$ is then used to guide refinement of the new estimate. Importantly, the preceding parameters are not used to initialize the current model. Information is transferred only after the current representation has been estimated independently.

Component matching. Independent IVA fits can recover equivalent sources in different orders and with opposite signs. Before transferring parameters, we therefore match the current sources to those learned at the preceding stage. We normalize the rows of the unmixing matrices using $\mathcal{N}(W)_{i:} = W_{i:}/(\|W_{i:}\|_2 + \epsilon)$, where $\epsilon > 0$ provides numerical stability. For each pair of current and preceding SCVs, we compute the absolute inner products of their normalized rows and sum these similarities over the inherited modalities.

We then use the Hungarian algorithm to find a single component permutation shared by all inherited modalities. After matching, we correct component signs separately for each modality. Let $\widetilde{W}_{m,t-1}$ denote the resulting aligned previous-stage matrix. These aligned matrices remain fixed during refinement. The matching procedure is detailed in Appendix A.2.

Representation transfer. We use an anchor penalty term to control how much information is transferred from previous stages to the current stage. This penalty encourages the current unmixing matrices to retain previously learned information by limiting their deviation from the aligned references. Specifically, we define

$$\mathcal{R}_t(\Theta_t) = \frac{1}{2K|\mathcal{M}_{t-1}|} \sum_{m \in \mathcal{M}_{t-1}} \left\| \mathcal{N}(W_{m,t}) - \mathcal{N}(\widetilde{W}_{m,t-1}) \right\|_F^2. \tag{3}$$

Here, $\|\cdot\|_F$ denotes the Frobenius norm. Comparing normalized rows encourages agreement between corresponding source directions without directly constraining their magnitudes. The penalty is averaged over the K SCVs and the inherited modalities.

The progressive objective at stage $t > 1$ is

Algorithm 1: Progressive Multimodal Fusion (PROGMF)

Input: Observations and subject identifiers $\{X_m, \mathcal{I}_m\}_{m=1}^{M}$; modality order $\omega = (m_1, \ldots, m_M)$; number of SCVs K; transfer coefficients $\{\lambda_t\}_{t=2}^{T}$; minimum scale $c_{\min}$

Output: Final estimate $\widehat{\Theta}_T$

$T \leftarrow M - 1$
$\mathcal{M}_1 \leftarrow \{m_1, m_2\}$
$\mathcal{C}_1 \leftarrow \{s \mid s \in \mathcal{I}_{m_1} \text{ and } s \in \mathcal{I}_{m_2}\}$
$\mathcal{X}_1 \leftarrow \{X_{m,1}\}_{m \in \mathcal{M}_1}$ `// Align subjects in` $\mathcal{C}_1$
$\widehat{\Theta}_1 \leftarrow \mathrm{FitIVA}(\mathcal{X}_1, K)$
for $t = 2, \ldots, T$ **do**
  $\mathcal{M}_t \leftarrow \{m_1, \ldots, m_{t+1}\}$
  $\mathcal{C}_t \leftarrow \{s \in \mathcal{C}_{t-1} : s \in \mathcal{I}_{m_{t+1}}\}$
  $\mathcal{X}_t \leftarrow \{X_{m,t}\}_{m \in \mathcal{M}_t}$ `// Align subjects in` $\mathcal{C}_t$
  $\bar{\Theta}_t \leftarrow \mathrm{FitIVA}(\mathcal{X}_t, K)$ `// Fresh initialization`
  $\widetilde{\Theta}_{t-1} \leftarrow \mathrm{Align}(\widehat{\Theta}_{t-1}, \bar{\Theta}_t)$
  $c_t \leftarrow \max\left(\left|\mathcal{L}_{\mathrm{IVA},t}(\bar{\Theta}_t)\right|, c_{\min}\right)$
  $\widehat{\Theta}_t \leftarrow \mathrm{Refine}(\bar{\Theta}_t, \widetilde{\Theta}_{t-1}, \lambda_t c_t)$
return $\widehat{\Theta}_T$

$$\mathcal{L}_{\mathrm{ProgMF},t}(\Theta_t) = \mathcal{L}_{\mathrm{IVA},t}(\Theta_t) + \lambda_t c_t \mathcal{R}_t(\Theta_t). \tag{4}$$

The coefficient $\lambda_t \geq 0$ controls the strength of representation transfer. When $\lambda_t = 0$, the anchor penalty is removed, providing a no-transfer configuration. To account for differences in the numerical magnitude of the IVA objective, we use $c_t = \max(|\mathcal{L}_{\mathrm{IVA},t}(\bar{\Theta}_t)|, c_{\min})$, where $c_{\min} > 0$ prevents the scale from vanishing. This is a practical scaling convention, not a guarantee of identical effective transfer strength across stages. Starting from the independent estimate $\bar{\Theta}_t$, we minimize Eq. equation 4 to obtain $\widehat{\Theta}_t$. The aligned references and c_t are computed once and held fixed during refinement. Only inherited modalities receive a direct transfer penalty, but the unmixing matrices of all active modalities are optimized jointly.

3 EXPERIMENTAL SETUP

We evaluate PROGMF using controlled simulations and multimodal neuroimaging data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). The simulations examine source recovery under shrinking multimodal overlap, the contribution of earlier-only subjects, sensitivity to fusion order, and robustness to observation noise. The ADNI experiment evaluates the downstream predictive utility of the learned representations.

3.1 SIMULATION DATA

We simulate four neuroimaging-inspired modalities representing gray matter (GM), functional network connectivity (FNC), fractional anisotropy (FA), and positron emission tomography (PET). GM, FA, and PET each contain $64 \times 64 = 4096$ spatial features, while FNC contains the 1378 unique off-diagonal entries of a symmetric 53×53 connectivity matrix.

Observations follow $X_m = A_m S_m + E_m$, where A_m is the modality-specific mixing matrix, S_m contains the latent sources, and E_m is additive Gaussian noise. We generate $K = 10$ SCVs from multivariate Laplace distributions. Modality-specific mixing patterns are constructed using SimTB templates and structured spatial or connectivity patterns (Erhardt et al., 2012). The default signal-to-noise ratio (SNR) is $10\,\mathrm{dB}$.

We vary the initial cohort size over $N \in \{500, 700, 1000, 2000, 3000\}$ and the cumulative overlap reduction over $r \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$. Three nested cohorts are constructed with $N_1 = N$, $N_2 = N(1 - r/2)$, and $N_3 = N(1 - r)$. The primary fusion order is GM + FNC $\rightarrow$ FA $\rightarrow$ PET, with FA introduced at Stage 2 and PET at Stage 3.

We use this fusion order throughout the simulation experiments unless otherwise stated. Each experiment is repeated using 20 independently generated datasets. Within each dataset, all methods use identical latent sources, mixing matrices, noise, and subject cohorts. Additional generation and preprocessing details are provided in Appendix B.

3.2 Baselines and Evaluation Metrics

Baselines and imputation. We compare ProgMF against complete-view IVA and three imputation methods: Soft Imputation, Iterative SVD, and Matrix Factorization. Complete-view IVA uses only subjects observed across all modalities. Soft Imputation (Mazumder et al., 2010; Csala et al., 2024) performs nuclear-norm-regularized matrix completion, Iterative SVD (Troyanskaya et al., 2001; Csala et al., 2024) repeatedly applies low-rank approximation, and Matrix Factorization (Koren et al., 2009; Ma et al., 2026) reconstructs missing values using low-rank factors.

All three imputation methods preserve observed values and reconstruct missing modality blocks before fitting IVA. The IVA models use the same objective and optimization settings across methods. ProgMF additionally refines its independent stage estimates using the proposed transfer objective. Further details on imputation and optimization are provided in Appendices B and C, respectively.

Simulation evaluation. We evaluate source recovery using joint intersymbol interference (JISI) (Anderson et al., 2012). For each modality m, let $G_m = \widehat{W}_{m,T} A_m^{\text{eff}}$ denote the gain matrix, where $\widehat{W}_{m,T}$ is the final-stage unmixing matrix and A_m^{eff} is the ground-truth mixing matrix after applying the same feature transformation used for the observations. The average absolute gain is $\bar{g}_{ij} = M^{-1} \sum_{m=1}^{M} |(G_m)_{ij}|$. We compute JISI $= (J_{\text{row}} + J_{\text{col}})/[2K(K-1)]$, where $J_{\text{row}} = \sum_i (\frac{\sum_j \bar{g}_{ij}}{\max_p \bar{g}_{ip}} - 1)$ and $J_{\text{col}} = \sum_j (\frac{\sum_i \bar{g}_{ij}}{\max_p \bar{g}_{pj}} - 1)$. JISI measures source recovery while accounting for permutation and scaling ambiguities. Lower values indicate better recovery, with zero corresponding to perfect recovery.

Hyperparameter selection. We evaluate the stage-specific transfer coefficients over $\lambda_2, \lambda_3 \in \{0.001, 0.003, 0.01, 0.03, 0.1\}$. For the primary shrinking-overlap experiment, transfer coefficients are selected using ground-truth JISI on separate synthetic calibration datasets and fixed before evaluation on 20 independently generated datasets. The evaluation datasets are not used for coefficient selection.

For the fusion-order experiment, we use five independent calibration datasets to select transfer coefficients on the canonical fusion order and hold them fixed across all 12 orders within each overlap condition. Within each initial-pair coupling group, we display the schedule with the lowest mean calibration JISI. The selected schedules are evaluated on 20 independent datasets. The noise-robustness experiment reuses coefficients selected at $10\,\text{dB}$ without noise-specific retuning.

Real-data evaluation. We evaluate ProgMF using ADNI gray matter (GM), functional network connectivity (FNC), and amyloid PET (Veitch et al., 2024; Jagust et al., 2024). After matching imaging visits within 90 days, 746 subjects have both GM and FNC, while only 293 have all three modalities. A complete-view model would therefore exclude GM and FNC observations from 453 additional subjects.

We reserve 20% of the fully matched subjects as an independent test set shared by all methods. We perform five-fold cross-validation on the remaining 80% for model selection. Within each fold, validation and test subjects are excluded from representation learning and preprocessing at every stage. We compare ProgMF, using $\text{GM} + \text{FNC} \rightarrow \text{PET}$, against complete-view IVA and Soft Imputation. All methods estimate 20 SCVs.

After model selection, the models are retrained on the full development set and evaluated on the same held-out test subjects. We use class-weighted logistic regression for CN versus dementia classification and report accuracy and F1 Score. Accuracy is calculated as the mean of the class-specific recalls, while F1 Score is calculated as the mean of the class-specific F1 values. Further details are provided in Appendix D.

4 RESULTS

We evaluate PROGMF using controlled simulations and real ADNI neuroimaging data. We first examine source recovery under shrinking subject overlap in Section 4.1 and the contribution of additional early-stage subjects in Section 4.2. We then investigate fusion order in Section 4.3 and robustness to observation noise in Section 4.4. Finally, we evaluate the learned representations on ADNI in Section 4.5.

4.1 RECOVERY UNDER SHRINKING SUBJECT OVERLAP

We examine how source recovery changes as fewer subjects remain available across all modalities. As shown in Fig. 2, PROGMF and the imputation methods achieve similar recovery when subject overlap is high. As the final cohort becomes smaller, PROGMF generally achieves lower JISI, particularly under substantial overlap reduction.

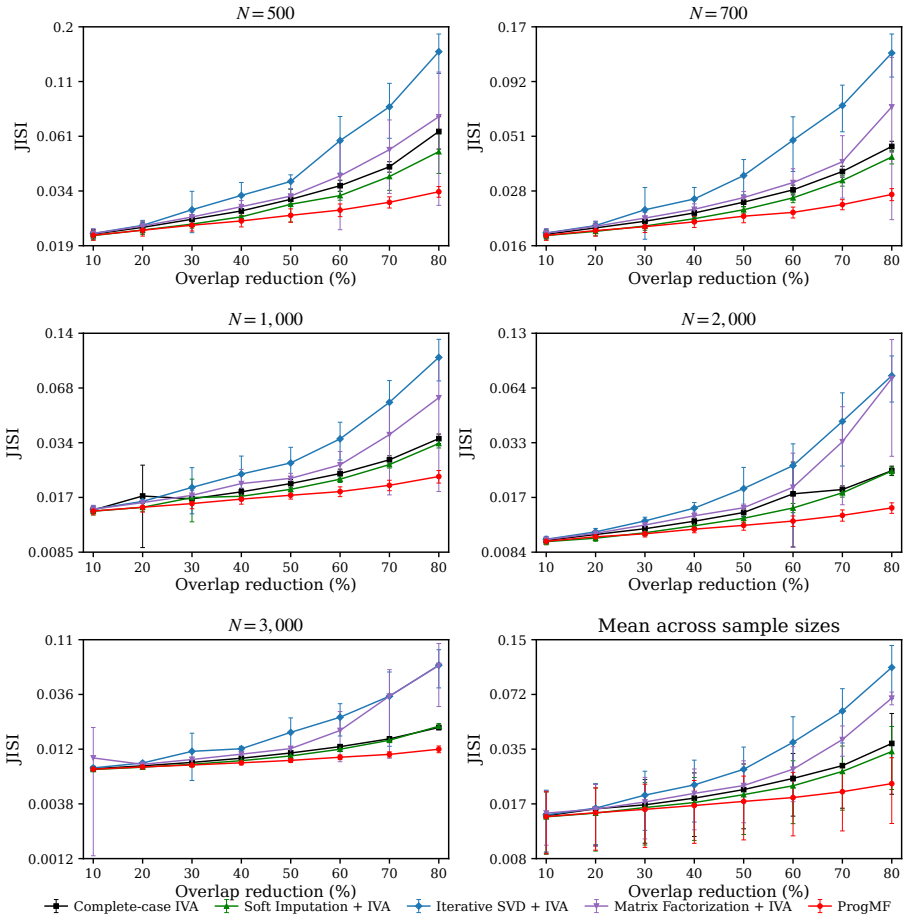

Figure 2: **Source recovery under shrinking subject overlap.** Mean JISI across five initial cohort sizes and overlap reductions from 10% to 80%. Error bars in the individual cohort-size panels show standard deviations across 20 independent datasets. The final panel shows the mean across cohort sizes, with standard deviations across the five cohort-size means. Lower JISI is better.

The imputation methods show increasing recovery error as the final cohort contracts, while complete-view IVA is limited to the subjects observed across all modalities. In contrast, PROGMF carries information learned from larger earlier cohorts into subsequent stages. These results show that progressive transfer can improve source recovery when the number of fully observed subjects is limited.

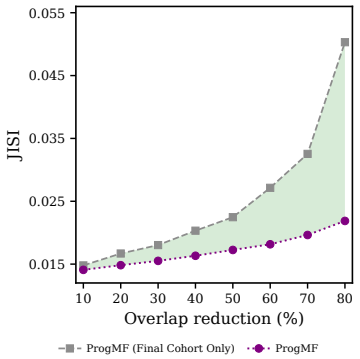

Figure 3: **Effect of larger early-stage cohorts on source recovery.** Mean JISI for PROGMF and its final-cohort variant, averaged across five initial cohort sizes. Shading indicates the difference between the mean curves. Lower JISI is better.

4.2 EFFECT OF LARGER EARLY-STAGE COHORTS

To assess the benefit of additional subjects during early fusion, we compare PROGMF with a final-cohort variant that uses only the final complete cohort at every stage. Both variants use the same progressive training procedure, modality order, preprocessing, initialization protocol, and final-stage observations. The final-cohort variant also uses the transfer coefficients selected for PROGMF, without retuning.

Figure 3 shows that both variants achieve similar recovery under mild overlap reduction. As overlap decreases, their recovery errors increasingly diverge. PROGMF exhibits a more gradual increase in recovery error, whereas the final-cohort variant shows substantially greater degradation as the final cohort contracts.

Because both variants follow the same progressive training procedure, the widening gap provides evidence that additional subjects available during early fusion improve subsequent source recovery. This comparison demonstrates the contribution of larger early-stage cohorts under fixed transfer coefficients, without independently retuning the final-cohort variant.

4.3 SENSITIVITY TO FUSION ORDER

We examine how fusion order affects source recovery using a simulation with strong GM–FNC coupling, moderate FA–PET coupling, and weaker coupling between the remaining modality pairs. We evaluate all 12 fusion orders using $N = 1000$, including complete-view case and overlap reductions from 10% to 80%. Transfer coefficients are selected using five independent calibration datasets on the canonical fusion order and held fixed across all orders within each overlap condition. Evaluation uses 20 additional independent datasets shared across all orders, with no evaluation data used for transfer-coefficient selection.

Figure 4 shows that schedules beginning with the strongly coupled GM–FNC pair generally achieve lower recovery error. Differences also appear under complete subject overlap, indicating that fusion order can affect recovery even when all modalities are available for every subject.

These findings demonstrate sensitivity to fusion order. However, because coupling strength and modality identity are not varied independently, they do not establish a general rule for selecting the initial modality pair.

4.4 ROBUSTNESS TO OBSERVATION NOISE

We examine whether progressive transfer remains effective when shrinking subject overlap is combined with observation noise. Using $N = 1000$, we evaluate four SNR levels $\{30, 10, 3, 0.46\}$ dB across eight overlap-reduction levels. Within each dataset, the sources, mixing matrices, subject groups, and Gaussian noise patterns remain fixed; only the noise magnitude changes. PROGMF retains its transfer coefficients selected at 10 dB, without additional tuning for each SNR.

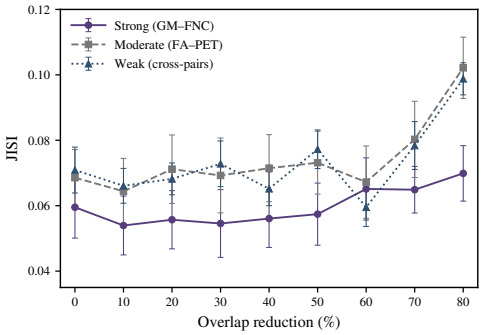

Figure 4: **Effect of fusion order on source recovery.** Mean final JISI for the selected schedule within each initial-pair coupling group (strong, moderate, and weak). Error bars indicate 95% confidence intervals across 20 independent evaluation datasets. Lower JISI is better.

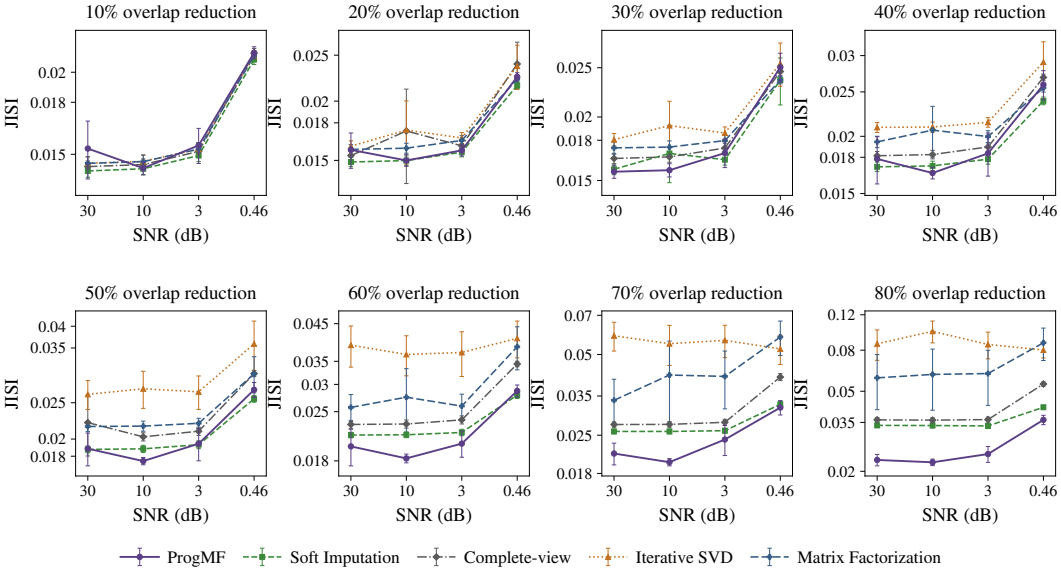

Figure 5: **Recovery under observation noise and shrinking subject overlap.** Final JISI across four SNRs and eight levels of overlap reduction. Curves show means over 20 datasets, with 95% confidence intervals. Each panel uses a logarithmic vertical scale with its own axis limits. Lower JISI indicates better recovery.

Figure 5 shows relatively small differences between methods under mild overlap reduction. As the final cohort becomes smaller, the differences become more pronounced. PROGMF generally maintains lower mean JISI under severe overlap reduction across the evaluated noise levels.

The improvement is less consistent at intermediate overlap levels. Nevertheless, the recovery advantage under severe overlap reduction persists as observation noise increases, despite using transfer coefficients selected at a single SNR. These results indicate that progressive transfer can remain effective under varying noise conditions, particularly when the final complete cohort is small.

4.5 EVALUATION ON ADNI NEUROIMAGING

We evaluate the learned representations for cognitively normal (CN) versus dementia classification using ADNI. All methods use the same independent 20% test set, with five-fold cross-validation performed on the remaining subjects for model selection.

Figure 6 shows that PROGMF achieves higher accuracy and F1 Score than complete-view IVA and Soft Imputation on the independent test set. These results suggest that the learned multimodal representations provide useful information for downstream diagnosis classification.

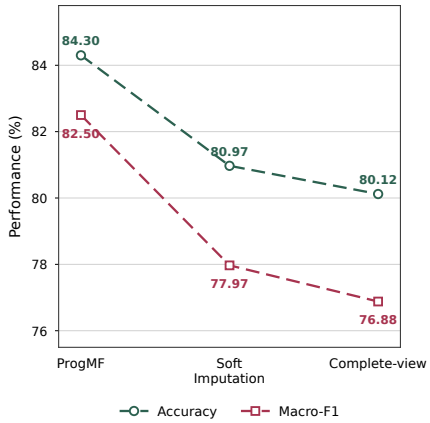

Figure 6: **ADNI diagnosis classification.** Accuracy and F1 Score on the independent held-out test set following five-fold cross-validation for model selection.

5 Discussion and Conclusion

We introduced Progressive Multimodal Fusion (ProgMF), a framework that incorporates modalities progressively while transferring representations learned from larger earlier-stage cohorts. Using IVA and SCV-aligned anchors, the framework enables successive fusion without imputing unavailable observations.

Controlled simulations show improved source recovery as subject overlap decreases, while the cohort ablation demonstrates the contribution of additional early-stage subjects under fixed transfer coefficients. Experiments using independently calibrated transfer coefficients also reveal sensitivity to fusion order and robustness across observation noise levels. On ADNI, ProgMF achieves higher accuracy and F1 Score than the evaluated baselines on an independent held-out test set.

Acknowledgments

To be added upon acceptance.

References

Julián N. Acosta, Guido J. Falcone, Pranav Rajpurkar, and Eric J. Topol. Multimodal biomedical ai. *Nature Medicine*, 28:1773–1784, 2022. doi: 10.1038/s41591-022-01981-2.

Matthew Anderson, Tülay Adalı, and Xi-Lin Li. Joint blind source separation with multivariate gaussian model: Algorithms and performance analysis. *IEEE Transactions on Signal Processing*, 60(4):1672–1683, 2012. doi: 10.1109/TSP.2011.2181836.

John Ashburner and Karl J. Friston. Unified segmentation. *NeuroImage*, 26(3):839–851, 2005. doi: 10.1016/j.neuroimage.2005.02.018.

Hunor Csala, Omid Amili, Roshan M. D'Souza, and Amirhossein Arzani. A comparison of machine learning methods for recovering noisy and missing 4D flow MRI data. *International Journal for Numerical Methods in Biomedical Engineering*, 40(11):e3858, 2024. doi: 10.1002/cnm.3858.

Yuhui Du, Zening Fu, Jing Sui, Shuang Gao, Yuhui Xing, Dan Lin, Mustafa Salman, Anees Abrol, Md Mahfuzur Rahaman, Jiayu Chen, et al. Neuromark: An automated and adaptive ICA-based pipeline to identify reproducible fMRI markers of brain disorders. *NeuroImage: Clinical*, 28: 102375, 2020. doi: 10.1016/j.nicl.2020.102375.

Erik B. Erhardt, Elena A. Allen, Yonghua Wei, Tom Eichele, and Vince D. Calhoun. SimTB, a simulation toolbox for fMRI data under a model of spatiotemporal separability. *NeuroImage*, 59 (4):4160–4167, 2012. doi: 10.1016/j.neuroimage.2011.11.088.

Nasir Hayat, Krzysztof J. Geras, and Farah E. Shamout. MedFuse: Multi-modal fusion with clinical time-series data and chest x-ray images. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pp. 479–503. PMLR, 2022. URL https://proceedings.mlr.press/v182/hayat22a.html.

William J. Jagust, Susan M. Landau, Robert A. Koeppe, Eric M. Reiman, Kewei Chen, Chester A. Mathis, Julie C. Price, Norman L. Foster, and Arthur Y. Wang. The alzheimer's disease neuroimaging initiative 2 PET core: 2015. *Alzheimer's & Dementia*, 11(7):757–771, 2015. doi: 10.1016/j.jalz.2015.05.001.

William J. Jagust, Robert A. Koeppe, Gil D. Rabinovici, Victor L. Villemagne, Theresa M. Harrison, and Susan M. Landau. The ADNI PET core at 20. *Alzheimer's & Dementia*, 20(10):7340–7349, 2024. doi: 10.1002/alz.14165.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009. doi: 10.1109/MC.2009.263.

Susan M. Landau, Anthony Fero, Suzanne L. Baker, Robert Koeppe, Mark Mintun, Kewei Chen, Eric M. Reiman, and William J. Jagust. Measurement of longitudinal β-amyloid change with ^{18}f-florbetapir PET and standardized uptake value ratios. *Journal of Nuclear Medicine*, 56(4): 567–574, 2015. doi: 10.2967/jnumed.114.148981.

Tao Ma, Jinfu Nie, Jian Huang, Yong-Biao Zhang, Joanna M. Biernacka, and Liguo Wang. Multi-output learning for systematic missing value imputation in DNA methylation arrays. *Bioinformatics Advances*, 6(1):vbag052, 2026. doi: 10.1093/bioadv/vbag052.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11(80):2287–2322, 2010. URL https://www.jmlr.org/papers/v11/mazumder10a.html.

Rogers F. Silva, Sergey M. Plis, Tülay Adali, Marios S. Pattichis, and Vince D. Calhoun. Multidataset independent subspace analysis with application to multimodal fusion. *IEEE Transactions on Image Processing*, 30:588–602, 2021. doi: 10.1109/TIP.2020.3028452. URL https://doi.org/10.1109/TIP.2020.3028452.

Rogers F. Silva, Eswar Damaraju, Xinhui Li, Peter Kochunov, Judith M. Ford, Daniel H. Mathalon, Jessica A. Turner, Theo G. M. van Erp, Tülay Adali, and Vince D. Calhoun. A method for multimodal IVA fusion within a MISA unified model reveals markers of age, sex, cognition, and schizophrenia in large neuroimaging studies. *Human Brain Mapping*, 45(17):e70037, 2024. doi: 10.1002/hbm.70037. URL https://doi.org/10.1002/hbm.70037.

Victor Solomon, Zening Fu, Rafal Angryk, Vince D. Calhoun, and Jingyu Liu. MSR-IVA: Masked structural residual independent vector analysis for state-aware fusion of structural MRI and dynamic functional network connectivity, 2026. URL https://arxiv.org/abs/2608.24978. Accepted at IEEE MLSP 2026.

Jing Sui, Tülay Adali, Qingbao Yu, Jiayu Chen, and Vince D. Calhoun. A review of multivariate methods for multimodal fusion of brain imaging data. *Journal of Neuroscience Methods*, 204(1): 68–81, 2012. doi: 10.1016/j.jneumeth.2011.10.031.

Hong Tao, Chenping Hou, Dongyun Yi, Jubo Zhu, and Dewen Hu. Joint embedding learning and low-rank approximation: A framework for incomplete multiview learning. *IEEE Transactions on Cybernetics*, 51(3):1690–1703, 2021. doi: 10.1109/TCYB.2019.2953564. URL https://doi.org/10.1109/TCYB.2019.2953564.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001. doi: 10.1093/bioinformatics/17.6.520.

Dallas P. Veitch, Michael W. Weiner, Melanie Miller, et al. The alzheimer's disease neuroimaging initiative in the era of alzheimer's disease treatment: A review of adni studies from 2021 to 2022. *Alzheimer's & Dementia*, 20(1):652–694, 2024. doi: 10.1002/alz.13449.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15878–15887, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Wang_Multi-Modal_Learning_With_Missing_Modality_via_Shared-Specific_Feature_Modelling_CVPR_2023_paper.html.

Qi Wang, Liang Zhan, Paul M. Thompson, and Jiayu Zhou. Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1828–1838, 2020. doi: 10.1145/3394486.3403234. URL https://doi.org/10.1145/3394486.3403234.

APPENDIX

Table of Contents

A Additional Method Details **14**

A.1 Scale-Controlled Kotz Contrast 14

A.2 Cross-Stage Component Alignment 14

B Synthetic Data and Calibration **14**

B.1 Source and Mixing-Matrix Generation 14

B.2 Simulation Preprocessing and Imputation 15

B.3 Transfer Calibration 15

C Experimental Implementation **15**

D ADNI Data and Evaluation **16**

D.1 Cohort Construction 16

D.2 Gray Matter MRI Preprocessing 16

D.3 Resting-State fMRI and FNC53 16

D.4 AV45 Amyloid PET Preprocessing 16

D.5 Fixed Feature Transformations 17

D.6 Cohort Characteristics 17

D.7 Data Splitting and Classification 17

E Use of Large Language Models **18**

A Additional Method Details

A.1 Scale-Controlled Kotz Contrast

We provide the numerical details of the IVA objective in Eq. equation 2. Let $d_t = |\mathcal{M}_t|$ denote the number of active modalities at stage t. For each SCV k, the sample covariance of the centered source estimates is $\widehat{\Sigma}_{k,t}$. Writing $V_{k,t} = \mathrm{diag}(\widehat{\Sigma}_{k,t})$, we estimate its cross-modal correlation matrix as $R_{k,t} = V_{k,t}^{-1/2}\widehat{\Sigma}_{k,t}V_{k,t}^{-1/2}$. The matrix is symmetrized and diagonally regularized to obtain $\widetilde{R}_{k,t}$.

With the Kotz parameters $\beta = 0.5$, $\lambda_{\mathrm{K}} = 1$, and $\eta = 1$, the SCV density becomes $p_{k,t}(\mathbf{y}) = C_{d_t}|\widetilde{R}_{k,t}|^{-1/2}\exp(-\sqrt{\mathbf{y}^\top \widetilde{R}_{k,t}^{-1}\mathbf{y}})$, where $C_{d_t} = \Gamma(d_t/2)/[2\pi^{d_t/2}\Gamma(d_t)]$.

Defining $q_{k,t,n} = (\mathbf{y}_{k,t}^{(n)})^\top \widetilde{R}_{k,t}^{-1}\mathbf{y}_{k,t}^{(n)}$, the implemented contrast is

$$\begin{aligned}\mathcal{L}_{\mathrm{IVA},t} = &\sum_{k=1}^{K}\left[\frac{1}{2}\log\det\widetilde{R}_{k,t} - \log C_{d_t} + \frac{1}{N_t}\sum_{n=1}^{N_t}\sqrt{q_{k,t,n}}\right] \\ &- \frac{1}{2}\sum_{m\in\mathcal{M}_t}\log\det(W_{m,t}W_{m,t}^\top + \delta I_K).\end{aligned} \tag{5}$$

The correlation matrices are updated from the current source estimates during optimization. The Kotz normalization constants are retained because the numerical IVA objective is also used to determine the transfer scale c_t.

A.2 Cross-Stage Component Alignment

For current component i and preceding component j, similarity is computed across the inherited modalities as

$$S_{ij}^{(t)} = \sum_{m\in\mathcal{M}_{t-1}}\left|\left\langle \mathcal{N}(\bar{W}_{m,t})_{i:}, \mathcal{N}(\widehat{W}_{m,t-1})_{j:}\right\rangle\right|. \tag{6}$$

The Hungarian algorithm finds a single permutation $\pi_t^\star$ that maximizes $\sum_i S_{i,\pi(i)}^{(t)}$. For each inherited modality, matched component signs are determined independently from their normalized row inner products.

The aligned reference is $\widetilde{W}_{m,t-1,i:} = a_{m,i}^{(t)}\widehat{W}_{m,t-1,\pi_t^\star(i):}$, where $a_{m,i}^{(t)}$ is the corresponding sign. The permutation and signs are determined once after independent stage estimation. All aligned references remain fixed during refinement.

B Synthetic Data and Calibration

B.1 Source and Mixing-Matrix Generation

The simulation uses four neuroimaging-inspired modalities. GM and PET employ distinct SimTB-derived spatial patterns (Erhardt et al., 2012), FA uses smooth tract-like patterns, and FNC uses structured signed connectivity patterns. The columns of each mixing matrix are normalized to unit Euclidean norm.

The default source model contains 10 independent SCVs, with cross-modal dependence within SCV k specified by

$$R_k = (1-\rho_k)I_4 + \rho_k\mathbf{1}\mathbf{1}^\top,$$

where

$$(\rho_1, \ldots, \rho_{10}) = (0.25, 0.30, 0.35, 0.45, 0.50, 0.55, 0.60, 0.70, 0.75, 0.80).$$

The fusion-order experiment uses a separate heterogeneous dependence structure to produce strong GM–FNC coupling, moderate FA–PET coupling, and weaker dependence between the remaining modality pairs.

Within each simulation realization, all compared methods use identical sources, mixing matrices, noise realizations, and subject subsets. Gaussian noise is scaled to the specified signal-to-noise ratio.

B.2 Simulation Preprocessing and Imputation

Each modality is centered, reduced using PCA, and whitened. The modality-specific transformation is estimated from the final complete cohort and reused across all fusion stages and comparison methods. The ground-truth mixing matrices undergo the corresponding feature transformations to obtain A_m^{eff}, as used in the JISI calculation.

For the imputation baselines, missing modality blocks are reconstructed before fitting IVA. Soft Imputation uses nuclear-norm-regularized matrix completion, Iterative SVD repeatedly applies truncated singular-value decomposition, and Matrix Factorization estimates low-rank factors.

Observed values remain unchanged. Completion hyperparameters are selected using held-out block reconstruction error rather than ground-truth source recovery. All completed datasets are subsequently evaluated using the same IVA fitting procedure.

B.3 Transfer Calibration

The primary shrinking-overlap experiment selects its stage-specific transfer coefficients from the strictly positive grid described in the main paper. Selection uses ground-truth JISI on independent synthetic calibration datasets. The coefficients are fixed before evaluation on the 20 independent datasets for each condition.

The earlier-stage cohort ablation uses the same selected coefficients for both variants without independently retuning the final-cohort variant.

For the fusion-order experiment, coefficients are calibrated on the canonical order using five independent datasets and held fixed across all 12 schedules within each overlap condition.

The schedules are grouped by their initial modality pair: two begin with GM–FNC, two begin with FA–PET, and eight begin with one of the remaining modality pairs. The schedule with the lowest mean calibration JISI within each group is selected for the main figure. The selected schedules are evaluated on 20 additional independent datasets.

The noise-robustness experiment reuses coefficients selected at $10\,\mathrm{dB}$ without SNR-specific retuning.

C Experimental Implementation

IVA models use FastICA-based initialization and L-BFGS-B optimization with automatic differentiation. Each progressive stage is independently initialized. After fitting the current-stage IVA model, component alignment is performed once, followed by optimization under the fixed structural anchor.

The IVA baselines and independent ProgMF stage estimates use the same base objective and optimization settings. Progressive refinement additionally minimizes the transfer objective in Eq. equation 4.

The implementation uses Python, PyTorch, NumPy, and SciPy. Simulation conditions are executed using SLURM-based parallel processing.

D ADNI DATA AND EVALUATION

D.1 COHORT CONSTRUCTION

We use three modalities from the Alzheimer's Disease Neuroimaging Initiative (ADNI): gray matter structural MRI, resting-state functional network connectivity (FNC), and florbetapir (AV45) amyloid PET (Veitch et al., 2024; Jagust et al., 2024).

Imaging records are matched by participant identifier and acquisition date within 90 days. A common clinical assessment is associated with each selected multimodal observation. When multiple imaging records are available, matched records are selected consistently so that inherited observations are identical across progressive stages.

The evaluated progression is $\mathrm{GM} + \mathrm{FNC} \rightarrow \mathrm{PET}$. The initial GM–FNC cohort contains 746 subjects, of whom 293 have all three modalities. The remaining 453 subjects contribute only at the first fusion stage. These counts describe the matched cohort before development and test subjects are separated.

D.2 GRAY MATTER MRI PREPROCESSING

T1-weighted structural MRI is represented using gray matter maps derived through voxel-based morphometry in SPM12. The preprocessing pipeline includes bias-field correction, tissue segmentation, spatial normalization to Montreal Neurological Institute space, modulation to preserve local tissue volume, and spatial smoothing (Ashburner & Friston, 2005; Sui et al., 2012).

The resulting spatially aligned gray matter maps are restricted to a common analysis mask and vectorized before dimensionality reduction. A consistent voxel ordering is maintained across subjects and fusion stages.

D.3 RESTING-STATE FMRI AND FNC53

Resting-state fMRI is represented using the NeuroMark framework implemented in GIFT (Du et al., 2020). The framework estimates subject-specific intrinsic connectivity networks and their corresponding time courses using spatially constrained independent component analysis.

The 53 networks span seven functional domains: subcortical, auditory, sensorimotor, visual, cognitive control, default mode, and cerebellar.

After temporal preprocessing, including detrending, despiking, motion-related correction, and frequency filtering, static FNC is estimated using pairwise Pearson correlations between the network time courses.

The unique off-diagonal entries of the symmetric 53×53 correlation matrix are retained, yielding 1,378 FNC features per subject. The same network ordering is used for all subjects.

D.4 AV45 AMYLOID PET PREPROCESSING

We use preprocessed ^{18}F-florbetapir (AV45) PET images, which measure fibrillar amyloid-β deposition. The PET processing pipeline includes motion correction, co-registration, frame averaging, spatial standardization, and smoothing (Jagust et al., 2015; Landau et al., 2015).

Voxel intensities are expressed as standardized uptake value ratios (SUVR), using the cerebellar cortex as the reference region:

$$\mathrm{SUVR}(v) = \frac{\mathrm{SUV}(v)}{\overline{\mathrm{SUV}}_{\text{cerebellar cortex}}}.$$

The processed volumes are restricted to cortical gray matter using a common spatial mask before vectorization and dimensionality reduction. This restriction excludes the cerebellar reference region and noncortical tissue from the PET representation.

Table 1: **ADNI cohort characteristics.** Demographic and clinical characteristics of the 293-subject complete GM–FNC–PET cohort. Continuous variables are reported as mean ± standard deviation.

Characteristic	Overall	CN	MCI	Dementia
Participants, n	293	146	85	62
Age, years	75.37 ± 7.57	74.44 ± 7.81	76.06 ± 6.95	76.64 ± 7.64
Education, years	16.39 ± 2.51	16.90 ± 2.30	16.16 ± 2.73	15.48 ± 2.39
APOE $\varepsilon 4$ carrier, n (%)	112 (39.9)	45 (31.9)	31 (38.8)	36 (60.0)
MMSE	27.42 ± 3.37	29.11 ± 1.09	27.73 ± 2.37	23.11 ± 4.17
ADAS-Cog 13	14.84 ± 11.28	8.53 ± 4.25	14.20 ± 7.54	30.30 ± 11.92
CDR-SB	1.40 ± 2.22	0.13 ± 0.37	1.28 ± 1.19	4.46 ± 2.75

APOE $\varepsilon 4$ status was available for 281 participants: 141 CN, 80 MCI, and 60 dementia. Carrier percentages use participants with available APOE status as the denominator. Clinical summaries use available measurements at the matched assessment.
CN, cognitively normal; MCI, mild cognitive impairment; MMSE, Mini-Mental State Examination; ADAS-Cog 13, Alzheimer's Disease Assessment Scale–Cognitive Subscale; CDR-SB, Clinical Dementia Rating Sum of Boxes.

D.5 Fixed Feature Transformations

Cross-stage transfer requires the inherited modality representations to remain in the same feature coordinates. We therefore use one fixed modality-specific PCA and whitening transformation within each training run.

Let $X_{m,t}^{\mathrm{raw}} \in \mathbb{R}^{V_m \times N_t}$ denote the vectorized imaging observations before dimensionality reduction. With feature mean μ_m and fixed PCA-whitening matrix $P_m \in \mathbb{R}^{D_m \times V_m}$, the inputs to IVA are

$$X_{m,t} = P_m \left(X_{m,t}^{\mathrm{raw}} - \mu_m \mathbf{1}^\top \right) \in \mathbb{R}^{D_m \times N_t}.$$

Here D_m is the post-transformation feature dimension used by the IVA model. PCA retains components according to an 85% explained-variance threshold, followed by whitening.

For the held-out evaluation, the feature transformations are estimated using training subjects only and reused across progressive stages and comparison methods. The common model order is $K = 20$ SCVs for all three modalities.

D.6 Cohort Characteristics

Table 1 summarizes the characteristics to be reported for the current 293-subject three-modality cohort. Diagnostic groups are defined using the current clinical diagnosis associated with the matched imaging visit rather than baseline diagnosis.

D.7 Data Splitting and Classification

We evaluate cognitively normal versus dementia classification using the current diagnosis associated with the matched imaging visit. Participants with MCI or other diagnoses are excluded from the binary classification task.

A common independent test set is formed from 20% of the fully matched subjects. Five-fold cross-validation on the remaining development subjects is used for model selection. Within each fold, validation and test subjects are excluded from preprocessing, representation learning, and classifier fitting at all stages.

Following model selection, the representation models and a class-weighted logistic regression classifier are retrained on the development set and evaluated on the same independent test subjects for all compared methods.

Accuracy is calculated as the mean of the two class-specific recalls. F1 Score is calculated as the mean of the two class-specific F1 values.

E USE OF LARGE LANGUAGE MODELS

Large language models (LLMs) were used solely for language polishing and editorial assistance. They did not contribute to the theoretical development, methodological design, experimental results, or scientific conclusions of this work.