LEARNING CONSISTENT MULTIMODAL REPRESENTATIONS UNDER MISSING-NOT-AT-RANDOM SHIFTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Missing modalities are ubiquitous in real-world multimodal applications. Although substantial progress has been made in learning from incomplete multimodal data, most existing methods implicitly assume that modality missingness occurs at random and is independent of the prediction target. In practice, however, modality availability can be correlated with the label, giving rise to missing-not-at-random (MNAR) phenomenon. Under such conditions, multimodal models may exploit missingness patterns as predictive shortcuts rather than learning the underlying multimodal semantics, leading to severe performance degradation when missingness patterns shift at inference time. We observe that such shortcut learning induces representation inconsistency, where the representation of the same instance changes with the set of available modalities. We theoretically show that the performance degradation under missingness-pattern shift is bounded by two factors: the discrepancy between source and target missingness patterns and the model's representation inconsistency. This analysis suggests that reducing representation inconsistency can improve robustness to missingness-pattern shifts without requiring prior knowledge of the target missingness distribution. Motivated by this insight, we propose AURA, a multimodal learning framework that explicitly encourages representation consistency across different modality-availability patterns. To enable comprehensive evaluation under MNAR conditions, we introduce Controlled MNAR Shift Evaluation protocol, that systematically controls both the dependence between missingness and labels and the degree of source-target missingness-pattern shift. Extensive experiments on multiple large-scale multimodal datasets demonstrate that AURA consistently improves robustness under MNAR shifts while maintaining strong performance on standard missing-modality benchmarks. On stroke type prediction, AURA's relative improvement in balanced accuracy over the strongest competing method increases from 20.4% under milder MNAR shifts to 60.2% under more severe shifts. Our code is available at anonymous.4open.science.

1 INTRODUCTION

Multimodal models integrate information from multiple modalities to capture complementary aspects of the same sample and improve predictive accuracy (Ngiam et al., 2011; Baltrušaitis et al., 2019). A major challenge in multimodal learning is missing modalities where not all modalities are available for every sample. This challenge has motivated substantial research on learning from incomplete multimodal inputs (Wu et al., 2026; Zhan et al., 2025). Yun et al. (2024), Hemker et al. (2024), and Swamy et al. (2023) adapt the fusion process to different sets of available modalities, while Wu et al. (2025) models cross-modal distributions to generate representations for missing modalities. Other methods introduce additional learning objectives to improve robustness (Woo et al., 2023; Li et al., 2025; Chen et al., 2026a).

However, much of this work focuses on randomly missing modalities. In practice, missingness can be nonrandom and associated with the prediction target (Swamy et al., 2023; Liang et al., 2025). This distinction is particularly important in healthcare, where the availability of tests and imaging data depends on a patient's condition, clinical decisions, and care workflows. We observe this association in a stroke cohort derived from MIMIC-IV. As shown in Fig. 1, modality-availability patterns provide predictive signals for stroke types and mortality outcomes, even without using the

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.

observed modality content. Such correlations create a potential shortcut for multimodal models, which may rely on which modalities are present rather than on the underlying clinical evidence. This reliance can become problematic when changes in data collection sites or procedures alter the relationship between modality availability and the prediction target. We focus on robustness to such missing-not-at-random (MNAR) shifts.

We study this problem through the lens of **representation consistency**. Intuitively, representations of the same sample should remain stable across changes in modality availability while retaining information relevant to the prediction task. We derive a theoretical bound on performance degradation under missingness-pattern shifts in terms of the shift magnitude and the expected worst-case representation variation across availability patterns. This result provides a simple principle for improving MNAR robustness, that is reducing the sensitivity of learned representations to modality-availability patterns while preserving their predictive utility.

Motivated by this insight, we introduce AURA, a multimodal learning framework that promotes representation consistency. AURA first constructs modality-specific features and then refines them using a selective cross-modal attention (SCMA) mechanism. During training, it creates augmented views by randomly removing subsets of the available modalities and explicitly aligns the resulting representation with that of the observed view. Together, cross-modal refinement and representation alignment encourage the model to capture relationships between modalities while mitigating reliance on modality-availability shortcuts.

To assess robustness to MNAR shifts, we introduce the **Controlled MNAR Shift Evaluation** protocol. This protocol systematically varies both the dependence between modality missingness and labels and the degree of train-test missingness-pattern shift. It enables evaluation across a range of settings to examine how models respond when the relationship between modality availability and labels changes at test time.

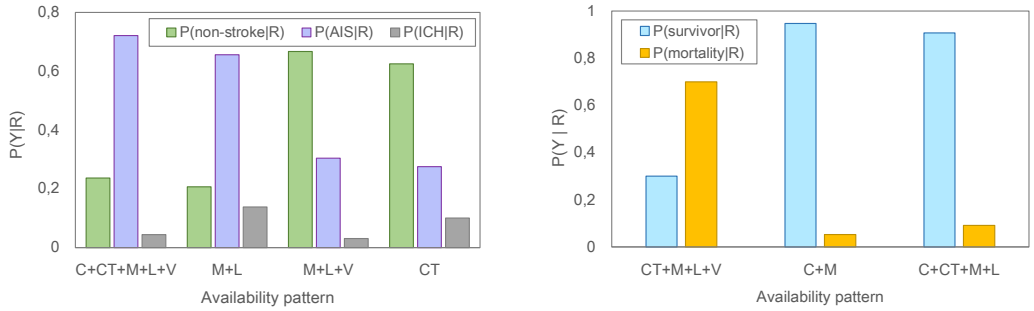
Our contributions are summarized as follows:

- We study robustness to MNAR shifts and introduce **representation consistency** as a guiding principle. We derive a theoretical bound linking performance degradation to the magnitude of the missingness shift and the expected worst-case representation variation across availability patterns.
- Building on this theoretical insight, we propose **AURA**, a multimodal learning framework that promotes representation consistency across different modality-availability patterns. Experiments show that AURA outperforms state-of-the-art baselines in the evaluated missing-modality settings, including MNAR shifts.
- We introduce the **Controlled MNAR Shift Evaluation** protocol, enabling systematic evaluation of multimodal models across different strengths of missingness-label dependence and degrees of train-test missingness-pattern shift.

2 Related Work

Multimodal learning with missing modalities has been studied through flexible fusion, missing-modality representations, and additional robustness objectives (Yun et al., 2024; Hemker et al., 2024; Swamy et al., 2023; Wu et al., 2025; Woo et al., 2023; Li et al., 2025). Several works also study representation changes across modality subsets (Lin & Hu, 2023; Chen et al., 2024; Reza et al., 2025), with empirical evidence suggesting that more robust models tend to exhibit smaller representation variation across missingness patterns. Our work builds on these directions by studying missingness-mechanism shift and theoretically relating representation inconsistency across availability patterns to changes in prediction loss.

Related work on non-random missingness further shows that modality availability can be informative (Liang et al., 2025) and that models may fail when label–missingness relationships change (Swamy et al., 2023). Existing stress tests, however, change this relationship together with the amount of removed modality information, making the two effects difficult to disentangle. We therefore introduce a controlled evaluation that changes the label–missingness relationship while matching per-modality artificial-removal rates. A broader discussion is provided in Appendix A.



(a) Stroke type prediction

(b) Mortality prediction

Figure 1: Modality availability patterns are predictive of clinical outcomes. C, CT, M, L, and V denote clinical text, CT, MRI reports, laboratory measurements, and vital signs, respectively. (a) Stroke-type differs across patterns, with some patterns more strongly associated with AIS and others with non-stroke cases. (b) Mortality varies across patterns, with higher mortality for CT+M+L+V than for C+M or C+CT+M+L.

3 Problem Formulation and Representation Consistency

3.1 Multimodal Data and Missingness

Our formulation follows the standard full-data view of missing data (Rubin, 1976). Consider a multimodal learning problem with modality set $\mathcal{M} = \{1, \ldots, M\}$. Each sample is associated with an underlying latent state S. Let $X = (X_1, \ldots, X_M)$ denote the multimodal observation, where X_m is the observation from modality m, and let Y denote the prediction target

$$X_m \sim P(X_m \mid S), \qquad Y \sim P(Y \mid S). \tag{1}$$

We denote the modality-availability pattern by $R = (R_1, \ldots, R_M) \in \{0,1\}^M$, where $R_m = 1$ indicates that modality m is observed and $R_m = 0$ that it is missing. For a realized pattern r, let

$$X^{(r)} = \{X_m : m \in \mathcal{M},\ r_m = 1\}. \tag{2}$$

A multimodal model F_θ maps the available inputs $X^{(r)}$ and their availability pattern r to a representation Z_r:

$$Z_r = F_\theta(X^{(r)}, r), \qquad \hat{Y}_r = h(Z_r), \tag{3}$$

where h denotes the prediction head and $\hat{Y}_r$ is the prediction under availability pattern r. Thus, a sample may produce different representations Z_r depending on which modalities are available.

3.2 Missingness-Mechanism Shift

We describe the process that determines modality availability by a missingness mechanism

$$q(r \mid S, X, Y) = P(R = r \mid S, X, Y). \tag{4}$$

Under missing-not-at-random (MNAR) missingness, modality availability can depend on the prediction target. Thus, the availability pattern R may become predictive of Y, allowing a model to exploit a shortcut from availability to the target rather than relying on the observed modality content.

Let P_{tr} and P_{te} denote the data distributions in the training and test environments, respectively, with corresponding missingness mechanisms q_{tr} and q_{te}. We consider settings where the data distribution remains unchanged between training and test, but the missingness mechanism changes

$$P_{\mathrm{tr}}(S, X, Y) = P_{\mathrm{te}}(S, X, Y), \qquad q_{\mathrm{tr}}(r \mid S, X, Y) \neq q_{\mathrm{te}}(r \mid S, X, Y). \tag{5}$$

Thus, a model that relies strongly on the relationship between R and Y during training may suffer a large performance drop when this relationship changes at test time. We refer to robustness against such changes as *robustness to missingness-mechanism shift*.

3.3 Representation Consistency and Robustness Bound

To reduce sensitivity to missingness-mechanism shifts, we consider the principle of *representation consistency*. Let $\mathcal{R}$ denote the set of admissible availability patterns for a sample, and let $r_0 \in \mathcal{R}$ denote a reference pattern, such as the original view. We define the sample-level representation inconsistency as the largest representation discrepancy from the reference pattern:

$$\delta(S, X, Y) = \max_{r \in \mathcal{R}} \|Z_r - Z_{r_0}\|, \tag{6}$$

where $\|\cdot\|$ denotes the Euclidean norm. Let $\ell(\hat{Y}, Y)$ denote the task-specific loss, and let $\mathcal{L}_q$ denote the expected task loss under missingness mechanism q. Here, $\circ$ denotes function composition.

Theorem (Missingness-shift sensitivity bound). Suppose that the data distribution is unchanged between training and test, $P_{\mathrm{tr}}(S, X, Y) = P_{\mathrm{te}}(S, X, Y)$, while the missingness mechanism changes from q_{tr} to q_{te}. If $\ell \circ h$ is L-Lipschitz with respect to the final representation, then:

$$|\mathcal{L}_{q_{\mathrm{te}}} - \mathcal{L}_{q_{\mathrm{tr}}}| \le 2L\, \mathbb{E}_{(S,X,Y)\sim P} \left[\mathrm{TV}(q_{\mathrm{te}}(\cdot \mid S, X, Y), q_{\mathrm{tr}}(\cdot \mid S, X, Y))\, \delta(S, X, Y) \right], \tag{7}$$

where TV denotes total variation distance. The full proof is provided in Appendix B. The result shows that the expected loss gap is bounded by the **magnitude of the missingness shift** and the **expected worst-case representation variation** across availability patterns. This result provides a theoretical basis for promoting representation consistency for MNAR shifts.

4 AURA

4.1 Overview

Guided by the representation-consistency principle established in Section 3, we introduce AURA, a multimodal learning framework that promotes representation stability across modality-availability patterns. As illustrated in Figure 2, AURA consists of three stages. First, the Modality-Specific Representation Learning stage encodes each modality into a dedicated representation. Second, the Cross-Modal Refinement stage selectively incorporates information from other available modalities to refine these representations. Finally, during training, AURA constructs augmented views by randomly masking subsets of the observed modalities and applies an alignment loss to encourage consistency between the representations of the original and augmented views.

4.2 Modality-Specific Representation Learning

Outputs from modality-specific encoders may differ in their semantics, feature statistics, and sequence lengths. Passing all encoder tokens directly to the fusion module therefore requires it to handle features from heterogeneous modalities while also learning their interactions. Instead, AURA first constructs a compact set of modality-specific feature tokens before cross-modal fusion. For modality m, a modality-specific encoder produces a sequence of tokens $E_m = \mathrm{Enc}_m(X_m)$. We then summarize E_m into s_m latent slots using a Perceiver-style cross-attention mechanism (Jaegle et al., 2021), which we refer to as Modality-Specific Attention (MSA):

$$H_m = \mathrm{MSA}_m(Q_m, E_m, E_m), \qquad H_m \in \mathbb{R}^{s_m \times d},$$

where $Q_m \in \mathbb{R}^{s_m \times d}$ contains learnable modality-specific latent queries. The number of slots s_m can be chosen separately for each modality, allowing long encoder sequences of tokens to be distilled into a bounded number of tokens before fusion.

When modality m is unavailable, its s_m feature token slots are replaced by a learnable modality-level missing representation $B_m \in \mathbb{R}^{s_m \times d}$:

$$\widetilde{H}_m = \begin{cases} H_m, & R_m = 1, \\ B_m, & R_m = 0. \end{cases}$$

Unlike combination-conditioned missing banks, which associate missing representations with combinations of observed modalities (Yun et al., 2024), we use a single missing representation for each modality. As a result, the bank grows linearly rather than combinatorially with the number of modalities. This simpler design also complements the augmented-view training introduced below. Because observed modalities are deliberately dropped when constructing augmented views, B_m can receive training signal even when modality m is rarely or never, naturally missing in the training data.

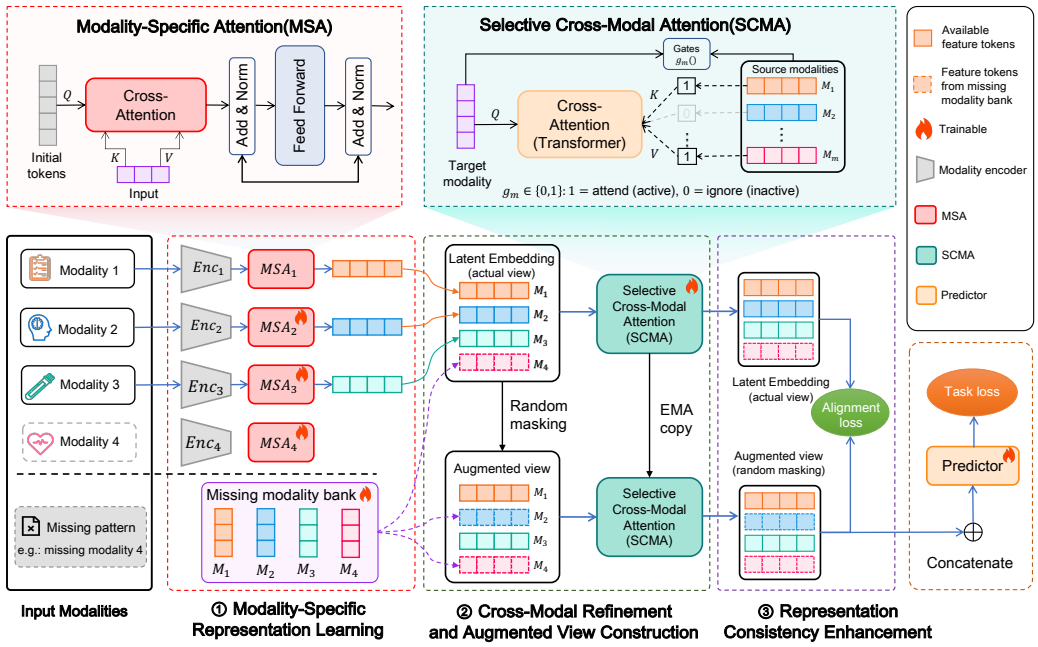


Figure 2: AURA (1) constructs modality-specific representations and then (2) selectively refines them using cross-modal context. During training, an augmented view is created by masking a subset of observed modalities, and its representation is aligned with that of the observed view to promote representation consistency (3).

4.3 Selective Cross-Modal Refinement

After constructing a compact representation for each modality, AURA allows these modality-specific features to exchange information through *Selective Cross-Modal Attention* (SCMA). SCMA treats each modality in turn as a target and selectively incorporates context from a subset of the remaining source modalities. For a target modality m, let $\mathcal{G}_m$ denote the set of source modalities selected for its refinement. SCMA can be written as:

$$H_m^{\text{ref}} = \text{SCMA}_m\left(\widetilde{H}_m, \{\widetilde{H}_j : j \in \mathcal{G}_m\}\right),$$

where $\widetilde{H}_m$ denotes the modality representation constructed as described in Section 4.2. Active source modalities provide keys and values for cross-attention, whereas inactive sources are ignored. Details of the gating mechanism $\mathcal{G}_m$ and SCMA implementation are provided in Appendix C.

4.4 Representation-Consistency Alignment

AURA constructs an augmented view by randomly removing a subset of the observed modalities, based on the representation-consistency result in Section 3.

Let $\mathcal{D}$ denote the set of modalities artificially removed from the augmented view. The augmented view is processed by the student network, while the observed view is processed by an exponential-moving-average (EMA) (Grill et al., 2020) copy of the representation branch. Let H_m^{aug} and $\bar{H}_m^{\text{obs}}$ denote their refined representations after SCMA:

$$\mathcal{L}_{\text{align}} = \frac{1}{|\mathcal{D}|} \sum_{m \in \mathcal{D}} d\left(H_m^{\text{aug}}, \bar{H}_m^{\text{obs}}\right). \tag{8}$$

In practice, $d(\cdot,\cdot)$ can be instantiated as an ℓ_2 distance, a cosine distance over corresponding token positions, or a Wasserstein distance between normalized representations. In our experiments, we use the Wasserstein distance as the default alignment discrepancy (See appendix F.4).

4.5 Representation Consistency Training Objective

The refined representations of the augmented view are concatenated and passed to the prediction head. We compute the standard task loss

$$\mathcal{L}_{\text{task}} = \ell(h(Z^{\text{aug}}), Y)\,, \tag{9}$$

where Z^{aug} denotes the representation obtained from the artificially masked view and ℓ is the task-specific loss. The overall training objective combines predictive learning with the representation-consistency alignment objective:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}, \tag{10}$$

where λ_{align} controls the strength of representation alignment. AURA does not force the model to ignore modality availability, because missingness can reflect the data-acquisition process and itself provide informative context (Liang et al., 2025). Instead, it discourages availability from becoming a dominant shortcut. The task loss allows the model to use helpful availability cues, while the alignment loss discourages overreliance on them.

5 Controlled MNAR Shift Evaluation Protocol

5.1 The Need for Controlled MNAR Evaluation

Swamy et al. (2023) evaluates MNAR robustness by masking one modality for a selected class with probability p_{mask} and changing the targeted class at test time. However, p_{mask} controls the amount of masking rather than the resulting label–missingness dependence, which can remain weak in multiclass settings (Appendix D). Increasing p_{mask} also removes more modality information, so the resulting performance drop can reflect both missingness-mechanism shift and information loss. We therefore propose a controlled protocol that varies the missingness mechanism while matching per-modality artificial-removal rates.

5.2 Controlled MNAR Environment Construction

Let $R \in \{0,1\}^M$ denote the original modality-availability pattern, and let $\hat{R}$ denote the pattern after artificial corruption. For each class y, we define a stochastic transform

$$T_y(\hat{r} \mid r) = P(\hat{R} = \hat{r} \mid R = r, Y = y), \tag{11}$$

where artificial corruption is drop-only, i.e., $\hat{r} \preceq r$, so that the transform can remove an observed modality. We additionally exclude the all-missing pattern $\hat{r} = \mathbf{0}$.

MNAR severity. We quantify the strength of the label–missingness dependence by the normalized mutual information

$$S_{\text{MNAR}}(T) = \frac{I_T(\hat{R}; Y)}{H(Y)}, \tag{12}$$

where the mutual information is computed under $P_T(\hat{R}, Y)$. This quantity measures how much label information is encoded in the corrupted availability pattern. We use a target value τ of S_{MNAR} to control the MNAR severity of the training dataset.

Artificial-removal cost. To quantify the amount of corruption caused by T, for each modality m, we define the fraction of observed instances that are removed as $D_m(T) = P_T(\hat{R}_m = 0 \mid R_m = 1)$.

We then define the overall artificial-removal cost as a weighted average:

$$D(T) = \frac{\sum_{m=1}^{M} w_m D_m(T)}{\sum_{m=1}^{M} w_m}, \tag{13}$$

where $w_m > 0$ denotes the removal cost associated with modality m. We assign larger weights to modalities with lower natural availability, so that removing a rarer modality incurs a larger penalty.

Training transformation. For a MNAR severity τ, we construct training transformation T_{train}:

$$T_{train} = \arg\min_{T} D(T) \quad \text{s.t.} \quad S_{\text{MNAR}}(T) = \tau. \tag{14}$$

Thus, among feasible mechanisms with the same MNAR severity, T_{train} minimizes the specified artificial-removal cost.

Shifted transformation. We next construct a second transformation T_{test} that matches the artificial-removal rate of each modality while changing the label–missingness distribution:

$$\text{Dist}(T_{train}, T) = \sum_{y} p(y)\ \text{JSD}\left(q_{train}(\hat{R} \mid y) \,\|\, q_T(\hat{R} \mid y)\right), \tag{15}$$

where q_{train} denotes the distribution induced by T_{train}. We then solve

$$T_{test} = \arg\max_{T} \text{Dist}(T_{train}, T) \quad \text{s.t.} \qquad D_m(T) = D_m(T_{train}), \qquad \forall m. \tag{16}$$

Consequently, T_{train} and T_{test} have matching modality artificial-removal rates but different label–missingness distributions. Complete details of the optimization are provided in Appendix E.

Evaluation protocol. For each training MNAR severity τ, we corrupt the training and validation sets using T_{train} and evaluate on the same test cohort under TEST-SAME (T_{train}) and TEST-SHIFT (T_{test}). The performance gap reflects sensitivity to changes in the label–missingness relationship.

6 PERFORMANCE EVALUATION

6.1 EXPERIMENTAL SETTINGS

Datasets and Tasks. We evaluate AURA on four prediction tasks across MIMIC-IV (Johnson et al., 2024; 2023) and ADNI (Mueller et al., 2005). The MIMIC-IV experiments include three-class stroke type prediction, 100-day mortality prediction, and length-of-stay regression, using clinical notes, radiology reports, laboratory measurements, vital signs, and admission ICD codes. Full details of cohort construction, task definitions, and preprocessing are provided in Appendix G. We additionally evaluate three-class disease-stage prediction on ADNI, following the experimental setting of Flex-MoE (Yun et al., 2024). We evaluate MNAR robustness on the stroke type and mortality tasks using the MultiModN protocol (Swamy et al., 2023), and further evaluate stroke type prediction using our Controlled MNAR Shift Evaluation protocol.

Baselines. We compare AURA against multimodal learning and missing-modality methods: MultiModN (Swamy et al., 2023), Flex-MoE (Yun et al., 2024), HEALNet (Hemker et al., 2024), AMC (Peng et al., 2025), and MASAM (Chen et al., 2026b). We additionally compare against the copula-based CM^2 framework (Wu et al., 2025). Since extending its copula-based modelling to our higher-modality settings incurs computational and implementation overhead, CM^2 is evaluated on the three-modality mortality task only.

Evaluation and Implementation details. We evaluate classification using balanced accuracy (BAcc), macro F1-score, and Cohen's κ, with AUROC and AUPRC additionally reported for binary tasks; regression is evaluated using MSE, MAE, and R^2. For ablations, we additionally measure representation inconsistency (Rep. Incons.; Appendix I.3). All results are aggregated over five patient-level cross-validation folds, stratified by the target label and observed modality-availability pattern. Normalization statistics and hyperparameters are determined using only the training and validation partitions, respectively. Further details are provided in Appendix G

6.2 OVERALL PREDICTIVE PERFORMANCE

Tables 1 and 2 summarize the predictive performance across the tasks. AURA consistently achieves the best performance across all reported metrics. On the three MIMIC-IV tasks, over the strongest competing methods, AURA improves Kappa by 3.12% on stroke type prediction, reduces MSE by 5.8% on length-of-stay prediction, and improves AUROC and AUPRC by 1.04% and 4.25%, respectively, on mortality prediction.

Table 1: Results on stroke type prediction and length-of-stay prediction, reported as mean ± standard deviation over five folds. Best results are shown in **bold** and second-best results are underlined.

	Stroke type prediction			Length-of-stay prediction		
Method	**BAcc (%) ↑**	**F1 (%) ↑**	**Kappa ↑**	**MSE ↓**	**MAE ↓**	R^2 ↑
HEALNet	$71.99_{\pm 0.57}$	$72.56_{\pm 0.70}$	$.5863_{\pm .0109}$	$.3413_{\pm .0157}$	$.4540_{\pm .0110}$	$.5221_{\pm .0235}$
Flex-MoE	$70.90_{\pm 0.57}$	$71.19_{\pm 0.57}$	$.5706_{\pm .0090}$	$.3923_{\pm .0159}$	$.4898_{\pm .0101}$	$.4509_{\pm .0216}$
MultiModN	$69.22_{\pm 1.04}$	$69.98_{\pm 0.83}$	$.5551_{\pm .0121}$	$.4427_{\pm .0156}$	$.5230_{\pm .0091}$	$.3804_{\pm .0204}$
MASAM	$72.16_{\pm 0.38}$	$72.39_{\pm 0.88}$	$.5859_{\pm .0077}$	$.4066_{\pm .0139}$	$.4919_{\pm .0075}$	$.4310_{\pm .0183}$
AMC	$62.69_{\pm 0.90}$	$63.14_{\pm 0.78}$	$.4606_{\pm .0086}$	$.4070_{\pm .0138}$	$.4996_{\pm .0079}$	$.4303_{\pm .0204}$
AURA	$\mathbf{73.49_{\pm 0.33}}$	$\mathbf{73.77_{\pm 0.75}}$	$\mathbf{.6046_{\pm .0057}}$	$\mathbf{.3216_{\pm .0060}}$	$\mathbf{.4400_{\pm .0022}}$	$\mathbf{.5498_{\pm .0105}}$

Table 2: Results on mortality prediction and ADNI disease-stage prediction.

	Mortality prediction			ADNI disease-stage prediction		
Method	**BAcc (%) ↑**	**AUROC (%) ↑**	**AUPRC (%) ↑**	**BAcc (%) ↑**	**F1 (%) ↑**	**Kappa ↑**
HEALNet	$72.61_{\pm 0.84}$	$85.65_{\pm 0.11}$	$66.98_{\pm 1.93}$	$57.29_{\pm 2.49}$	$57.20_{\pm 2.11}$	$.3517_{\pm .0232}$
Flex-MoE	$74.31_{\pm 1.39}$	$84.03_{\pm 0.65}$	$63.44_{\pm 2.56}$	$53.02_{\pm 2.26}$	$53.62_{\pm 2.24}$	$.3011_{\pm .0305}$
MultiModN	$72.21_{\pm 0.52}$	$85.97_{\pm 0.60}$	$67.05_{\pm 1.25}$	$53.05_{\pm 2.90}$	$52.70_{\pm 2.75}$	$.2925_{\pm .0392}$
CM^2	$70.29_{\pm 3.10}$	$84.81_{\pm 1.03}$	$65.27_{\pm 3.09}$	–	–	–
MASAM	$70.61_{\pm 1.30}$	$82.78_{\pm 0.87}$	$63.51_{\pm 1.62}$	$57.68_{\pm 2.77}$	$54.66_{\pm 2.74}$	$.3262_{\pm .0290}$
AMC	$67.62_{\pm 2.08}$	$80.58_{\pm 1.48}$	$58.80_{\pm 3.56}$	$49.75_{\pm 2.46}$	$44.82_{\pm 2.86}$	$.2386_{\pm .0243}$
AURA	$\mathbf{74.69_{\pm 1.67}}$	$\mathbf{86.86_{\pm 0.69}}$	$\mathbf{69.90_{\pm 1.19}}$	$\mathbf{61.65_{\pm 1.70}}$	$\mathbf{60.84_{\pm 1.21}}$	$\mathbf{.3901_{\pm .0204}}$

On ADNI, it improves BAcc from 0.5768 to 0.6165 over the strongest competing method and achieves the best macro F1 and Cohen's κ. Overall, these results indicate that promoting robustness to MNAR does not require sacrificing predictive performance under the standard setting.

6.3 Robustness under MNAR Evaluation

Results on MultiModN-style MNAR Evaluation. Figure 3 reports the performance degradation under flipped MNAR relative to each method's standard-setting performance. AURA exhibits smaller degradation than the other evaluated methods across masking rates. For mortality prediction (Table 3), AURA achieves the highest performance under both standard and shifted missingness across all reported metrics. Its degradation remains small from 7.3% in BAcc to 20.7% in AUPRC, whereas competing methods suffer larger drops, particularly in AUPRC.

Results on Controlled MNAR Shift Evaluation protocol (ours) Under the Controlled MNAR Shift Evaluation protocol (Section 5.2), Figure 3 shows that competing methods degrade sharply when the missingness mechanism changes, despite strong Test-Same performance. This failure is much clearer at higher training MNAR severities ($\tau \geq 0.4$): as reported in Appendix I, several competing methods achieve near-zero or even negative Cohen's κ under Test-Shift. In contrast, AURA retains substantially positive κ and exhibits a markedly smaller Test-Same–Test-Shift gap across all evaluated severity levels.

6.4 Ablation Study

We ablate the main components of AURA on stroke type prediction; full results are provided in Appendix I.4. Removing SCMA mainly reduces standard predictive performance, with Cohen's κ decreasing from 0.6046 to 0.5773, while Rep. Incons. remains close (0.3975 vs. 0.3846). In contrast, removing the alignment loss or modality dropout increases Rep. Incons. to 0.8890 and 0.9268, respectively, and increases the Test-Same–Test-Shift κ drop from 40.6% to 64.9% and 87.6%.

We also evaluate a full-alignment variant that aligns all modality representations rather than only the artificially removed ones. Full alignment reduces Rep. Incons. from 0.3975 to 0.1923 and slightly

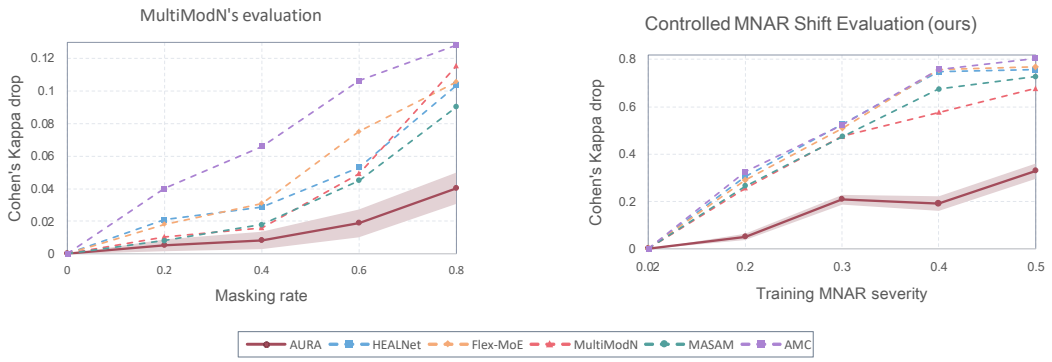


Figure 3: (a) Cohen's κ drop relative to standard performance when evaluated under flipped MNAR. (b) Cohen's κ drop from Test-Same T_{train} to Test-Shift T_{test}.

Table 3: Robustness on 100-day mortality prediction under high MNAR severity. Each cell reports the standard-setting performance on the first line and performance under flipped MNAR on the second line, with the relative change shown in **bold**.

Method	BACC (%) ↑	F1 (%) ↑	AUROC (%) ↑	PR-AUC (%) ↑
HEALNet	72.61 ±0.84 59.07 ±1.49 **(-18.7%)**	74.05 ±0.72 60.06 ±1.85 **(-18.9%)**	85.65 ±0.11 74.44 ±2.65 **(-13.1%)**	66.98 ±1.93 46.37 ±3.18 **(-30.8%)**
Flex-MoE	74.31 ±1.39 55.16 ±2.90 **(-25.8%)**	72.88 ±1.19 55.13 ±3.19 **(-24.4%)**	84.03 ±0.65 66.01 ±6.49 **(-21.5%)**	63.44 ±2.56 36.91 ±2.52 **(-41.8%)**
MultiModN	72.21 ±0.52 58.14 ±3.89 **(-19.5%)**	73.89 ±0.68 58.84 ±4.62 **(-20.4%)**	85.97 ±0.60 67.91 ±5.23 **(-21.0%)**	67.05 ±1.25 38.83 ±6.14 **(-42.1%)**
CM^2	70.29 ±3.10 60.03 ±1.04 **(-14.6%)**	72.31 ±2.54 61.06 ±0.98 **(-15.6%)**	84.81 ±1.03 74.57 ±2.07 **(-12.1%)**	65.27 ±3.09 48.24 ±2.99 **(-26.1%)**
MASAM	70.61 ±1.30 54.06 ±0.45 **(-23.4%)**	72.25 ±2.70 52.10 ±1.00 **(-27.9%)**	82.78 ±0.87 59.27 ±0.51 **(-28.4%)**	63.51 ±1.62 32.42 ±0.52 **(-49.0%)**
AMC	67.62 ±2.08 50.71 ±1.58 **(-25.0%)**	68.96 ±1.37 45.02 ±3.54 **(-34.7%)**	80.58 ±1.48 54.49 ±8.71 **(-32.4%)**	58.80 ±3.56 27.27 ±6.34 **(-53.6%)**
AURA	74.69 ±1.67 69.26 ±1.62 **(-7.3%)**	74.72 ±0.72 68.67 ±1.34 **(-8.1%)**	86.86 ±0.69 79.84 ±1.29 **(-8.1%)**	69.90 ±1.19 55.41 ±3.19 **(-20.7%)**

reduces the Test-Same–Test-Shift drop from 40.6% to 39.1%. However, it also lowers standard κ from 0.6046 to 0.5703, and Test-Shift κ from 0.4236 to 0.3987. This suggests that full alignment makes the representations more stable, but can also remove information useful for prediction.

7 Conclusion

We study multimodal learning under shifts in missing-not-at-random missingness, where availability patterns can become predictive shortcuts. We introduce representation consistency as a principle for robust multimodal learning and show theoretically that representation inconsistency across modality patterns controls sensitivity to missingness-mechanism shifts. Motivated by this result, we propose AURA, a multimodal learning framework that promotes representation consistency across different modality-availability patterns. We further introduce the Controlled MNAR Shift Evaluation protocol to evaluate robustness to missingness changes. Across the MIMIC-IV and ADNI tasks, AURA achieves strong predictive performance. On the MNAR-shift settings, it also shows smaller degradation than the compared methods. Together, these results highlight representation consistency as a promising direction for robust multimodal learning under realistic changes in modality availability.

AI USE STATEMENT

We used generative AI tools to provide feedback on the presentation of the research methodology and experiments, and to assist with drafting, editing, and improving the readability of parts of the manuscript. Generative AI tools were also used to support summarization. All research decisions, experimental designs, mathematical claims, implementations, analyses, and reported results were reviewed and verified by the authors. We take full responsibility for the final content of this work, including text and claims produced with the aid of generative AI.

ETHICS STATEMENT

This study uses existing clinical research datasets, including MIMIC-IV, MIMIC-IV-Note, and ADNI, under their respective data-access requirements. We do not attempt to identify individual participants. Our experiments are intended to study multimodal learning under missing-data distribution shifts rather than to develop a system for direct clinical deployment. Models trained on retrospective clinical data may inherit biases in data collection, modality availability, and patient populations, and their performance may not transfer to other institutions or clinical workflows. Any clinical use would therefore require additional validation, including evaluation of safety, fairness, and distribution shift in the intended deployment setting.

REPRODUCIBILITY STATEMENT

We provide detailed information to support reproducibility throughout the appendix. Appendix B contains the complete proof of the theoretical result; Appendix C describes the SCMA module; Appendix E specifies the Controlled MNAR Shift construction and numerical optimization; Appendix F provides details of modality dropout and representation alignment; and Appendix G describes dataset construction, preprocessing, and evaluation splits. Complete MNAR robustness results and component ablations are reported in Appendix I.

REFERENCES

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

Jiajun Chen, Sai Cheng, Yuan Yutao, YiRui Zhang, Haitao Yuan, Peng Peng, and Yi Zhong. Promise: Prompt-attentive hierarchical contrastive learning for robust cross-modal representation with missing modalities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(24):20076–20082, March 2026a. ISSN 2159-5399. doi: 10.1609/aaai.v40i24.39093. URL http://dx.doi.org/10.1609/aaai.v40i24.39093.

Mengxi Chen, Fei Zhang, Zihua Zhao, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Probabilistic conformal distillation for enhancing missing modality robustness. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 36218–36242. Curran Associates, Inc., 2024. doi: 10.52202/079017-1142. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/3fb97beb678a1c46bf25f40b3a8c6f98-Paper-Conference.pdf.

Zijie Chen, Kejing Yin, Wenfang Yao, William K. Cheung, and Jing Qin. MASAM: Multimodal adaptive sharpness-aware minimization for heterogeneous data fusion. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=AUKeDukcUi.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2681–2690. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/feydy19a.html.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020. URL https://arxiv.org/abs/2006.07733.

Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. HEALNet: Multimodal fusion for heterogeneous biomedical data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=HUxtJcQpDS.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jaegle21a.html.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV-Note: Deidentified free-text clinical notes. *PhysioNet*, January 2023. doi: 10.13026/1n74-ne17. URL https://doi.org/10.13026/1n74-ne17. Version 2.2.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. *PhysioNet*, October 2024. doi: 10.13026/kpb9-mt58. URL https://doi.org/10.13026/kpb9-mt58. Version 3.1.

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, October 2025.

Zihan Liang, Ziwen Pan, and Ruoxuan Xiong. Causal representation learning from multimodal clinical records under non-random modality missingness, 2025. URL https://arxiv.org/abs/2509.17228.

Ronghao Lin and Haifeng Hu. MissModal: Increasing robustness to missing modality in multimodal sentiment analysis. *Transactions of the Association for Computational Linguistics*, 11: 1686–1702, 2023. doi: 10.1162/tacl_a_00628. URL https://aclanthology.org/2023.tacl-1.94/.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1614–1623, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/martins16.html.

Susanne G. Mueller, Michael W. Weiner, Leon J. Thal, Ronald C. Petersen, Clifford R. Jack, William Jagust, John Q. Trojanowski, Arthur W. Toga, and Laurel Beckett. Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's & Dementia*, 1(1):55–66, 2005. doi: 10.1016/j.jalz.2005.06.003.

Jiquan Ngiam, Mingjie Khosla, Aditya Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pp. 689–696, 2011.

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 *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=MzQed6QEmS.

Md Kaykobad Reza, Ashley Prater-Bennette, and M. Salman Asif. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):742–754, 2025. doi: 10.1109/TPAMI.2024.3476487.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 12 1976. ISSN 0006-3444. doi: 10.1093/biomet/63.3.581. URL https://doi.org/10.1093/biomet/63.3.581.

Vinitra Swamy, Malika Satayeva, Jibril Frej, Thierry Bossy, Thijs Vogels, Martin Jaggi, Tanja Käser, and Mary-Anne Hartley. MultimoDN—multimodal, multi-task, interpretable modular networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=iB3Ew6z4WL.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert-Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272, February 2020. ISSN 1548-7105. doi: 10.1038/s41592-019-0686-2. URL http://dx.doi.org/10.1038/s41592-019-0686-2.

Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Towards good practices for missing modality robust action recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3):2776–2784, 2023. ISSN 2159-5399. doi: 10.1609/aaai.v37i3.25378. URL http://dx.doi.org/10.1609/aaai.v37i3.25378.

Feng Wu, Tsai Hor Chan, Fuying Wang, Guosheng Yin, and Lequan Yu. Cross-modal alignment via variational copula modelling. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=6pNp7to8kW.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=tc7RFcx4hT. Survey Certification.

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 *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=ihEHCbqZEx.

Yifan Zhan, Rui Yang, Junxian You, Mengjie Huang, Weibo Liu, and Xiaohui Liu. A systematic literature review on incomplete multimodal learning: techniques and challenges. *Systems Science & Control Engineering*, 13(1):2467083, 2025. doi: 10.1080/21642583.2025.2467083. URL https://doi.org/10.1080/21642583.2025.2467083.

A Extended Related Work

Multimodal learning with missing modalities. A broad line of work studies how multimodal models can operate when only a subset of modalities is available, recent surveys providing comprehensive overviews (Wu et al., 2026; Zhan et al., 2025). Hemker et al. (2024) and Swamy et al. (2023) adapt model computation to the available inputs: HEALNet skips unavailable modality updates during iterative fusion, while MultiModN uses a modular architecture that can bypass missing modalities. Flex-MoE (Yun et al., 2024) models arbitrary modality combinations using a missing-modality bank together with combination-dependent experts, allowing the model to specialize its behavior to different availability patterns. CM^2 (Wu et al., 2025) takes a probabilistic approach, using variational copula modelling to capture cross-modal dependence and construct representations for unavailable modalities. This provides a principled formulation of cross-modal completion, although extending the copula-based modelling to larger numbers of modalities introduces additional computational and implementation complexity. Other approaches improve robustness through additional learning objectives (Woo et al., 2023; Li et al., 2025). These works primarily address how to learn effectively from incomplete inputs; our focus is on robustness when the mechanism determining modality availability itself changes.

Representation consistency under missing modalities. Several works study how learned representations behave across different modality subsets (Lin & Hu, 2023; Chen et al., 2024). Reza et al. (2025) further provide empirical evidence that models that are more robust to missing modalities tend to exhibit smaller representation changes across missingness patterns. Our work studies this connection specifically under missingness-mechanism shift and theoretically relates changes in prediction loss to representation inconsistency across availability patterns.

Informative and non-random modality missingness. A smaller body of work considers settings in which modality availability is itself informative. Liang et al. (2025) treats non-random modality missingness as useful structure that can contribute to prediction, highlighting that missingness information is not inherently undesirable. In contrast, Swamy et al. (2023) show that models can exploit label–missingness relationships and may degrade when these relationships change at inference time. These observations are complementary: availability may contain useful information about the data-acquisition process, but it can also become an unstable predictive shortcut. Accordingly, AURA does not aim to remove all information carried by missingness, but to reduce excessive dependence on availability patterns when their relationship with the target changes across environments.

Evaluation under missingness-mechanism shift. Existing MNAR stress tests commonly induce label-dependent missingness by masking a selected modality for samples from particular classes. In the evaluation of Swamy et al. (2023), increasing the masking probability also increases the amount of modality information removed, while the masking probability itself does not directly measure the label–missingness dependence. Its evaluation, however, entangles the performance drop caused by changes in the label–missingness relationship with that caused by removing more modality information.

B Proof of the Representation-Consistency Bound

We define the sample-level representation inconsistency as the largest representation discrepancy from the reference pattern:

$$\delta(S, X, Y) = \max_{r \in \mathcal{R}} \|Z_r - Z_{r_0}\| . \tag{17}$$

The reference pattern r_0 can be sample-specific. In our setting, we use each sample's natural availability pattern as r_0.

Accordingly, the expected representation inconsistency is

$$\mathbb{E}_{S,X,Y}\left[\delta(S,X,Y)\right] = \mathbb{E}_{S,X,Y}\left[\max_{r\in\mathcal{R}} \|Z_r - Z_{r_0}\|\right]. \tag{18}$$

We next define the task loss under pattern r as

$$a_r(S,X,Y) = \ell(h(Z_r),Y) = \ell\Big(h(F_\theta(X^{(r)},r)),Y\Big). \tag{19}$$

For a missingness mechanism $q(r \mid S,X,Y)$, the expected task loss is

$$\mathcal{L}_q = \mathbb{E}_{S,X,Y}\left[\sum_{r\in\mathcal{R}} q(r \mid S,X,Y)\, a_r(S,X,Y)\right]. \tag{20}$$

We assume that the composition of the prediction head and task loss, $\ell \circ h$, is L-Lipschitz with respect to the final representation:

$$|\ell(h(z),y) - \ell(h(z'),y)| \le L\,\|z - z'\|. \tag{21}$$

This assumption states that a small change in the learned representation cannot cause an arbitrarily large change in the task loss.

For each (S,X,Y) and pattern $r \in \mathcal{R}$, define

$$\Delta q_r(S,X,Y) = q_{\mathrm{te}}(r \mid S,X,Y) - q_{\mathrm{tr}}(r \mid S,X,Y). \tag{22}$$

Because $q_{\mathrm{te}}(\cdot \mid S,X,Y)$ and $q_{\mathrm{tr}}(\cdot \mid S,X,Y)$ are probability distributions over $\mathcal{R}$,

$$\sum_{r\in\mathcal{R}} \Delta q_r(S,X,Y) = 0 \qquad \text{for every } (S,X,Y). \tag{23}$$

Using the definition of the expected task loss,

$$\mathcal{L}_{q_{\mathrm{te}}} - \mathcal{L}_{q_{\mathrm{tr}}} = \mathbb{E}_{S,X,Y}\left[\sum_{r\in\mathcal{R}} \Delta q_r(S,X,Y)\, a_r(S,X,Y)\right] \tag{24}$$

$$= \mathbb{E}_{S,X,Y}\left[\sum_{r\in\mathcal{R}} \Delta q_r(S,X,Y)\,(a_r(S,X,Y) - a_{r_0}(S,X,Y))\right], \tag{25}$$

where the second equality follows from equation 23.

Applying the triangle inequality gives

$$|\mathcal{L}_{q_{\mathrm{te}}} - \mathcal{L}_{q_{\mathrm{tr}}}| \le \mathbb{E}_{S,X,Y}\left[\sum_{r\in\mathcal{R}} |\Delta q_r(S,X,Y)|\,|a_r(S,X,Y) - a_{r_0}(S,X,Y)|\right]. \tag{26}$$

By the L-Lipschitz assumption in equation 21,

$$|a_r(S,X,Y) - a_{r_0}(S,X,Y)| \le L\,\|Z_r - Z_{r_0}\|. \tag{27}$$

Therefore,

$$|\mathcal{L}_{q_{\mathrm{te}}} - \mathcal{L}_{q_{\mathrm{tr}}}| \le L\,\mathbb{E}_{S,X,Y}\left[\sum_{r\in\mathcal{R}} |\Delta q_r(S,X,Y)|\,\|Z_r - Z_{r_0}\|\right] \tag{28}$$

$$\le L\,\mathbb{E}_{S,X,Y}\left[A(S,X,Y)\,\delta(S,X,Y)\right], \tag{29}$$

where

$$A(S,X,Y) = \sum_{r\in\mathcal{R}} |\Delta q_r(S,X,Y)| = 2\mathrm{TV}(q_{te}(\cdot|S,X,Y), q_{tr}(\cdot|S,X,Y)) \tag{30}$$

and

$$\delta(S,X,Y) = \max_{r\in\mathcal{R}} \|Z_r - Z_{r_0}\|. \tag{31}$$

This completes the proof of equation 7.

Moreover, applying Hölder's inequality with respect to the shared distribution $P(S, X, Y)$,

$$\mathbb{E}_{S,X,Y}\left[A(S,X,Y)\,\delta(S,X,Y)\right] \le \|A\|_{L^\infty(P)}\,\|\delta\|_{L^1(P)}\,. \tag{32}$$

For every (S, X, Y),

$$A(S,X,Y) = \sum_{r\in\mathcal{R}} |q_{\mathrm{te}}(r \mid S,X,Y) - q_{\mathrm{tr}}(r \mid S,X,Y)| \tag{33}$$

$$\le \sum_{r\in\mathcal{R}} q_{\mathrm{te}}(r \mid S,X,Y) + \sum_{r\in\mathcal{R}} q_{\mathrm{tr}}(r \mid S,X,Y) \tag{34}$$

$$= 2. \tag{35}$$

Hence,

$$\|A\|_{L^\infty(P)} \le 2. \tag{36}$$

Since $\delta(S, X, Y) \ge 0$,

$$\|\delta\|_{L^1(P)} = \mathbb{E}_{S,X,Y}\left[\delta(S,X,Y)\right]. \tag{37}$$

To make the shift magnitude explicit, define

$$\epsilon_{\mathrm{shift}} = \operatorname*{ess\,sup}_{S,X,Y} \mathrm{TV}(q_{\mathrm{te}}(\cdot \mid S,X,Y), q_{\mathrm{tr}}(\cdot \mid S,X,Y))\,.$$

By equation 30, $\|A\|_{L^\infty(P)} = 2\epsilon_{\mathrm{shift}}$. Therefore, equation 32 gives

$$|\mathcal{L}_{q_{\mathrm{te}}} - \mathcal{L}_{q_{\mathrm{tr}}}| \le 2L\epsilon_{\mathrm{shift}}\mathbb{E}_{S,X,Y}\left[\delta(S,X,Y)\right].$$

Since $\epsilon_{\mathrm{shift}} \le 1$, we further obtain the more conservative, mechanism-independent bound

$$|\mathcal{L}_{q_{\mathrm{te}}} - \mathcal{L}_{q_{\mathrm{tr}}}| \le 2L\,\mathbb{E}_{S,X,Y}\left[\delta(S,X,Y)\right]. \tag{38}$$

These theoretical results motivate reducing variation in the final representations across availability patterns. The assumption in Eq. 21 may not hold globally for squared-error regression with unbounded predictions or targets. It can hold when predictions and targets are bounded, or with losses such as MAE or Huber and a Lipschitz prediction head. We therefore use the bound mainly as a theoretical guide for representation-consistency learning.

AURA follows this principle by using a practical modality-level alignment objective on artificially perturbed views as a surrogate. This objective does not directly optimize the worst-case Euclidean discrepancy over the final representation used in the theoretical bound.

C Selective Cross-Modal Attention

This section provides additional details of the Selective Cross-Modal Attention (SCMA) module introduced in Section 4.3. SCMA performs target-specific cross-modal refinement by selecting a sparse subset of source modalities for each target modality and using their representations as cross-modal context.

Target-specific source selection. Let

$$\widetilde{H}_m \in \mathbb{R}^{s_m \times d}$$

denote the s_m feature slots of modality m produced by the modality-specific evidence construction stage. We omit the batch index for clarity. To determine which source modalities should contribute to a target modality m, we first summarize each modality by mean-pooling its feature slots:

$$c_m = \frac{1}{s_m}\sum_{k=1}^{s_m} \widetilde{H}_{m,k}, \qquad c_m \in \mathbb{R}^d.$$

For each target modality m, a target-specific router computes the compatibility between the target representation c_m and each candidate source c_j:

$$e_{m,j} = \frac{\langle W_m^q c_m,\, W_m^k c_j \rangle}{\sqrt{d}}, \qquad j \neq m,$$

where W_m^q and W_m^k are learnable projections associated with target modality m. We exclude the target itself from its candidate source set.

The compatibility scores are transformed using Sparsemax (Martins & Astudillo, 2016):

$$a_m = \mathrm{Sparsemax}\left(\{e_{m,j}\}_{j \neq m}\right),$$

where $a_{m,j} \geq 0$ denotes the resulting routing score of source modality j for target modality m. Since Sparsemax can assign exact zeros, we define the active source set as

$$\mathcal{G}_m = \{j \in \mathcal{M} \setminus \{m\} : a_{m,j} > 0\}.$$

Only modalities in $\mathcal{G}_m$ participate in the subsequent cross-modal refinement. Thus, the routing mechanism performs target-specific source selection without requiring a fixed subset of modalities to interact for every sample.

Cross-modal refinement. Given the active source set $\mathcal{G}_m$, we concatenate their feature slots to form the source sequence

$$H_m^{\mathrm{src}} = \underset{j \in \mathcal{G}_m}{\mathrm{Concat}}\, \widetilde{H}_j.$$

The routing score $a_{m,j}$ is shared by all feature slots originating from source modality j. SCMA uses the target slots as queries and the selected source slots as keys and values. In addition to determining which modalities are active, the routing scores modulate their value representations.

More precisely, let

$$\bar{Q}_m = \mathrm{Linear}_q(\widetilde{H}_m),$$

and define the key and value sequences as

$$K_m = \underset{j \in \mathcal{G}_m}{\mathrm{Concat}}\, \mathrm{Linear}_k(\widetilde{H}_j),$$

$$V_m = \underset{j \in \mathcal{G}_m}{\mathrm{Concat}}\, a_{m,j}\ \mathrm{Linear}_v(\widetilde{H}_j).$$

The target representation is then refined using a standard multi-head cross-attention Transformer block (Vaswani et al., 2017):

$$H_m^{\mathrm{ref}} = \mathrm{CrossAttnBlock}\left(\bar{Q}_m, K_m, V_m\right).$$

In SCMA, the sparse routing distribution $a_{m,\cdot}$ is used not only to define the selected neighborhood $\mathcal{G}_m$, but also to scale the value tokens passed to the cross-attention block. This design is important for optimization. By multiplying each selected source value by its routing weight $a_{m,j}$, the output of the SCMA block becomes dependent on the sparse routing distribution. This creates a differentiable path from the task loss back to the router. Therefore, the router can be trained end-to-end to assign larger weights to useful source modalities and suppress irrelevant ones.

Stacked refinement. Multiple SCMA layers can be stacked to progressively refine the modality representations. Let

$$\mathcal{H}^{(\ell-1)} = \left\{H_1^{(\ell-1)}, \ldots, H_M^{(\ell-1)}\right\}$$

denote the input to refinement layer ℓ. Each target modality is updated from the same input representation set:

$$H_m^{(\ell)} = \mathrm{SCMA}_m\left(H_m^{(\ell-1)}, \mathcal{H}^{(\ell-1)}\right), \qquad m = 1, \ldots, M.$$

The resulting set

$$\mathcal{H}^{(\ell)} = \left\{H_1^{(\ell)}, \ldots, H_M^{(\ell)}\right\}$$

is then passed to the next refinement layer.

Table 4: Normalized label–missingness mutual information induced by one-class masking with $p_{\text{mask}} = 0.8$ in the balanced setting

Number of classes K	**2**	**3**	**4**	**5**	**10**
$\tau_K(0.8)$	0.610	0.376	0.271	0.211	0.099

D Masking Probability Does Not Directly Measure MNAR Severity

We provide a simple analysis illustrating why the masking probability p_{mask} does not directly determine the strength of the induced label–missingness association, particularly in multiclass problems.

Effect of the number of classes. Consider a balanced K-class classification problem with $Y \in \{1, \dots, K\}$. Following a one-class masking construction, suppose a selected modality is artificially removed with probability p only for samples from class $Y = 1$, while it is always retained for the remaining classes. Let $A \in \{0, 1\}$ denote the artificial-missingness indicator, where $A = 1$ means that the modality is removed. Then

$$P(A = 1 \mid Y = 1) = p, \qquad P(A = 1 \mid Y = k) = 0, \quad k \neq 1. \tag{39}$$

Since the classes are balanced,

$$P(A = 1) = \frac{p}{K}. \tag{40}$$

Let

$$h(u) = -u \log u - (1 - u) \log(1 - u) \tag{41}$$

denote the binary entropy function. The entropy of the missingness indicator is therefore

$$H(A) = h\left(\frac{p}{K}\right), \tag{42}$$

while its conditional entropy given the label is

$$H(A \mid Y) = \frac{1}{K} h(p), \tag{43}$$

because artificial masking is stochastic only for the targeted class. Hence, the mutual information between the induced missingness and the label is

$$I(A; Y) = h\left(\frac{p}{K}\right) - \frac{1}{K} h(p). \tag{44}$$

Normalizing by the label entropy $H(Y) = \log K$ gives

$$\tau_K(p) = \frac{I(A; Y)}{H(Y)} = \frac{h\left(p/K\right) - \frac{1}{K} h(p)}{\log K}. \tag{45}$$

Equation 45 shows that the same masking probability does not correspond to a fixed MNAR severity across classification problems.

Intuitively, one-class masking directly modifies the missingness distribution of only one class. As the number of classes increases, this class constitutes an increasingly smaller fraction of the label space, and the resulting label–missingness dependence vanishes. For example, with $p = 0.8$, the normalized mutual information decreases substantially as K increases:

Empirical severity in the stroke prediction setting. The preceding derivation intentionally considers an idealized balanced setting without natural missingness in order to isolate the effect of the number of classes. In real multimodal data, the resulting MNAR severity also depends on class prevalence, natural modality availability, and the joint missingness patterns already present in the dataset. We therefore additionally measure the realized normalized mutual information in our three-class stroke type prediction setting under the same one-class masking construction.

Table 5: Normalized label–missingness mutual information under increasing one-class masking probabilities in the three-class stroke type prediction setting

p_{mask}	**0.0**	**0.2**	**0.4**	**0.6**	**0.8**
$I(R;Y)/H(Y)$	0.026	0.042	0.074	0.125	0.202

Although the masking probability increases to 0.8, the resulting normalized label–missingness mutual information reaches only 0.202 in this setting. Together with the analytical result above, this illustrates that p_{mask} characterizes the corruption procedure rather than the strength of the resulting MNAR dependence, motivating a severity measure defined directly from the joint distribution of the missingness pattern and the label.

E Additional Details of the Controlled MNAR Shift Evaluation protocol

E.1 Pattern-Level Corruption and Feasibility

Let $\mathcal{M}_{\text{MNAR}}$ denote the set of modalities involved in the controlled corruption process, and let $R \in \{0,1\}^{|\mathcal{M}_{\text{MNAR}}|}$ denote their natural availability pattern. For each class y, we define a stochastic pattern-level transform

$$T_y(\hat{r} \mid r) = P(\hat{R} = \hat{r} \mid R = r, Y = y), \tag{46}$$

where $\hat{R}$ is the availability pattern after artificial corruption.

Artificial corruption is restricted to removing naturally observed modalities. Thus, the set of feasible descendants of a pattern r is

$$\mathcal{F}(r) = \{\hat{r} : \hat{r} \preceq r,\ \hat{r} \neq \mathbf{0}\}, \tag{47}$$

where $\hat{r} \preceq r$ denotes element-wise inequality. Accordingly,

$$T_y(\hat{r} \mid r) = 0, \qquad \hat{r} \notin \mathcal{F}(r), \tag{48}$$

and each active transition row satisfies

$$T_y(\hat{r} \mid r) \geq 0, \qquad \sum_{\hat{r} \in \mathcal{F}(r)} T_y(\hat{r} \mid r) = 1. \tag{49}$$

Natural patterns with zero empirical probability for a given class do not introduce optimization variables. Samples for which none of the modalities in $\mathcal{M}_{\text{MNAR}}$ are naturally available are excluded from the controlled corruption process, since no valid drop-only descendant exists under equation 47.

E.2 Induced Distribution and MNAR Severity

Let

$$p(r, y) = P(R = r, Y = y) \tag{50}$$

denote the empirical joint distribution before artificial corruption. For a transform T, the post-corruption joint distribution is

$$q_T(\hat{r}, y) = \sum_r p(r, y) T_y(\hat{r} \mid r). \tag{51}$$

Its corresponding marginals are

$$q_T(\hat{r}) = \sum_y q_T(\hat{r}, y), \qquad p(y) = \sum_{\hat{r}} q_T(\hat{r}, y). \tag{52}$$

Since the transform only changes modality availability, the label marginal $p(y)$ remains unchanged.

We measure MNAR severity by the normalized mutual information between the post-corruption availability pattern and the label:

$$S_{\mathrm{MNAR}}(T) = \frac{I_T(\hat{R}; Y)}{H(Y)}, \tag{53}$$

where

$$I_T(\hat{R}; Y) = \sum_{\hat{r}, y} q_T(\hat{r}, y) \log \frac{q_T(\hat{r}, y)}{q_T(\hat{r}) p(y)}, \tag{54}$$

and

$$H(Y) = -\sum_{y} p(y) \log p(y). \tag{55}$$

E.3 Availability-Aware Artificial-Removal Cost

For modality $m \in \mathcal{M}_{\mathrm{MNAR}}$, we define its artificial-drop rate as

$$D_m(T) = P_T(\hat{R}_m = 0 \mid R_m = 1). \tag{56}$$

Let

$$p_m = P(R_m = 1) \tag{57}$$

denote the natural availability of modality m. We assign the availability-aware weight

$$w_m = \frac{1}{p_m}, \tag{58}$$

and define the overall removal cost as

$$D(T) = \frac{\sum_{m \in \mathcal{M}_{\mathrm{MNAR}}} w_m D_m(T)}{\sum_{m \in \mathcal{M}_{\mathrm{MNAR}}} w_m}. \tag{59}$$

Hence, artificially removing a naturally scarce modality incurs a larger penalty than removing a frequently observed one. This discourages the optimizer from achieving a desired missingness–label dependence by disproportionately corrupting modalities that are already rarely available.

E.4 Transformation Construction

Training transformation. For each training MNAR severity τ, we first construct the training transformation T_{train} by solving

$$\begin{aligned} T_{train} = \arg\min_{T} \quad & D(T) \\ \text{s.t.} \quad & S_{\mathrm{MNAR}}(T) = \tau, \\ & T_y(\hat{r} \mid r) \geq 0, \\ & \sum_{\hat{r} \in \mathcal{F}(r)} T_y(\hat{r} \mid r) = 1, \\ & T_y(\hat{r} \mid r) = 0 \quad \text{for } \hat{r} \notin \mathcal{F}(r). \end{aligned} \tag{60}$$

Thus, T_{train} achieves the prescribed training MNAR severity using the smallest availability-aware artificial-removal cost.

Shifted transformation. Given T_{train}, we seek a second transform with a large missingness-mechanism discrepancy while preserving the artificial-removal rate of every controlled modality. We measure the discrepancy between two environments using the Jensen–Shannon divergence between their post-corruption joint distributions:

$$\mathrm{Dist}(T_{train}, T) = \mathrm{JSD}\left(q_{T_{train}}(\hat{R}, Y) \,\|\, q_T(\hat{R}, Y)\right). \tag{61}$$

Because both transforms preserve the same label marginal $p(y)$, Equation 61 is equivalently

$$\mathrm{Dist}(T_{train}, T) = \sum_{y} p(y)\, \mathrm{JSD}\left(q_{T_{train}}(\hat{R} \mid y) \,\|\, q_T(\hat{R} \mid y)\right). \tag{62}$$

Table 6: Hyperparameters used to construct the controlled MNAR environments.

Hyperparameter	**Value**
Training MNAR severities τ	$\{0.2, 0.3, 0.4, 0.5\}$
High-severity initialization runs	24
T_{train} optimization starts	10
T_{test} LP candidates	120
T_{test} SLSQP refinement starts	12
Maximum solver iterations	2000
SLSQP function tolerance	10^{-11}
Accepted T_{train} severity error	2×10^{-4}
Accepted simplex error	2×10^{-6}
Accepted per-modality D_m mismatch	2×10^{-7}

The shifted environment is then obtained from

$$\begin{aligned} T_{test} = \arg\max_{T} \quad & \text{Dist}(T_{train}, T) \\ \text{s.t.} \quad & D_m(T) = D_m(T_{train}), \qquad \forall m \in \mathcal{M}_{\text{MNAR}}, \\ & T_y(\hat{r} \mid r) \geq 0, \\ & \sum_{\hat{r} \in \mathcal{F}(r)} T_y(\hat{r} \mid r) = 1, \\ & T_y(\hat{r} \mid r) = 0 \quad \text{for } \hat{r} \notin \mathcal{F}(r). \end{aligned} \tag{63}$$

Matching each D_m individually ensures that T_{train} and T_{test} remove the same fraction of naturally available samples from every controlled modality while inducing different label–missingness distributions.

E.5 Numerical Optimization

We solve the constrained problems using `SciPy` (Virtanen et al., 2020). For T_{train}, we use Sequential Least Squares Programming (SLSQP) with multiple initializations and retain the feasible solution with the smallest removal cost. For T_{test}, we first generate feasible extreme points under the per-modality removal constraints using `scipy.optimize.linprog` with the HiGHS backend. The candidates with the largest Jensen–Shannon divergence are then refined using SLSQP, and the best feasible solution is retained.

Because the cross-validation folds are stratified jointly by missingness pattern and label, the training and test sets have closely matched empirical label–missingness distributions. We use this empirical distribution to construct both the training and shifted transformations.

The hyperparameters used in our experiments are summarized in Table 6.

Table 7: Realized statistics of the Controlled MNAR Shift environments. $S_{\text{MNAR}} = I(\hat{R}; Y)/H(Y)$ measures the dependence between the corrupted availability pattern and the label. JSD measures the discrepancy between the post-corruption label–missingness distributions induced by T_{train} and T_{test}. $\bar{D}$ denotes the mean realized artificial-removal rate across the controlled modalities.

Target τ	$S_{\text{MNAR}}(T_{\text{train}})$	$S_{\text{MNAR}}(T_{\text{test}})$	JSD	$\bar{D}$
0.2	0.2001	0.2440	0.1543	0.0552
0.3	0.3000	0.3187	0.2704	0.1004
0.4	0.4001	0.3593	0.3484	0.1165
0.5	0.5001	0.3905	0.3854	0.1558

Table 7 reports the statistics after materializing the transformations on the controlled-MNAR cohort. The realized severity of T_{train} closely matches its prescribed target τ. In contrast, τ does not constrain T_{test}: the shifted transformation is optimized to maximize the discrepancy from T_{train} while

matching the per-modality artificial-removal rates. Therefore, $S_{\mathrm{MNAR}}(T_{\mathrm{test}})$ can be either higher or lower than $S_{\mathrm{MNAR}}(T_{\mathrm{train}})$.

The JSD column reports the realized discrepancy between the two post-corruption joint distributions. The last column summarizes the amount of artificial corruption. The per-modality removal rates are matched between T_{train} and T_{test}; after materialization, the maximum observed difference in any D_m is below 7×10^{-5}.

F Additional Details of Representation-Consistency Alignment

F.1 Artificial Modality Dropout

For each training sample i, let $\mathcal{O}_i$ denote the set of naturally observed modalities. When $|\mathcal{O}_i| > 1$, the number of artificially removed modalities is sampled uniformly as

$$k_i \sim \mathrm{Uniform}\{1, \ldots, |\mathcal{O}_i| - 1\}. \tag{64}$$

Conditioned on k_i, we uniformly sample one of the $\binom{|\mathcal{O}_i|}{k_i}$ subsets of size k_i:

$$\mathcal{D}_i \sim \mathrm{Uniform}\left\{D \subset \mathcal{O}_i : |D| = k_i\right\}. \tag{65}$$

Consequently, artificial dropout is applied only to modalities that are actually observed for that sample. Naturally missing modalities are never included in $\mathcal{D}_i$, and the augmented view always retains at least one observed modality. If $|\mathcal{O}_i| \leq 1$, no additional modality is removed and $\mathcal{D}_i = \varnothing$.

F.2 Wasserstein Alignment

For modality m, let $H^{\mathrm{aug}}_{i,m}, \bar{H}^{\mathrm{obs}}_{i,m} \in \mathbb{R}^{s_m \times d}$ denote the student augmented-view representation and the EMA observed-view representation, respectively.

For $s_m > 1$, each slot is first normalized independently along its feature dimension:

$$\widehat{H}_{i,m,t} = \frac{H_{i,m,t}}{\|H_{i,m,t}\|_2}. \tag{66}$$

We then treat the s_m slots as an empirical distribution with uniform mass and compute a Sinkhorn approximation of the 1-Wasserstein (Cuturi, 2013) distance between the two slot sets:

$$d_{i,m} = W_1^{\mathrm{Sink}}\left(\widehat{H}^{\mathrm{aug}}_{i,m}, \mathrm{sg}\left(\widehat{\bar{H}}^{\mathrm{obs}}_{i,m}\right)\right). \tag{67}$$

Our implementation uses `GeomLoss` (Feydy et al., 2019) with `loss="sinkhorn"`, $p = 1$, `blur=0.05`, and the `tensorized` backend. No explicit token-position correspondence is imposed: Wasserstein alignment compares the two modality representations as unordered slot sets.

For modalities represented by a single slot ($s_m = 1$), we use cosine distance,

$$d_{i,m} = 1 - \frac{\left(H^{\mathrm{aug}}_{i,m,1}\right)^{\top} \bar{H}^{\mathrm{obs}}_{i,m,1}}{\|H^{\mathrm{aug}}_{i,m,1}\|_2 \|\bar{H}^{\mathrm{obs}}_{i,m,1}\|_2}. \tag{68}$$

For each modality, the discrepancy is averaged over all samples in the minibatch for which that modality was artificially removed. The final alignment loss is the mean of these modality-level losses. Modalities that are not artificially removed for any sample in the minibatch do not contribute to the loss. If no modality is removed in the minibatch, $\mathcal{L}_{\mathrm{align}} = 0$.

We align only the artificially removed modalities. The goal is to encourage the model to recover information about the missing modality from the modalities that remain observed, rather than forcing the representations of already observed modalities to stay unchanged across views. This is a practical training choice, rather than a direct consequence of the theoretical bound.

Table 8: Effect of different alignment discrepancies on stroke type prediction under the standard setting. Best results are shown in bold.

Alignment loss	BAcc (%) ↑	F1 (%) ↑	Kappa ↑
Wasserstein	$\mathbf{73.49} \pm 0.33$	$\mathbf{73.77} \pm 0.75$	$\mathbf{.6046} \pm .0057$
Cosine	73.38 ± 0.56	73.70 ± 0.46	$.6045 \pm .0078$
MSE	73.07 ± 1.97	73.61 ± 0.21	$.6026 \pm .0035$

Table 9: Effect of different alignment discrepancies on length-of-stay prediction under the standard setting. Best results are shown in bold.

Alignment discrepancy	MSE ↓	MAE ↓	R^2 ↑
Wasserstein	$\mathbf{.3216} \pm .0060$	$\mathbf{.4400} \pm .0022$	$\mathbf{.5498} \pm .0105$
Cosine	$.3328 \pm .0124$	$.4473 \pm .0071$	$.5342 \pm .0172$
MSE	$.3337 \pm .0119$	$.4478 \pm .0070$	$.5339 \pm .0172$

F.3 EMA Target Branch

The observed-view target is produced by an exponential-moving-average copy of the representation branch. After each student optimization step, the EMA parameters are updated according to

$$\bar{\theta} \leftarrow 0.99\,\bar{\theta} + 0.01\,\theta. \tag{69}$$

The EMA branch is used only to construct the stop-gradient alignment target and is removed at inference time.

F.4 Effect of the Alignment Discrepancy

We further examine the choice of discrepancy function used for representation-consistency alignment. We compare the Sinkhorn-approximated Wasserstein distance used by AURA against token-wise cosine distance and mean squared error (MSE). We first evaluate the three choices under the standard setting, and then examine their behavior under the Controlled MNAR Shift Evaluation protocol.

Under the standard setting, the three discrepancy functions yield relatively similar performance on stroke type prediction, although Wasserstein alignment achieves the highest result on all three metrics (Table 8). The difference becomes more pronounced on the length-of-stay task, where Wasserstein alignment reduces MSE from 0.3328 with cosine distance to 0.3216 and improves R^2 from 0.5342 to 0.5498 (Table 9).

Evaluation under the Controlled MNAR Shift Evaluation protocol further distinguishes the alignment objectives. Wasserstein alignment achieves the highest absolute Test-Shift performance across BACC, F1, and Cohen's Kappa. In particular, its degradation is smaller than cosine alignment for all three metrics. MSE alignment is generally weaker in absolute performance, despite exhibiting a smaller relative F1 drop in this experiment.

Based on its strong standard performance and higher absolute performance under missingness-mechanism shift, we use the Sinkhorn-approximated Wasserstein distance as the default alignment discrepancy in the main experiments. This is an empirical design choice rather than being determined by the theoretical bound

G Dataset Construction and Preprocessing

G.1 Data Sources

We constructed the study cohort from MIMIC-IV v3.1 and MIMIC-IV-Note v2.2. Structured modalities, including ICD-9/10 codes, laboratory measurements, and vital signs, were extracted from

Table 10: Effect of different alignment discrepancies under the Controlled MNAR Shift Evaluation protocol with $\tau = 0.5$ on the STP task. Each cell reports Test-Same performance on the first row and Test-Shift performance on the second row. Parentheses denote the relative change from Test-Same to Test-Shift.

Alignment discrepancy	BAcc (%) ↑	F1 (%) ↑	Kappa ↑
Wasserstein	80.83 ±0.69 59.29 ±1.89 **(-26.6%)**	81.80 ±0.73 59.25 ±1.86 **(-27.6%)**	.7130 ±.0111 .4236 ±.0125 **(-40.6%)**
Cosine	82.15 ±0.38 56.98 ±1.46 **(-30.6%)**	83.11 ±0.47 57.01 ±1.50 **(-31.4%)**	.7328 ±.0066 .3948 ±.0249 **(-46.1%)**
MSE	76.28 ±2.14 54.03 ±3.25 **(-29.2%)**	77.54 ±1.84 58.69 ±7.16 **(-24.3%)**	.6527 ±.0253 .3540 ±.0405 **(-45.8%)**

MIMIC-IV, whereas textual modalities, including clinical notes and head CT/MRI radiology reports, were extracted from MIMIC-IV-Note.

To simulate early clinical decision-making for early stroke classification task, we restricted input information to an early admission observation window. Unless otherwise stated, clinical, laboratory, vital-sign, and radiology features were extracted from information available within the first 24 hours after hospital admission. This restriction was designed to reduce the risk of using post-outcome or discharge-time information when constructing early prediction tasks.

G.2 Cohort Selection

We focused on stroke-related clinical prediction problems. Stroke admissions were identified using ICD-9/10 diagnosis codes.

For the early stroke type classification task, we additionally collected stroke-mimic admissions from clinically relevant non-stroke conditions that may present with stroke-like symptoms. These included seizure, migraine, and hypoglycemia.

G.3 Prediction Tasks

We constructed three clinical prediction tasks covering both classification and regression settings.

Early stroke type prediction. The first task is a three-class classification task that distinguishes AIS, ICH, and stroke-mimic admissions. Each admission is represented using five modalities: clinical text, head CT reports, head MRI reports, laboratory measurements, and vital signs. The task is designed to reflect an early stroke assessment scenario, where the model must integrate heterogeneous clinical evidence under naturally incomplete modality availability.

100-day mortality prediction. The second task is a binary classification task that predicts whether a patient dies within 100 days after hospital admission. The label is defined using the recorded date of death: an admission is assigned a positive label if the death date occurs within 100 days after the admission time, and a negative label otherwise. The input modalities for this task include clinical text, admission ICD-9/10 diagnosis codes, and laboratory measurements. To enable a fair comparison with unimodal baselines, we retained only admissions for which all three modalities were available. To mitigate data leakage, we used only ICD codes recorded at admission.

Length-of-stay prediction. The third task is a regression task that predicts hospital length of stay for stroke patients. Length of stay was computed as the time difference between discharge time and admission time and measured in days. The input modalities include clinical text, head CT reports, head MRI reports, laboratory measurements, and vital signs. Following common practice for skewed clinical duration targets, we used the log-transformed length of stay as the regression target:

$$y_{\mathrm{LOS}} = \log(1 + \mathrm{LOS}_{\mathrm{days}}).$$

For evaluation, we report regression metrics on the transformed target space.

ADNI disease-stage prediction. We additionally consider three-class disease-stage prediction on ADNI, where the target is to distinguish cognitively normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD). Each sample may contain four modalities: MRI, clinical variables, biospecimen measurements, and genetic features. We follow the ADNI cohort construction and preprocessing of Flex-MoE (Yun et al., 2024); our adaptation to the cross-validation setting is described in Appendix G.

G.4 Modality Extraction

For each hospital admission, we extracted heterogeneous modalities from structured and unstructured clinical records. Table 11 summarizes the modalities used in this study.

Table 11: Summary of extracted modalities.

Modality	Source	Representation
Clinical text	Clinical notes	Token-level Clinical-Longformer embeddings
CT report	Head CT radiology reports	Token-level BioClinical-ModernBERT embeddings
MRI report	Head MRI radiology reports	Token-level BioClinical-ModernBERT embeddings
Laboratory values	Blood laboratory measurements	Standardized 23-dimensional vector
Vital signs	ICU chart events and monitoring variables	Standardized 22-dimensional vector
ICD codes	ICD-9/10 diagnosis codes	Token-level SapBERT embeddings

Clinical text. For clinical text, we extracted admission-related sections from clinical notes, including Chief Complaint, History of Present Illness, and Past Medical History when available. These sections capture the reason for admission, the clinical history before hospitalization, and relevant comorbidities. When multiple text segments were available for the same admission, we concatenated them using semicolon separators. The resulting clinical text was encoded using a frozen Clinical-Longformer encoder (`yikuan8/Clinical-Longformer`). We used the last hidden states as the clinical text representation, retaining both the `[CLS]` token representation and the remaining token-level representations.

Radiology reports. For radiology modalities, we extracted head CT and head MRI reports from radiology notes. To reduce potential label leakage from explicit diagnostic conclusions, we used the Technique and Findings sections and excluded Impression or Conclusion sections. CT and MRI reports were encoded using a frozen BioClinical-ModernBERT encoder (`thomas-sounack/BioClinical-ModernBERT-base`). Similar to the clinical text modality, we retained the last hidden states as token-level radiology representations.

Laboratory values. Laboratory measurements were selected based on commonly available blood-test variables in stroke-related admissions. We used 23 laboratory variables: Anion Gap, Bicarbonate, Calcium Total, Chloride, Creatinine, Glucose, Hematocrit, Hemoglobin, INR(PT), MCH, MCHC, MCV, Magnesium, PT, PTT, Phosphate, Platelet Count, Potassium, RDW, Red Blood Cells, Sodium, Urea Nitrogen, and White Blood Cells. For each admission, laboratory values within the observation window were aggregated into a fixed-dimensional vector.

Vital signs. Vital-sign and monitoring variables were extracted from ICU chart events. We used 22 variables: Admission Weight, Heart Rate, O2 saturation pulse oximetry, GCS Eye Opening, GCS Motor Response, GCS Verbal Response, Respiratory Rate, Alarms On, Heart Rate Alarm High, Heart Rate Alarm Low, O2 Saturation Pulseoxymetry Alarm Low, Resp Alarm High, Resp Alarm Low, O2 Saturation Pulseoxymetry Alarm High, Parameters Checked, SpO2 Desat Limit, Braden Sensory Perception, Braden Mobility, Braden Moisture, Braden Activity, Braden Nutrition, and Braden Friction/Shear. For each admission, repeated measurements within the observation window were aggregated by their mean value.

Admission ICD codes. For this modality, we used ICD-9/10 diagnosis codes recorded at admission. To reduce the risk of fine-grained diagnosis-code leakage in the 100-day mortal-

ity prediction task, we first truncated each code to its coarse diagnostic category before mapping it to text. Specifically, ICD-10 codes were reduced to their three-character category prefixes (e.g., `I63412` → `I63`), while ICD-9 codes were reduced to their first three digits before any subcategory specification (e.g., `43311` → `433`). These truncated codes were then mapped to their textual descriptions and hierarchical path descriptions. The resulting ICD text sequence was concatenated using semicolon separators and encoded using a frozen SapBERT encoder (`cambridgeltl/SapBERT-from-PubMedBERT-fulltext`). We used the last hidden states as the ICD-code representation, retaining both the `[CLS]` token representation and the remaining token-level representations.

G.5 Tabular Preprocessing

For laboratory and vital-sign modalities, all preprocessing steps were performed within each cross-validation run to avoid information leakage. Specifically, for each run, we computed the mean and standard deviation of each tabular feature using only the training split. The validation and test splits were then standardized using the same training-set statistics:

$$\tilde{x}_{i,j} = \frac{x_{i,j} - \mu_j^{\text{train}}}{\sigma_j^{\text{train}} + \epsilon},$$

where $x_{i,j}$ denotes feature j for sample i, and μ_j^{train} and σ_j^{train} denote the training-set mean and standard deviation of feature j. Missing tabular entries were filled with zero after standardization. Since zero corresponds to the training-set mean in the standardized feature space, this is equivalent to mean imputation using training statistics in the original feature space.

G.6 Encoder Freezing

All pretrained text encoders were frozen during model training. This design choice reduces computational cost and ensures that comparisons among fusion models focus primarily on the multimodal fusion mechanism rather than on differences in end-to-end encoder fine-tuning. The frozen encoders were used to produce modality-specific token representations, which were then passed into the fusion models.

G.7 Cross-Validation Splits

We evaluated all models using a 5-fold cross-validation protocol with patient-level splits. For classification tasks, the folds were stratified jointly by the target label and observed missing-modality pattern. For each run, one fold was held out as the test set, and the remaining four folds were used for model development. The development portion was further divided into training and validation sets with an 8:2 ratio, and model selection was performed only on the validation split.

All preprocessing operations that require estimating data statistics, such as tabular standardization, were fit only on the training split of each run and then applied to the validation and test splits. This prevents information from the validation or test data from influencing feature normalization.

G.8 Dataset Statistics

Table 12: Dataset statistics for the three prediction tasks.

Task	#Admissions	Target type	Modalities
Stroke type prediction	20749	3-class classification	C, CT, MRI, L, V
100-day mortality prediction	7446	Binary classification	C, ICD, L
Length-of-stay prediction	12474	Regression	C, CT, MRI, L, V

Table 12 summarizes the final dataset statistics for each task. For classification tasks, Table 13 reports the class distribution.

Table 13: Class distribution for classification tasks.

Task	Class	Count	Percentage
Stroke type prediction	AIS	7908	38.11%
Stroke type prediction	ICH	4238	20.43%
Stroke type prediction	Stroke mimic	8603	41.46%
100-day mortality prediction	Survived beyond 100 days	5717	76.78%
100-day mortality prediction	Died within 100 days	1729	23.22%

Table 14: Availability rate of each modality across labels

Label	Clinical note	CT report	MRI report	Lab	Vital
AIS	74.5%	38.5%	36.1%	85.4%	46.9%
ICH	72.3%	41.0%	25.5%	87.8%	70.0%
Stroke mimic	72.3%	18.6%	20.1%	85.4%	46.9%

H Task-wise Training Hyperparameters

Tables 15 and 16 summarize the main architectural and optimization settings used to train AURA on each task.

Table 15: Task-wise architectural hyperparameters of AURA. STP denotes stroke type prediction, MP denotes 100-day mortality prediction, and LOS denotes length-of-stay prediction.

Hyperparameter	STP	MP	LOS	ADNI
d_{model}	128	64	128	64
Attention heads	8	8	8	4
Attention dropout	0.2	0.1	0.1	0.2
MLP ratio	2.0	2.0	2.0	2.0
MSA depth	2	2	2	1
SCMA depth	2	1	2	2
Shared weights	Yes	Yes	Yes	Yes
EMA decay	0.95	0.99	0.99	0.99
Modality slots	$[16, 8, 8, 1, 1]$	$[16, 4, 1]$	$[16, 8, 8, 1, 1]$	$[4, 4, 1, 12]$

I Complete MNAR Robustness Results

I.1 Complete Results under MultiModN-style MNAR Evaluation

We report the complete results under the MultiModN-style MNAR evaluation in Tables 17–19. All results are reported as mean $\pm$ standard deviation over five folds. Best results are shown in **bold** and second-best results are underlined.

I.2 Complete Results under the Controlled MNAR Shift Evaluation protocol

Figure 4 shows the complete Controlled MNAR Shift results on the stroke type prediction task.

The complete absolute results underlying Figure 4 are reported in Tables 20–22. All results are reported as mean $\pm$ standard deviation over five folds. Best results are shown in **bold** and second-best results are underlined separately within each test environment.

Because the feasible set of T is defined by linear constraints and $q_T(\hat{R}, Y)$ is linear in T, maximizing the JSD can favor boundary or extreme feasible transformations. Such solutions can produce

Table 16: Task-wise optimization hyperparameters used to train AURA.

Hyperparameter	STP	MP	LOS	ADNI
Batch size	128	64	64	32
Maximum epochs	30	30	30	30
Optimizer	AdamW	AdamW	AdamW	AdamW
Learning rate	1×10^{-4}	1×10^{-4}	1×10^{-4}	3×10^{-4}
Weight decay	5×10^{-5}	3×10^{-4}	2×10^{-5}	5×10^{-4}
LR scheduler	Cosine	Cosine	Cosine	Cosine
Alignment weight	0.1	0.1	0.1	0.1
Task loss	Cross-entropy	Binary cross-entropy	MSE	Cross-entropy
Checkpoint metric	Val. κ	Val. F1	Val. R^2	Val. balanced accuracy

Table 17: Balanced accuracy (%) under the MultiModN-style MNAR evaluation.

Method	Standard	Masking probability p_{mask}			
		0.2	0.4	0.6	0.8
HEALNet	71.99 ± 0.57	70.71 ± 0.21	70.24 ± 0.64	68.44 ± 0.58	64.88 ± 0.88
Flex-MoE	70.90 ± 0.57	69.33 ± 0.98	68.51 ± 0.43	66.65 ± 0.73	62.77 ± 0.66
MultiModN	69.22 ± 1.04	69.71 ± 1.06	68.55 ± 1.03	66.06 ± 1.10	61.11 ± 0.93
MASAM	72.16 ± 0.38	71.58 ± 0.52	70.95 ± 0.90	69.24 ± 2.00	65.61 ± 1.22
AMC	62.69 ± 0.90	59.66 ± 0.75	57.79 ± 1.60	54.71 ± 0.81	52.81 ± 0.97
AURA	$\mathbf{73.49 \pm 0.33}$	$\mathbf{72.94 \pm 0.70}$	$\mathbf{72.52 \pm 0.35}$	$\mathbf{71.45 \pm 0.66}$	$\mathbf{69.81 \pm 0.49}$

Table 18: Macro F1 (%) under the MultiModN-style MNAR evaluation.

Method	Standard	Masking probability p_{mask}			
		0.2	0.4	0.6	0.8
HEALNet	72.56 ± 0.70	70.95 ± 0.38	70.29 ± 0.28	68.39 ± 0.90	64.57 ± 0.82
Flex-MoE	71.19 ± 0.57	69.79 ± 0.87	68.89 ± 0.51	66.92 ± 0.80	62.84 ± 0.57
MultiModN	69.98 ± 0.83	70.00 ± 1.01	68.76 ± 0.99	66.15 ± 1.06	60.90 ± 1.04
MASAM	72.39 ± 0.88	71.72 ± 0.57	70.95 ± 1.03	68.91 ± 2.27	65.30 ± 1.72
AMC	63.14 ± 0.78	60.02 ± 0.71	57.93 ± 1.60	54.68 ± 0.82	52.69 ± 1.05
AURA	$\mathbf{73.77 \pm 0.75}$	$\mathbf{73.32 \pm 0.66}$	$\mathbf{73.03 \pm 0.40}$	$\mathbf{71.82 \pm 0.74}$	$\mathbf{69.85 \pm 0.34}$

Table 19: Cohen's κ under the MultiModN-style MNAR evaluation.

Method	Standard	Masking probability p_{mask}			
		0.2	0.4	0.6	0.8
HEALNet	$.5863 \pm .0109$	$.5653 \pm .0066$	$.5575 \pm .0043$	$.5332 \pm .0150$	$.4828 \pm .0147$
Flex-MoE	$.5706 \pm .0090$	$.5524 \pm .0153$	$.5397 \pm .0108$	$.5155 \pm .0131$	$.4648 \pm .0090$
MultiModN	$.5551 \pm .0121$	$.5446 \pm .0169$	$.5391 \pm .0172$	$.5057 \pm .0180$	$.4395 \pm .0177$
MASAM	$.5859 \pm .0077$	$.5777 \pm .0084$	$.5679 \pm .0134$	$.5409 \pm .0321$	$.4956 \pm .0270$
AMC	$.4606 \pm .0086$	$.4204 \pm .0094$	$.3946 \pm .0168$	$.3543 \pm .0076$	$.3324 \pm .0178$
AURA	$\mathbf{.6046 \pm .0057}$	$\mathbf{.5991 \pm .0117}$	$\mathbf{.5961 \pm .0063}$	$\mathbf{.5815 \pm .0141}$	$\mathbf{.5641 \pm .0036}$

Table 20: Balanced accuracy (%) under the Controlled MNAR Shift Evaluation protocol.

Method	Evaluation	Standard	Training MNAR severity τ			
			0.2	0.3	0.4	0.5
HEALNet	Test-Same	71.99 ± 0.57	77.22 ± 0.68	79.32 ± 0.71	79.92 ± 0.63	86.75 ± 0.42
	Test-Shift		53.75 ± 1.75	46.83 ± 1.75	30.97 ± 0.55	33.18 ± 1.05
Flex-MoE	Test-Same	70.90 ± 0.57	75.02 ± 0.48	78.33 ± 0.67	79.13 ± 0.73	86.17 ± 0.18
	Test-Shift		52.70 ± 3.42	45.63 ± 1.06	30.11 ± 1.13	34.37 ± 0.41
MultiModN	Test-Same	69.22 ± 1.04	75.73 ± 0.52	78.13 ± 0.52	78.46 ± 0.70	84.45 ± 0.45
	Test-Shift		56.39 ± 0.83	48.18 ± 0.87	$\underline{39.23 \pm 1.92}$	$\underline{37.01 \pm 0.85}$
MASAM	Test-Same	$\underline{72.16 \pm 0.38}$	77.07 ± 0.68	79.36 ± 0.43	80.15 ± 0.47	86.50 ± 0.27
	Test-Shift		$\underline{57.22 \pm 1.41}$	$\underline{49.85 \pm 0.78}$	35.53 ± 2.56	34.87 ± 0.97
AMC	Test-Same	62.69 ± 0.90	72.20 ± 0.62	74.55 ± 1.36	77.01 ± 0.95	84.36 ± 0.42
	Test-Shift		46.69 ± 1.96	42.02 ± 2.19	26.44 ± 0.78	26.46 ± 0.73
AURA	Test-Same	$\mathbf{73.49 \pm 0.33}$	72.81 ± 1.39	75.15 ± 0.38	76.08 ± 0.79	80.83 ± 0.69
	Test-Shift		$\mathbf{68.89 \pm 1.12}$	$\mathbf{61.31 \pm 0.73}$	$\mathbf{62.43 \pm 0.71}$	$\mathbf{59.29 \pm 1.89}$

Table 21: Macro F1 (%) under the Controlled MNAR Shift Evaluation protocol.

Method	Evaluation	Standard	Training MNAR severity τ			
			0.2	0.3	0.4	0.5
HEALNet	Test-Same	$\underline{72.56 \pm 0.70}$	78.05 ± 0.44	80.01 ± 0.69	80.71 ± 0.60	87.26 ± 0.19
	Test-Shift		53.51 ± 1.74	47.19 ± 1.60	30.78 ± 0.62	35.00 ± 0.99
Flex-MoE	Test-Same	71.19 ± 0.57	75.82 ± 0.36	77.88 ± 0.39	80.02 ± 0.73	86.40 ± 0.15
	Test-Shift		52.62 ± 3.59	45.44 ± 1.10	30.05 ± 1.32	35.03 ± 0.41
MultiModN	Test-Same	69.98 ± 0.83	75.96 ± 0.59	77.94 ± 0.62	78.63 ± 0.91	84.98 ± 0.48
	Test-Shift		56.02 ± 0.81	47.67 ± 0.59	$\underline{39.20 \pm 1.84}$	$\underline{37.74 \pm 0.77}$
MASAM	Test-Same	72.39 ± 0.88	77.51 ± 0.68	79.61 ± 0.31	80.40 ± 0.59	86.78 ± 0.33
	Test-Shift		$\underline{57.04 \pm 1.36}$	$\underline{49.88 \pm 0.76}$	35.44 ± 2.56	35.67 ± 0.89
AMC	Test-Same	63.14 ± 0.78	73.04 ± 0.48	75.28 ± 1.04	77.45 ± 0.90	84.29 ± 0.55
	Test-Shift		46.24 ± 2.20	42.09 ± 2.18	26.39 ± 0.93	27.80 ± 0.74
AURA	Test-Same	$\mathbf{73.77 \pm 0.75}$	73.60 ± 1.13	75.79 ± 0.38	76.72 ± 0.57	81.80 ± 0.73
	Test-Shift		$\mathbf{68.97 \pm 1.02}$	$\mathbf{62.26 \pm 0.83}$	$\mathbf{62.43 \pm 0.69}$	$\mathbf{59.25 \pm 1.86}$

Table 22: Cohen's κ under the Controlled MNAR Shift Evaluation protocol.

Method	Evaluation	Standard	Training MNAR severity τ			
			0.2	0.3	0.4	0.5
HEALNet	Test-Same	$\underline{.5863 \pm .0109}$	$.6633 \pm .0080$	$.7041 \pm .0105$	$.7113 \pm .0088$	$.7942 \pm .0045$
	Test-Shift		$.3566 \pm .0201$	$.1792 \pm .0229$	$-.0371 \pm .0081$	$.0380 \pm .0135$
Flex-MoE	Test-Same	$.5706 \pm .0090$	$.6278 \pm .0052$	$.6719 \pm .0063$	$.6976 \pm .0102$	$.7793 \pm .0026$
	Test-Shift		$.3372 \pm .0389$	$.1628 \pm .0111$	$-.0596 \pm .0210$	$.0505 \pm .0107$
MultiModN	Test-Same	$.5551 \pm .0121$	$.6350 \pm .0115$	$.6760 \pm .0121$	$.6783 \pm .0138$	$.7661 \pm .0073$
	Test-Shift		$.3774 \pm .0095$	$.2004 \pm .0091$	$\underline{.1014 \pm .0278}$	$\underline{.0886 \pm .0096}$
MASAM	Test-Same	$.5859 \pm .0077$	$.6575 \pm .0132$	$.6981 \pm .0059$	$.7065 \pm .0094$	$.7892 \pm .0059$
	Test-Shift		$\underline{.3923 \pm .0192}$	$\underline{.2246 \pm .0095}$	$.0317 \pm .0388$	$.0618 \pm .0147$
AMC	Test-Same	$.4606 \pm .0086$	$.5937 \pm .0064$	$.6393 \pm .0131$	$.6655 \pm .0132$	$.7506 \pm .0074$
	Test-Shift		$.2674 \pm .0257$	$.1168 \pm .0342$	$-.0919 \pm .0141$	$-.0541 \pm .0120$
AURA	Test-Same	$\mathbf{.6046 \pm .0057}$	$.5981 \pm .0182$	$.6388 \pm .0082$	$.6426 \pm .0115$	$.7130 \pm .0111$
	Test-Shift		$\mathbf{.5476 \pm .0167}$	$\mathbf{.4304 \pm .0123}$	$\mathbf{.4512 \pm .0117}$	$\mathbf{.4236 \pm .0125}$

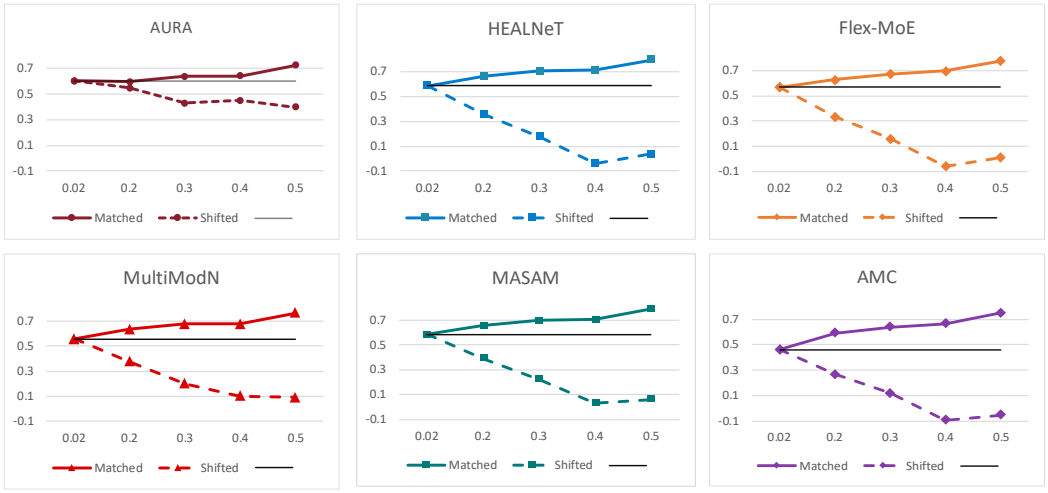


Figure 4: Controlled MNAR Shift Evaluation results on stroke type prediction. The horizontal axis shows the training MNAR severity τ, and the vertical axis shows Cohen's κ. For each method, the solid curve denotes Test-Same performance, the dashed curve denotes Test-Shift performance, and the horizontal reference line denotes the corresponding standard-setting performance.

strongly different class-conditional availability-pattern distributions from T_{train}, which is consistent with the severe degradation observed for some baselines under Test-Shift.

I.3 Representation Inconsistency Evaluation

To empirically examine representation consistency across modality-availability patterns, we measure how much the learned representation of the same sample changes when different subsets of its naturally observed modalities are retained. This evaluation follows the worst-case intuition of the representation inconsistency in Eq. 6, but is computed at the modality-specific representation level using the same discrepancy adopted by the default alignment objective.

For sample i, let

$$O_i \subseteq \mathcal{M}$$

denote its set of naturally observed modalities. We use the natural availability pattern O_i as the reference view.

For each such sample, we enumerate all proper subsets of the naturally observed modalities that retain at least two modalities:

$$\mathcal{S}_i = \{S \subsetneq O_i : |S| \neq \varnothing\}.$$

Thus, the evaluation introduces only additional modality removal and never makes a naturally missing modality available.

Let $H_{i,m}^{O_i}$ denote the refined representation of modality m when the sample is processed under its natural availability pattern, and let $H_{i,m}^{S}$ denote the corresponding representation when only modalities in S are retained. For each alternative pattern S, we compute the discrepancy separately for each modality and average across modalities:

$$d_i(S) = \frac{1}{M} \sum_{m=1}^{M} d\left(H_{i,m}^{O_i}, H_{i,m}^{S}\right),$$

We then define the sample-level representation inconsistency as the largest discrepancy over all admissible modality subsets:

$$\delta_i^{\text{emp}} = \max_{S \in \mathcal{S}_i} d_i(S).$$

Table 23: Ablation results on the STP task under the Controlled MNAR Shift Evaluation protocol. Within each cell, the three rows report Standard, TEST-SAME, and TEST-SHIFT, respectively. Parentheses denote the relative performance drop from TEST-SAME to TEST-SHIFT. Rep. Incons. denotes representation inconsistency across modality-availability subsets.

Variant	BAcc (%) ↑	F1 (%) ↑	Kappa ↑	Rep. Incons.
AURA	73.49 ±0.33 80.83 ±0.69 59.29 ±1.89 **(-26.6%)**	73.77 ±0.75 81.80 ±0.73 59.25 ±1.86 **(-27.6%)**	.6046 ±.0057 .7130 ±.0111 .4236 ±.0125 **(-40.6%)**	.3975 ±.0385
w/o SCMA	71.63 ±0.60 81.87 ±0.99 57.85 ±2.30 **(-29.3%)**	71.87 ±0.72 82.81 ±1.01 57.85 ±2.23 **(-30.1%)**	.5773 ±.0090 .7271 ±.0162 .4042 ±.0314 **(-44.4%)**	.3846 ±.0121
w/o Alignment loss	72.52 ±0.49 83.66 ±0.88 46.28 ±1.15 **(-44.7%)**	72.88 ±0.44 84.54 ±0.66 46.42 ±1.08 **(-45.1%)**	.5914 ±.0090 .7531 ±.0111 .2642 ±.0160 **(-64.9%)**	.8890 ±.0178
w/o Modality Dropout	72.76 ±0.53 86.81 ±0.37 36.79 ±1.86 **(-57.6%)**	73.00 ±0.37 87.27 ±0.43 37.54 ±1.79 **(-57.0%)**	.5936 ±.0052 .7951 ±.0065 .0985 ±.0279 **(-87.6%)**	.9268 ±.0126
Full alignment	71.80 ±0.89 74.22 ±0.42 56.59 ±0.89 **(-23.8%)**	71.94 ±0.82 75.21 ±0.29 56.58 ±0.94 **(-24.8%)**	.5703 ±.0142 .6548 ±.0081 .3987 ±.0171 **(-39.1%)**	.1923 ±.0118

The representation inconsistency reported in the ablation study is the average of this worst-case quantity over all eligible test samples:

$$\mathrm{RepIncons} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \delta_i^{\mathrm{emp}}, \qquad \mathcal{I} = \{i : |O_i| \geq 2\}.$$

We compute this quantity independently for each cross-validation fold and report the mean and standard deviation across folds. Lower values indicate that the model produces more stable modality-specific representations when the availability pattern changes.

We emphasize that $\mathrm{RepIncons}$ is an empirical proxy for, rather than a direct estimate of, the theoretical quantity in Eq. 6. Nevertheless, both quantities capture the same property of interest: the sensitivity of learned representations to changes in modality availability.

I.4 COMPONENT ABLATIONS

We provide the complete ablation results for the main components of AURA. The experiments separate the roles of selective cross-modal refinement, artificial modality dropout, and representation alignment. Table 23 reports standard predictive performance, robustness under the Controlled MNAR Shift Evaluation protocol, and representation inconsistency for each variant.

w/o SCMA. This variant replaces SCMA with standard self-attention while keeping the rest of the AURA architecture and training procedure unchanged. It isolates the contribution of selective cross-modal refinement relative to a standard attention-based fusion mechanism.

w/o Alignment loss. This variant retains the artificial modality dropout used by the full model, so an augmented view is still constructed by removing a subset of the observed modalities and the supervised task loss is computed on this view. However, we remove the representation-alignment objective and the EMA target branch. This comparison isolates the contribution of explicitly aligning representations across availability patterns beyond the robustness obtained from exposure to randomly masked views alone.

w/o Modality Dropout. This variant removes the artificial modality-dropout procedure used to construct augmented views. Since no augmented view is created, the associated representation-alignment objective and EMA target branch are also removed.

Full alignment. This variant applies the alignment loss to all modality representations, rather than only to the artificially removed modalities.

All variants use the same data splits and evaluation protocol as the full model.