

A Broader Statement of Impact

This research addresses the challenge of heterogeneous missing data in multimodal federated learning. Our novel design and theoretical analysis help bridge gaps between incomplete multimodal clients in fragmented systems by effectively handling diverse missing data patterns. This enables practical applications in privacy-sensitive multimodal settings with highly incomplete data. While the potential real-world use of our methods could raise ethical concerns, these are indirect and unpredictable consequences beyond the scope of this work. Our experiments rely solely on publicly available datasets, and no ethical issues arise from our evaluation process.

B Missing Modality Simulation

This section details how we simulate missing modality in a comprehensive and controllable way. Following [34], we define two types of ratio in missing modality, denoted as p_s and p_m . First, p_s , namely sample ratio, is the ratio of samples with missing modalities over a given dataset. Second, p_m is modality ratio, and used as the ratio of missing modalities within those samples. For simplicity, a pair of (p_m, p_s) can be called *missing statistics*, since it reflects statistics of modality missing in both detailed and overall views (see Fig. 5). The *missing degree* is then defined as $p_m \times p_s$, representing the overall proportion of instances with missing modalities. These missing statistic can remodel the an arbitrary dataset \mathcal{D} via a missing matrix:

$$\phi(\mathcal{D}) = \begin{bmatrix} b_1^1 & \dots & b_1^{|\mathcal{M}|} \\ \vdots & \ddots & \vdots \\ b_{|\mathcal{D}|}^1 & \dots & b_{|\mathcal{D}|}^{|\mathcal{M}|} \end{bmatrix}, \quad (10)$$

where $b_{dm} \in \{0, 1\}$ indicates whether modality m is missing (0) or available (1) for the d -th sample. Here, $|\mathcal{M}|$ is the cardinality of \mathcal{M} , and $|\mathcal{D}|$ is the number of samples. The incomplete dataset $\hat{\mathcal{D}}$ can be obtained by multiplying \mathcal{D} by the missing matrix:

$$\hat{x}_i = [x_{d1}, \dots, x_{d|\mathcal{M}|}] \odot [b_{d1}, \dots, b_{d|\mathcal{M}|}], \quad (11)$$

where \odot represents element-wise multiplication. Examples of incomplete datasets are shown in Fig. 5. In this work, we apply the same (p_m/p_s) pairs for all clients in our experiments.

C Implementation Details

Dataset Preparation. All baselines use data from the PTBXL and EDF datasets. The PTBXL dataset contains 3,963 clinical samples across five classes. Each sample includes 12 modalities, corresponding to electrocardiogram (ECG) recordings, and is labeled with a single class. Details can be found in [39]. The EDF dataset consists of 197 full-night polysomnographic (PSG) recordings with five key modalities (excluding rectal temperature and biomarkers). Each recording is segmented into multiple sleep stages, including Wake and stages S1–S4. For this work, we relabel S1 and S2 as N1 and N2, and merge S3 and S4 into N3, resulting in a 5-class classification problem [20]. We segment all sleep recordings into individual signals, each representing a sleep pattern, creating a unified dataset of 8,755 signals. This unified dataset is used for all experiments. Both datasets are divided into training and testing sets with ratio 80/20. The testing are used for evaluation on the server side, while the training sets are split to all clients following IID or NonIID settings. For NonIID setting, we use Dirichlet distribution with $\alpha = 0.5$ to distribute training data points.

Hyperparameter Settings. All methods in this work use an Inception Network as the modality encoder, following [39]. Experiments are run on an A6000 GPU with 48GB of memory. For classification, we use Cross Entropy Loss for \mathcal{L}_{task} . The embedding dimension is set to $C = 128$. There are $K = 32$ clients in total, with 10 clients randomly selected to participate in each training round. Each selected client trains the model for $E = 3$ epochs per round. Optimization is done using Stochastic Gradient Descent (SGD) [42]. Communication with the server occurs over $T = 1000$ rounds. Both the alignment contrastive weight (λ) and the relevance regularization weight (η) are set to 0.1 for all experiments. However, λ is increased to 0.2 when $p_m \in \{0.8, 1.0\}$, corresponding to extreme missing modality scenarios that require stronger alignment. Detailed hyperparameter settings are listed in Tab. 4. Unless otherwise specified, we use the original configurations from the referenced papers.

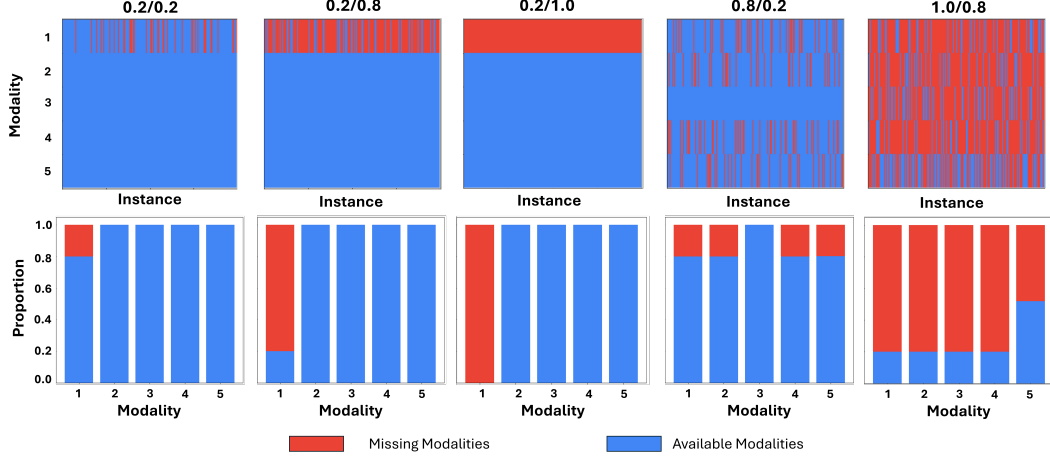


Figure 5: Examples of incomplete datasets $\hat{\mathcal{D}}$ with varying missing statistics (p_m/p_s). By controlling these missing statistics, we create diverse evaluation scenarios that reflect real-world conditions.

Table 4: Hyperparameter setting for all baselines and our PEPsy

Dataset	Method	p_m	Batch Size	Communication Round (T)	Eps. in Local Training (E)	Contrastive Weight (λ)	Optimizer & Learning Rate	Total Clients (K)	Sampled Clients
PTBXL	FedProx	0.2	32	1000	3	0.1	SGD lr: 0.01	32	10
	MIFL	0.4	32	1000	3	0.1	SGD lr: 0.01	32	10
	FedInMM	0.6	32	1000	3	0.1	SGD lr: 0.01	32	10
	FedMSplit	0.8	32	1000	3	0.2	SGD lr: 0.01	32	10
	FedMAC PEPSY	1.0	32	1000	3	0.2	SGD lr: 0.01	32	10
EDF	FedProx	0.2	128	500	3	0.1	SGD lr: 0.1	32	10
	MIFL	0.4	128	500	3	0.1	SGD lr: 0.1	32	10
	FedInMM	0.6	128	500	3	0.1	SGD lr: 0.1	32	10
	FedMSplit	0.8	128	500	3	0.2	SGD lr: 0.1	32	10
	FedMAC PEPSY	1.0	128	500	3	0.2	SGD lr: 0.1	32	10

D Theorem Proof

D.1 Theorem Setup

This section provides the initial setup for our proof for Theorem 3.1. From now on, we remove the subscript indicating instance index in our notation for simplicity. Following notations in Section 1, our proposed method described in Section 2 can be expressed as composition of two internal functions: $\hat{\mathbf{y}} = f_p(\{\mathbf{w}_i\}_{i=1}^{|\mathcal{M}|}) = f_p(\{f_e(\mathbf{x}_i)\}_{i=1}^{|\mathcal{M}|})$. Here, $f_p(\cdot)$ and $f_e(\cdot)$ are post-process head and feature extractor, respectively. In specific, $f_e(\cdot)$ takes each modality \mathbf{x}_i as input and generates a modality representation \mathbf{w}_i (as shown in Section 2) by concatenating three types of information including modality-specific ($\mathbf{w}_i^{\text{mod}}$), data-specific ($\mathbf{w}_i^{\text{ins}}$) and missing-pattern ($\mathbf{w}_i^{\text{mis}}$) features, i.e., $\mathbf{w}_i = [\mathbf{w}_i^{\text{mod}} \circ \mathbf{w}_i^{\text{ins}} \circ \mathbf{w}_i^{\text{con}}]$. In addition, to make the proof easy to follow, we denote \mathbf{h}_i and \mathbf{u} as extraction for present modalities and imputation for missing modalities, respectively, as described in Section 2.2.1, leading to follow-up notation of modality representations \mathbf{w}_i and $\mathbf{w}_i(\mathbf{u})$. To clarify, if the notation \mathbf{h}_i is used for missing modality, i.e., $i \in \mathcal{S}$, it means that \mathbf{h}_i here is the "true" feature if that modality presents. We use this notation in our proof from now on.

Assumption D.1 The post-processing head f_p is Lipschitz continuous with respect to the input vector \mathbf{x} , i.e., there exists a constant $L > 0$ such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the following condition holds:

$$\|f_p(\mathbf{x}_i) - f_p(\mathbf{x}_j)\| \leq L\|\mathbf{x}_i - \mathbf{x}_j\|,$$

where $f_p : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is the post-processing head, $\|\cdot\|$ denotes the chosen norm (here the ℓ_2 -norm), and L is a Lipschitz constant.

Assumption D.2 During test time, all parameters of the proposed framework are bounded. Specifically, for any weight matrix A , we have:

$$\epsilon_A^- \leq \|A\| \leq \epsilon_A^+,$$

where $\|\cdot\|$ denotes the ℓ_2 -norm and ϵ_A^- and ϵ_A^+ are positive constants that bound the spectral norm of A . This assumption similarly applies to the output representations that are transformed by the learned weight matrices.

In Assumption D.1, we assume that the neural network used as the post-processing head in our proposed design is Lipschitz continuous. This assumption is widely accepted in the machine learning community due to its relevance in ensuring stable and smooth behavior of the model.

Assumption D.2 states that the learned parameters of the network are bounded during test time. This assumption is reasonable and holds true in most real-world scenarios, where the model parameters are deterministic and constrained within known ranges during inference. Such bounds are typically enforced either through explicit regularization during training or through implicit constraints imposed by the training process itself (e.g., gradient clipping or weight normalization). Therefore, this assumption is not only theoretically sound but also consistent with common practices in machine learning.

Remark D.3 (Bounded Extracted Representations) In our data-specific representation extraction, each output feature \mathbf{h}_i of modality i is normalized to zero mean and unit variance (via Batch Normalization layer), followed by a learned scaling (γ) and shift (β) parameters. When Assumption D.2 holds, we have $\epsilon_\gamma^- \leq \|\gamma\| \leq \epsilon_\gamma^+$ and $\epsilon_\beta^- \leq \|\beta\| \leq \epsilon_\beta^+$ and derive:

$$\|\mathbf{h}_i\| = \|\gamma \bar{\mathbf{h}}_i + \beta\| \leq \|\gamma\| \cdot \|\bar{\mathbf{h}}_i\| + \|\beta\|. \quad (12)$$

where $\bar{\mathbf{h}}_i$ is batch-normalized \mathbf{h}_i . Since the normalized term has unit variance, its norm is bounded by \sqrt{C} , where C is the feature dimension. Hence,

$$\max(\epsilon_\gamma^- \sqrt{C} - \epsilon_\beta^+, 0) \leq \|\mathbf{h}_i\| \leq \epsilon_\gamma^+ \sqrt{C} + \epsilon_\beta^+, \quad (13)$$

Let $\epsilon_{\gamma\beta}^- \triangleq \max(\epsilon_\gamma^- \sqrt{C} - \epsilon_\beta^+, 0)$ and $\epsilon_{\gamma\beta}^+ \triangleq \epsilon_\gamma^+ \sqrt{C} + \epsilon_\beta^+$, Eq. 13 shows that $\|\mathbf{h}_i\|$ is bounded within a deterministic range. Consequently, the imputation feature derived by taking average of available modalities is bounded for the same reason.

D.2 Theoretical Analysis in Simple Case

In this section, we first investigate the behavior of PEPsy in a simple case of missing modality before further generalization. Let consider the deviations of our proposal when feeding full-modality input and one missing the first $|\mathcal{S}|$ out of \mathcal{M} modalities, i.e., $\mathcal{S}_f = \{1, \dots, |\mathcal{S}|\}$ as follows:

$$\|\mathbf{y}^{\mathcal{S}} - \mathbf{y}^\emptyset\| \quad (14)$$

$$= \left\| f_p\left(\{\mathbf{w}_i\}_{i=|\mathcal{S}|+1}^{|\mathcal{M}|}, \{\mathbf{w}_j(\mathbf{u})\}_{j=1}^{|\mathcal{S}|}\right) - f_p\left(\{\mathbf{w}_i\}_{i=1}^{|\mathcal{M}|}\right) \right\| \quad (15)$$

$$= \left\| \frac{1}{|\mathcal{S}|} \left(\|\mathbf{w}_1(\mathbf{u}) - \mathbf{w}_1\| \nabla_{\mathbf{w}_1(\mathbf{u})} f_p(\mathbf{w}_1) + \dots + \|\mathbf{w}_{|\mathcal{M}|}(\mathbf{u}) - \mathbf{w}_{|\mathcal{M}|}\| \nabla_{\mathbf{w}_{|\mathcal{M}|}(\mathbf{u})} f_p(\mathbf{w}_{|\mathcal{M}|}) \right) \right\| \quad (16)$$

$$= \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\mathbf{w}_i(\mathbf{u}) - \mathbf{w}_i\| \nabla_{\mathbf{w}_i(\mathbf{u})} f_p(\mathbf{w}_i) \quad (17)$$

Here, we use first-order Taylor approximation $|\mathcal{S}|$ times to transform Eq. 15 to Eq. 16. Since $f_p(\cdot)$ is L -Lipschitz (see Assumption D.1), Eq 17 can be transformed as:

$$\|\mathbf{y}^{\mathcal{S}_f} - \mathbf{y}^\emptyset\| \quad (18)$$

$$\leq L \sum_{i=1}^{|\mathcal{M}|} \|\mathbf{w}_i(\mathbf{u}) - \mathbf{w}_i\| \quad (19)$$

$$\leq L \sum_{i=1}^{|\mathcal{M}|} \|\left[\mathbf{w}_i^{\text{mod}} \circ \mathbf{u} \circ \mathbf{w}_i^{\text{con}}\right] - \left[\mathbf{w}_i^{\text{mod}} \circ \mathbf{h}_i \circ \mathbf{w}_i^{\text{con}}\right]\| \quad (20)$$

$$= L \sum_{i=1}^{|\mathcal{M}|} \left\| 0 \circ (\mathbf{u} - \mathbf{h}_i) \circ \underset{\psi_p}{\operatorname{argmax}} \left(e(\mathbf{q}(\mathbf{w}_i^{\text{mod}} \circ \mathbf{u}), \psi_p) \right) - \underset{\psi_p}{\operatorname{argmax}} \left(e(\mathbf{q}(\mathbf{w}_i^{\text{mod}} \circ \mathbf{h}_i), \psi_p) \right) \right\| \quad (21)$$

952 where $\mathbf{w}_i(\mathbf{u})$ is the imputed representation for modality i , obtained using the imputation data-specific
 953 feature \mathbf{u}^{ins} , and \mathbf{w}_i is the original modality feature. Here, we represent the query-key matching
 954 function $\underset{\psi_p}{\operatorname{argmax}} e(\mathbf{q}(\mathbf{w}_i^{\text{mod}} \circ \mathbf{w}_i^{\text{ins}}, \psi_p))$ as an approximate attention selecting the ψ_p with the

955 highest weight, by using softmax function $\sigma(\cdot, \cdot) \triangleq \operatorname{softmax}(e(\mathbf{q}(\cdot), \cdot))$. For simplicity, we use $\tilde{\sigma}$
 956 as a Lipschitz constant of this approximated similarity function. Considering individual modality
 957 component $\|\mathbf{w}_i(\mathbf{u}) - \mathbf{w}_i\|$, these lead to the following derivations:

$$\|\mathbf{w}_i(\mathbf{u}) - \mathbf{w}_i\| \quad (22)$$

$$\approx \left\| 0 \circ (\mathbf{u} - \mathbf{h}_i) \circ \left\{ \sum_{p=1}^{\tau} [\sigma(\mathbf{w}_i^{\text{mod}} \circ \mathbf{u}, \psi_p) - \sigma(\mathbf{w}_i^{\text{mod}} \circ \mathbf{h}_i, \psi_p)] \odot \psi_p \right\} \right\| \quad (23)$$

$$\leq \|\mathbf{u} - \mathbf{h}_i\| + \sum_{p=1}^{\tau} \|\sigma(\mathbf{w}_i^{\text{mod}} \circ \mathbf{u}, \psi_p) - \sigma(\mathbf{w}_i^{\text{mod}} \circ \mathbf{h}_i, \psi_p)\| \odot \|\psi_p\|. \quad (24)$$

$$\leq \|\mathbf{u} - \mathbf{h}_i\| + \tilde{\sigma} \sum_{p=1}^{\tau} \|\mathbf{u} - \mathbf{h}_i\| \times \|\psi_p\| \quad (25)$$

$$\leq (1 + \tilde{\sigma} \tau \max_{\psi_p}(\epsilon_{\psi_p}^+)) \|\mathbf{u} - \mathbf{h}_i\| \quad (26)$$

$$\leq \mu \sqrt{\|\mathbf{u}\|^2 + \|\mathbf{h}_i\|^2 - 2\mathbf{u}\mathbf{h}_i^\top}. \quad (27)$$

958 where $\mu = 1 + \tilde{\sigma} \tau \max_{\psi_p}(\epsilon_{\psi_p}^+)$ and $\epsilon_{\psi_p}^+$ denotes upperbound of embedding controls, which is fixed in
 959 test time. Taking the summation over all i , we obtain:

$$\sum_{i=1}^{|\mathcal{S}|} \|\mathbf{w}_i(\mathbf{u}) - \mathbf{w}_i\| \quad (28)$$

$$\leq \mu \sum_{i=1}^{|\mathcal{S}|} \sqrt{\|\mathbf{u}\|^2 + \|\mathbf{h}_i\|^2 - 2\mathbf{u}\mathbf{h}_i^\top} \quad (29)$$

$$\leq \mu \sum_{i=1}^{|\mathcal{S}|} \left(\left\| \frac{1}{|\mathcal{M}| - |\mathcal{S}|} \sum_{j=|\mathcal{S}|+1}^{|\mathcal{M}|} \mathbf{h}_j \right\|^2 + \|\mathbf{h}_i\|^2 - \frac{2}{|\mathcal{M}| - |\mathcal{S}|} \sum_{j=|\mathcal{S}|+1}^{|\mathcal{M}|} \mathbf{h}_j \mathbf{h}_i^\top \right)^{\frac{1}{2}}. \quad (30)$$

960 Here, \mathbf{u} represents the imputed data-specific representation, computed as the mean of corresponding
 961 features from the available modalities (see Section 2.2.1). This justifies the transformation from
 962 Eq. 29 to Eq. 30. Based on Remark D.3, we have:

$$\sum_{i=1}^{|\mathcal{S}|} \|\mathbf{w}_i(\mathbf{u}) - \mathbf{w}_i\| \quad (31)$$

$$\leq \mu \sum_{i=1}^{|\mathcal{S}|} \left(\frac{2}{|\mathcal{M}| - |\mathcal{S}|} (|\mathcal{M}| - |\mathcal{S}|) \epsilon_{\gamma\beta}^{+2} + \epsilon_{\gamma\beta}^{+2} - \frac{2}{|\mathcal{M}| - |\mathcal{S}|} \sum_{j=|\mathcal{S}|+1}^{|\mathcal{M}|} \mathbf{h}_j \mathbf{h}_i^\top \right)^{\frac{1}{2}}. \quad (32)$$

$$\leq \mu \sum_{i=1}^{|\mathcal{S}|} \left(3\epsilon_{\gamma\beta}^{+2} - \frac{2}{|\mathcal{M}| - |\mathcal{S}|} \sum_{j=|\mathcal{S}|+1}^{|\mathcal{M}|} \mathbf{h}_j \mathbf{h}_i^\top \right)^{\frac{1}{2}} \quad (33)$$

$$\leq \mu \sum_{i=1}^{|\mathcal{S}|} \left(3\epsilon_{\gamma\beta}^{+2} - \frac{2}{|\mathcal{M}| - |\mathcal{S}|} \sum_{j=|\mathcal{S}|+1}^{|\mathcal{M}|} \mathbf{h}_j \mathbf{h}_i^\top \right)^{\frac{1}{2}} \quad (34)$$

$$\leq \mu \sum_{i=1}^S \left(\frac{1}{|\mathcal{M}| - |\mathcal{S}|} \sum_{j=|f|+1}^{|\mathcal{M}|} 3\epsilon_{\gamma\beta}^{+2} - \frac{1}{|\mathcal{M}| - |\mathcal{S}|} \sum_{j=|\mathcal{S}|+1}^{\mathcal{M}} 2\mathbf{h}_j \mathbf{h}_i^\top \right)^{\frac{1}{2}} \quad (35)$$

$$\leq \sqrt{\frac{\mu^2}{|\mathcal{M}| - |\mathcal{S}|}} \sum_{i=1}^S \left(\sum_{j=|\mathcal{S}|+1}^{\mathcal{M}} 3\epsilon_{\gamma\beta}^{+2} - 2\mathbf{h}_j \mathbf{h}_i^\top \right)^{\frac{1}{2}} \quad (36)$$

Here, the bound on $\|\mathbf{w}_i(\mathbf{u}) - \mathbf{w}_i\|$ highlights how the interaction terms between \mathbf{h}_j and \mathbf{h}_i contribute to the overall norm. Furthermore, the right-handed side of 36 is non-negative showing the validity of this transformation. If we further substitute Eq. 36 in Eq. 19, we obtain an intermediate inequality:

$$\|\mathbf{y}^{S_f} - \mathbf{y}^\emptyset\| \leq \sqrt{\frac{\mu^2}{|\mathcal{M}| - |\mathcal{S}|}} \sum_{i=1}^{|\mathcal{S}|} \left(\sum_{j=|\mathcal{S}|+1}^{|\mathcal{M}|} 3\epsilon_{\gamma\beta}^2 - 2\mathbf{h}_j \mathbf{h}_i^\top \right)^{\frac{1}{2}} \quad (37)$$

where we restate $\mu^2 \leftarrow \mu^2 L$ without loss of generalization since both μ and L are constant.

D.3 Theoretical Analysis Generalization

In this section, we extend the bound in Eq. 37, originally derived assuming the first $|\mathcal{S}|$ modalities out of \mathcal{M} are missing. The current bound assumes the missing modalities are the first $|\mathcal{S}|$ in order. We generalize this to the case where any subset $\mathcal{S} \subset \mathcal{M}$ of size $|\mathcal{S}|$ is missing. To do this, we generalize bound in Eq. 37 over missing modality set \mathcal{S} , and over all instances of an arbitrary dataset \mathcal{D} .

D.3.1 Generalize over Missing Modality Set.

Given \mathcal{M} is the set of all modalities, with cardinality $|\mathcal{M}|$, we define $\mathcal{S} \subseteq \mathcal{M}$ as a subset representing the missing modalities, with cardinality of $|\mathcal{S}|$. For each missing modality $i \in \mathcal{S}$, we define a random variable Z_S^i as follows:

$$\mathbf{z}_S^i = \sum_{j \notin \mathcal{S}} (3\epsilon_{\gamma\beta}^2 - 2\mathbf{h}_j \mathbf{h}_i^\top), \quad (38)$$

The expected value of $\sqrt{\mathbf{Z}_S^i}$, averaged over all possible missing subsets \mathcal{S} , is then computed as the following equation:

$$\mathbb{E} \left[\sqrt{\mathbf{Z}_S^i} \right] = \frac{1}{|\mathcal{S}| \binom{|\mathcal{M}|}{|\mathcal{S}|}} \sum_{\mathcal{S} \subseteq \mathcal{M}} \sum_{i \in \mathcal{S}} \sqrt{\mathbf{z}_S^i} \quad (39)$$

$$\leq \sqrt{\frac{1}{|\mathcal{S}| \binom{|\mathcal{M}|}{|\mathcal{S}|}} \sum_{\mathcal{S} \subseteq \mathcal{M}} \sum_{i \in \mathcal{S}} \sum_{j \notin \mathcal{S}} (3\epsilon_{\gamma\beta}^2 - 2\mathbf{h}_j \mathbf{h}_i^\top)} \quad (40)$$

where $\binom{|\mathcal{M}|}{|\mathcal{S}|}$ denotes the number of ways to choose $|\mathcal{S}|$ elements from \mathcal{M} . To derive Eq.40 from Eq. 39, we apply the Jensen's inequality due to the concavity of square root function.

Observation. The term $\sum_{\mathcal{S} \subseteq \mathcal{M}} \sum_{i \in \mathcal{S}} \sum_{j \notin \mathcal{S}} (3\epsilon_{\gamma\beta}^2 - 2\mathbf{h}_j \mathbf{h}_i^\top)$ means that we are summing over all subsets $\mathcal{S} \subseteq \mathcal{M}$ of fixed size $|\mathcal{S}|$. For each subset, we sum over all ordered pairs (i, j) where $i \in \mathcal{S}$ and $j \notin \mathcal{S}$. For a fixed pair (i, j) with $i \neq j$, the number of subsets \mathcal{S} that include i and exclude j depends only on i and j . In other words, once i and j are fixed, the remaining $|\mathcal{S}| - 1$ elements of \mathcal{S} must be chosen from the remaining $|\mathcal{M}| - 2$ elements (excluding i and j), giving exactly $\binom{|\mathcal{M}| - 2}{|\mathcal{S}| - 1}$ subsets. Therefore, each term $(3\epsilon_{\gamma\beta}^2 - 2\mathbf{h}_j \mathbf{h}_i^\top)$ appears precisely $\binom{|\mathcal{M}| - 2}{|\mathcal{S}| - 1}$ times in the sum. This lets us rewrite the original triple sum as a double sum over all ordered pairs (i, j) with $i \neq j$, multiplied by the constant $\binom{|\mathcal{M}| - 2}{|\mathcal{S}| - 1}$, simplifying into:

$$\sum_{\mathcal{S} \subseteq \mathcal{M}} \sum_{i \in \mathcal{S}} \sum_{j \notin \mathcal{S}} (3\epsilon_{\gamma\beta}^{+2} - 2\mathbf{h}_j \mathbf{h}_i^\top) = \sum_{i=1}^{|\mathcal{M}|} \sum_{\substack{j=1 \\ j \neq i}}^{|\mathcal{M}|} \binom{|\mathcal{M}| - 2}{|\mathcal{S}| - 1} (3\epsilon^2 - 2\mathbf{h}_i \mathbf{h}_j^\top). \quad (41)$$

988 Substituting Eq. 41 into Eq. 40, we obtain:

$$\mathbb{E}_{i,S} \left[\sqrt{\mathbf{Z}_S^i} \right] \quad (42)$$

$$\leq \sqrt{\frac{1}{|\mathcal{S}| \binom{|\mathcal{M}|}{|\mathcal{S}|}} \sum_{i=1}^{|\mathcal{M}|} \sum_{\substack{j=1 \\ j \neq i}}^{|\mathcal{M}|} \binom{|\mathcal{M}| - 2}{|\mathcal{S}| - 1} (3\epsilon_{\gamma\beta}^{+2} - 2\mathbf{h}_i \mathbf{h}_j^\top)} \quad (43)$$

$$= \sqrt{\frac{|\mathcal{S}|! (|\mathcal{M}| - |\mathcal{S}|)!}{|\mathcal{M}|! |\mathcal{S}|} \times \frac{(|\mathcal{M}| - 2)!}{(|\mathcal{S}| - 1)! (|\mathcal{M}| - |\mathcal{S}| - 1)!} \sum_{i=1}^{|\mathcal{M}|} \sum_{\substack{j=1 \\ j \neq i}}^{|\mathcal{M}|} (3\epsilon_{\gamma\beta}^2 - 2\mathbf{h}_i \mathbf{h}_j^\top)} \quad (44)$$

$$= \sqrt{\frac{|\mathcal{M}| - |\mathcal{S}|}{|\mathcal{M}| (|\mathcal{M}| - 1)} \sum_{i=1}^{|\mathcal{M}|} \sum_{\substack{j=1 \\ j \neq i}}^{|\mathcal{M}|} (3\epsilon_{\gamma\beta}^2 - 2\mathbf{h}_i \mathbf{h}_j^\top)} \quad (45)$$

$$= \sqrt{\frac{|\mathcal{M}| - |\mathcal{S}|}{|\mathcal{M}| (|\mathcal{M}| - 1)}} \times \sqrt{\sum_{i=1}^{|\mathcal{M}|} \sum_{\substack{j=1 \\ j \neq i}}^{|\mathcal{M}|} (3\epsilon_{\gamma\beta}^{+2} - 2\mathbf{h}_i \mathbf{h}_j^\top)}. \quad (46)$$

989 We now bound the expectation of Eq. 37 over all possible missing modality patterns (\mathcal{S}) as follows:

$$\mathbb{E}_{\mathcal{S}} \left[\|\mathbf{y}^{\mathcal{S}} - \mathbf{y}^{\emptyset}\| \right] = \frac{1}{\binom{|\mathcal{M}|}{|\mathcal{S}|}} \sum_{\mathcal{S} \subseteq \mathcal{M}} \|\mathbf{y}^{\mathcal{S}} - \mathbf{y}^{\emptyset}\| \quad (47)$$

$$\leq \sqrt{\frac{\mu^2 |\mathcal{S}|^2}{|\mathcal{M}| - |\mathcal{S}|}} \frac{1}{|\mathcal{S}| \binom{|\mathcal{M}|}{|\mathcal{S}|}} \sum_{\mathcal{S} \subseteq \mathcal{M}} \left[\sum_{i \in \mathcal{S}} \left(\sum_{j \notin \mathcal{S}} (3\epsilon_{\gamma\beta}^{+2} - \mathbf{h}_i \mathbf{h}_j^\top) \right)^{\frac{1}{2}} \right] \quad (48)$$

$$\leq \sqrt{\frac{\mu^2 |\mathcal{S}|^2}{|\mathcal{M}| - |\mathcal{S}|}} \mathbb{E}_{i,S} \left[\sqrt{\mathbf{Z}_S^i} \right] \quad (49)$$

$$\leq \sqrt{\frac{\mu^2 |\mathcal{S}|^2}{|\mathcal{M}| - |\mathcal{S}|}} \sqrt{\frac{|\mathcal{M}| - |\mathcal{S}|}{|\mathcal{M}| (|\mathcal{M}| - 1)}} \sqrt{\sum_{i=1}^{|\mathcal{M}|} \sum_{\substack{j=1 \\ j \neq i}}^{|\mathcal{M}|} (3\epsilon_{\gamma\beta}^{+2} - 2\mathbf{h}_i \mathbf{h}_j^\top)} \quad (50)$$

$$\leq \sqrt{\frac{\mu^2 |\mathcal{S}|^2}{|\mathcal{M}| (|\mathcal{M}| - 1)}} \sqrt{\sum_{i=1}^{|\mathcal{M}|} \sum_{\substack{j=1 \\ j \neq i}}^{|\mathcal{M}|} (3\epsilon_{\gamma\beta}^{+2} - 2\mathbf{h}_i \mathbf{h}_j^\top)}. \quad (51)$$

990 In summary, in this section, we derive an upper bound for the expected outcome deviation in missing-
991 and full-modality scenarios over the missing scenarios (\mathcal{S}) as:

$$\mathbb{E}_{\mathcal{S}} \left[\|\mathbf{y}^{\mathcal{S}} - \mathbf{y}^{\emptyset}\| \right] \leq \sqrt{\frac{\mu^2 |\mathcal{S}|^2}{|\mathcal{M}| (|\mathcal{M}| - 1)}} \sqrt{\sum_{i=1}^{|\mathcal{M}|} \sum_{\substack{j=1 \\ j \neq i}}^{|\mathcal{M}|} (3\epsilon_{\gamma\beta}^{+2} - 2\mathbf{h}_i \mathbf{h}_j^\top)} \quad (52)$$

992 D.3.2 Generalize over Instances

993 This section describes how we generalize the bound in Eq. 52 to batch- or dataset-level. Furthermore,
994 we reveal the connection between our theoretical bound and the training loss function that we propose,
995 indicating the effectiveness of training loss in our proposal. To address this, we start by considering

996 the mean difference over a dataset \mathcal{D} with cardinality $|\mathcal{D}|$:

$$\frac{1}{|\mathcal{D}|} \sum_{d=1}^{|\mathcal{D}|} \mathbb{E}_S \left[\|\mathbf{y}_{\mathbf{x}_d}^S - \mathbf{y}_{\mathbf{x}_d}^\emptyset\| \right] \leq \sqrt{\frac{\mu^2 |\mathcal{S}|^2}{|\mathcal{M}|(|\mathcal{M}| - 1)}} \frac{1}{|\mathcal{D}|} \sum_d \sqrt{\sum_{i=1}^{|\mathcal{M}|} \sum_{\substack{j=1 \\ j \neq i}}^{|\mathcal{M}|} (3\epsilon_{\gamma\beta}^{+2} - 2\mathbf{h}_{di} \mathbf{h}_{dj}^\top)} \quad (53)$$

$$\leq \frac{\sqrt{|\mathcal{D}|} \mu |\mathcal{S}|}{|\mathcal{D}| \sqrt{|\mathcal{M}|(|\mathcal{M}| - 1)}} \sqrt{\sum_{d=1}^{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{M}|} \sum_{\substack{j=1 \\ j \neq i}}^{|\mathcal{M}|} (3\epsilon_{\gamma\beta}^{+2} - 2\mathbf{h}_{di} \mathbf{h}_{dj}^\top)} \quad (54)$$

997 in which Eq. 53 is transformed to Eq. 54 by using triangle inequality. To avoid confusion, we analyze
 998 the right-hand term separately, as it plays a central role in the transformation process. Let $\tilde{\mathbf{h}}_{di}$ denote
 999 the ℓ_2 -normalized feature, i.e., $\tilde{\mathbf{h}}_{di} = \mathbf{h}_{di} / \|\mathbf{h}_{di}\|$.

$$\sum_{d=1}^{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{M}|} \sum_{\substack{j=1 \\ j \neq i}}^{|\mathcal{M}|} (3\epsilon_{\gamma\beta}^{+2} - 2\mathbf{h}_{di} \mathbf{h}_{dj}^\top) \quad (55)$$

$$\leq \epsilon_{\gamma\beta}^{-2} \sum_d \sum_{i=1}^{|\mathcal{M}|} \sum_{\substack{j=1 \\ j \neq i}}^{|\mathcal{M}|} (3\frac{\epsilon_{\gamma\beta}^{+2}}{\epsilon_{\gamma\beta}^{-2}} - 2\tilde{\mathbf{h}}_{di} \tilde{\mathbf{h}}_{dj}^\top) \quad (56)$$

$$\leq 3|\mathcal{D}| |\mathcal{M}| (|\mathcal{M}| - 1) \epsilon_{\gamma\beta}^{+2} + 2\epsilon_{\gamma\beta}^{-2} \sum_{d=1}^{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{M}|} \sum_{\substack{j=1 \\ j \neq i}}^{|\mathcal{M}|} \left(-\tilde{\mathbf{h}}_{di} \tilde{\mathbf{h}}_{dj}^\top + \log(|\mathcal{D}|(|\mathcal{D}| - 1)|\mathcal{M}|^2) - \frac{\epsilon_{\gamma\beta}^{+2}}{\epsilon_{\gamma\beta}^{-2}} + \frac{\epsilon_{\gamma\beta}^{+2}}{\epsilon_{\gamma\beta}^{-2}} \right) \quad (57)$$

$$\leq 5|\mathcal{D}| |\mathcal{M}| (|\mathcal{M}| - 1) \epsilon_{\gamma\beta}^{+2} + 2\epsilon_{\gamma\beta}^{-2} \sum_{d,i,j \neq i} \left(-\tilde{\mathbf{h}}_{di} \tilde{\mathbf{h}}_{dj}^\top + \log(|\mathcal{D}|(|\mathcal{D}| - 1)|\mathcal{M}|^2) - \frac{\epsilon_{\gamma\beta}^{+2}}{\epsilon_{\gamma\beta}^{-2}} \right) \quad (58)$$

$$\leq 5|\mathcal{D}| |\mathcal{M}| (|\mathcal{M}| - 1) \epsilon_{\gamma\beta}^{+2} + 2\epsilon_{\gamma\beta}^{-2} \sum_{d,i,j \neq i} \left(-\tilde{\mathbf{h}}_{di} \tilde{\mathbf{h}}_{dj}^\top + \log(|\mathcal{D}|(|\mathcal{D}| - 1)|\mathcal{M}|^2 \exp(-(\frac{\epsilon_{\gamma\beta}^{+2}}{\epsilon_{\gamma\beta}^{-2}})^2)) \right) \quad (59)$$

$$\leq 5|\mathcal{D}| |\mathcal{M}| (|\mathcal{M}| - 1) \epsilon_{\gamma\beta}^{+2} + 2\epsilon_{\gamma\beta}^{-2} \sum_{d,i,j \neq i} \left(-\log \exp(\tilde{\mathbf{h}}_{di} \tilde{\mathbf{h}}_{dj}^\top) + \log \left(\sum_{d_1}^{|\mathcal{D}|} \sum_{d_2 \neq d_1}^{|\mathcal{D}|} \sum_{k_1}^{|\mathcal{M}|} \sum_{k_2}^{|\mathcal{M}|} \exp(-(\frac{\epsilon_{\gamma\beta}^{+2}}{\epsilon_{\gamma\beta}^{-2}})^2) \right) \right) \quad (60)$$

$$\leq 5|\mathcal{D}| |\mathcal{M}| (|\mathcal{M}| - 1) \epsilon_{\gamma\beta}^{+2} + 2\epsilon_{\gamma\beta}^{-2} \sum_{d,i,j \neq i} -\log \frac{\exp(\tilde{\mathbf{h}}_{di} \tilde{\mathbf{h}}_{dj}^\top)}{\sum_{d_1}^{|\mathcal{D}|} \sum_{d_2 \neq d_1}^{|\mathcal{D}|} \sum_{k_1}^{|\mathcal{M}|} \sum_{k_2}^{|\mathcal{M}|} \exp(-(\frac{\epsilon_{\gamma\beta}^{+2}}{\epsilon_{\gamma\beta}^{-2}})^2)} \quad (61)$$

$$\leq 5|\mathcal{D}| |\mathcal{M}| (|\mathcal{M}| - 1) \epsilon_{\gamma\beta}^{+2} + 2\epsilon_{\gamma\beta}^{-2} \sum_{d,i,j \neq i} -\log \frac{\exp(\tilde{\mathbf{h}}_{di} \tilde{\mathbf{h}}_{dj}^\top)}{\sum_{d_1}^{|\mathcal{D}|} \sum_{d_2 \neq d_1}^{|\mathcal{D}|} \sum_{k_1}^{\mathcal{M}} \sum_{k_2}^{\mathcal{M}} \exp(\tilde{\mathbf{h}}_{d_1 k_1} \tilde{\mathbf{h}}_{d_2 k_2}^\top)} \quad (62)$$

$$\leq 5|\mathcal{D}| |\mathcal{M}| (|\mathcal{M}| - 1) \epsilon_{\gamma\beta}^{+2} + 2\epsilon_{\gamma\beta}^{-2} |\mathcal{D}| \mathcal{L}_{ds}(\mathbf{x}_d, \emptyset) \quad (63)$$

1000 where $\mathcal{L}_{ds}(\cdot, \cdot)$ is defined in Section 2.2.1). Substitute Eq. 63 into Eq. 54, we have:

$$\frac{1}{|\mathcal{D}|} \sum_{d=1}^{|\mathcal{D}|} \mathbb{E}_S \left[\|\mathbf{y}_{\mathbf{x}_d}^S - \mathbf{y}_{\mathbf{x}_d}^\emptyset\| \right] \leq \mu |\mathcal{S}| \sqrt{5\epsilon_{\gamma\beta}^{+2} + \frac{2\epsilon_{\gamma\beta}^{-2}}{|\mathcal{D}| |\mathcal{M}| (|\mathcal{M}| - 1)} \sum_{d=1}^{|\mathcal{D}|} \mathcal{L}_{ds}(\mathbf{x}_d, \emptyset)} \quad (64)$$

1001 We now investigate how the presence of missing modalities impacts the bound, and consequently, the
 1002 effectiveness of our approach. Assume each instance $\mathbf{x}_d \in \mathcal{D}$ has a missing set \mathcal{S}_d with the same

1003 cardinality $|\mathcal{S}|$, i.e., $\mathcal{S} \subset \mathcal{M}$, $|\mathcal{S}_d| = |\mathcal{S}| \forall d$. Hence,

$$\sum_{d=1}^{|\mathcal{D}|} \mathcal{L}_{ds}(\mathbf{x}_d, \emptyset) = \sum_{d,i,j \neq i} -\log \frac{\exp(\tilde{\mathbf{h}}_{di} \tilde{\mathbf{h}}_{dj}^\top)}{\sum_{d_1}^{|\mathcal{D}|} \sum_{d_2 \neq d_1}^{|\mathcal{D}|} \sum_{d_1}^{|\mathcal{M}|} \sum_{d_2}^{|\mathcal{M}|} \exp(\tilde{\mathbf{h}}_{d_1 k_1} \tilde{\mathbf{h}}_{d_2 k_2}^\top)} \quad (65)$$

$$= \sum_{d,i,j \neq i} \log \exp(-\tilde{\mathbf{h}}_{di} \tilde{\mathbf{h}}_{dj}^\top) + \log \left(\sum_{d_1}^{|\mathcal{D}|} \sum_{d_2 \neq d_1}^{|\mathcal{D}|} \sum_{k_1}^{|\mathcal{M}|} \sum_{k_2}^{|\mathcal{M}|} \exp(\tilde{\mathbf{h}}_{d_1 k_1} \tilde{\mathbf{h}}_{d_2 k_2}^\top) \right) \quad (66)$$

1004 Let $A_1 \triangleq \sum_{d,i,j \neq i} -\tilde{\mathbf{h}}_{di} \tilde{\mathbf{h}}_{dj}^\top$ and $A_2 \triangleq \sum_{d_1}^{|\mathcal{D}|} \sum_{d_2 \neq d_1}^{|\mathcal{D}|} \sum_{k_1}^{|\mathcal{M}|} \sum_{k_2}^{|\mathcal{M}|} \exp(\tilde{\mathbf{h}}_{d_1 k_1} \tilde{\mathbf{h}}_{d_2 k_2}^\top)$, we now
 1005 further expand each term as follows:

1006 Consider A_1 :

$$A_1 = \sum_d \sum_{i,j \neq i} -\tilde{\mathbf{h}}_{di} \tilde{\mathbf{h}}_{dj}^\top \quad (67)$$

$$= \frac{|\mathcal{M}|(|\mathcal{M}| - 1)}{(|\mathcal{M}| - |\mathcal{S}|)|\mathcal{S}|} \frac{(|\mathcal{M}| - |\mathcal{S}|)|\mathcal{S}|}{|\mathcal{M}|(|\mathcal{M}| - 1)} \sum_d \sum_{i,j \neq i} -\tilde{\mathbf{h}}_{di} \tilde{\mathbf{h}}_{dj}^\top \quad (68)$$

$$= \frac{|\mathcal{M}|(|\mathcal{M}| - 1)}{(|\mathcal{M}| - |\mathcal{S}|)|\mathcal{S}|} \frac{(|\mathcal{M}| - |\mathcal{S}|)!|\mathcal{S}|!}{|\mathcal{M}|!} \frac{(|\mathcal{M}| - 2)!}{(|\mathcal{S}| - 1)!(|\mathcal{M}| - |\mathcal{S}| - 1)!} \sum_d \sum_{i,j \neq i} -\tilde{\mathbf{h}}_{di} \tilde{\mathbf{h}}_{dj}^\top \quad (69)$$

$$= \frac{|\mathcal{M}|(|\mathcal{M}| - 1)}{(|\mathcal{M}| - |\mathcal{S}|)|\mathcal{S}|} \frac{1}{\binom{|\mathcal{M}|}{|\mathcal{S}|}} \binom{|\mathcal{M}| - 2}{|\mathcal{S}| - 1} \sum_d \sum_{i,j \neq i} -\tilde{\mathbf{h}}_{di} \tilde{\mathbf{h}}_{dj}^\top \quad (70)$$

$$= \frac{|\mathcal{M}|(|\mathcal{M}| - 1)}{(|\mathcal{M}| - |\mathcal{S}|)|\mathcal{S}|} \mathbb{E}_{\mathcal{S}_d} \left[\sum_d \sum_{i \in \mathcal{S}_d} \sum_{j \notin \mathcal{S}_d} -\tilde{\mathbf{h}}_{di} \tilde{\mathbf{h}}_{dj}^\top \right] \quad (71)$$

1007 Under missing modality scenarios, i.e., $\mathcal{S}_d \neq \emptyset$, $\tilde{\mathbf{h}}_{di}, \forall i \in \mathcal{S}_d$ is approximated as $\frac{1}{|\mathcal{M}| - |\mathcal{S}|} \sum_{j \notin \mathcal{S}_d} \tilde{\mathbf{h}}_{dj}$.

1008 In other words, we can express Eq. 71 as:

$$A_1 = \frac{|\mathcal{M}|(|\mathcal{M}| - 1)}{(|\mathcal{M}| - |\mathcal{S}|)|\mathcal{S}|} \mathbb{E}_{\mathcal{S}_d} \left[\sum_d \sum_{i \in \mathcal{S}_d} \sum_{j \notin \mathcal{S}_d} -\tilde{\mathbf{h}}_{di} \tilde{\mathbf{h}}_{dj}^\top \right] \quad (72)$$

$$= \frac{|\mathcal{M}|(|\mathcal{M}| - 1)}{(|\mathcal{M}| - |\mathcal{S}|)|\mathcal{S}|} \mathbb{E}_{\mathcal{S}_d} \left[\sum_d \sum_{i \in \mathcal{S}_d} \sum_{j \notin \mathcal{S}_d} \frac{1}{|\mathcal{M}| - |\mathcal{S}|} \sum_{k \notin \mathcal{S}_d} -\tilde{\mathbf{h}}_{dk} \tilde{\mathbf{h}}_{dj}^\top \right] \quad (73)$$

$$\leq \frac{|\mathcal{M}|(|\mathcal{M}| - 1)}{(|\mathcal{M}| - |\mathcal{S}|)^2 |\mathcal{S}|} \mathbb{E}_{\mathcal{S}_d} \left[- \sum_d \sum_{i \in \mathcal{S}_d} \sum_{j \notin \mathcal{S}_d} \sum_{k \notin \mathcal{S}_d} \tilde{\mathbf{h}}_{dk} \tilde{\mathbf{h}}_{dj}^\top \right] \quad (74)$$

$$\leq \frac{|\mathcal{M}|(|\mathcal{M}| - 1)}{(|\mathcal{M}| - |\mathcal{S}|)^2} \mathbb{E}_{\mathcal{S}_d} \left[- \sum_d \sum_{j \notin \mathcal{S}_d} \sum_{k \notin \mathcal{S}_d} \tilde{\mathbf{h}}_{dk} \tilde{\mathbf{h}}_{dj}^\top \right] \quad (75)$$

$$(76)$$

1009 Consider A_2 :

$$A_2 = \sum_{d_1}^{|\mathcal{D}|} \sum_{d_2}^{|\mathcal{D}|} \sum_{k_1}^{|\mathcal{M}|} \sum_{k_2}^{|\mathcal{M}|} \exp(\tilde{\mathbf{h}}_{d_1 k_1} \tilde{\mathbf{h}}_{d_2 k_2}^\top) \quad (77)$$

$$= \sum_{d_1, d_2 \neq d_1}^{|\mathcal{D}|} \sum_{k_1}^{|\mathcal{M}|} \left[\sum_{j_2 \notin \mathcal{S}_{d_2}} \exp(\tilde{\mathbf{h}}_{d_1 k_1} \tilde{\mathbf{h}}_{d_2 j_2}^\top) + \sum_{i_2 \in \mathcal{S}_{d_2}} \exp(\tilde{\mathbf{h}}_{d_1 k_1} \tilde{\mathbf{h}}_{d_2 i_2}^\top) \right] \quad (78)$$

$$\begin{aligned}
&= \sum_{d_1, d_2 \neq d_1}^{|\mathcal{D}|} \left\{ \sum_{j_1 \notin \mathcal{S}_{d_1}} \left[\sum_{j_2 \notin \mathcal{S}_{d_2}} \exp(\tilde{\mathbf{h}}_{d_1 j_1} \tilde{\mathbf{h}}_{d_2 j_2}^\top) + \sum_{i_2 \in \mathcal{S}_{d_2}} \exp(\tilde{\mathbf{h}}_{d_1 j_1} \tilde{\mathbf{h}}_{d_2 i_2}^\top) \right] \right. \\
&\quad \left. + \sum_{i_1 \in \mathcal{S}_{d_1}} \left[\sum_{j_2 \notin \mathcal{S}_{d_2}} \exp(\tilde{\mathbf{h}}_{d_1 i_1} \tilde{\mathbf{h}}_{d_2 j_2}^\top) + \sum_{i_2 \in \mathcal{S}_{d_2}} \exp(\tilde{\mathbf{h}}_{d_1 i_1} \tilde{\mathbf{h}}_{d_2 i_2}^\top) \right] \right\} \quad (79)
\end{aligned}$$

$$\begin{aligned}
&= \sum_{d_1, d_2 \neq d_1}^{|\mathcal{D}|} \left\{ \sum_{j_1 \notin \mathcal{S}_{d_1}} \sum_{j_2 \notin \mathcal{S}_{d_2}} \exp(\tilde{\mathbf{h}}_{d_1 j_1} \tilde{\mathbf{h}}_{d_2 j_2}^\top) + \sum_{j_1 \notin \mathcal{S}_{d_1}} \sum_{i_2 \in \mathcal{S}_{d_2}} \exp(\tilde{\mathbf{h}}_{d_1 j_1} \tilde{\mathbf{h}}_{d_2 i_2}^\top) \right. \\
&\quad \left. + \sum_{i_1 \in \mathcal{S}_{d_1}} \sum_{j_2 \notin \mathcal{S}_{d_2}} \exp(\tilde{\mathbf{h}}_{d_1 i_1} \tilde{\mathbf{h}}_{d_2 j_2}^\top) + \sum_{i_1 \in \mathcal{S}_{d_1}} \sum_{i_2 \in \mathcal{S}_{d_2}} \exp(\tilde{\mathbf{h}}_{d_1 i_1} \tilde{\mathbf{h}}_{d_2 i_2}^\top) \right\} \quad (80)
\end{aligned}$$

$$\begin{aligned}
&= \sum_{d_1, d_2 \neq d_1}^{|\mathcal{D}|} \left\{ \sum_{\substack{j_1 \notin \mathcal{S}_{d_1} \\ j_2 \notin \mathcal{S}_{d_2}}} \exp(\tilde{\mathbf{h}}_{d_1 j_1} \tilde{\mathbf{h}}_{d_2 j_2}^\top) + \sum_{\substack{j_1 \notin \mathcal{S}_{d_1} \\ i_2 \in \mathcal{S}_{d_2}}} \exp(\tilde{\mathbf{h}}_{d_1 j_1} \frac{1}{|\mathcal{M}| - |\mathcal{S}|} \sum_{j_2 \notin \mathcal{S}_{d_2}} \tilde{\mathbf{h}}_{d_2 j_2}) \right. \\
&\quad + \sum_{\substack{i_1 \in \mathcal{S}_{d_1} \\ j_2 \notin \mathcal{S}_{d_2}}} \exp(\frac{1}{|\mathcal{M}| - |\mathcal{S}|} \sum_{j_1 \notin \mathcal{S}_{d_1}} \tilde{\mathbf{h}}_{d_1 j_1} \tilde{\mathbf{h}}_{d_2 j_2}^\top) \\
&\quad \left. + \sum_{\substack{i_1 \in \mathcal{S}_{d_1} \\ i_2 \in \mathcal{S}_{d_2}}} \exp(\frac{1}{(|\mathcal{M}| - |\mathcal{S}|)^2} \sum_{j_1 \notin \mathcal{S}_{d_1}} \sum_{j_2 \notin \mathcal{S}_{d_2}} \tilde{\mathbf{h}}_{d_1 j_1} \tilde{\mathbf{h}}_{d_2 j_2}^\top) \right\} \quad (81)
\end{aligned}$$

$$\begin{aligned}
&= \sum_{d_1, d_2 \neq d_1}^{|\mathcal{D}|} \left\{ \sum_{\substack{j_1 \notin \mathcal{S}_{d_1} \\ j_2 \notin \mathcal{S}_{d_2}}} \exp(\tilde{\mathbf{h}}_{d_1 j_1} \tilde{\mathbf{h}}_{d_2 j_2}^\top) + |\mathcal{S}| \sum_{j_1 \notin \mathcal{S}_{d_1}} \exp(\tilde{\mathbf{h}}_{d_1 j_1} \frac{1}{|\mathcal{M}| - |\mathcal{S}|} \sum_{j_2 \notin \mathcal{S}_{d_2}} \tilde{\mathbf{h}}_{d_2 j_2}) \right. \\
&\quad + |\mathcal{S}| \sum_{j_2 \notin \mathcal{S}_{d_2}} \exp(\frac{1}{|\mathcal{M}| - |\mathcal{S}|} \sum_{j_1 \notin \mathcal{S}_{d_1}} \tilde{\mathbf{h}}_{d_1 j_1} \tilde{\mathbf{h}}_{d_2 j_2}^\top) \\
&\quad \left. + |\mathcal{S}|^2 \exp(\frac{1}{(|\mathcal{M}| - |\mathcal{S}|)^2} \sum_{j_1 \notin \mathcal{S}_{d_1}} \sum_{j_2 \notin \mathcal{S}_{d_2}} \tilde{\mathbf{h}}_{d_1 j_1} \tilde{\mathbf{h}}_{d_2 j_2}^\top) \right\} \quad (82)
\end{aligned}$$

1010 **Observation.** Exponential is a convex function, hence we apply Jensen's inequality to the followings:

$$1011 \quad \mathbf{1.} \quad |\mathcal{S}|^2 \exp(\frac{1}{(|\mathcal{M}| - |\mathcal{S}|)^2} \sum_{\substack{j_1 \notin \mathcal{S}_{d_1} \\ j_2 \notin \mathcal{S}_{d_2}}} \tilde{\mathbf{h}}_{d_1 j_1} \tilde{\mathbf{h}}_{d_2 j_2}^\top) \leq \frac{1}{(|\mathcal{M}| - |\mathcal{S}|)^2} \sum_{\substack{j_1 \notin \mathcal{S}_{d_1} \\ j_2 \notin \mathcal{S}_{d_2}}} \exp(\tilde{\mathbf{h}}_{d_1 j_1} \tilde{\mathbf{h}}_{d_2 j_2}^\top)$$

$$1012 \quad \mathbf{2.} \quad |\mathcal{S}| \sum_{j_1 \notin \mathcal{S}_{d_1}} \exp(\tilde{\mathbf{h}}_{d_1 j_1} \frac{1}{|\mathcal{M}| - |\mathcal{S}|} \sum_{j_2 \notin \mathcal{S}_{d_2}} \tilde{\mathbf{h}}_{d_2 j_2}^\top) \leq \frac{1}{|\mathcal{M}| - |\mathcal{S}|} \sum_{\substack{j_1 \notin \mathcal{S}_{d_1} \\ i_2 \notin \mathcal{S}_{d_2}}} \exp(\tilde{\mathbf{h}}_{d_1 j_1} \tilde{\mathbf{h}}_{d_2 j_2}^\top)$$

$$1013 \quad \mathbf{3.} \quad |\mathcal{S}| \sum_{j_2 \notin \mathcal{S}_{d_2}} \exp(\frac{1}{|\mathcal{M}| - |\mathcal{S}|} \sum_{j_1 \notin \mathcal{S}_{d_1}} \tilde{\mathbf{h}}_{d_1 j_1} \tilde{\mathbf{h}}_{d_2 j_2}^\top) \leq \frac{|\mathcal{S}|}{|\mathcal{M}| - |\mathcal{S}|} \sum_{\substack{j_1 \notin \mathcal{S}_{d_1} \\ j_2 \notin \mathcal{S}_{d_2}}} \exp(\tilde{\mathbf{h}}_{d_1 j_1} \tilde{\mathbf{h}}_{d_2 j_2}^\top)$$

1014 which derive Eq. 82 into:

$$A_2 \leq \left[1 + 2 \frac{|\mathcal{S}|}{|\mathcal{M}| - |\mathcal{S}|} + \left(\frac{|\mathcal{S}|}{|\mathcal{M}| - |\mathcal{S}|} \right)^2 \right] \sum_{d_1, d_2 \neq d_1}^{|\mathcal{D}|} \sum_{\substack{j_1 \notin \mathcal{S}_{d_1} \\ j_2 \notin \mathcal{S}_{d_2}}} \exp(\tilde{\mathbf{h}}_{d_1 j_1} \tilde{\mathbf{h}}_{d_2 j_2}^\top) \quad (83)$$

$$\leq \left(\frac{|\mathcal{S}|}{|\mathcal{M}| - |\mathcal{S}|} + 1 \right)^2 \sum_{d_1, d_2 \neq d_1}^{|\mathcal{D}|} \sum_{\substack{j_1 \notin \mathcal{S}_{d_1} \\ j_2 \notin \mathcal{S}_{d_2}}} \exp(\tilde{\mathbf{h}}_{d_1 j_1} \tilde{\mathbf{h}}_{d_2 j_2}^\top) \quad (84)$$

$$\leq \left(\frac{|\mathcal{M}|}{|\mathcal{M}| - |\mathcal{S}|} \right)^2 \sum_{d_1, d_2 \neq d_1}^{|\mathcal{D}|} \sum_{\substack{j_1 \notin \mathcal{S}_{d_1} \\ j_2 \notin \mathcal{S}_{d_2}}} \exp(\tilde{\mathbf{h}}_{d_1 j_1} \tilde{\mathbf{h}}_{d_2 j_2}^\top) \quad (85)$$

1015 Substitute Eq. 71 and 85 into Eq. 66, we have:

$$\sum_{d=1}^{|\mathcal{D}|} \mathcal{L}_{ds}(\mathbf{x}_d, \emptyset) \quad (86)$$

$$\leq \frac{|\mathcal{M}|(|\mathcal{M}| - 1)}{(|\mathcal{M}| - |\mathcal{S}|)^2} \mathbb{E}_{\mathcal{S}_d} \left[- \sum_d^{|\mathcal{D}|} \sum_{j \notin \mathcal{S}_d} \sum_{k \notin \mathcal{S}_d} \tilde{\mathbf{h}}_{dk} \tilde{\mathbf{h}}_{dj} \right] + \sum_{d, i, j \neq i} \log \left[\left(\frac{|\mathcal{M}|}{|\mathcal{M}| - |\mathcal{S}|} \right)^2 \sum_{d_1, d_2 \neq d_1}^{|\mathcal{D}|} \sum_{\substack{j_1 \notin \mathcal{S}_{d_1} \\ j_2 \notin \mathcal{S}_{d_2}}} \exp(\tilde{\mathbf{h}}_{d_1 j_1} \tilde{\mathbf{h}}_{d_2 j_2}^\top) \right] \quad (87)$$

$$\leq \frac{|\mathcal{M}|(|\mathcal{M}| - 1)}{(|\mathcal{M}| - |\mathcal{S}|)^2} \mathbb{E}_{\mathcal{S}_d} \left[- \sum_d^{|\mathcal{D}|} \sum_{j \notin \mathcal{S}_d} \sum_{k \notin \mathcal{S}_d} \tilde{\mathbf{h}}_{dk} \tilde{\mathbf{h}}_{dj} \right] + |\mathcal{M}|(|\mathcal{M}| - 1) \sum_d^{|\mathcal{D}|} \log \left[\left(\frac{|\mathcal{M}|}{|\mathcal{M}| - |\mathcal{S}|} \right)^2 \sum_{d_1, d_2 \neq d_1}^{|\mathcal{D}|} \sum_{\substack{j_1 \notin \mathcal{S}_{d_1} \\ j_2 \notin \mathcal{S}_{d_2}}} \exp(\tilde{\mathbf{h}}_{d_1 j_1} \tilde{\mathbf{h}}_{d_2 j_2}^\top) \right] \quad (88)$$

$$\leq \frac{|\mathcal{M}|(|\mathcal{M}| - 1)}{(|\mathcal{M}| - |\mathcal{S}|)^2} \sum_d^{|\mathcal{D}|} \left\{ \mathbb{E}_{\mathcal{S}_d} \sum_{j \notin \mathcal{S}_d} \sum_{k \notin \mathcal{S}_d} \left[- \log \exp \tilde{\mathbf{h}}_{dk} \tilde{\mathbf{h}}_{dj} + \log \left[\left(\frac{|\mathcal{M}|}{|\mathcal{M}| - |\mathcal{S}|} \right)^2 \sum_{d_1, d_2 \neq d_1}^{|\mathcal{D}|} \sum_{\substack{j_1 \notin \mathcal{S}_{d_1} \\ j_2 \notin \mathcal{S}_{d_2}}} \exp(\tilde{\mathbf{h}}_{d_1 j_1} \tilde{\mathbf{h}}_{d_2 j_2}^\top) \right] \right] \right\} \quad (89)$$

$$\leq \frac{|\mathcal{M}|(|\mathcal{M}| - 1)}{(|\mathcal{M}| - |\mathcal{S}|)^2} \sum_d^{|\mathcal{D}|} \left\{ \mathbb{E}_{\mathcal{S}_d} \left[- \sum_{j \notin \mathcal{S}_d} \sum_{k \notin \mathcal{S}_d} \log \frac{\exp \tilde{\mathbf{h}}_{dk} \tilde{\mathbf{h}}_{dj}}{\sum_{d_1, d_2 \neq d_1}^{|\mathcal{D}|} \sum_{\substack{j_1 \notin \mathcal{S}_{d_1} \\ j_2 \notin \mathcal{S}_{d_2}}} \exp(\tilde{\mathbf{h}}_{d_1 j_1} \tilde{\mathbf{h}}_{d_2 j_2}^\top)} \right] + \sum_{j \notin \mathcal{S}_d} \sum_{k \notin \mathcal{S}_d} \log \left(\frac{|\mathcal{M}|}{|\mathcal{M}| - |\mathcal{S}|} \right)^2 \right] \right\} \quad (90)$$

$$\leq \frac{|\mathcal{M}|(|\mathcal{M}| - 1)}{(|\mathcal{M}| - |\mathcal{S}|)^2} \sum_d^{|\mathcal{D}|} \left\{ \mathbb{E}_{\mathcal{S}_d} \left[\mathcal{L}_{ds}(\mathbf{x}_d, \mathcal{S}_d) \right] + (|\mathcal{M}| - |\mathcal{S}|)^2 \log \left(\frac{|\mathcal{M}|}{|\mathcal{M}| - |\mathcal{S}|} \right)^2 \right\} \quad (91)$$

1016 Substitute Eq. 91 in Eq. 97, we have:

$$\frac{1}{|\mathcal{D}|} \sum_{d=1}^{|\mathcal{D}|} \mathbb{E}_{\mathcal{S}} \left[\|\mathbf{y}_{\mathbf{x}_d}^{\mathcal{S}} - \mathbf{y}_{\mathbf{x}_d}^{\emptyset}\| \right] \quad (92)$$

$$\leq \mu |\mathcal{S}| \sqrt{5\epsilon_{\gamma\beta}^2 + \frac{2\epsilon_{\gamma\beta}^{-2}}{|\mathcal{D}| |\mathcal{M}| (|\mathcal{M}| - 1)} \sum_{d=1}^{|\mathcal{D}|} \mathcal{L}_{ds}(\mathbf{x}_d, \emptyset)} \quad (93)$$

$$\leq \mu |\mathcal{S}| \sqrt{5\epsilon_{\gamma\beta}^2 + \frac{2\epsilon_{\gamma\beta}^{-2}}{(|\mathcal{M}| - |\mathcal{S}|)^2} \left\{ \frac{1}{|\mathcal{D}|} \sum_d^{|\mathcal{D}|} \mathbb{E}_{\mathcal{S}_d} \left[\mathcal{L}_{ds}(\mathbf{x}_d, \mathcal{S}_d) \right] \right\} + 2\epsilon_{\gamma\beta}^{-2} \log \frac{|\mathcal{S}|^2}{(|\mathcal{M}| - |\mathcal{S}|)^2}} \quad (94)$$

1017 which is equivalent to:

$$\mathbb{E}_{\mathbf{x}, \mathcal{S}} \left[\|\mathbf{y}_{\mathbf{x}}^{\mathcal{S}} - \mathbf{y}_{\mathbf{x}}^{\emptyset}\| \right] \quad (95)$$

$$\leq \mu|\mathcal{S}|\sqrt{5\epsilon_{\gamma\beta}^2 + \frac{2\epsilon_{\gamma\beta}^{-2}}{(|\mathcal{M}| - |\mathcal{S}|)^2} \mathbb{E}_{\mathbf{x}, \mathcal{S}} \left[\mathcal{L}_{ds}(\mathbf{x}, \mathcal{S}) \right]} + 2\epsilon_{\gamma\beta}^{-2} \log \frac{|\mathcal{M}|^2}{(|\mathcal{M}| - |\mathcal{S}|)^2} \quad (96)$$

$$\leq \mathcal{O} \left(\mu|\mathcal{S}|\sqrt{\frac{\mathbb{E}_{\mathbf{x}, \mathcal{S}}[\mathcal{L}_{ds}(\mathbf{x}, \mathcal{S})]}{(|\mathcal{M}| - |\mathcal{S}|)^2} + \log \frac{|\mathcal{M}|^2}{(|\mathcal{M}| - |\mathcal{S}|)^2}} \right) \quad (97)$$

1018 E Additional Experimental Results

1019 E.1 Additional Comparison with Baselines

1020 **Extensive Main Results.** We evaluate the performance of PEPSY in comparison to baseline methods
 1021 through an extensive set of experiments conducted on both PTBXL and EDF datasets. The results,
 1022 reported in Tab. 6, demonstrates that PEPSY consistently outperform baselines in all tested scenarios
 1023 with different missing modality settings. Notably, a similar observation that PEPSY remains stable
 1024 when the missing degree increases, e.g., $p_m = 0.2$ to $p_m = 0.8$, while the other drop significantly,
 1025 can be seen from the reported results. The second-best methods vary across testing scenarios but
 1026 are most frequently FedMAC. Interestingly, the standard deviation of all methods when dealing with
 1027 data heterogeneity (in NonIID setting) are higher than those of IID setting, indicating the impact
 1028 of data heterogeneity on federated performance. To summarize, the extended results highlights the
 1029 superiority of our proposal against baseline approaches in varying missing scenarios, along with its
 1030 stability when missing degree raises.

1031 **Extensive Missing Scenarios Analysis.** In addition to the results in the main text, we conducted
 1032 further experiments comparing the performance of PEPSY (our method) with baselines under more
 1033 varied missing modality scenarios. Specifically, we expanded the values of p_m and p_s to include 0.4,
 1034 0.6, and 1.0, covering a range from 0.2 to 1.0. The results are shown in Tab. 7 and Tab. 8.

1035 As can be seen in these tables, PEPSY consistently outperforms all baselines across all testing
 1036 scenarios. For the PTBXL dataset (see Tab. 7), the performance gap is small (3% - 4%) when the
 1037 missing degree is low, e.g., $p_m = 0.2$. However, as the missing degree increases (e.g., $p_m = 0.8$
 1038 and $p_m = 1.0$), PEPSY maintains a clear advantage over other methods in both IID and NonIID
 1039 settings, with a significant gap of approximate 11% in accuracy. Similarly, for the EDF dataset,
 1040 PEPSY outperforms baselines by a significant margin - up to nearly 10% - across additional missing
 1041 modality scenarios. This demonstrates the effectiveness and robustness of our approach to missing
 1042 modalities in federated learning systems, regardless of data heterogeneity.

1043 **Modality Alignment Analysis.** Fig. 6 compares modality alignment of our proposed PEPSY and
 1044 two other baselines, namely FedProx and FedMAC, which correspond to traditional FL method and
 1045 second-best approach in most evaluation experiments. Intuitively, to achieve high performance
 1046 regardless of available modalities, an optimal solution should align modalities well in a representation
 1047 space, which hence discards reliance on present modalities. As can be seen from Fig. 6, FedProx and
 1048 FedMAC fail to align different modalities, indicating their strong dependence on different available
 1049 modality sets. This is because FedProx does not have a mechanism for modality alignment, while
 1050 FedMAC discards modality-specific information. In contrast, our proposed PEPSY integrates both
 1051 modality- and data-specific information, which are further reconfigured by a shareable data-missing
 1052 profile leading to less reliance on modalities. The figures show how all modalities are aligned after
 1053 PEPSY's training, highlighting effectiveness of the proposal under missing modality scenarios.

1054 E.2 Additional Ablation Studies

1055 In this section, we conduct additional ablation studies on two crucial components in our design: data-
 1056 missing profile, along with the relevance loss term, and modality fusion, along with the reconfiguration
 1057 regularization. Correspondingly, we introduce two variants of PEPSY, namely PEPSY-NP (No Profile)
 1058 and PEPSY-NR (No Reconfiguration). To evaluate their contributions in our proposal, we analyse
 1059 both quantitative and qualitative results.

1060 **Quantitative Results.** Tab. 5 shows impacts of different components on the final components.
 1061 First, when we remove data-missing profile (see PEPSY-NP variant), the performance drops from
 1062 0.2% to 4%, indicating the importance data-missing profile to stabilize output performance. In this

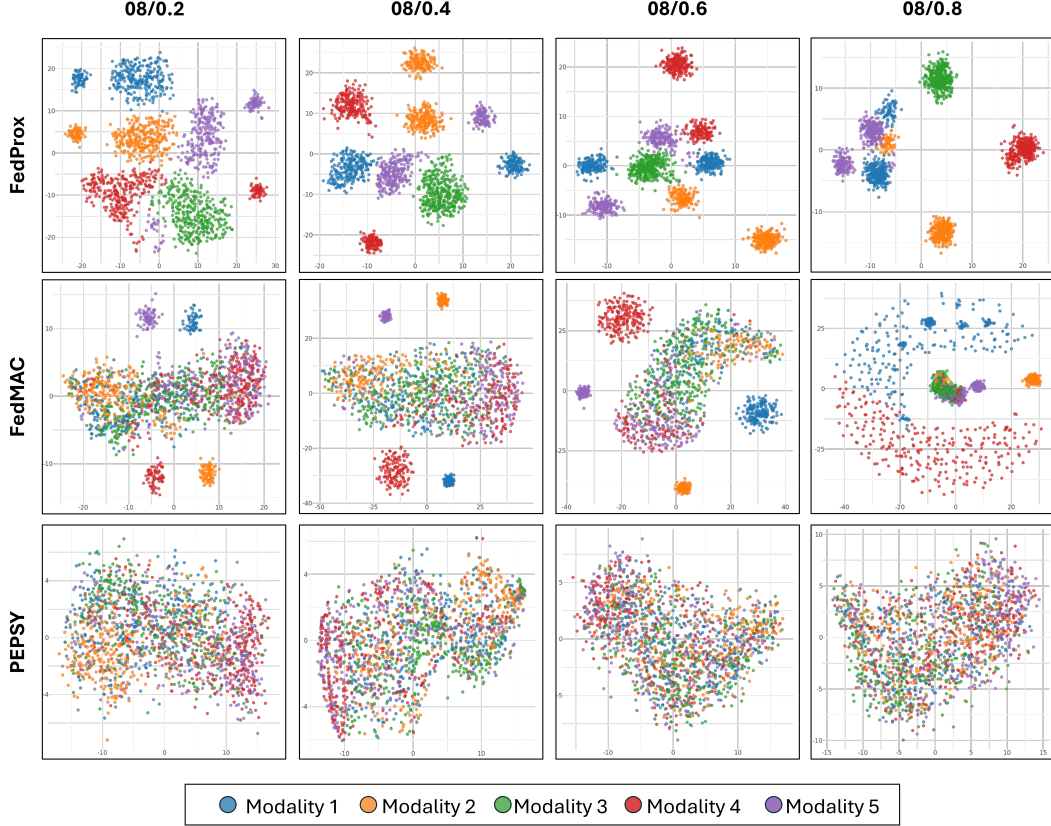


Figure 6: Modality representations of different methods under multiple missing scenarios. We train and provide t-SNE 2D visualizations of modality representations constructed by three methods, including our proposal, in different p_m/p_s settings. All experiments are conducted on EDF dataset, nonIID setting.

Table 5: Ablation studies on crucial components of PEPSY under different missing statistics (p_m/p_s). We report top-1 accuracy across multiple experiments on the EDF dataset, in NonIID setting.

Method	0.8/0.2	0.8/0.4	0.8/0.6	0.8/0.8	0.8/1.0
PEPSY-NP	46.49%	47.92%	52.42%	52.08%	43.98%
PEPSY-NR	43.30%	43.47%	43.47%	43.58%	19.97%
PEPSY	51.80%	51.06%	55.05%	52.25%	46.09%

variant, the reconfiguration supervision signal, a contrastive alignment - based loss, is preserved, hence ensuring modalities are aligned, which are eventually similar to modality fusion in previous works [46, 34]. On the other hand, omitting reconfiguration signal and modality fusion, which results in PEPSY-NR variant, worsen final performance by a larger margin, up to more than 26%. This is because without the reconfiguration signal, the data-missing profile lacks guidance to reconfigure the biased information generated from raw data into complete ones, hence failing to handle missing modalities efficiently. In summary, both components are crucial in our design to ensure robust and stable performance in multimodal federated learning.

Qualitative Results. We further visualize representations that each PEPSY variant constructs for an individual modality under different missing scenarios, given the same trained backbone. In particular, each variant is trained on a specific missing statistic $p_m/p_s = 0.8/0.8$ in NonIID setting and tested on handcrafted missing tests, including: Miss 1 (modality 1 is missed); Miss 1, 4; Missing 1, 4, 5; Miss 1, 3, 4, 5; Full modality. Intuitively, a representation constructed for modality 1 should remain closely aligned across all tests. As can be seen in Fig. 7, while two ablated variants PEPSY-NP and PEPSY-NR fail to ensure this stability, our proposed PEPSY can construct closely aligned representations in all settings, highlighting its stable feature construction. This is because our data-missing profile

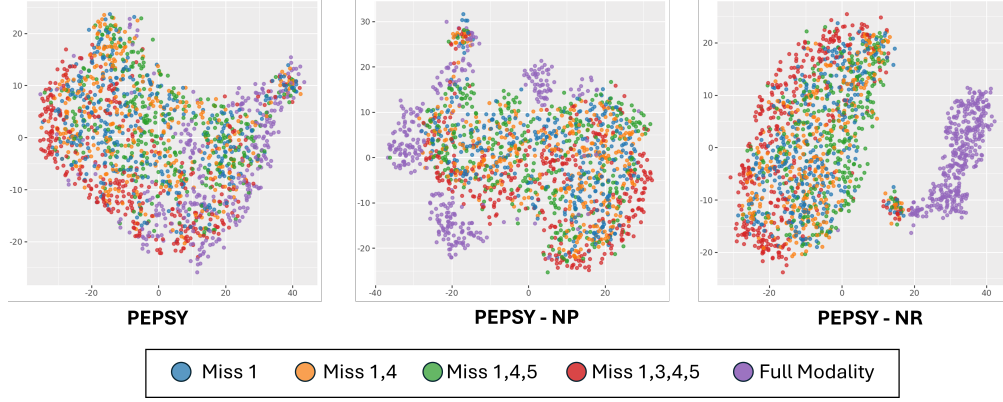


Figure 7: Stability of modality representations under different missing modality scenarios. Ideally, a modality’s representation should remain stable regardless of which other modalities are missing. This stability is not achieved when either the data-missing profile is removed (-NP version) or the reconfiguration signal is omitted (-NR version) from our proposed PEPsy.

Table 6: Performance of baselines on the PTBXL and EDF datasets under various missing patterns in train and test sets, for both IID and Non-IID scenarios. The best and second-best results are highlighted in **bold red** and **blue**, respectively. Results are based on more runs than those reported in Tab. 1.

Dataset	$p_{m p_s}$	Method	IID					Non-IID				
			0.2	0.4	0.6	0.8	1.0	0.2	0.4	0.6	0.8	1.0
PTBXL	0.2	FedProx [27]	73.43 ± 0.38	73.64 ± 1.01	71.42 ± 1.18	71.37 ± 2.50	69.93 ± 4.61	54.01 ± 3.66	51.15 ± 5.30	50.06 ± 12.22	54.89 ± 1.54	44.17 ± 1.31
		MITL [39]	73.52 ± 1.45	70.95 ± 1.90	71.41 ± 1.46	56.66 ± 22.68	69.99 ± 3.05	50.99 ± 2.38	47.16 ± 3.16	49.39 ± 1.75	51.37 ± 2.55	50.78 ± 4.76
		FedInM [56]	69.78 ± 5.16	69.27 ± 3.21	66.16 ± 3.01	65.49 ± 2.25	65.45 ± 2.70	34.17 ± 6.82	40.48 ± 10.87	41.23 ± 11.34	40.52 ± 11.20	40.31 ± 10.70
		FedMSplit [7]	54.84 ± 22.31	53.63 ± 21.72	52.12 ± 21.55	52.50 ± 21.52	55.84 ± 13.22	42.75 ± 3.56	42.58 ± 6.07	41.62 ± 6.06	40.27 ± 3.09	39.39 ± 1.66
		FedMAC [34]	78.56 ± 0.47	77.30 ± 0.81	76.25 ± 0.49	75.49 ± 1.07	74.70 ± 0.83	58.26 ± 4.81	58.55 ± 3.02	54.98 ± 7.74	50.94 ± 1.25	48.38 ± 0.59
	0.8	PEPSY	78.81 ± 0.72	77.43 ± 0.88	76.75 ± 1.47	76.13 ± 0.25	75.41 ± 0.82	71.45 ± 0.39	69.70 ± 2.08	66.92 ± 2.83	68.26 ± 2.56	66.75 ± 5.32
		FedProx [27]	72.76 ± 0.57	70.24 ± 1.61	68.77 ± 2.30	65.24 ± 4.94	33.79 ± 3.39	48.43 ± 1.25	42.08 ± 0.53	34.17 ± 3.14	27.32 ± 1.67	29.97 ± 1.31
		MITL [39]	69.90 ± 1.14	65.36 ± 2.12	55.44 ± 6.44	50.61 ± 14.99	35.39 ± 6.90	44.26 ± 3.87	37.75 ± 12.67	32.67 ± 8.82	28.12 ± 6.03	29.67 ± 2.54
		FedInM [56]	63.10 ± 2.77	61.92 ± 1.53	60.36 ± 0.16	56.95 ± 2.13	35.31 ± 13.56	49.81 ± 17.45	46.41 ± 14.99	42.95 ± 12.72	42.37 ± 12.21	36.70 ± 14.23
		FedMSplit [7]	54.77 ± 20.66	49.56 ± 18.20	45.82 ± 16.29	43.97 ± 15.87	23.91 ± 2.18	51.03 ± 2.09	44.51 ± 0.77	38.25 ± 4.49	29.91 ± 6.11	28.33 ± 2.26
EDF	0.2	FedMAC [34]	74.25 ± 0.48	73.06 ± 0.65	70.36 ± 0.75	67.17 ± 2.98	41.51 ± 6.64	53.05 ± 0.41	51.03 ± 3.19	36.95 ± 0.18	45.90 ± 4.45	43.29 ± 1.54
		PEPSY	76.25 ± 0.77	75.96 ± 1.82	76.42 ± 0.98	75.08 ± 1.65	45.07 ± 0.26	63.01 ± 3.95	65.40 ± 1.01	69.19 ± 0.16	60.40 ± 7.11	53.07 ± 2.66
		FedProx [27]	44.08 ± 0.59	43.54 ± 0.62	43.99 ± 0.57	35.65 ± 12.22	34.02 ± 14.46	34.58 ± 13.80	44.61 ± 0.63	44.02 ± 0.30	32.25 ± 11.94	44.27 ± 0.34
		MITL [39]	44.19 ± 0.73	44.27 ± 0.96	43.15 ± 0.83	43.32 ± 2.19	43.54 ± 0.27	43.17 ± 1.76	43.35 ± 2.26	44.05 ± 0.35	32.74 ± 15.73	44.42 ± 0.33
		FedInM [56]	40.39 ± 0.14	40.39 ± 0.09	40.24 ± 0.11	40.33 ± 0.12	40.37 ± 0.21	40.99 ± 0.98	40.73 ± 0.57	40.46 ± 0.24	40.87 ± 0.94	40.43 ± 0.26
	0.8	FedMSplit [7]	41.91 ± 2.31	36.47 ± 11.44	43.09 ± 2.20	43.77 ± 1.47	41.42 ± 2.80	42.95 ± 1.37	33.98 ± 14.43	42.88 ± 1.15	26.08 ± 13.54	43.43 ± 1.11
		FedMAC [34]	39.00 ± 12.45	40.43 ± 10.29	41.85 ± 7.58	43.58 ± 5.47	43.01 ± 1.39	38.60 ± 12.32	39.44 ± 9.62	41.04 ± 6.87	43.13 ± 4.66	43.96 ± 1.80
		PEPSY	48.76 ± 5.41	49.37 ± 4.43	48.70 ± 4.03	49.27 ± 3.30	46.87 ± 2.46	54.84 ± 3.32	50.28 ± 4.11	54.50 ± 0.14	51.07 ± 5.24	53.35 ± 6.13
		FedProx [27]	41.49 ± 3.69	31.15 ± 11.57	33.73 ± 4.92	19.72 ± 6.91	33.53 ± 14.10	43.87 ± 0.44	24.34 ± 14.02	34.56 ± 13.11	34.56 ± 12.99	34.17 ± 11.53
		MITL [39]	44.51 ± 0.45	42.25 ± 1.67	42.99 ± 0.91	41.07 ± 0.61	42.40 ± 1.65	43.42 ± 1.41	43.83 ± 0.90	43.01 ± 0.99	42.99 ± 1.00	42.40 ± 0.70
		FedInM [56]	40.31 ± 0.13	40.29 ± 0.11	40.26 ± 0.14	40.25 ± 0.02	40.22 ± 0.01	40.84 ± 0.77	40.81 ± 0.79	40.50 ± 0.37	40.31 ± 0.14	40.36 ± 0.22
		FedMSplit [7]	41.44 ± 3.16	32.99 ± 13.22	42.21 ± 1.42	36.64 ± 6.21	43.02 ± 0.47	35.71 ± 10.75	42.75 ± 1.64	33.54 ± 13.70	41.87 ± 1.94	43.38 ± 0.50
		FedMAC [34]	43.77 ± 1.52	42.54 ± 2.39	41.51 ± 0.73	41.80 ± 2.14	26.33 ± 1.47	46.01 ± 0.98	45.73 ± 0.99	45.66 ± 0.49	46.22 ± 0.84	34.21 ± 8.87
		PEPSY	54.02 ± 1.41	49.02 ± 0.38	49.23 ± 1.47	52.78 ± 4.49	46.91 ± 3.70	48.95 ± 2.14	51.52 ± 0.60	50.97 ± 0.44	50.96 ± 1.99	46.07 ± 0.82

effectively distills data-missing information from raw data, which are used later for reconfiguration. These visualizations further emphasize completeness of our design.

F Limitations

Although PEPsy outperforms prior methods in handling heterogeneous data-missing patterns in multimodal federated learning, it may face challenges when downstream task domains vary significantly. Large domain shifts can create distinct, domain-specific missing data profiles that require more trainable embeddings for effective adaptation. A key open question is whether we can quantify these shifts and bound the number of embeddings needed for reconfiguration—an issue beyond this work’s scope but important for future research, especially in federated settings with clients operating in diverse domains and missing data patterns. Moreover, this study relies on training models from scratch and does not leverage pretrained foundation models. Future efforts could explore incorporating pretrained encoders to build shareable missing data profiles, improving representation learning efficiency and effectiveness.

Table 7: Performance of baselines on the PTBXL dataset under various missing patterns in train and test sets, for both IID and Non-IID scenarios. The best and second-best results are highlighted in **bold red** and **blue**, respectively. We use a hyphen (–) to denote $p_m/p_s = 1.0/1.0$, indicating that all modalities are missing and these cases are excluded from evaluation.

pm\ps	Method	IID					NonIID				
		0.2	0.4	0.6	0.8	1.0	0.2	0.4	0.6	0.8	1.0
0.4	FedProx	71.63%	63.81%	65.57%	64.69%	45.76%	47.79%	45.27%	39.97%	33.67%	37.58%
	MIFL	71.37%	65.95%	66.46%	45.02%	53.85%	52.59%	39.22%	37.33%	38.08%	37.20%
	FedInMM	69.61%	68.35%	64.69%	63.43%	64.19%	63.43%	66.33%	62.29%	61.66%	59.52%
	FedMSplit	70.62%	62.93%	60.28%	60.66%	38.97%	53.97%	48.17%	43.17%	46.27%	34.30%
	FedMAC	75.79%	74.02%	73.52%	73.64%	67.84%	69.48%	52.21%	45.65%	43.76%	47.41%
	PEPSY	78.44%	77.55%	76.04%	76.29%	71.37%	71.12%	71.12%	68.10%	70.87%	70.62%
0.6	FedProx	72.38%	69.74%	65.07%	63.18%	47.41%	44.01%	38.08%	37.45%	28.75%	29.00%
	MIFL	70.99%	67.59%	55.61%	49.81%	25.47%	56.75%	43.76%	43.00%	35.69%	25.60%
	FedInMM	67.21%	61.79%	59.14%	58.26%	25.60%	62.42%	59.14%	49.56%	56.36%	49.43%
	FedMSplit	69.10%	63.81%	51.45%	40.48%	37.07%	40.73%	47.29%	38.71%	35.43%	26.48%
	FedMAC	75.28%	74.02%	73.52%	73.64%	56.75%	51.45%	50.44%	50.06%	27.87%	46.15%
	PEPSY	76.55%	74.53%	74.15%	74.15%	57.63%	70.87%	69.23%	68.47%	68.98%	58.76%
1.0	FedProx	75.03%	72.63%	68.73%	58.51%	–	61.03%	51.57%	42.62%	33.29%	–
	MIFL	73.52%	71.37%	66.09%	47.54%	–	59.64%	50.44%	39.60%	33.67%	–
	FedInMM	62.80%	62.42%	53.97%	49.68%	–	59.02%	54.85%	50.06%	41.86%	–
	FedMSplit	72.13%	68.10%	66.46%	54.48%	–	57.25%	52.08%	45.02%	33.92%	–
	FedMAC	75.16%	74.40%	72.38%	69.74%	–	59.52%	44.51%	51.32%	41.74%	–
	PEPSY	76.04%	77.05%	75.03%	72.76%	–	71.25%	67.21%	68.60%	59.14%	–

Table 8: Performance of baselines on the EDF dataset under various missing patterns in train and test sets, for both IID and Non-IID scenarios. The best and second-best results are highlighted in **bold red** and **blue**, respectively. We use a hyphen (–) to denote $p_m/p_s = 1.0/1.0$, indicating that all modalities are missing and these cases are excluded from evaluation.

pm\ps	Method	IID					NonIID				
		0.2	0.4	0.6	0.8	1.0	0.2	0.4	0.6	0.8	1.0
0.4	FedProx	44.38%	44.25%	43.70%	44.95%	43.07%	45.00%	44.55%	44.61%	44.55%	44.72%
	MIFL	43.35%	44.72%	43.72%	44.89%	44.66%	44.61%	44.67%	44.72%	44.49%	40.27%
	FedInMM	40.50%	40.50%	40.67%	40.56%	40.90%	40.62%	42.38%	40.50%	40.45%	41.19%
	FedMSplit	44.95%	45.10%	44.61%	44.61%	44.67%	44.43%	44.38%	44.61%	44.10%	44.21%
	FedMAC	50.49%	48.26%	48.09%	50.03%	41.93%	49.80%	46.49%	46.66%	44.72%	46.83%
	PEPSY	55.68%	55.33%	54.54%	55.45%	49.91%	58.02%	52.54%	49.80%	48.32%	51.97%
0.6	FedProx	34.91%	34.23%	33.14%	29.89%	42.61%	41.24%	42.50%	42.56%	43.18%	40.45%
	MIFL	44.32%	42.84%	43.98%	44.78%	44.61%	45.18%	44.38%	44.38%	44.10%	44.27%
	FedInMM	40.67%	40.44%	40.56%	40.62%	40.45%	41.47%	41.7%	40.73%	40.62%	40.67%
	FedMSplit	44.38%	44.55%	44.61%	44.44%	43.47%	44.15%	44.55%	44.15%	42.27%	43.53%
	FedMAC	50.99%	49.40%	48.66%	48.20%	16.71%	47.80%	47.46%	45.58%	43.64%	38.62%
	PEPSY	51.28%	50.54%	50.26%	50.60%	44.66%	48.66%	51.12%	49.67%	51.85%	45.07%
1.0	FedProx	36.22%	35.14%	33.89%	31.72%	–	44.38%	44.67%	44.44%	43.75%	–
	MIFL	42.56%	42.90%	41.19%	41.47%	–	44.15%	43.75%	44.27%	44.21%	–
	FedInMM	40.45%	40.56%	40.50%	40.22%	–	40.56%	40.39%	40.38%	40.27%	–
	FedMSplit	43.47%	43.47%	42.56%	41.42%	–	42.44%	43.98%	43.70%	44.89%	–
	FedMAC	40.22%	40.45%	40.96%	38.11%	–	47.22%	46.83%	46.44%	46.15%	–
	PEPSY	54.93%	52.48%	48.49%	45.41%	–	50.09%	48.26%	49.67%	49.96%	–