Disentangled Cross-Modal Representation Learning with Enhanced Mutual Supervision

Lu Gao*
School of Computer Science and Engineering
Central South University
Changsha, China
244711035@csu.edu.cn

Wenlan Chen*
School of Computer Science and Engineering
Central South University
Changsha, China
244701041@csu.edu.cn

Daoyuan Wang
School of Computer Science and Engineering
Central South University
Changsha, China
244701040@csu.edu.cn

Fei Guo†
School of Computer Science and Engineering
Central South University
Changsha, China
guofei@csu.edu.cn

Cheng Liang†
School of Information Science and Engineering
Shandong Normal University
Jinan, China
alcs417@sdnu.edu.cn

Abstract

Cross-modal representation learning aims to extract semantically aligned representations from heterogeneous modalities such as images and text. Existing multimodal VAE-based models often suffer from limited capability to align heterogeneous modalities or lack sufficient structural constraints to clearly separate the modality-specific and shared factors. In this work, we propose a novel framework, termed **D**isentangled **C**ross-**M**odal Representation Learning with **E**nhanced **M**utual Supervision (DCMEM). Specifically, our model disentangles the common and distinct information across modalities and regularizes the shared representation learned from each modality in a mutually supervised manner. Moreover, we incorporate the information bottleneck principle into our model to ensure that the shared and modality-specific factors encode exclusive yet complementary information. Notably, our model is designed to be trainable on both complete and partial multimodal datasets with a valid Evidence Lower Bound. Extensive experimental results demonstrate significant improvements of our model over existing methods on various tasks including cross-modal generation, clustering and classification.

1 Introduction

Cross-modal representation learning aims to bridge the semantic gap between heterogeneous data from different modalities, such as images, text, audio and video [1, 2, 3]. The key challenge of the task lies in capturing both the modality-specific features and the shared semantic structure across modalities, despite their inherent differences in format and statistical properties [4]. Disentangled representation

*Both authors contributed equally to this work and are joint first authors.
†Corresponding authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

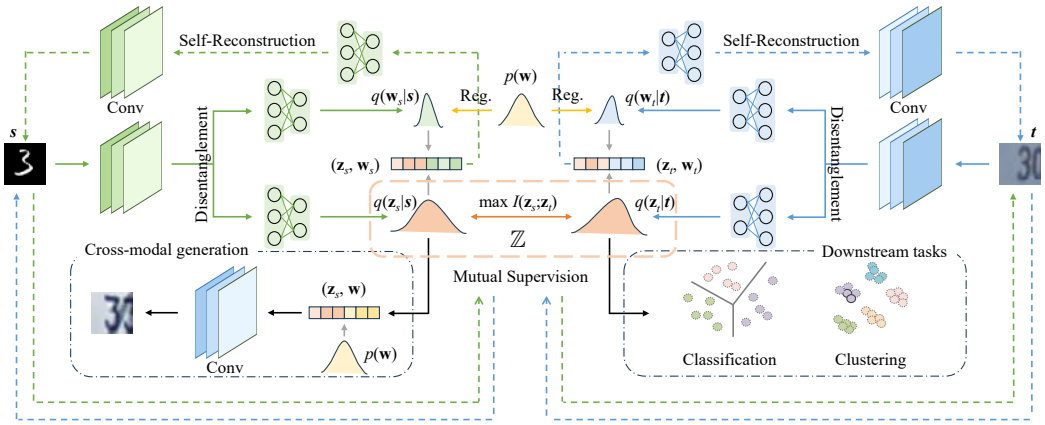


Figure 1: The architecture of DCMEM. For each modality, the encoder extracts shared $(\boldsymbol{z}_s, \boldsymbol{z}_t)$ and modality-specific $(\boldsymbol{w}_s, \boldsymbol{w}_t)$ latent variables. Self-reconstruction and KL regularization (denoted as "Reg.") ensure disentanglement, where $\boldsymbol{w}$ follows a standard Gaussian prior and shared variables serve as priors for each other. Shared representations lie in a common latent space $\mathbb{Z}$, enabling mutual supervision via cross-modal paths $\boldsymbol{s} \rightarrow \mathbb{Z} \rightarrow \boldsymbol{t}$ and $\boldsymbol{t} \rightarrow \mathbb{Z} \rightarrow \boldsymbol{s}$. We further align $\boldsymbol{z}_s$ and $\boldsymbol{z}_t$ by maximizing their mutual information. DCMEM supports cross-modal generation, classification and clustering.

learning has thus emerged as a promising approach to address this challenge by explicitly separating modality-invariant (shared) factors from modality-specific (private) ones, enabling more interpretable and robust cross-modal representations [5]. Variational Autoencoders (VAEs) have been a prevailing framework for disentanglement with their probabilistic formulation and ability to learn structured latent representations. In particular, multimodal VAEs have gained significant attention for their ability to jointly encode and decode information from multiple modalities in a unified probabilistic space [6]. From a methodological perspective, multimodal VAEs can be broadly divided into two categories: models that focus on learning shared latent variables and those that incorporate modality-specific private variables. The former (e.g., MVAE [7], MMVAE [8], MoPoE [9], MEME [10]) aim to capture the common semantic structure across modalities, while the latter (e.g., MMVAE+ [11], CMVAE [12], IIAE [13], Multi-VAE [14]) explicitly disentangle shared and modality-specific information by introducing separate latent spaces. However, these methods either suffer from limited capability to align heterogeneous modalities or lack sufficient structural constraints between modalities, which hinders the clear separation of informative shared and modality-specific factors. In this work, we propose a novel multimodal framework, termed **D**isentangled **C**ross-**M**odal representation learning with **E**nhanced **M**utual supervision (DCMEM), to address these challenges. As illustrated in Figure 1, our model disentangles the common and distinct information across modalities by extracting shared and modality-specific latent representations using multiple VAEs. Specifically, we regularize the shared representation learned from one modality using that from the other in a mutually supervised manner. Moreover, we incorporate the information bottleneck principle into our model to ensure that the shared and modality-specific factors encode exclusive yet complementary information for reconstructing each modality. Finally, we apply an alignment constraint in the shared latent space to promote consistent semantic and inter-modality coherence. Notably, our model is designed to be trainable simultaneously on both complete and partial multimodal datasets with a valid Evidence Lower Bound (ELBO). The main contributions of this paper are summarized as follows:

- We propose DCMEM, a novel multimodal framework that disentangles effectively shared and modality-specific factors. Our model leverages enhanced mutual supervision to improve cross-modal alignment and semantic consistency, while simultaneously enforcing structured representation learning through the information bottleneck principle.
- Our model is inherently designed to be trainable on both complete and partially missing modalities simultaneously. This improves the robustness of our model and broadens its applicability to real-world scenarios where incomplete modalities are prevalent.
- Extensive experiments on three diverse datasets demonstrate that our method produces more coherent and informative embeddings compared to existing multimodal VAE-based approaches

in terms of various evaluation metrics, including generation coherence, clustering accuracy and classification accuracy.

2 Related Work

2.1 Multimodal VAEs

Multimodal generation tasks have gained significant attention, driving research on various generative models. Among them, multimodal VAEs have become particularly popular due to their impressive performance. Early multimodal VAEs [15, 16, 17] typically employ joint encoders over concatenated inputs, often requiring auxiliary components or processes to support cross-modal generation. Wu *et al.* [7] propose a scalable method based on the Product of Experts (PoE) to address this problem. However, subsequent research [8, 9, 18, 19] identifies several issues with PoE, such as calibration errors. To address these limitations, Shi *et al.* propose MMVAE [8], using Mixture of Experts (MoE) to model the joint posterior as a mixture of unimodal posteriors. The MoPoE model [9] combines PoE and MoE to balance semantic coherence with effective joint distribution learning. Several studies [11, 20] highlight a trade-off between generation quality and coherence in multimodal VAEs. PoE-based models typically achieve lower coherence, while MoE-based models suffer from lower generation quality [11, 21]. To address these issues, various models incorporate different regularizers to improve performance. For instance, MVTCAE [22] uses mutual information theory for regularization, mmJSD [23] employs a dynamic prior to combine modality information, MEME [10] enhances performance through mutual supervision and MMVM [24] regularizes learned posterior approximations with a data-dependent prior. Moreover, Palumbo *et al.* introduce MMVAE+ [11], which incorporates modality-specific subspaces to improve cross-modal likelihood estimation [22, 25]. Building on MMVAE+, CMVAE [12] introduces clustering variables and a mixture distribution for model categories [26, 27], extending multimodal generation to include clustering capabilities. Recently, Gao *et al.* introduce MVP [28] which proposes an informational prior based on cyclic permutations, enabling both generative and clustering tasks.

2.2 Information Bottleneck

Information bottleneck plays a crucial role in multimodal clustering and representation learning, enabling models to extract robust and interpretable latent representations by capturing shared and complementary information across different modalities. Wang *et al.* [29] propose a supervised method that maximizes mutual information between joint representations and labels while filtering out irrelevant data from original views. Federici *et al.* introduce MIB [30], which applies the information bottleneck to identify shared and view-specific information between views. Hwang *et al.* [13] develop a cross-domain generative model to enable image-to-image translation, while Lin *et al.* propose COMPLETER [31], which maximizes shared mutual information and minimizes conditional entropy to recover missing views. In addition, CMIB-Nets [32] balances the consistency and complementarity of multimodal views by extracting shared and view-specific information. Wan *et al.* [33] focus on self-supervised learning to regularize mutual information and improve clustering performance. Hu *et al.* [34] introduce a propagation information bottleneck to facilitate the transition from representation learning to clustering structure learning. Huang *et al.* [35] identify key requirements for effective multi-view learning and propose a model that integrates representation learning with clustering. Mao *et al.* [36] enhance alignment in clustering by distinguishing between consistency and redundancy through mutual information maximization, while Yan *et al.* [37] introduce a differentiable information bottleneck for deterministic multi-view clustering.

3 Methods

Problem Statement. We consider a cross-modal learning scenario where the data consists of two modalities, $\boldsymbol{s}$ and $\boldsymbol{t}$, with some observations potentially missing one of the modalities. In this context, we represent the data containing only modality $\boldsymbol{s}$ as $\mathcal{D}_s$, the data containing only modality $\boldsymbol{t}$ as $\mathcal{D}_t$ and the data containing both modalities paired as $\mathcal{D}_{s,t}$. Our goal is to learn meaningful latent representations from the combined datasets $\mathcal{D} = \mathcal{D}_s \cup \mathcal{D}_t \cup \mathcal{D}_{s,t}$ for various downstream tasks such as clustering and classification, while maintaining high-quality cross-modal generation and coherence.

3.1 Disentangled Mutual Supervision

Given the paired data setting $(\boldsymbol{s}, \boldsymbol{t})$, the information flow $\boldsymbol{s} \leftrightarrow \boldsymbol{z} \leftrightarrow \boldsymbol{t}$ is often employed in cross-modal representation learning, aiming to extract a latent representation $\boldsymbol{z}$ that encodes the information shared between both modalities [10]. However, such a design lacks a separate space for modality-specific information, which causes the shared latent representation $\boldsymbol{z}$ to inevitably contain entangled noise through mutual supervision under both information flows. Therefore, we propose to learn a modality specific factor $\boldsymbol{w}$ along with the shared latent $\boldsymbol{z}$ to decompose the exclusive and common features of each modality. Specifically, we present the main concept by considering the information flow from $\boldsymbol{s}$ to $\boldsymbol{t}$ and the derivation for the opposite direction is similar. According to our assumption, we define the generative process of our model as follows: $p_{\theta,\psi_z}(\boldsymbol{t}, \boldsymbol{s}, \boldsymbol{z}_s, \boldsymbol{w}_s) = p(\boldsymbol{t})\, p_{\psi_z}(\boldsymbol{z}_s \mid \boldsymbol{t})\, p(\boldsymbol{w}_s)\, p_\theta(\boldsymbol{s} \mid \boldsymbol{z}_s, \boldsymbol{w}_s)$, where $\boldsymbol{w}_s$ captures modality-specific content from $\boldsymbol{s}$, $\boldsymbol{z}_s$ represents the shared information necessary for generating $\boldsymbol{t}$. The prior of $\boldsymbol{w}_s$ is assumed to follow a standard normal distribution, i.e., $p(\boldsymbol{w}_s) \sim \mathcal{N}(0, \mathbf{I})$. Based on the generative model, the marginal log-likelihood $p_{\theta,\psi_z}(\boldsymbol{t}, \boldsymbol{s})$ requires integration over the latent variables $\boldsymbol{z}_s$ and $\boldsymbol{z}_w$, which is generally intractable. We then seek a variational posterior $q_{\boldsymbol{\phi},\varphi}(\boldsymbol{z}_s, \boldsymbol{w}_s|\boldsymbol{s}, \boldsymbol{t})$ to overcome this intractability and derive the ELBO via Jensen's inequality:

$$\log p_{\theta,\psi_z}(\boldsymbol{t}, \boldsymbol{s}) = \log \int p_{\theta,\psi_z}(\boldsymbol{t}, \boldsymbol{s}, \boldsymbol{z}_s, \boldsymbol{w}_s)\, d\boldsymbol{z}_s d\boldsymbol{w}_s \geq \mathbb{E}_{q_{\boldsymbol{\phi},\varphi}(\boldsymbol{z}_s, \boldsymbol{w}_s|\boldsymbol{s},\boldsymbol{t})} \log \frac{p_{\theta,\psi_z}(\boldsymbol{t}, \boldsymbol{s}, \boldsymbol{z}_s, \boldsymbol{w}_s)}{q_{\boldsymbol{\phi},\varphi}(\boldsymbol{z}_s, \boldsymbol{w}_s \mid \boldsymbol{s}, \boldsymbol{t})}. \tag{1}$$

In the course of variational inference, we derive a factorized form of the posterior distribution $q_{\boldsymbol{\phi},\varphi}(\boldsymbol{z}_s, \boldsymbol{w}_s, \boldsymbol{t} \mid \boldsymbol{s})$ under the assumptions that (i) the latent variables $\boldsymbol{z}_s$ and $\boldsymbol{w}_s$ are independent, and (ii) the observations $\boldsymbol{s}$ and $\boldsymbol{t}$ are conditionally independent given $\boldsymbol{z}_s$. Thus, the inference process is represented as:$q_{\boldsymbol{\phi},\varphi}(\boldsymbol{z}_s, \boldsymbol{w}_s, \boldsymbol{t} \mid \boldsymbol{s}) = q_{\phi_z}(\boldsymbol{z}_s \mid \boldsymbol{s})\, q_{\phi_w}(\boldsymbol{w}_s \mid \boldsymbol{s})\, q_\varphi(\boldsymbol{t} \mid \boldsymbol{z}_s)$. Once $q_{\boldsymbol{\phi},\varphi}(\boldsymbol{z}_s, \boldsymbol{w}_s, \boldsymbol{t} \mid \boldsymbol{s})$ is computed, we can apply Bayes' rule to obtain $q_{\boldsymbol{\phi},\varphi}(\boldsymbol{z}_s, \boldsymbol{w}_s \mid \boldsymbol{s}, \boldsymbol{t})$ according to the following decomposition: $q_{\boldsymbol{\phi},\varphi}(\boldsymbol{z}_s, \boldsymbol{w}_s|\boldsymbol{s}, \boldsymbol{t}) = \frac{q_{\boldsymbol{\phi},\varphi}(\boldsymbol{z}_s, \boldsymbol{w}_s, \boldsymbol{t}|\boldsymbol{s})}{q_{\phi_z,\varphi}(\boldsymbol{t}|\boldsymbol{s})} = \frac{q_{\phi_z}(\boldsymbol{z}_s|\boldsymbol{s}) q_{\phi_w}(\boldsymbol{w}_s|\boldsymbol{s}) q_\varphi(\boldsymbol{t}|\boldsymbol{z}_s)}{q_{\phi_z,\varphi}(\boldsymbol{t}|\boldsymbol{s})}$, where $q_{\phi_z,\varphi}(\boldsymbol{t} \mid \boldsymbol{s}) = \int q_{\phi_z}(\boldsymbol{z}_s \mid \boldsymbol{s})\, q_\varphi(\boldsymbol{t} \mid \boldsymbol{z}_s)\, d\boldsymbol{z}_s$. And $\boldsymbol{\phi} = \{\phi_z, \phi_w\}$ denotes the full set of parameters governing the variational distribution of $\boldsymbol{z}_s$ and $\boldsymbol{w}_s$, each modeled as a Gaussian distribution. By substituting inference process into Eq. (1), we obtain:

$$\begin{aligned} \log p_{\theta,\psi_z}(\boldsymbol{t}, \boldsymbol{s}) \geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_s, \boldsymbol{w}_s|\boldsymbol{s})} \left[\frac{q_\varphi(\boldsymbol{t} \mid \boldsymbol{z}_s)}{q_{\phi_z,\varphi}(\boldsymbol{t} \mid \boldsymbol{s})} \log \frac{p_{\psi_z}(\boldsymbol{z}_s \mid \boldsymbol{t})\, p(\boldsymbol{w}_s)\, p_\theta(\boldsymbol{s} \mid \boldsymbol{z}_s, \boldsymbol{w}_s)}{q_{\phi_z}(\boldsymbol{z}_s \mid \boldsymbol{s})\, q_{\phi_w}(\boldsymbol{w}_s \mid \boldsymbol{s})\, q_\varphi(\boldsymbol{t} \mid \boldsymbol{z}_s)} \right] \\ + \log q_{\phi_z,\varphi}(\boldsymbol{t}|\boldsymbol{s}) + \log p(\boldsymbol{t}). \end{aligned} \tag{2}$$

In Eq. (2), the ratio $\frac{q_\varphi(\boldsymbol{t}|\boldsymbol{z}_s)}{q_{\phi_z,\varphi}(\boldsymbol{t}|\boldsymbol{s})}$ serves as an importance weight that adjusts for the mismatch between the sampled latent-induced distribution and the actual conditional distribution. This adjustment improves the estimation accuracy and prevents the loss of pivotal information regarding $\boldsymbol{t}$ during sampling. Moreover, the log term measures the agreement between the generative process (through $p_{\psi_z}(\boldsymbol{z}_s \mid \boldsymbol{t})\, p(\boldsymbol{w}_s)\, p_\theta(\boldsymbol{s} \mid \boldsymbol{z}_s, \boldsymbol{w}_s)$) and the variational inference (through $q_{\phi_z}(\boldsymbol{z}_s \mid \boldsymbol{s})\, q_{\phi_w}(\boldsymbol{w}_s \mid \boldsymbol{s})\, q_\varphi(\boldsymbol{t} \mid \boldsymbol{z}_s)$). Maximizing this term encourages consistency between the latent structure inferred from data and the one induced by the generative process. The additional term $\log q_{\phi_z,\varphi}(\boldsymbol{t}|\boldsymbol{s})$ acts as a regularization term to encourage its alignment with the true distribution of $\boldsymbol{t}$ and balance the other components of the ELBO. Since $\log p(\boldsymbol{t})$ is constant with respect to the model parameters, it does not affect the optimization and can be ignored in the objective function.

Structured Representation Learning. The maximization of the ELBO alone presented in Eq. (2) is insufficient to ensure complete disentanglement of the latent variables $\boldsymbol{w}_s$ and $\boldsymbol{z}_s$, as any arbitrary mutually exclusive factorization may be equally favored in the absence of mechanisms that explicitly encourage information retention in the shared factor. Considering this, we add a mutual information maximization term $I(\boldsymbol{z}_s; \boldsymbol{t}; \boldsymbol{s})$ to enforce $\boldsymbol{z}_s$ containing all relevant information across modailties. We further incorporate another mutual information penalty $I(\boldsymbol{z}_s; \boldsymbol{w}_s)$ to encourage the decomposition of $\boldsymbol{w}_s$ and $\boldsymbol{z}_s$. The two information constraints can be unified in the following form:

$$\max I(\boldsymbol{z}_s; \boldsymbol{t}; \boldsymbol{s}) - I(\boldsymbol{z}_s; \boldsymbol{w}_s) = I(\boldsymbol{s}; \boldsymbol{z}_s, \boldsymbol{w}_s) - I(\boldsymbol{s}; \boldsymbol{w}_s) - I(\boldsymbol{z}_s; \boldsymbol{s} \mid \boldsymbol{t}) \tag{3}$$

The detailed derivations for Eq. (3) are provided in the Appendix A.1. Since direct optimization of mutual information is generally intractable, approximation methods such as variational inference or Monte Carlo sampling are often used to estimate these terms [38, 39, 40]. We subsequently derive computationally feasible approximations for each mutual information component as follows.

The first term $I(\boldsymbol{s}; \boldsymbol{z}_s, \boldsymbol{w}_s)$ quantifies the amount of information about the input modality $\boldsymbol{s}$ that is captured by the latent variables $\boldsymbol{z}_s$ and $\boldsymbol{w}_s$. Since its calculation involves $q(\boldsymbol{s} \mid \boldsymbol{w}_s, \boldsymbol{z}_s) = \frac{q(\boldsymbol{w}_s, \boldsymbol{z}_s|\boldsymbol{s})\, p_D(\boldsymbol{s})}{\int p_D(\boldsymbol{s})\, q(\boldsymbol{z}_s, \boldsymbol{w}_s|\boldsymbol{s})\, d\boldsymbol{s}}$, where $p_D(\boldsymbol{s})$ appears both as the empirical data distribution in the expectation and as an integral term in the denominator, making the computation intractable. We instead obtain a variational lower bound based on the generative distribution $p_\theta(\boldsymbol{s} \mid \boldsymbol{z}_s, \boldsymbol{w}_s)$:

$$\begin{aligned} I(\boldsymbol{z}_s, \boldsymbol{w}_s; \boldsymbol{s}) &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_s, \boldsymbol{w}_s|\boldsymbol{s}) p_D(\boldsymbol{s})} \log \frac{q(\boldsymbol{s} \mid \boldsymbol{z}_s, \boldsymbol{w}_s)}{p_D(\boldsymbol{s})} \\ &= H(\boldsymbol{s}) + \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_s, \boldsymbol{w}_s|\boldsymbol{s}) p_D(\boldsymbol{s})} \log p_\theta(\boldsymbol{s} \mid \boldsymbol{z}_s, \boldsymbol{w}_s) \\ &+ \mathbb{E}_{q(\boldsymbol{z}_s, \boldsymbol{w}_s)} \left[D_{KL}\left(q(\boldsymbol{s} \mid \boldsymbol{z}_s, \boldsymbol{w}_s) \| p_\theta(\boldsymbol{s} \mid \boldsymbol{z}_s, \boldsymbol{w}_s) \right) \right] \\ &\geq H(\boldsymbol{s}) + \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_s, \boldsymbol{w}_s|\boldsymbol{s}) p_D(\boldsymbol{s})} \log p_\theta(\boldsymbol{s}|\boldsymbol{z}_s, \boldsymbol{w}_s), \end{aligned} \tag{4}$$

where the entropy term $H(\boldsymbol{s})$ is treated as a constant. The remaining expectation term corresponds to a reconstruction loss under Gaussian distribution assumption and it indicates that maximizing $I(\boldsymbol{z}_s, \boldsymbol{w}_s; \boldsymbol{s})$ facilitates $\boldsymbol{w}_s$ and $\boldsymbol{z}_s$ to jointly contain all relevant information to modality $\boldsymbol{s}$. The second term $-I(\boldsymbol{s}; \boldsymbol{w}_s)$ can be omitted as a standalone constraint. A detailed explanation is provided in the Appendix A.3. The conditional mutual information term $-I(\boldsymbol{z}_s; \boldsymbol{s} \mid \boldsymbol{t})$ is minimized to suppress view-specific redundancy. The latent representation $\boldsymbol{z}_s$ is regularized by the counterpart modality $\boldsymbol{t}$, encouraging $\boldsymbol{z}_s$ to capture only the information accessible from both views. Specifically, by defining $\boldsymbol{z}_s$ and $\boldsymbol{z}_t$ over a shared latent space $\mathbb{Z}$, $I(\boldsymbol{z}_s; \boldsymbol{s} \mid \boldsymbol{t})$ can be variationally approximated as:

$$-I(\boldsymbol{z}_s; \boldsymbol{s} \mid \boldsymbol{t}) \geq -\mathbb{E}_{p_D(\boldsymbol{s}, \boldsymbol{t})} \left[D_{KL}\left(q_{\phi_z}(\boldsymbol{z}_s = \boldsymbol{z} \mid \boldsymbol{s}) \| q_{\psi_z}(\boldsymbol{z}_t = \boldsymbol{z} \mid \boldsymbol{t}) \right) \right]. \tag{5}$$

A complete derivation of Eq. (5) is provided in Appendix A.2. Consequently, by integrating these variationally tractable surrogate objectives, namely the reconstruction-based surrogate for $I(\boldsymbol{z}_s, \boldsymbol{w}_s; \boldsymbol{s})$ in Eq. (4), the KL divergence bound for $-I(\boldsymbol{s}; \boldsymbol{w}_s)$ and the cross-modal regularization for $I(\boldsymbol{z}_s; \boldsymbol{s} \mid \boldsymbol{t})$ in Eq. (5), the original ELBO in Eq. (2) becomes the following tractable optimization objective:

$$\begin{aligned} \mathcal{L}_{\{\boldsymbol{\Phi}, \boldsymbol{\Psi}\}}(\boldsymbol{s}, \boldsymbol{t}) &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_s, \boldsymbol{w}_s|\boldsymbol{s})} \left[\frac{q_\varphi(\boldsymbol{t} \mid \boldsymbol{z}_s)}{q_{\phi_z, \varphi}(\boldsymbol{t} \mid \boldsymbol{s})} \log \frac{p_{\psi_z}(\boldsymbol{z}_s \mid \boldsymbol{t})\, p(\boldsymbol{w}_s)\, p_\theta(\boldsymbol{s} \mid \boldsymbol{z}_s, \boldsymbol{w}_s)}{q_{\phi_z}(\boldsymbol{z}_s \mid \boldsymbol{s})\, q_{\phi_w}(\boldsymbol{w}_s \mid \boldsymbol{s})\, q_\varphi(\boldsymbol{t} \mid \boldsymbol{z}_s)} \right] \\ &+ \log q_{\phi_z, \varphi}(\boldsymbol{t}|\boldsymbol{s}) + \mathbb{E}_{q_\phi(\boldsymbol{z}_s, \boldsymbol{w}_s|\boldsymbol{s}) p_D(\boldsymbol{s})} \log p_\theta(\boldsymbol{s} \mid \boldsymbol{z}_s, \boldsymbol{w}_s) \\ &- \mathbb{E}_{p_D(\boldsymbol{s}, \boldsymbol{t})} \left[D_{KL}\left(q_{\phi_z}(\boldsymbol{z}_s = \boldsymbol{z} \mid \boldsymbol{s}) \| q_{\psi_z}(\boldsymbol{z}_t = \boldsymbol{z} \mid \boldsymbol{t}) \right) \right], \end{aligned} \tag{6}$$

where $\boldsymbol{\Phi} = \{\boldsymbol{\phi}, \varphi\}$ and $\boldsymbol{\Psi} = \{\boldsymbol{\psi}, \theta\}$ correspond to the encoder and decoder parameters, respectively.

Mutual Supervision. Mutual supervision leverages reciprocal guidance between modalities to learn a semantically consistent shared latent space. Unlike explicit integration methods like PoE or MoE, it offers greater flexibility and robustness without requiring strict alignment or direct fusion. Building on the notion that each modality both informs and constrains the other, we formulate the objective in the case where $\boldsymbol{t}$ represents the source data and $\boldsymbol{s}$ the target data, i.e. $\boldsymbol{t} \to \boldsymbol{z}_t \to \boldsymbol{s}$, as follows:

$$\begin{aligned} \mathcal{L}_{\{\boldsymbol{\Psi}, \boldsymbol{\Phi}\}}(\boldsymbol{s}, \boldsymbol{t}) &= \mathbb{E}_{q_{\boldsymbol{\psi}}(\boldsymbol{z}_t, \boldsymbol{w}_t|\boldsymbol{t})} \left[\frac{q_\theta(\boldsymbol{s} \mid \boldsymbol{z}_t)}{q_{\theta, \psi_z}(\boldsymbol{s} \mid \boldsymbol{t})} \log \frac{p_{\phi_z}(\boldsymbol{z}_t \mid \boldsymbol{s})\, p(\boldsymbol{w}_t)\, p_\varphi(\boldsymbol{t} \mid \boldsymbol{z}_t, \boldsymbol{w}_t)}{q_{\psi_z}(\boldsymbol{z}_t \mid \boldsymbol{t})\, q_{\psi_w}(\boldsymbol{w}_t \mid \boldsymbol{t})\, q_\theta(\boldsymbol{s} \mid \boldsymbol{z}_t)} \right] \\ &+ \log q_{\theta, \psi}(\boldsymbol{s}|\boldsymbol{t}) + E_{q_\psi(\boldsymbol{z}_t, \boldsymbol{w}_t|\boldsymbol{t}) p_D(\boldsymbol{t})} \log p_\varphi(\boldsymbol{t} \mid \boldsymbol{z}_t, \boldsymbol{w}_t) \\ &- \mathbb{E}_{p_D(\boldsymbol{s}, \boldsymbol{t})} \left[D_{KL}\left(q_{\psi_z}(\boldsymbol{z}_t = \boldsymbol{z} \mid \boldsymbol{t}) \| q_{\phi_z}(\boldsymbol{z}_s = \boldsymbol{z} \mid \boldsymbol{s}) \right) \right]. \end{aligned} \tag{7}$$

This formulation mirrors the standard direction $\mathcal{L}_{\{\boldsymbol{\Phi}, \boldsymbol{\Psi}\}}(\boldsymbol{s}, \boldsymbol{t})$ and captures the reverse information flow. To instantiate this symmetric structure, we implement the model by swapping the roles of the generative and inference networks, where we exchange the parameter sets $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$. To integrate both directions of information flow, we combine the contributions from both $\mathcal{L}_{\{\boldsymbol{\Phi}, \boldsymbol{\Psi}\}}(\boldsymbol{s}, \boldsymbol{t})$ and $\mathcal{L}_{\{\boldsymbol{\Psi}, \boldsymbol{\Phi}\}}(\boldsymbol{s}, \boldsymbol{t})$, yielding the objective function:

$$\mathcal{L}_{Bi}(\boldsymbol{s}, \boldsymbol{t}) = \frac{1}{2} \left(\mathcal{L}_{\{\boldsymbol{\Phi}, \boldsymbol{\Psi}\}}(\boldsymbol{s}, \boldsymbol{t}) + \mathcal{L}_{\{\boldsymbol{\Psi}, \boldsymbol{\Phi}\}}(\boldsymbol{s}, \boldsymbol{t}) \right). \tag{8}$$

Shared Representations Alignment. We leverage $I(\boldsymbol{z}_s, \boldsymbol{z}_t)$ to directly align the modality-invariant latent variables, preventing them from diverging or encoding discrepant semantics for the same content. This ensures that mutual supervision is not only reflected in the generative reconstruction paths, but also enforced through semantic alignment at the latent representation level. We incorporate

$I(\boldsymbol{z}_s, \boldsymbol{z}_t)$ as a regularization component in addition to Eq. (8) and derive the following training objective in the paired-data scenario:

$$\mathcal{L}_{Bi}\left(\boldsymbol{s}, \boldsymbol{t}\right) = \frac{1}{2}\left(\mathcal{L}_{\{\boldsymbol{\Phi},\boldsymbol{\Psi}\}}(\boldsymbol{s}, \boldsymbol{t}) + \mathcal{L}_{\{\boldsymbol{\Psi},\boldsymbol{\Phi}\}}(\boldsymbol{s}, \boldsymbol{t})\right) + \alpha I\left(\boldsymbol{z}_s; \boldsymbol{z}_t\right). \tag{9}$$

where α serves as a regularization coefficient that balances the contribution of the mutual information term $I\left(\boldsymbol{z}_s; \boldsymbol{z}_t\right)$. Empirically, we estimate $I\left(\boldsymbol{z}_s; \boldsymbol{z}_t\right)$ using contrastive learning, which has been confirmed as an effective way to solve for the mutual information maximization. Specifically, we align paired latent features while distinguishing unpaired ones, computing pairwise cosine similarities within each batch and applying cross-entropy without extra projection layers.

3.2 Partial Observations Scenario

Our model naturally extends to single-view scenarios due to its autoregressive cross-modal generative structure. Given an observed modality, the shared latent representation can be inferred and used to approximate the missing modality via learned generative paths. First considering that the $\boldsymbol{s}$-mode data is available, we can derive a variational approximation for $\log p_{\theta,\psi}\left(\boldsymbol{s}\right)$ by marginalizing over the unobserved modality $\boldsymbol{t}$, resulting in the following lower bound:

$$\begin{aligned}\log p_{\theta,\psi_z}\left(\boldsymbol{s}\right) &= \log \int p\left(\boldsymbol{t}\right) p_{\psi_z}\left(\boldsymbol{z}_s \mid \boldsymbol{t}\right) p\left(\boldsymbol{w}_s\right) p_\theta\left(\boldsymbol{s} \mid \boldsymbol{z}_s, \boldsymbol{w}_s\right) d\boldsymbol{t} d\boldsymbol{z}_s d\boldsymbol{w}_s \\ &\geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_s, \boldsymbol{w}_s \mid \boldsymbol{s})} \log \frac{p_\theta\left(\boldsymbol{s} \mid \boldsymbol{z}_s, \boldsymbol{w}_s\right) p\left(\boldsymbol{w}_s\right) p_{u^t}\left(\boldsymbol{z}_s\right)}{q_{\phi_z}\left(\boldsymbol{z}_s \mid \boldsymbol{s}\right) q_{\phi_w}\left(\boldsymbol{w}_s \mid \boldsymbol{s}\right)},\end{aligned} \tag{10}$$

where $p_{u^t}\left(\boldsymbol{z}_s\right) = \int p\left(\boldsymbol{t}\right) p_{\psi_z}\left(\boldsymbol{z}_s \mid \boldsymbol{t}\right) d\boldsymbol{t}$. Notably, even in the absence of paired $\boldsymbol{t}$-observations, the model can still regularize the latent representation $\boldsymbol{z}_s$ through a shared prior derived from the distribution of $\boldsymbol{t}$. Inspired by VampPrior [41], we define a batch-dependent prior over B representative anchors $\{\boldsymbol{u}_i^t\}_{i=1}^B$ sampled from the $\boldsymbol{t}$-modality, yielding: $p_{u^t}\left(\boldsymbol{z}_s\right) = \frac{1}{B}\sum_{i=1}^B p_{\psi_z}\left(\boldsymbol{z}_s | \boldsymbol{u}_i^t\right)$, where dynamic resampling ensures the prior adapts to the evolving latent structure, improving stability and expressiveness. Eq. (10) can then be rewritten as:

$$\begin{aligned}\mathcal{L}_s\left(\boldsymbol{s}\right) &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_s, \boldsymbol{w}_s \mid \boldsymbol{s})} \log \frac{p_\theta\left(\boldsymbol{s} \mid \boldsymbol{z}_s, \boldsymbol{w}_s\right) p\left(\boldsymbol{w}_s\right) \frac{1}{B}\sum_{i=1}^B p_{\psi_z}\left(\boldsymbol{z}_s \mid \boldsymbol{u}_i^t\right)}{q_{\phi_z}\left(\boldsymbol{z}_s \mid \boldsymbol{s}\right) q_{\phi_w}\left(\boldsymbol{w}_s \mid \boldsymbol{s}\right)} \\ &= \mathbb{E}_{q_\phi(\boldsymbol{z}_s, \boldsymbol{w}_s \mid \boldsymbol{s})} \log p_\theta\left(\boldsymbol{s} \mid \boldsymbol{z}_s, \boldsymbol{w}_s\right) - D_{KL}\left(q_{\phi_z}\left(\boldsymbol{z}_s \mid \boldsymbol{s}\right) \| \frac{1}{B}\sum_{i=1}^B p_{\psi_z}\left(\boldsymbol{z}_s \mid \boldsymbol{u}_i^t\right)\right) \\ &- D_{KL}\left(q_{\phi_w}\left(\boldsymbol{w}_s \mid \boldsymbol{s}\right) || p\left(\boldsymbol{w}_s\right)\right).\end{aligned} \tag{11}$$

With this design, we can still leverage the information from the $\boldsymbol{t}$-modality to effectively constrain and guide the model. In a comparable manner, when the $\boldsymbol{s}$-modality is missing and $\boldsymbol{t}$-modality is available, the objective is defined as:

$$\begin{aligned}\mathcal{L}_t\left(\boldsymbol{t}\right) &= \mathbb{E}_{q_{\boldsymbol{\psi}}(\boldsymbol{z}_t, \boldsymbol{w}_t \mid \boldsymbol{t})} \log p_\varphi\left(\boldsymbol{t} \mid \boldsymbol{z}_t, \boldsymbol{w}_t\right) - D_{KL}\left(q_{\psi_z}\left(\boldsymbol{z}_t \mid \boldsymbol{t}\right) \| \frac{1}{B}\sum_{i=1}^B p_{\phi_z}(\boldsymbol{z}_t \mid \boldsymbol{u}_i^s)\right) \\ &- D_{KL}\left(q_{\psi_w}\left(\boldsymbol{w}_t | \boldsymbol{t}\right) \| p\left(\boldsymbol{w}_t\right)\right).\end{aligned} \tag{12}$$

The final objective function integrates the target loss functions for both paired and unpaired cases, encompassing three scenarios: paired samples, samples with only $\boldsymbol{s}$-modality, and samples with only $\boldsymbol{t}$-modality. It is expressed as follows and the training procedure is detailed in Appendix A.7.

$$\mathcal{L}\left(\mathcal{D}\right) = \sum_{s,t \in \mathcal{D}_{s,t}} \mathcal{L}_{Bi}\left(\boldsymbol{s}, \boldsymbol{t}\right) + \sum_{s \in \mathcal{D}_s} \mathcal{L}_s\left(\boldsymbol{s}\right) + \sum_{t \in \mathcal{D}_t} \mathcal{L}_t\left(\boldsymbol{t}\right). \tag{13}$$

4 Experiments

4.1 Experiments Setup

Datasets. Two widely used datasets, i.e. MNIST-SVHN and CUBICC, are adopted in our experiments. The MNIST-SVHN dataset consists of the MNIST and Street View House Numbers (SVHN) datasets, where the samples share digit labels (10 classes) but have different digit styles [8]. The

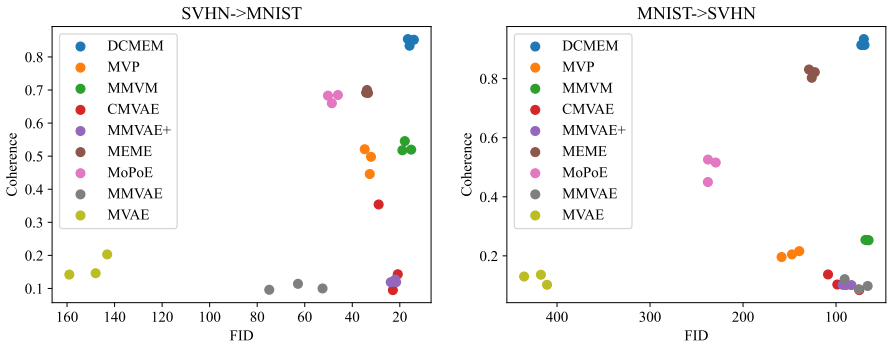


Figure 2: The cross-modal generation performance of DCMEM and existing multimodal VAEs on the MNIST-SVHN dataset. Each model was run independently three times. The best performance is located in the top-right corner of each figure.

CUB Image-Captions for Clustering (CUBICC) dataset, a variant of the CUB image caption dataset constructed by Palumbo *et al.* [12], consists of two modalities: bird images and their corresponding descriptive captions. The dataset is divided into 8 categories based on bird species. In addition, the Human Breast Cancer spatial transcriptomics dataset provides high-resolution spatial measurements across multiple modalities, including gene expression, spatial coordinates, and tissue morphological features [42]. The dataset comprises 20 distinct spatial label categories corresponding to different tissue or structural regions. These datasets are selected as representative benchmarks to evaluate the model's ability to disentangle and align multimodal latent spaces rather than to cover all possible modality combinations.

Baselines. To comprehensively evaluate the performance of the proposed method, we compare it with eight existing multimodal VAEs, including MVAE [7], MMVAE [8], MoPoE [9], MEME [10], MMVAE+ [11], CMVAE [12], MMVM [24] and MVP [28]. For all comparison methods, we adopt the model architectures proposed in their respective papers and use their default optimal parameters. To assess the model's capability in handling partially observed datasets, we construct a set of incomplete bimodal datasets by randomly removing one modality at missing rates of $\eta \in \{0.25, 0.5, 0.75\}$ and then train MVAE, MoPoE, MEME, MVP as well as our method on these modified datasets. Each method is run three times to ensure the reliability of the results. We provide the implementation details of our method in Appendix B.2.

Evaluation. For the generative task, we primarily evaluate generation coherence and generation quality as in previous work [11, 12, 24]. To evaluate generation coherence, we use a pretrained classifier to classify the generated samples and evaluate the generation coherence in terms of the classification accuracy. The generation quality is assessed using the FID metric [43]. Moreover, we investigate the effectiveness of the learned latent representations based on classification and clustering analyses. For classification, we follow previous work by training a linear classifier on the latent space and report the accuracy. For clustering, we apply K-means and evaluate the results using ACC, NMI and ARI. For models with both shared and modality-specific latent variables, all evaluations are conducted on the shared latent space, which captures modality-invariant semantics.

4.2 Comparison of Generation Performance

In this section, we compare the performance of the proposed DCMEM model with that of the aforementioned competitive multimodal VAEs in terms of cross-modal generation. We first conduct the cross-modal generation experiment on the fully paired MNIST-SVHN dataset. As shown in Figure 2, our model demonstrates superior performance, achieving both high generative coherence and quality. In Figures 3 and 7, we present the cross-modal generation results for each model. It is evident from these results that our model effectively captures the underlying digit labels and generates accurate cross-modal samples. In contrast, other models tend to misidentify similar digits, struggling with certain digit pairs. This discrepancy highlights the advantages of our model in handling the complexity of cross-modal generation tasks, maintaining both high fidelity and label consistency.

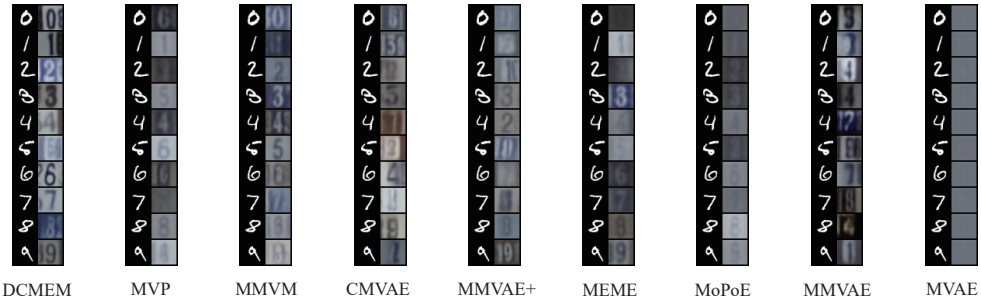


Figure 3: Qualitative results of cross-modal generation on the MNIST-SVHN dataset for each model. The left side shows the input samples, while the right side displays the cross-modal generated samples.

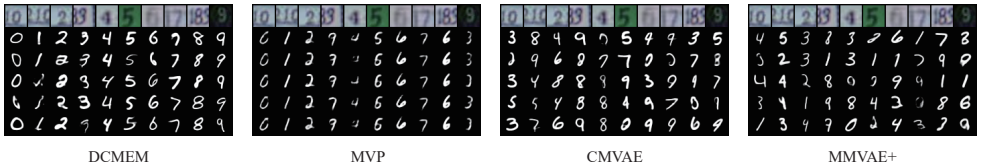


Figure 4: Five SVHN-to-MNIST samples are generated by varying only the modality-specific latent variables on the MNIST-SVHN dataset.

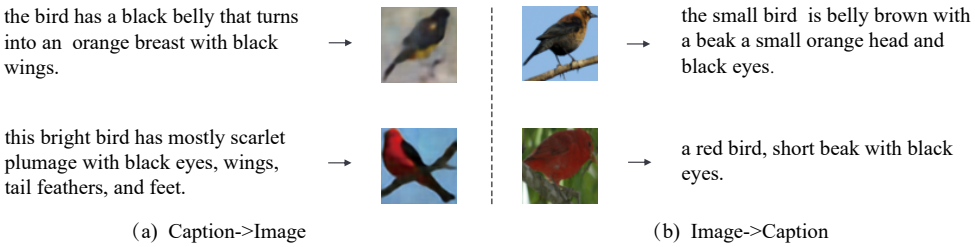


Figure 5: Qualitative results of cross-modal generation by DCMEM on the CUBICC dataset. The samples on the left side of the arrows represent the input samples, while the samples on the right side show the generated cross-modal samples.

Among the models, MMVAE+, CMVAE, MVP and our DCMEM all incorporate both shared and modality-specific latent variables. To further analyze the generation coherence and quality, we fix the shared latent variables and randomly sample five different modality-specific latent variables for cross-modal generation. This allows us to evaluate how well each model can generate diverse samples while maintaining consistency across modalities. The experimental results are illustrated in Figures 4 and 8. Our model demonstrates the ability to generate images that maintain the same digit class, while introducing variations in digit shape and color as the modality-specific variables are modified. In contrast, both MMVAE+ and CMVAE exhibit entanglement between class information and modality-specific variables, leading to inconsistencies in the generated samples. Meanwhile, MVP fails to introduce sufficient variation through its modality-specific latent variables, resulting in limited sample diversity. These comparisons further highlight the strength of our approach in producing diverse and high-quality cross-modal generations without sacrificing semantic consistency or class identity.

To further demonstrate the performance of our model in more complex scenarios, we conduct a cross-modal generation experiment on the CUBICC dataset and Figure 5 presents the generation results. As expected, our model consistently achieves high-quality cross-modal generation performance across different modalities. The generated samples align well with the corresponding modalities, maintaining a high level of coherence between the image and text representations. This showcases the model's ability to effectively handle multimodal data in more intricate and diverse settings, demonstrating its robustness in real-world applications. Moreover, the model achieves robust cross-

Table 1: Quantitative comparison of clustering performance for each model's latent representations on the CUBICC dataset. The best and second-best results are highlighted in bold and underlined, respectively.

Methods	Image Representation			Caption Representation			Joint Representation		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
MVAE	26.2	12.4	7.5	18.1	2.4	0.9	38.7	26.8	18.0
MMVAE	23.1	12.1	6.1	14.5	1.3	0.1	15.8	1.5	0.2
MoPoE	33.4	17.6	11.5	43.5	27.1	19.9	40.8	30.4	20.2
MEME	44.8	43.4	28.4	36.3	29.5	18.6	19.8	4.8	2.1
MMVAE+	27.7	11.9	7.1	48.7	36.4	26.8	64.4	52.6	44.1
CMVAE	<u>67.7</u>	<u>58.3</u>	<u>47.4</u>	<u>65.1</u>	**53.3**	<u>42.7</u>	<u>73.7</u>	<u>67.4</u>	<u>57.2</u>
MMVM	58.9	56.9	44.5	23.9	9.4	5.4	66.8	67.0	55.5
MVP	64.1	53.8	41.8	48.5	34.4	26.1	61.1	55.6	44.0
DCMEM	**86.9**	**77.4**	**72.4**	**69.7**	<u>52.2</u>	**44.2**	**86.3**	**76.8**	**71.5**

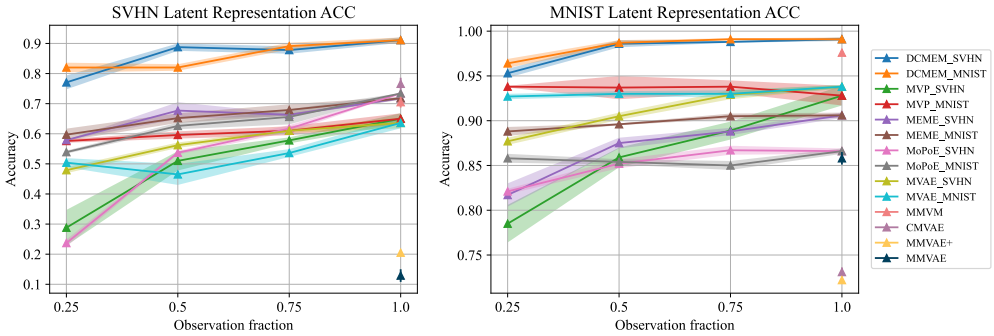


Figure 6: Classification accuracy under different missing rates on the MNIST-SVHN dataset. Shaded areas represent the standard deviation across multiple runs. The subscripts in method names indicate the observed modality. For example, DCMEM_SVHN (Observation fraction = 0.25) denotes that the training data consists of 25% paired samples and 75% unimodal SVHN samples.

modal generation under varying missing rates, effectively balancing generative quality and class-level semantic consistency (see Appendix C.1).

4.3 Latent Representation Analysis

In this section, we focus on evaluating the effectiveness of the latent representations learned by the model. All models are trained on the fully paired MNIST-SVHN and CUBICC datasets and the latent representations for the test set are extracted using the trained models. We also evaluate the models under partially paired (incomplete) scenarios, as detailed in Appendix C.2. We first perform clustering on the latent representations of each modality to assess their individual clustering performance, followed by clustering on the joint representations. For models such as MVAE and MoPoE, which produce joint latent representations, we directly use the joint representations for K-means clustering. CMVAE directly yields the clustering labels as it learns a clustering variable. For other models that provide neither the joint representations nor clustering factors, we concatenate the latent representations from each modality and then apply K-means clustering to the resulting joint representation. The clustering results are summarized in Tables 1 and 2.

Quantitative metrics clearly demonstrate that our model outperforms others in terms of clustering performance. The T-SNE [44] plots of our model's latent representations are presented in Figure 9. From these visualizations, it is evident that the latent representations for both modalities naturally form distinct clusters. Moreover, the latent representations of the same category across modalities align into a single cohesive cluster, indicating that our model effectively captures the shared information between modalities. In contrast, the latent representations of other models either fail to form well-defined clusters or exhibit separation between modalities, as shown in Figures 10 and 11. These results further emphasize the superiority of our model in learning coherent and meaningful latent representations, making it highly effective for clustering tasks involving multimodal data.

For the classification task, we present the results based solely on the classification accuracy metric, evaluating model performance under various missing data conditions. The classification results are shown in Figures 6 and 12. Our model consistently outperforms the other alternatives and demonstrates robust classification performance across varying missing rates. In contrast, other baseline models exhibit significant performance degradation at certain missing rates, indicating the limited capability in exploiting incomplete data. For example, MMVAE+ underperforms on the MNIST-SVHN dataset. This suboptimal performance is likely due to its inability to effectively capture class-discriminative features within the shared latent space. Moreover, although models such as MEME and MVP are designed to handle missing modalities, their performance degrades substantially as the missing rate increases. This suggests a limited robustness to incomplete data, highlighting the advantage of our method in maintaining high classification accuracy under varying degrees of missing information. Analyses in Appendix C.3 and Appendix C.4 further show that our model consistently preserves semantic and class-level alignment across modalities. Appendices C.5–C.7 provide detailed descriptions of the ablation study, parameter analysis, and computational resources.

4.4 Application to Human Breast Cancer Dataset

To evaluate the applicability of our model in other fields, we conduct an additional experiment on a spatial transcriptomics dataset of human breast cancer [42]. The dataset presents significant challenges for accurate spatial domain identification due to technical noise and inherent biological variability. We compare our model with eight multimodal VAE baselines and five spatial domain clustering methods, including Scanpy [45], STAGATE [46], GraphST [47], SiGra [48], and xSiGra [49]. For the VAE-based baselines lacking explicit spatial relationships modeling, we adopt the CoordConv strategy [50], encoding spatial (x, y) coordinates as two additional image channels. In contrast, spatial clustering methods are inherently designed to capture spatial dependencies through graph-based or attention mechanisms. As illustrated in Figure 13, DCMEM achieves the highest clustering performance, with an ARI of 55.1 and NMI of 69.7, substantially outperforming both VAE-based and spatially-awared approaches. Among multimodal VAEs, MMVAE achieves the best results, yet it still falls short compared to Scanpy, GraphST and our model. Methods such as MVAE and MMVM struggle to resolve fine-grained cell population structures. Although spatial methods like STAGATE and GraphST generate more coherent partitions, DCMEM produces sharper cluster boundaries and exhibits better alignment with ground truth labels. These results demonstrate the effectiveness of our method in integrating heterogeneous modalities and capturing informative spatial and molecular patterns, underscoring its strong potential for real-world clustering tasks.

5 Broader Impact & Limitations

This work focuses on learning disentangled cross-modal representations and enhancing generation quality and coherence by leveraging mutual supervision along with the information bottleneck principle. Our approach is applicable to a variety of tasks in scientific and engineering domains and holds the potential for positive societal impact. In the biomedical field, for example, our model can detect fine-grained cell populations with precise boundaries and identify potential biomarkers associated with tumor heterogeneity for subsequent clinical validation, by analyzing spatial transcriptomics data at the single-cell level. At the bulk cancer omics level, it can also assist in identifying novel disease subtypes and predicting the survival outcomes to inform better treatment strategy design. Although our model yields better representations and more coherent outputs, it is specifically designed for bimodal data scenarios and may require non-trivial effort to achieve competitve performance on general multimodal datasets.

6 Conclusion

In this work, we propose DCMEM, a variational framework that integrates disentanglement and mutual supervision to learn structured cross-modal representations. It separates shared and modality-specific information via dedicated latent spaces and promotes semantic alignment by maximizing mutual information between shared latent variables across modalities, achieving strong performance across diverse tasks and settings.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62372279, 62322215, 62532017), and the Natural Science Foundation of Shandong Province (ZR2025QB62, ZR2023MF119). This study was also supported in part by the High-Performance Computing Center of Central South University.

References

[1] Vinitra Swamy, Malika Satayeva, Jibril Frej, Thierry Bossy, Thijs Vogels, Martin Jaggi, Tanja Kaser, and Mary-Anne Hartley. Multimodn—multimodal, multi-task, interpretable modular networks. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 28115–28138. Curran Associates, Inc., 2023.

[2] Divyam Madaan, Taro Makino, Sumit Chopra, and Kyunghyun Cho. Jointly modeling inter- & intra-modality dependencies for multi-modal learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 116084–116105. Curran Associates, Inc., 2024.

[3] Zichang Liu, Zhiqiang Tang, Xingjian Shi, Aston Zhang, Mu Li, Anshumali Shrivastava, and Andrew Gordon Wilson. Learning multimodal data augmentation in feature space. In *The Eleventh International Conference on Learning Representations*, 2023.

[4] Zhou Lu. A theory of multimodal learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 57244–57255. Curran Associates, Inc., 2023.

[5] Xuanchi Ren, Tao Yang, Yuwang Wang, and Wenjun Zeng. Learning disentangled representation by exploiting pretrained generative models: A contrastive learning view. In *International Conference on Learning Representations*, 2022.

[6] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *In International Conference on Learning Representations*.

[7] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[8] Yuge Shi, N. Siddharth, Brooks Paige, and Philip H. S. Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. 2019.

[9] Thomas M Sutter, Imant Daunhawer, and Julia E Vogt. Generalized multimodal elbo. In *International Conference on Learning Representations*, 2021.

[10] Tom Joy, Yuge Shi, Philip Torr, Tom Rainforth, Sebastian M Schmon, and Siddharth N. Learning multimodal VAEs through mutual supervision. In *International Conference on Learning Representations*, 2022.

[11] Emanuele Palumbo, Imant Daunhawer, and Julia E Vogt. MMVAE+: Enhancing the generative quality of multimodal VAEs without compromises. In *The Eleventh International Conference on Learning Representations*, 2023.

[12] Emanuele Palumbo, Laura Manduchi, Sonia Laguna, Daphné Chopard, and Julia E Vogt. Deep generative clustering with multimodal diffusion variational autoencoders. In *The Twelfth International Conference on Learning Representations*, 2024.

[13] HyeongJoo Hwang, Geon-Hyeong Kim, Seunghoon Hong, and Kee-Eung Kim. Variational interaction information maximization for cross-domain disentanglement. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22479–22491. Curran Associates, Inc., 2020.

[14] Jie Xu, Yazhou Ren, Huayi Tang, Xiaorong Pu, Xiaofeng Zhu, Ming Zeng, and Lifang He. Multi-vae: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9214–9223, 2021.

[15] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep genera-tive models. *stat*, 1050:7, 2016.

[16] Ramakrishna Vedantam, Ian Fischer, Jonathan Huang, and Kevin Murphy. Generative models of visually grounded imagination. In *International Conference on Learning Representations*, 2018.

[17] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.

[18] Mike Wu and Noah Goodman. Multimodal generative models for compositional representation learning. *arXiv preprint arXiv:1912.05075*, 2019.

[19] Richard Kurle, Stephan Günnemann, and Patrick Van der Smagt. Multi-source neural variational inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4114–4121, 2019.

[20] Imant Daunhawer, Thomas M Sutter, Kieran Chin-Cheong, Emanuele Palumbo, and Julia E Vogt. On the limitations of multimodal vaes. In *The Tenth International Conference on Learning Representations (ICLR 2022)*, 2022.

[21] Jannik Wolff, Tassilo Klein, Moin Nabi, Rahul G Krishnan, and Shinichi Nakajima. Mixture-of-experts vaes can disregard variation in surjective multimodal data. *arXiv preprint arXiv:2204.05229*, 2022.

[22] Thomas Sutter, Imant Daunhawer, and Julia Vogt. Multimodal generative learning utilizing jensen-shannon-divergence. *Advances in neural information processing systems*, 33:6100–6110, 2020.

[23] HyeongJoo Hwang, Geon-Hyeong Kim, Seunghoon Hong, and Kee-Eung Kim. Multi-view representation learning via total correlation objective. *Advances in Neural Information Processing Systems*, 34:12194–12207, 2021.

[24] Thomas Sutter, Yang Meng, Andrea Agostini, Daphné Chopard, Norbert Fortin, Julia Vogt, Babak Shahbaba, and Stephan Mandt. Unity by diversity: Improved representation learning for multimodal vaes. *Advances in Neural Information Processing Systems*, 37:74262–74297, 2024.

[25] Mihee Lee and Vladimir Pavlovic. Private-shared disentangled multimodal vae for learning of latent representations. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 1692–1700, 2021.

[26] Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

[27] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised generative approach to clustering. In *26th International Joint Conference on Artificial Intelligence, IJCAI 2017*, pages 1965–1972. International Joint Conferences on Artificial Intelligence, 2017.

[28] Xin Gao and Jian Pu. Deep incomplete multi-view learning via cyclic permutation of vaes. In *The Thirteenth International Conference on Learning Representations*, 2025.

[29] Qi Wang, Claire Boudreau, Qixing Luo, Pang-Ning Tan, and Jiayu Zhou. Deep multi-view information bottleneck. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 37–45. SIAM, 2019.

[30] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *8th International Conference on Learning Representations*, 2020.

[31] Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. Completer: Incomplete multi-view clustering via contrastive prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11174–11183, 2021.

[32] Zhibin Wan, Changqing Zhang, Pengfei Zhu, and Qinghua Hu. Multi-view information-bottleneck representation learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10085–10092, 2021.

[33] Shiye Wang, Changsheng Li, Yanming Li, Ye Yuan, and Guoren Wang. Self-supervised information bottleneck for deep multi-view subspace clustering. *IEEE Transactions on Image Processing*, 32:1555–1567, 2023.

[34] Shizhe Hu, Zenglin Shi, Xiaoqiang Yan, Zhengzheng Lou, and Yangdong Ye. Multiview clustering with propagating information bottleneck. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[35] Weitian Huang, Sirui Yang, and Hongmin Cai. Generalized information-theoretic multi-view clustering. *Advances in neural information processing systems*, 36:58752–58764, 2023.

[36] Yiqiao Mao, Xiaoqiang Yan, Jiaming Liu, and Yangdong Ye. Congmc: Consistency-guided multimodal clustering via mutual information maximin. *IEEE Transactions on Multimedia*, 26:5131–5146, 2023.

[37] Xiaoqiang Yan, Zhixiang Jin, Fengshou Han, and Yangdong Ye. Differentiable information bottleneck for deterministic multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27435–27444, 2024.

[38] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations*, 2020.

[39] Shiye Wang, Changsheng Li, Yanming Li, Ye Yuan, and Guoren Wang. Self-supervised information bottleneck for deep multi-view subspace clustering. *IEEE Trans. Image Process.*, 32:1555–1567, 2023.

[40] HyeongJoo Hwang, Geon-Hyeong Kim, Seunghoon Hong, and Kee-Eung Kim. Variational interaction information maximization for cross-domain disentanglement. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22479–22491. Curran Associates, Inc., 2020.

[41] Jakub Tomczak and Max Welling. Vae with a vampprior. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1214–1223. PMLR, 09–11 Apr 2018.

[42] Sunny Z Wu, Ghamdan Al-Eryani, Daniel Lee Roden, Simon Junankar, Kate Harvey, Alma Andersson, Aatish Thennavan, Chenfei Wang, James R Torpy, Nenad Bartonicek, et al. A single-cell and spatially resolved atlas of human breast cancers. *Nature genetics*, 53(9):1334–1347, 2021.

[43] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[44] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[45] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

[46] Kangning Dong and Shihua Zhang. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature communications*, 13(1):1739, 2022.

[47] Yahui Long, Kok Siong Ang, Mengwei Li, Kian Long Kelvin Chong, Raman Sethi, Chengwei Zhong, Hang Xu, Zhiwei Ong, Karishma Sachaphibulkij, Ao Chen, et al. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with graphst. *Nature Communications*, 14(1):1155, 2023.

[48] Ziyang Tang, Zuotian Li, Tieying Hou, Tonglin Zhang, Baijian Yang, Jing Su, and Qianqian Song. Sigra: single-cell spatial elucidation through an image-augmented graph transformer. *Nature Communications*, 14(1):5618, 2023.

[49] Aishwarya Budhkar, Ziyang Tang, Xiang Liu, Xuhong Zhang, Jing Su, and Qianqian Song. xsigra: explainable model for single-cell spatial data elucidation. *Briefings in Bioinformatics*, 25(5):bbae388, 2024.

[50] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9628–9639, Red Hook, NY, USA, 2018. Curran Associates Inc.

[51] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[52] Tom Joy, Sebastian M Schmon, Philip HS Torr, N Siddharth, and Tom Rainforth. Capturing label characteristics in vaes. *arXiv preprint arXiv:2006.10102*, 2020.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the paper, we present a novel multimodal approach that outperforms existing multimodal methods on multiple datasets. The contributions are clearly stated in the abstract and introduction, and align with the theoretical and empirical results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the paper in the Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical results include clear assumptions and complete proofs, provided in the main text or appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe all the experiments in full detail in the appendix (see Appendix B) such that all the results on all datasets can be reproduced.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the complete source code along with instructions, including README files and environment settings in the supplementary material. All datasets used in our experiments are publicly available, and we include clear instructions for accessing and preprocessing the data.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide all the necessary information either in the main text and appendix (see Appendix B).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report the mean and standard deviation of the results obtained from experiments conducted using three different random seeds.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided details on the compute resources used, including type of workers, memory, and execution time, as well as the total compute required for the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and confirm that our research complies with all of its guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential broader impact of the proposed work alongside its limitations in Section 5.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper does not involve the release of models or datasets with high risks of misuse, such as pretrained language models, image generators, or scraped datasets. Therefore, there are no specific safeguards discussed in the paper.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and code used in our work are publicly available and properly cited in the paper. We have explicitly acknowledged the original sources and included license information where applicable.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have included the code as supplementary material with our submission to ensure reproducibility during the review process. Upon acceptance, we will make the code publicly available with comprehensive documentation.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Supplementary Technical Details

A.1 Mutual Information Decomposition for Structured Representation Learning

Define X, Y and Z be random variables. The the chain rule for mutual information is:

$$I(X;Y;Z) = I(X;Y) - I(X;Y|Z) = I(X;Z) - I(X;Z|Y) = I(Y;Z) - I(Y;Z|X) \quad (14)$$

As defined in the main text, the modality-invariant variable $\boldsymbol{z}_s$ and the modality-specific variable $\boldsymbol{w}_s$ have already been introduced. $I(\boldsymbol{z}_s;\boldsymbol{w}_s)$ and $I(\boldsymbol{s};\boldsymbol{t};\boldsymbol{z}_s)$ can be computed as follows:

$$\begin{aligned} I(\boldsymbol{z}_s;\boldsymbol{w}_s) &= I(\boldsymbol{z}_s;\boldsymbol{s}) - I(\boldsymbol{z}_s;\boldsymbol{s} \mid \boldsymbol{w}_s) + I(\boldsymbol{w}_s;\boldsymbol{z}_s \mid \boldsymbol{s}) \\ I(\boldsymbol{s};\boldsymbol{t};\boldsymbol{z}_s) &= I(\boldsymbol{z}_s;\boldsymbol{s}) - I(\boldsymbol{z}_s;\boldsymbol{s} \mid \boldsymbol{t}) \end{aligned} \quad (15)$$

Under the assumption that the modality-invariant variable $\boldsymbol{z}_s$ and the modality-specific variable $\boldsymbol{w}_s$ are conditionally independent given the input $\boldsymbol{s}$, we have: $q(\boldsymbol{w}_s \mid \boldsymbol{s}) = q(\boldsymbol{w}_s \mid \boldsymbol{s}, \boldsymbol{z}_s)$. Thus, the conditional mutual information $I(\boldsymbol{w}_s;\boldsymbol{z}_s \mid \boldsymbol{s})$ simplifies to $I(\boldsymbol{w}_s;\boldsymbol{z}_s \mid \boldsymbol{s}) = H(\boldsymbol{w}_s \mid \boldsymbol{s}) - H(\boldsymbol{w}_s \mid \boldsymbol{s}, \boldsymbol{z}_s) = 0$. As a result, $I(\boldsymbol{w}_s;\boldsymbol{z}_s)$ can be further decomposed as follows: $I(\boldsymbol{z}_s;\boldsymbol{w}_s) = I(\boldsymbol{z}_s;\boldsymbol{s}) - I(\boldsymbol{z}_s;\boldsymbol{s} \mid \boldsymbol{w}_s) = I(\boldsymbol{z}_s;\boldsymbol{s}) + I(\boldsymbol{s};\boldsymbol{w}_s) - I(\boldsymbol{s};\boldsymbol{z}_s,\boldsymbol{w}_s)$. In conclusion, the mutual information decomposition for disentangled representation learning is given as follows:

$$\begin{aligned} & I(\boldsymbol{z}_s;\boldsymbol{t};\boldsymbol{s}) - I(\boldsymbol{z}_s;\boldsymbol{w}_s) \\ &= \cancel{I(\boldsymbol{z}_s;\boldsymbol{s})} - I(\boldsymbol{z}_s;\boldsymbol{s} \mid \boldsymbol{t}) + I(\boldsymbol{s};\boldsymbol{z}_s,\boldsymbol{w}_s) - I(\boldsymbol{s};\boldsymbol{w}_s) - \cancel{I(\boldsymbol{z}_s;\boldsymbol{s})} \\ &= I(\boldsymbol{s};\boldsymbol{z}_s,\boldsymbol{w}_s) - I(\boldsymbol{s};\boldsymbol{w}_s) - I(\boldsymbol{z}_s;\boldsymbol{s} \mid \boldsymbol{t}) \end{aligned} \quad (16)$$

A.2 Evidence Lower Bound on the cross-modal regularization

Assuming that both $\boldsymbol{z}_t$ and $\boldsymbol{z}_s$ lie in the same latent space $\mathbb{Z}$:

$$\begin{aligned} -I(\boldsymbol{z}_s;\boldsymbol{s} \mid \boldsymbol{t}) &= -\mathbb{E}_{p_D(\boldsymbol{s},\boldsymbol{t})}\mathbb{E}_{q_{\phi_z}(z_s|s)}\left[\log \frac{q_{\phi_z}(\boldsymbol{z}_s = \boldsymbol{z} \mid \boldsymbol{s})}{p_{\psi_z}(\boldsymbol{z}_s = \boldsymbol{z} \mid \boldsymbol{t})}\right] \\ &= -\mathbb{E}_{p_D(\boldsymbol{s},\boldsymbol{t})}\mathbb{E}_{q_{\phi_z}(\boldsymbol{z}_s|\boldsymbol{s})}\left[\log \frac{q_{\phi_z}(\boldsymbol{z}_s = \boldsymbol{z} \mid \boldsymbol{s})}{q_{\psi_z}(\boldsymbol{z}_t = \boldsymbol{z} \mid \boldsymbol{t})} \frac{q_{\psi_z}(\boldsymbol{z}_t = \boldsymbol{z} \mid \boldsymbol{t})}{p_{\psi_z}(\boldsymbol{z}_s = \boldsymbol{z} \mid \boldsymbol{t})}\right] \\ &= -\mathbb{E}_{p_D(\boldsymbol{s},\boldsymbol{t})}\left[D_{KL}\left(q_{\phi_z}(\boldsymbol{z}_s = \boldsymbol{z} \mid \boldsymbol{s}) || q_{\psi_z}(\boldsymbol{z}_t = \boldsymbol{z} \mid \boldsymbol{t})\right)\right] \\ &\quad + \mathbb{E}_{p_D(\boldsymbol{t})}\left[D_{KL}\left(p_{\psi_z}(\boldsymbol{z}_s = \boldsymbol{z} \mid \boldsymbol{t}) || q_{\psi_z}(\boldsymbol{z}_t = \boldsymbol{z} \mid \boldsymbol{t})\right)\right] \\ &\geq -\mathbb{E}_{p_D(\boldsymbol{s},\boldsymbol{t})}\left[D_{KL}\left(q_{\phi_z}(\boldsymbol{z}_s = \boldsymbol{z} \mid \boldsymbol{s}) || q_{\psi_z}(\boldsymbol{z}_t = \boldsymbol{z} \mid \boldsymbol{t})\right)\right]. \end{aligned} \quad (17)$$

Similarly, $-I(\boldsymbol{z}_t;\boldsymbol{t} \mid \boldsymbol{s})$ can be computed as:

$$-I(\boldsymbol{z}_t;\boldsymbol{t} \mid \boldsymbol{s}) \geq -\mathbb{E}_{p_D(\boldsymbol{s},\boldsymbol{t})}\left[D_{KL}\left(q_{\psi_z}(\boldsymbol{z}_t = \boldsymbol{z} \mid \boldsymbol{t}) \| q_{\phi_z}(\boldsymbol{z}_s = \boldsymbol{z} \mid \boldsymbol{s})\right)\right]. \quad (18)$$

A.3 Derivation of the Mutual Information Term $I(\boldsymbol{s};\boldsymbol{w}_s)$

The second term $-I(\boldsymbol{s};\boldsymbol{w}_s)$ also poses computational challenges due to its dependence on the marginal distribution $p_D(\boldsymbol{s})$, where $q(\boldsymbol{w}_s) = \int q(\boldsymbol{w}_s \mid \boldsymbol{s}) p_D(\boldsymbol{s}) d\boldsymbol{s}$. In a similar vein, this term is approximated using its variational lower bound $-\mathbb{E}_{p_D(\boldsymbol{s})}\left[D_{KL}\left(q_{\boldsymbol{\phi_w}}(\boldsymbol{w}_s \mid \boldsymbol{s}) \| p(\boldsymbol{w}_s)\right)\right]$. This KL divergence term naturally appears in the ELBO objective (Eq. (2)) due to $\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_s,\boldsymbol{w}_s|\boldsymbol{s})}\left[\frac{q_\varphi(\boldsymbol{t}|\boldsymbol{z}_s)}{q_{\phi_z,\varphi}(\boldsymbol{t}|\boldsymbol{s})} \log \frac{p(\boldsymbol{w}_s)}{q_{\phi_w}(\boldsymbol{w}_s|\boldsymbol{s})}\right] = \mathbb{E}_{q_{\boldsymbol{\phi_w}}}(\boldsymbol{w}_s \mid \boldsymbol{s}) \log \frac{p(\boldsymbol{w}_s)}{q_{\phi_w}(\boldsymbol{w}_s|\boldsymbol{s})} = -\mathbb{E}_{p_D(\boldsymbol{s})}\left[D_{KL}\left(q_{\boldsymbol{\phi_w}}(\boldsymbol{w}_s \mid \boldsymbol{s}) \| p(\boldsymbol{w}_s)\right)\right]$. Hence, it is implicitly optimized through the ELBO and can be omitted as a standalone constraint.

A.4 Mutual Information Approximation

The mutual information $I(\boldsymbol{z}_s, \boldsymbol{z}_t)$ between the shared latent variables $\boldsymbol{z}_s$ and $\boldsymbol{z}_t$ is approximated using a contrastive loss that encourages alignment between representations of paired inputs while distinguishing those from non-paired ones [51]. Given a batch of B paired latent features $(\boldsymbol{h}_s, \boldsymbol{h}_t)$ sampled from $\boldsymbol{z}_s$ and $\boldsymbol{z}_t$, we concatenate them into a set of $2B$ vectors. Pairwise cosine similarities are computed and scaled by a fixed temperature $\tau = 0.5$ to construct the contrastive logits. Positive pairs are defined between the i-th feature in $\boldsymbol{h}_s$ and the i-th feature in $\boldsymbol{h}_t$, as they correspond to the paired input instances. All other pairs in the batch are treated as negatives. We apply a cross-entropy loss to encourage the model to assign higher similarity to positives than to negatives.

A.5 Cross-Modal Reconstruction Mechanism

In our framework, $q_\varphi(\boldsymbol{t} \mid \boldsymbol{z}_s)$ serves a key role in the mutual supervision mechanism. Specifically, $\boldsymbol{z}_s$ is the shared latent representation inferred from modality $\boldsymbol{s}$ and $q_\varphi(\boldsymbol{t} \mid \boldsymbol{z}_s)$ models the reconstruction of modality $\boldsymbol{t}$ based solely on this shared information. This setup is intentionally designed to exclude the private latent variable $\boldsymbol{w}$, since the goal is to assess what information is common and transferable across modalities. In practice, we implement $q_\varphi(\boldsymbol{t} \mid \boldsymbol{z}_s)$ by setting $\boldsymbol{w} = \mathbf{0}$ and passing the concatenation $[\boldsymbol{z}_s, \mathbf{0}]$ into the decoder of modality $\boldsymbol{t}$. This design offers two key advantages: (1) it avoids introducing an additional decoder by reusing $p_\varphi(\boldsymbol{t} \mid \boldsymbol{z}_s, \boldsymbol{w})$ with $\boldsymbol{w}$ set to zero, keeping the architecture compact; and (2) it promotes effective disentanglement, as reconstructing $\boldsymbol{t}$ from $\boldsymbol{z}_s$ alone forces $\boldsymbol{z}_s$ to capture modality-invariant, shared information.

A.6 Importance Sampling Stability

The importance sampling weight $\frac{q_\varphi(\boldsymbol{t}|\boldsymbol{z}_s)}{q_{\phi_z,\varphi}(\boldsymbol{t}|\boldsymbol{s})}$ in Equation 2 plays a critical role in our training objective, but its direct computation can introduce numerical instability due to high variance in gradient estimates. This arises because both the numerator and denominator are parameterized distributions that are learned during training, and their stochastic nature may result in noisy, unreliable updates when used in Monte Carlo estimation of the ELBO. To address this issue, we adopt a stop-gradient strategy to stabilize training without compromising the objective, following a rationale similar to that in prior work [52, 10]. Concretely, we prevent gradients from flowing through the $\frac{q_\varphi(\boldsymbol{t}|\boldsymbol{z}_s)}{q_{\phi_z,\varphi}(\boldsymbol{t}|\boldsymbol{s})}$, treating it as a fixed scalar during backpropagation: stop_gradient $\left(\frac{q_\varphi(\boldsymbol{t}|\boldsymbol{z}_s)}{q_{\phi_z,\varphi}(\boldsymbol{t}|\boldsymbol{s})}\right)$. This modification ensures that the parameters are optimized based on more stable signals, as it avoids amplifying gradient noise through the ratio. Importantly, this treatment does not alter the forward computation of the objective but improves the robustness and reliability of the training dynamics. In summary, this design provides a practical and effective solution to variance-induced instability in training, while still aligning with the theoretical intent of the original ELBO formulation.

A.7 Training procedure for DCMEM

Algorithm 1: Optimization Procedure of DCMEM

Input: Multimodal dataset: $\mathcal{D} = \mathcal{D}_s \cup \mathcal{D}_t \cup \mathcal{D}_{s,t}$; Training epochs number: M; Hyperparameters: α; Model parameters $\{\boldsymbol{\Phi}, \boldsymbol{\Psi}\}$

Output: Latent representation $\boldsymbol{z}_s$ and $\boldsymbol{z}_t$

```
Randomly initialize model parameters {Φ, Ψ};
for epoch < M do
    for each sample in paired data D_{s,t} do
        Compute L_{Φ,Ψ}(s, t) and L_{Ψ,Φ}(s, t) via Eqs. (6) and (7);
        Estimate enhanced mutual-supervised information regularization by Eq. (8);
        Compute the bidirectional lower bound L_Bi(s, t) by Eq. (9);
    for each sample in modality-specific data D_s do
        Compute L_s(s) via Eq. (11);
    for each sample in modality-specific data D_t do
        Compute L_t(t) via Eq. (12);
    Update parameters {Φ, Ψ} by maximizing the overall objective in Eq. (13);
Compute latent representations z_s = f_{φ_z}(s) and z_t = f_{ψ_z}(t) using the optimized parameters;
return Latent representationas z_s and z_t
```

B Dataset and Implementation Details

B.1 Dataset Licences

- MNIST_SVHN: originally published in [8], downloaded the data from http://yann.lecun.com/exdb/mnist, http://ufldl.stanford.edu/housenumbers and the code from https://github.com/iffsid/mmvae, licensed under GPL 3.0.

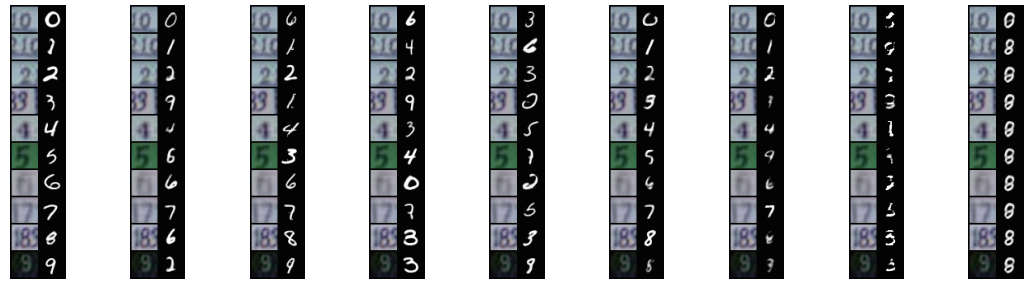


Figure 7: Supplementary qualitative results of cross-modal generation on the MNIST-SVHN dataset for each model.

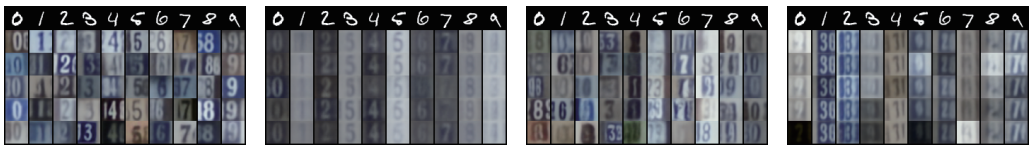


Figure 8: Five MNIST-to-SVHN samples are generated by varying only the modality-specific latent variables on the MNIST-SVHN dataset.

- CUBICC: originally published in [12], downloaded the data from https://polybox.ethz.ch/index.php/s/LRkTC2oa6YHHlUj/download, published under the MIT license.
- Human Breast Cancer: originally published in [42], downloaded the data from https://www.10xgenomics.com/datasets/human-breast-cancer-block-a-section-1-1-standard-1-0-0, published under the CC BY 4.0 license.

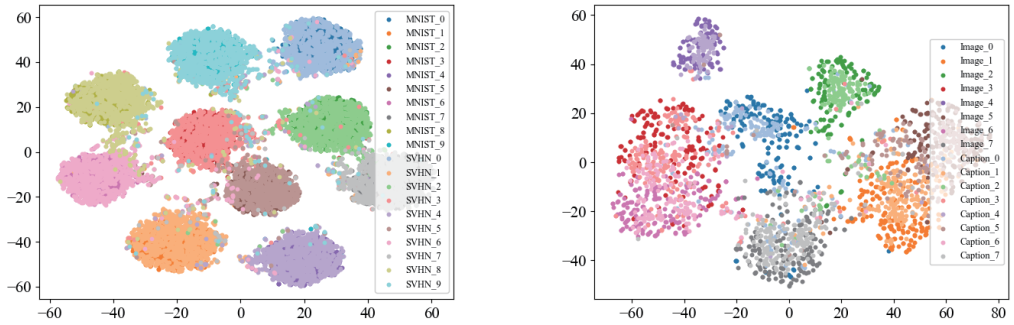
B.2 Implementation Details

To ensure a fair comparison, all models are run on a local server with an NVIDIA GeForce RTX 2080 Ti GPU, 64 GB of RAM running Ubuntu 18.04. For our model, we use a ResNet encoder and decoder for image data, and convolutional encoders and decoders for text data. The parameter α is set to 1. For the MNIST-SVHN dataset, the dimensions of the shared and specific latent spaces are set to 32. We use the Adam optimizer with a learning rate of 5e-4, a batch size of 64 and train the model for 100 epochs. For the CUBICC dataset, the dimensions of the shared and specific latent spaces are set to 48 and 16, respectively. The Adam optimizer is used with a learning rate of 1e-4, a batch size of 16 and training is conducted for 200 epochs. For the spatial transcriptomics dataset, we preprocess the gene expression data by selecting the top 3000 highly variable genes, followed by standard normalization and log-transformation. Both the encoder and decoder for this modality are implemented as fully connected neural networks. For the tissue morphology modality, input images are resized to 128×128 pixels. To incorporate spatial context, we adopt the CoordConv [50] technique by appending the 2D spatial coordinates (x, y) as two additional input channels, resulting in a 5-channel input. This modality is processed using convolutional neural networks for both encoding and decoding. Both the shared and specific latent dimensions are set to 32. Optimization is performed using Adam with a learning rate of 5e-4, a batch size of 64 and 100 training epochs.

C Additional Experimental Results

C.1 Additional Results on Cross-Modal Generation

To further evaluate the generative capability of our model, we conduct additional experiments under varying missing rates, focusing on both generative coherence and quality. The quantitative results



(a) MNIST-SVHN (b) CUBICC

Figure 9: T-SNE plot of the latent representations obtained by DCMEM on the MNIST-SVHN and CUBICC datasets. Here, MNIST_0 represents the data from the MNIST modality with the digit label 0, and similarly for other labels.

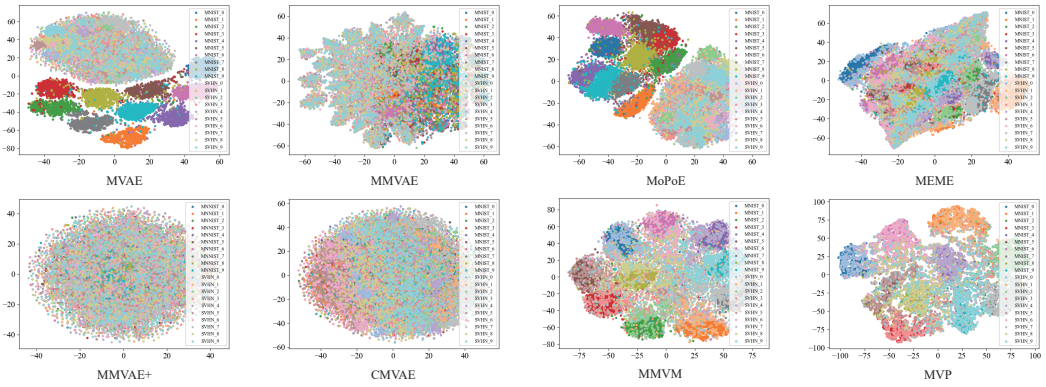


Figure 10: T-SNE plot of the latent representations obtained by baseline models on the MNIST-SVHN dataset.

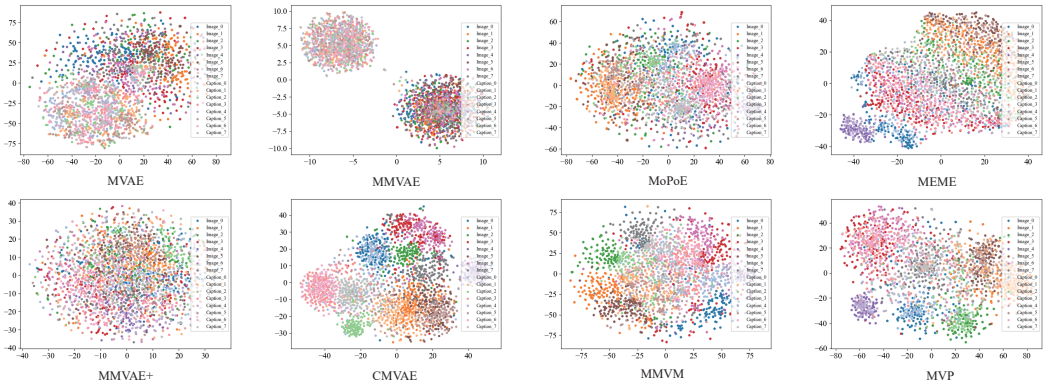


Figure 11: T-SNE plot of the latent representations obtained by baseline models on the CUBICC dataset.

on the MNIST-SVHN dataset are shown in Figures 14 and 15, while the qualitative results are presented in Figures 16–22. As shown in the figures, our model consistently achieves robust cross-modal generation under different levels of missing data. Although a slight degradation in generative consistency is observed as the missing rate increases, the model still maintains strong performance and demonstrates competitive stability compared to other baselines.

Table 2: Quantitative comparison of clustering performance for each model's latent representations on the MNIST-SVHN dataset. The best and second-best results are highlighted in bold and underlined, respectively.

Methods	SVHN Representation			MNIST Representation			Joint Representation		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
MVAE	27.9	16.0	13.1	79.2	65.5	62.6	42.7	35.3	24.5
MMVAE	22.0	10.4	10.1	21.8	10.3	10.1	22.6	10.7	10.1
MoPoE	37.9	27.2	18.5	50.5	45.6	33.0	64.1	60.5	50.7
MEME	21.9	10.3	10.0	36.5	32.1	20.4	22.4	10.6	10.1
MMVAE+	23.9	11.4	11.1	21.3	10.4	10.0	22.9	11.9	10.8
CMVAE	42.2	36.3	25.4	28.1	15.9	14.5	32.3	19.5	15.4
MMVM	42.2	27.1	20.7	<u>88.1</u>	<u>82.1</u>	<u>80.4</u>	77.5	72.2	67.5
MVP	<u>53.6</u>	<u>38.7</u>	<u>30.1</u>	81.4	79.6	73.6	<u>84.8</u>	<u>76.4</u>	<u>70.6</u>
DCMEM	**91.5**	**80.6**	**82.0**	**99.1**	**97.3**	**98.0**	**99.5**	**98.4**	**98.9**

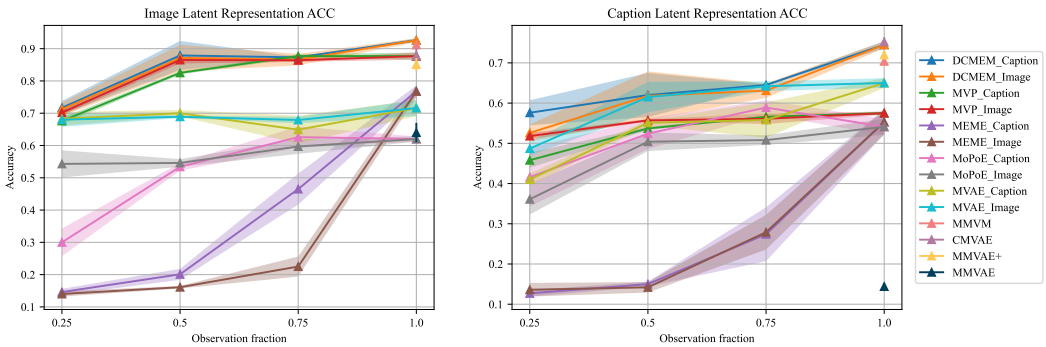


Figure 12: Classification accuracy under different missing rates on the CUBICC dataset. Shaded areas represent the standard deviation across multiple runs. The subscripts in method names indicate the observed modality. For example, DCMEM_Caption (Observation fraction = 0.25) denotes that the training data consists of 25% paired samples and 75% unimodal Caption samples.

Similarly, the cross-modal generation results on the CUBICC dataset are shown in Figures 23 and 24. Our model achieves the highest generative coherence across all missing conditions, indicating its effectiveness in preserving class-level semantics despite incomplete input. However, we observe a moderate decline in generation quality. This trade-off is primarily due to the inherent tension between the generation and clustering objectives: while generation benefits from latent representations that retain fine-grained modality-specific details, clustering prefers representations that focus on global class-level features. As a result, our model strategically balances these two objectives rather than optimizing solely for one, which inevitably limits performance in either direction when pursued independently.

C.2 Additional Results on Clustering

We evaluate the clustering performance of MVAE, MoPoE, MEME, MVP and DCMEM under varying missing scenarios on both the MNIST-SVHN and CUBICC datasets. As shown in Tables 3, 4, 5 and 6, we consider two settings for each dataset: one where the first modality is partially missing (e.g., MNIST or Image), and one where the second modality is partially missing (e.g., SVHN or Caption). For each setting, clustering is performed based on the latent representations learned from individual modalities as well as their joint embedding. DCMEM consistently achieves the best performance across all settings and representation types, demonstrating strong robustness to incomplete data. Even at low paired data fractions (e.g., 25%), DCMEM maintains high clustering accuracy, while baseline models such as MVAE, MoPoE, MEME and MVP exhibit significant performance degradation. This is especially evident in the Caption modality of the CUBICC dataset, where several baselines struggle to learn meaningful representations under high missing-view rates. Overall, these results highlight the effectiveness of DCMEM in learning coherent and discriminative

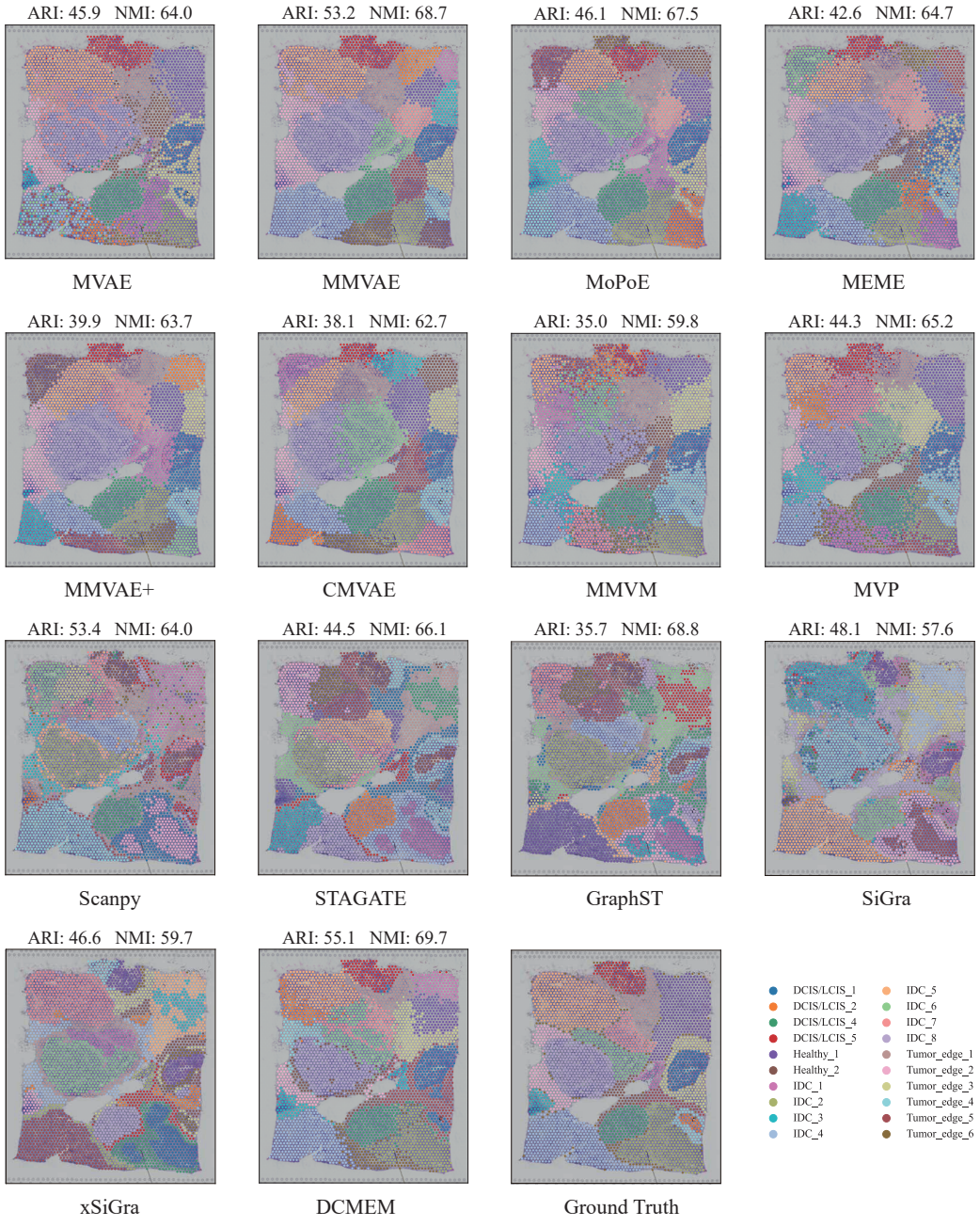


Figure 13: Visualization of clustering results on the human breast cancer dataset. Each subplot shows the clustering output of a different method, with colors indicating predicted clusters. Each method is run three times, and the mean ARI and NMI scores are reported above each plot.

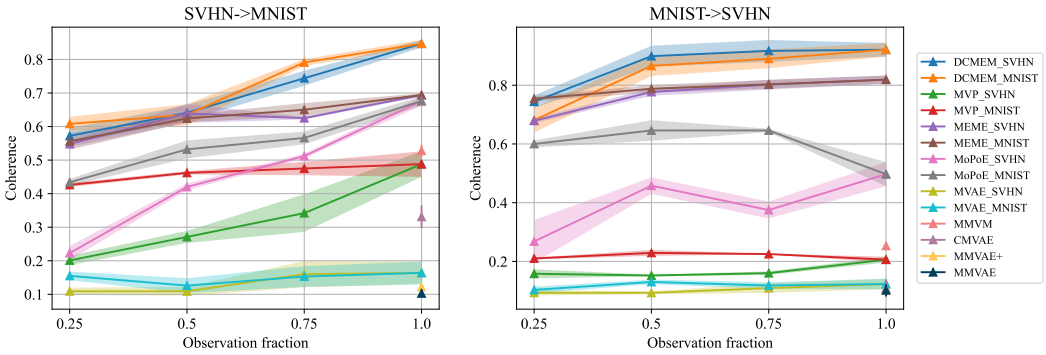


Figure 14: Classification accuracy of cross-modal generations under different missing rates on the MNIST-SVHN dataset. Shaded areas represent the standard deviation across multiple runs. The subscripts in method names indicate the observed modality. For example, DCMEM_SVHN (Observation fraction = 0.25) denotes that the training data consists of 25% paired samples and 75% unimodal SVHN samples.

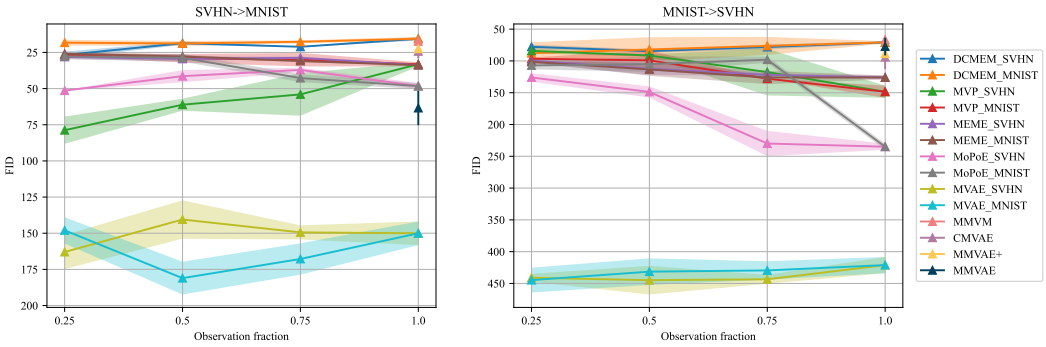


Figure 15: FID score of cross-modal generations under different missing rates on the MNIST-SVHN dataset. Shaded areas represent the standard deviation across multiple runs. The subscripts in method names indicate the observed modality. For example, DCMEM_SVHN (Observation fraction = 0.25) denotes that the training data consists of 25% paired samples and 75% unimodal SVHN samples.

latent spaces across different datasets and under various levels of modality incompleteness. Its ability to leverage both paired and unimodal data allows it to maintain superior clustering performance, setting it apart from existing multimodal VAE approaches.

C.3 Semantic Relatedness in the Latent Space

Semantic relatedness refers to the notion that semantically aligned multimodal inputs should yield more similar latent distributions than unrelated pairs. To investigate whether our models as well as the baselines exhibit this behavior, we adopt the 2-Wasserstein distance as a measure of semantic similarity between latent distributions. This metric is well-suited for comparing Gaussian distributions due to its closed-form expression in such cases. In our experiment, we compute pairwise 2-Wasserstein distances between all combinations of latent distributions within a mini-batch. We then visualize the resulting distances using histograms, color-coded to distinguish paired samples from unpaired samples. A clear separation between the two groups in the histogram indicates that the model captures meaningful semantic alignment across modalities. Figure 25 illustrates the relatedness histograms produced by our model on the MNIST-SVHN and CUBICC datasets. Figures 26 and 27 show the results for the baseline models.

As shown in Figure 25, our model exhibits consistently lower 2-Wasserstein distances for paired samples, while unpaired samples yield significantly higher distances. This clear separation demonstrates that our model effectively captures semantic alignment across modalities. In contrast, baseline models

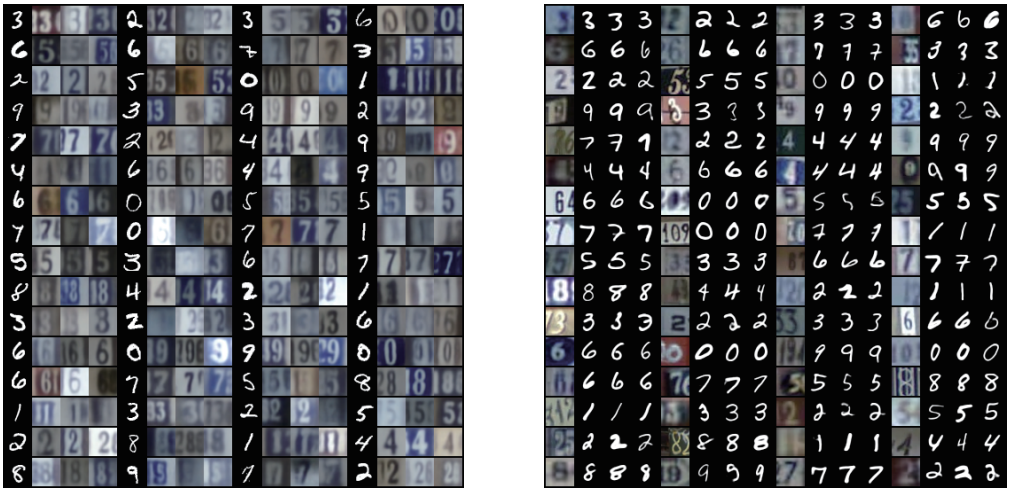


Figure 16: MNIST->SVHN (Left) and SVHN->MNIST (Right), for the fully observed case.

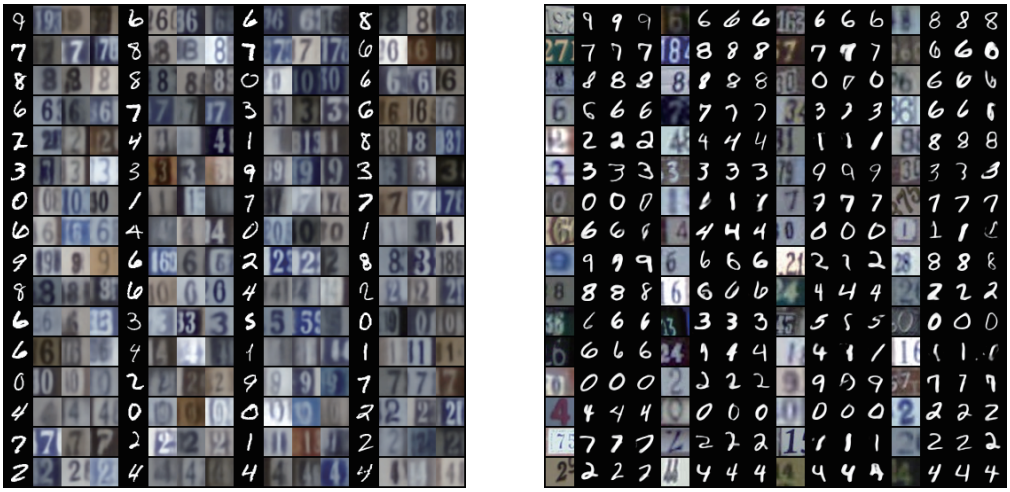


Figure 17: MNIST->SVHN (Left) and SVHN->MNIST (Right), when MNIST is observed 75% of the time.

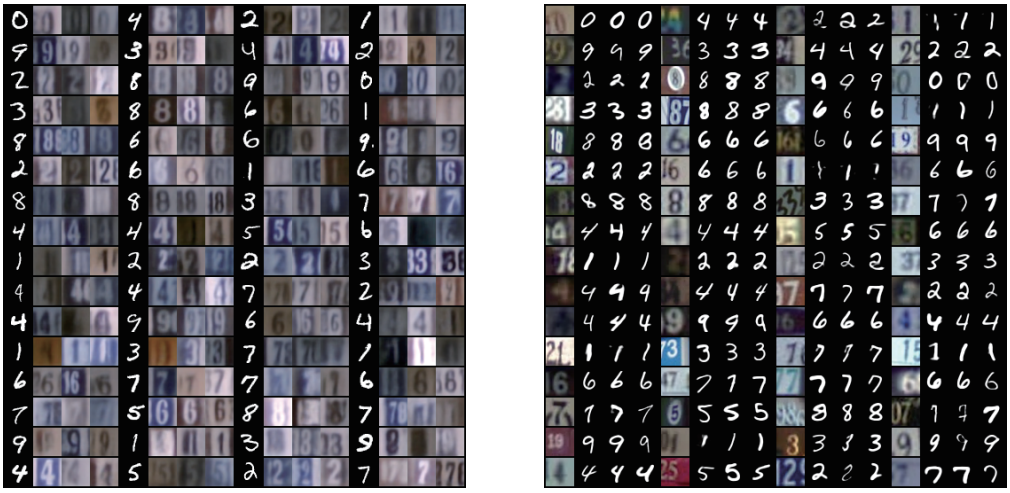


Figure 18: MNIST->SVHN (Left) and SVHN->MNIST (Right), when SVHN is observed 75% of the time.

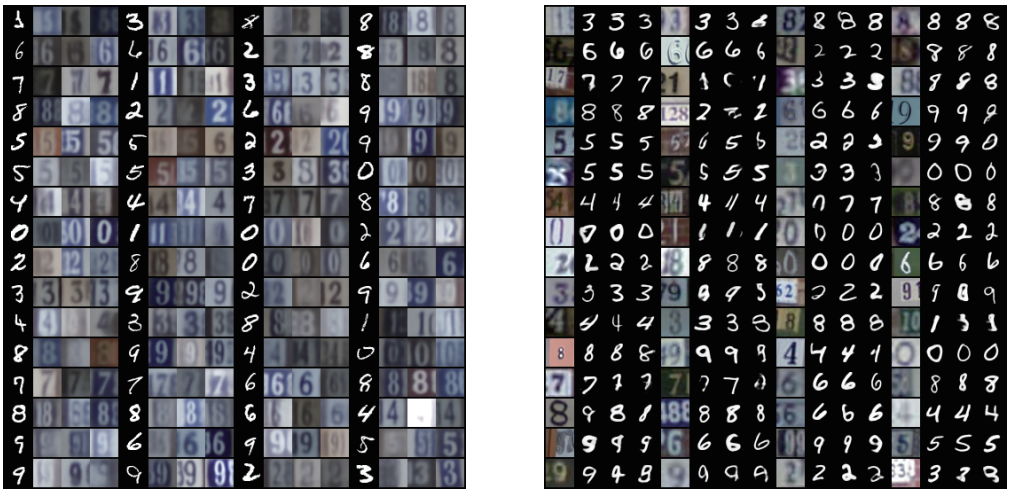


Figure 19: MNIST->SVHN (Left) and SVHN->MNIST (Right), when MNIST is observed 50% of the time.

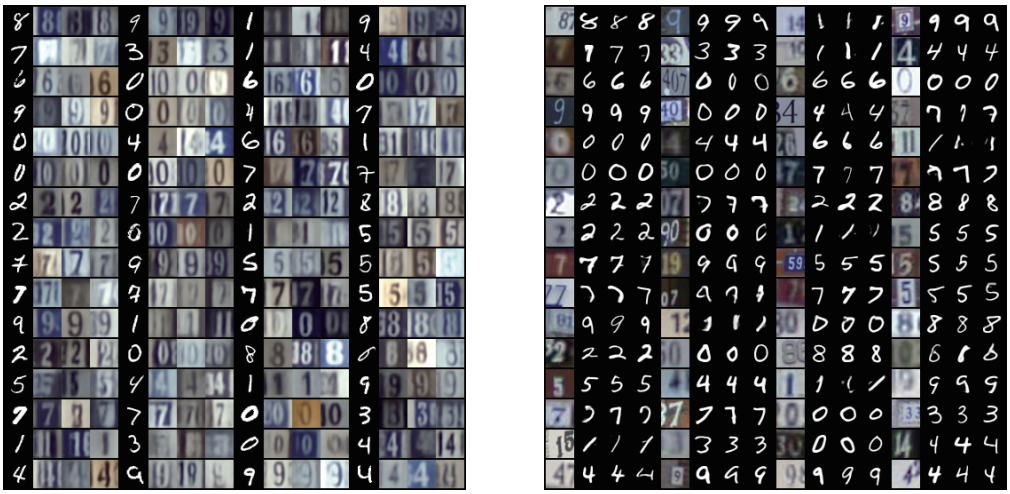


Figure 20: MNIST->SVHN (Left) and SVHN->MNIST (Right), when SVHN is observed 50% of the time.

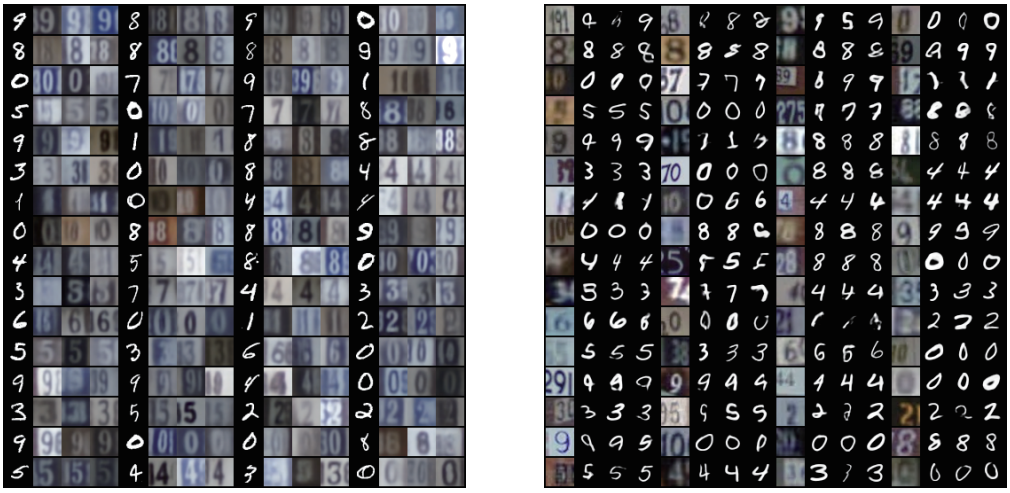


Figure 21: MNIST->SVHN (Left) and SVHN->MNIST (Right), when MNIST is observed 25% of the time.

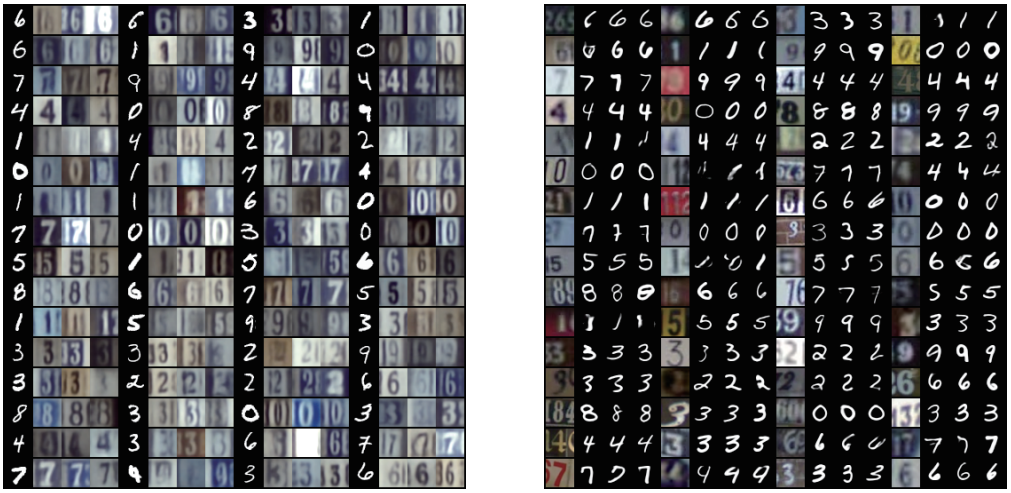


Figure 22: MNIST->SVHN (Left) and SVHN->MNIST (Right), when SVHN is observed 25% of the time.

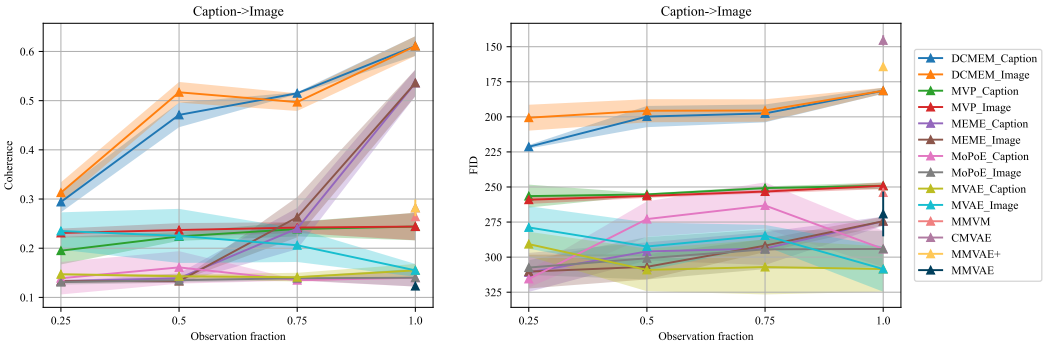


Figure 23: Classification accuracy and FID score of cross-modal generations under different missing rates on the CUBICC dataset. Shaded areas represent the standard deviation across multiple runs. The subscripts in method names indicate the observed modality. For example, DCMEM_Caption (Observation fraction = 0.25) denotes that the training data consists of 25% paired samples and 75% unimodal Caption samples.

such as MVAE and MMVAE display similar distance distributions for both paired and unpaired data, suggesting that their latent representations fail to encode meaningful semantic information. Although MoPoE and MEME capture a certain degree of semantic relatedness, as evidenced by the relatively small gap between paired and unpaired distributions, they achieve weaker semantic alignment compared to our model. MMVAE+, CMVAE and MMVM are only able to capture semantic differences on a single dataset with limited generalization capability. Although MVP shows a noticeable separation, the contrast between paired and unpaired distances is less pronounced than in our model. These comparisons further highlight the superior ability of our approach to learn semantically structured and modality-aligned latent representations.

C.4 Class-Contextual Relatedness in the Latent Space

To evaluate whether our model captures class-level semantic alignment across modalities, we conduct a class-contextual relatedness analysis on both the MNIST-SVHN and CUBICC datasets. Following the methodology proposed in MEME [10], we compute a class-conditioned distance matrix $K \in \mathbb{R}^{C\times C}$, where C is the number of classes in the dataset. Each entry K_{ij} represents the average 2-Wasserstein distance between the latent distributions of class i from one modality and class j from the other. Ideally, if the model successfully aligns class-level semantics across modalities, we expect the matrix to exhibit low distances along the diagonal (representing matched classes) and higher

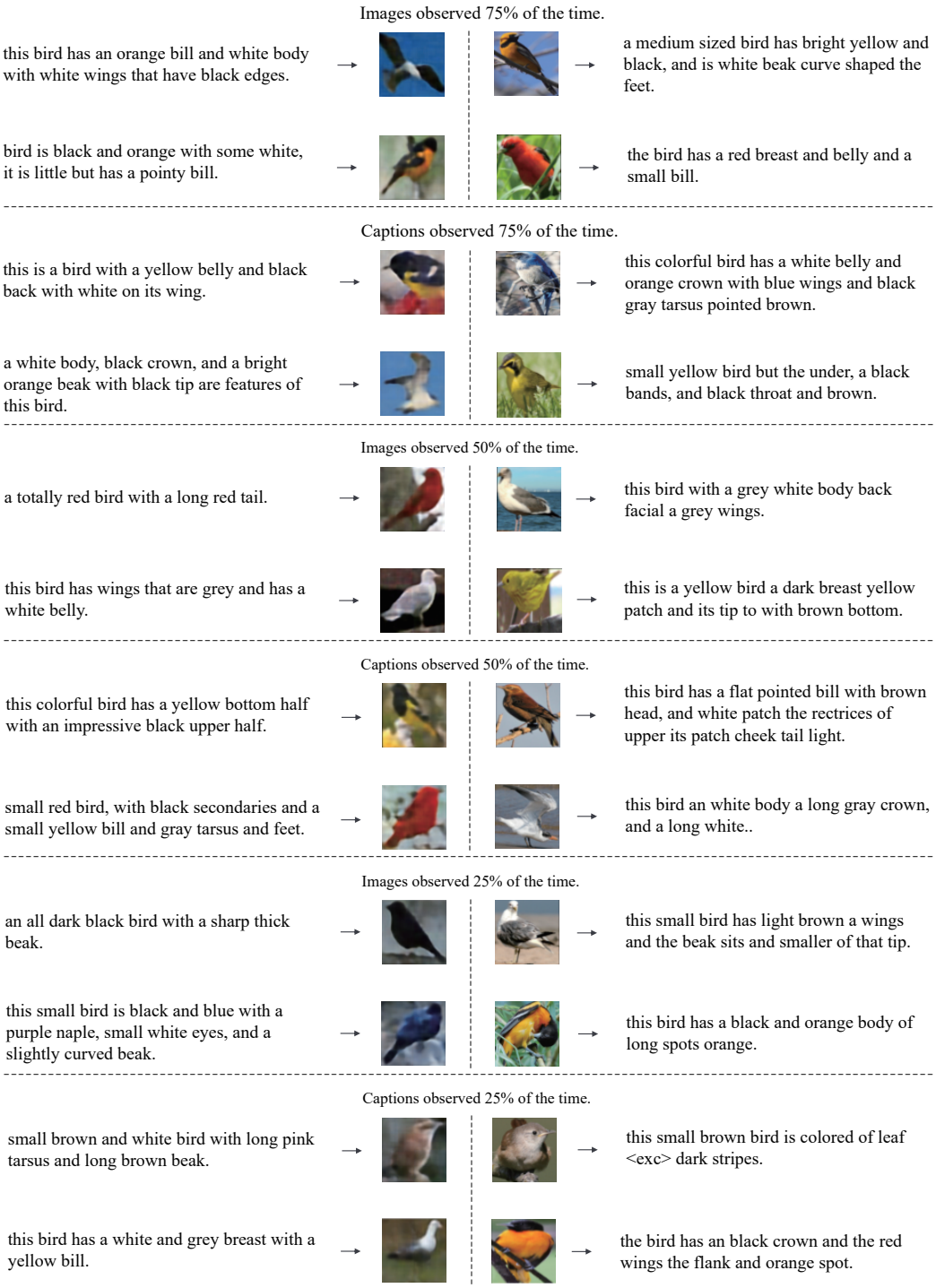
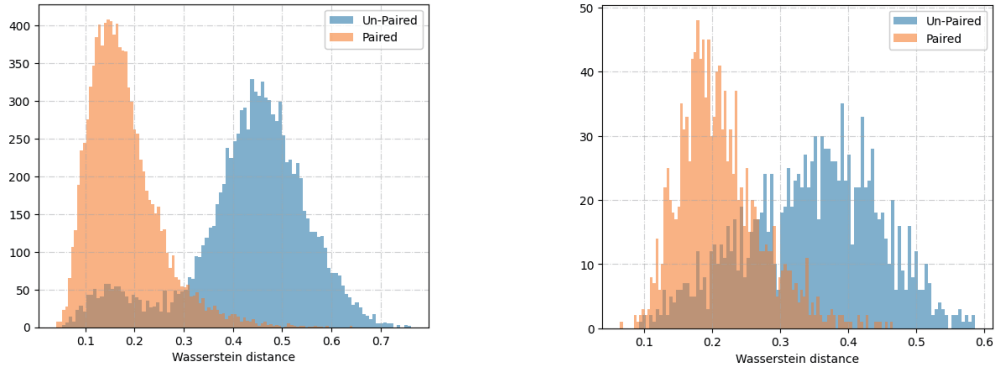


Figure 24: Cross-modal generations on the CUBICC dataset by DCMEM.



(a) MNIST-SVHN (b) CUBICC

Figure 25: Histograms of 2-Wasserstein distances between latent distributions for paired and unpaired multimodal samples obtained by DCMEM on the MNIST-SVHN and CUBICC datasets.

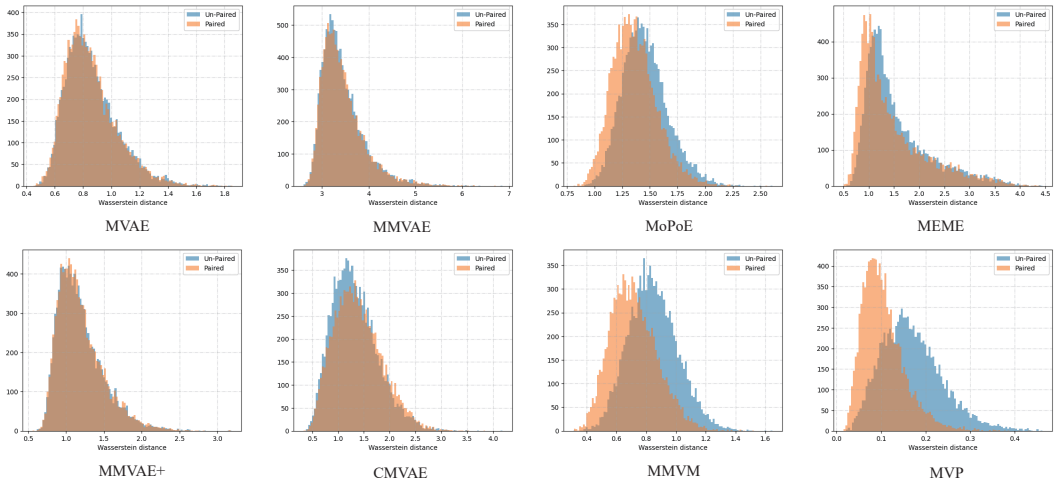


Figure 26: Histograms of 2-Wasserstein distances between latent distributions for paired and unpaired multimodal samples obtained by baseline models on the MNIST-SVHN dataset.

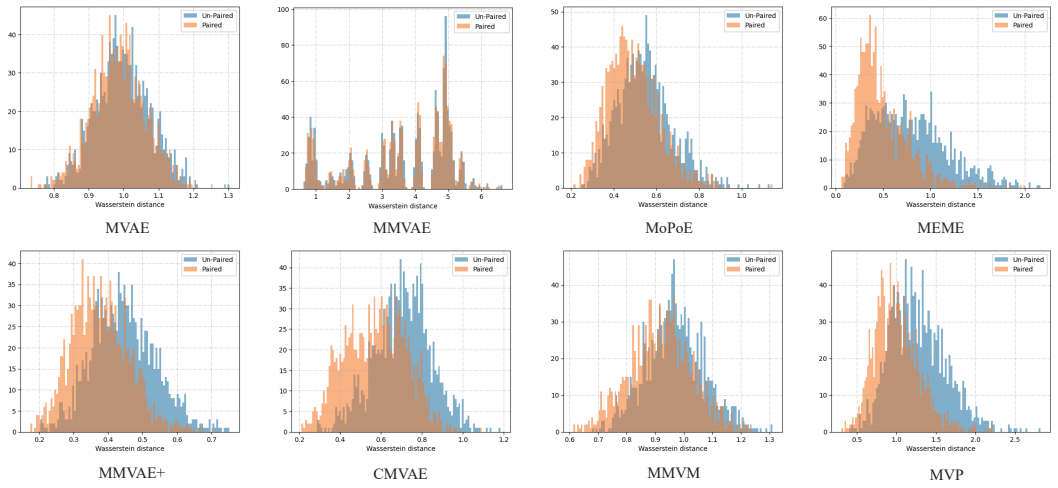
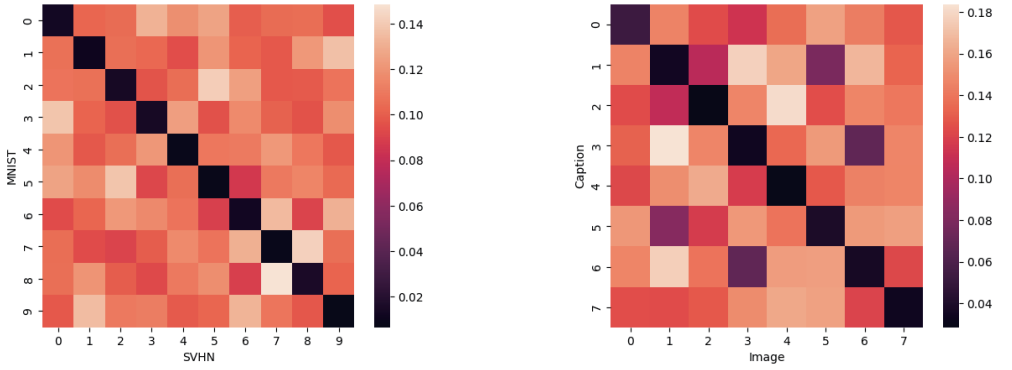


Figure 27: Histograms of 2-Wasserstein distances between latent distributions for paired and unpaired multimodal samples obtained by baseline models on the CUBICC dataset.



(a) MNIST-SVHN (b) CUBICC

Figure 28: Heatmaps of class-conditioned 2-Wasserstein distances between latent distributions obtained by DCMEM on the MNIST-SVHN and CUBICC datasets.

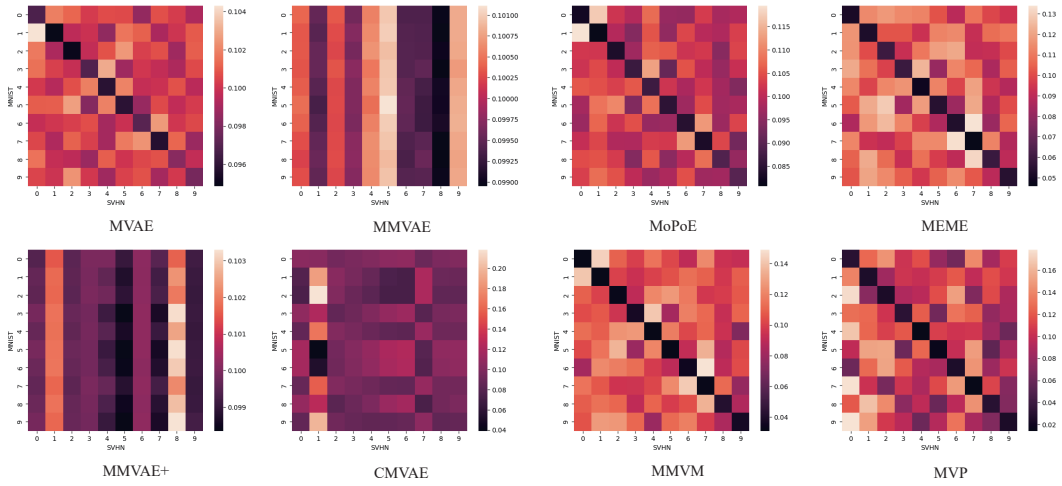


Figure 29: Heatmaps of class-conditioned 2-Wasserstein distances between latent distributions obtained by baseline models on the MNIST-SVHN dataset.

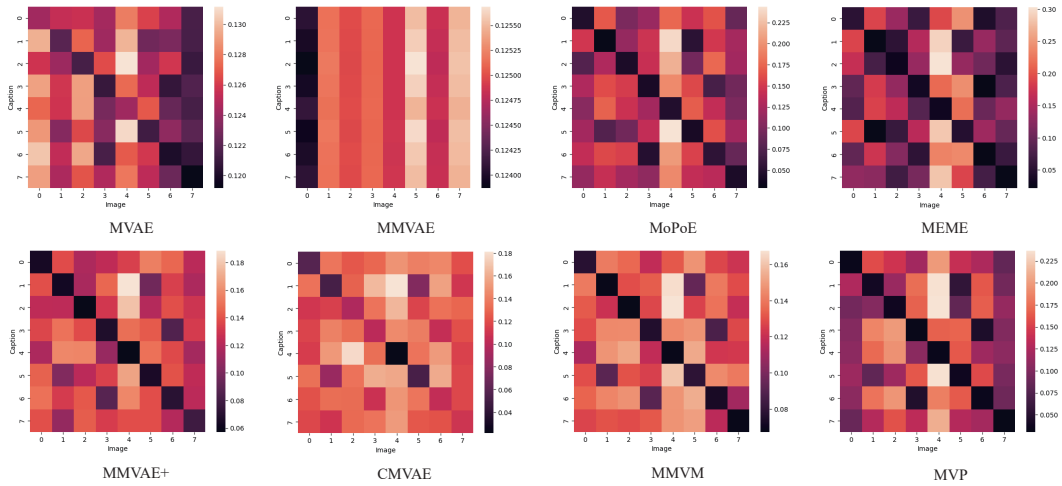


Figure 30: Heatmaps of class-conditioned 2-Wasserstein distances between latent distributions obtained by baseline models on the CUBICC dataset.

Table 3: Quantitative comparison of clustering performance based on latent representations under different missing rates in the MNIST modality on the MNIST-SVHN dataset. Fraction indicates the proportion of paired samples relative to the full training set. Each model is trained on a dataset consisting of paired data at a proportion of Fraction and unimodal SVHN data at a proportion of 1-Fraction, and evaluated on the complete test set.

Fraction	Methods	SVHN Representation			MNIST Representation			Joint Representation		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
0.75	MVAE	18.0	6.0	3.1	77.5	63.4	59.8	43.3	41.9	28.1
	MoPoE	27.0	13.4	8.3	50.3	41.4	29.0	61.8	52.4	42.7
	MEME	11.8	0.3	0.1	41.6	35.9	24.2	11.7	0.4	0.1
	MVP	56.4	54.6	39.9	61.3	60.3	45.7	37.8	37.6	22.1
	DCMEM	**87.4**	**73.8**	**74.0**	**98.7**	**96.3**	**97.2**	**99.5**	**98.4**	**98.9**
0.5	MVAE	18.0	6.0	3.2	70.6	56.0	52.1	43.3	41.8	27.7
	MoPoE	22.5	10.0	5.8	52.4	45.8	34.9	52.8	45.4	34.3
	MEME	11.7	0.3	0.1	41.8	35.8	24.6	11.6	0.3	0.1
	MVP	29.0	16.0	10.2	50.6	48.8	33.2	42.9	39.0	26.7
	DCMEM	**87.5**	**76.1**	**75.8**	**98.3**	**95.2**	**96.2**	**98.4**	**96.6**	**97.5**
0.25	MVAE	17.9	5.9	3.1	68.3	56.5	50.4	39.3	33.7	22.4
	MoPoE	18.6	6.1	3.3	54.3	43.5	31.8	42.5	35.9	23.9
	MEME	11.6	0.3	0.1	45.8	36.9	27.0	11.7	0.3	0.1
	MVP	12.3	0.6	0.1	39.1	42.0	24.8	12.3	0.6	0.1
	DCMEM	**62.0**	**47.2**	**38.4**	**92.3**	**83.4**	**90.3**	**92.3**	**83.4**	**84.0**

Table 4: Quantitative comparison of clustering performance based on latent representations under different missing rates in the SVHN modality on the MNIST-SVHN dataset. Fraction indicates the proportion of paired samples relative to the full training set. Each model is trained on a dataset consisting of paired data at a proportion of Fraction and unimodal MNIST data at a proportion of 1-Fraction, and evaluated on the complete test set.

Fraction	Methods	SVHN Representation			MNIST Representation			Joint Representation		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
0.75	MVAE	17.2	4.8	2.6	82.4	68.0	65.7	40.3	34.6	23.9
	MoPoE	32.1	21.6	13.4	53.8	47.2	35.5	61.1	52.4	42.9
	MEME	13.0	3.8	2.2	39.3	36.8	24.1	13.4	6.6	3.3
	MVP	49.0	31.0	23.9	78.1	71.0	64.9	75.9	70.9	64.0
	DCMEM	**89.0**	**76.2**	**77.2**	**99.1**	**97.3**	**98.0**	**99.7**	**99.1**	**99.4**
0.5	MVAE	16.4	5.5	2.4	79.6	65.4	62.5	38.6	38.4	24.7
	MoPoE	31.1	18.5	11.6	54.8	47.2	36.3	59.4	49.6	41.9
	MEME	14.5	3.3	1.6	39.6	33.5	22.1	12.3	0.6	0.2
	MVP	49.7	33.1	24.5	83.3	72.9	69.9	77.4	73.5	67.3
	DCMEM	**85.5**	**72.4**	**72.7**	**98.7**	**96.3**	**97.1**	**99.6**	**98.8**	**99.1**
0.25	MVAE	17.1	5.4	2.7	76.9	63.7	60.6	38.3	31.1	20.7
	MoPoE	25.7	13.0	7.4	65.3	53.0	46.3	54.6	44.0	34.3
	MEME	15.8	4.9	2.6	51.1	43.0	34.4	20.5	13.0	7.7
	MVP	47.6	49.5	33.5	85.7	71.3	68.4	87.2	76.9	76.6
	DCMEM	**82.6**	**65.6**	**65.4**	**95.2**	**88.4**	**89.6**	**98.5**	**96.0**	**96.8**

values off-diagonal (mismatched classes). To visualize this, we present the resulting matrices as heatmaps in Figure 28, where darker colors indicate smaller distances. The baseline results are shown in Figures 29 and 30.

As shown in Figure 28, our model produces a clear diagonal structure in the class-conditioned distance matrices, indicating that it effectively aligns semantically corresponding classes across modalities. This pattern is consistently observed on both the MNIST-SVHN and CUBICC datasets, suggesting robust class-level semantic alignment in the learned latent space. In comparison, baseline models such as MVAE, MMVAE+ and CMVAE fail to exhibit a clear diagonal on at least one of the datasets, revealing their limited ability to consistently model class-level correspondence. Other models, including MEME and MVP, either produce diagonals with less pronounced contrast or show

Table 5: Quantitative comparison of clustering performance based on latent representations under different missing rates in the Image modality on the CUBICC dataset. Fraction indicates the proportion of paired samples relative to the full training set. Each model is trained on a dataset consisting of paired data at a proportion of Fraction and unimodal Caption data at a proportion of 1-Fraction, and evaluated on the complete test set.

Fraction	Methods	Image Representation			Caption Representation			Joint Representation		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
0.75	MVAE	27.2	11.8	7.0	17.8	2.1	0.7	36.9	27.7	18.7
	MoPoE	38.1	28.4	17.4	50.8	37.0	27.5	58.5	46.6	35.2
	MEME	45.8	42.1	25.0	27.6	16.3	8.5	45.0	37.1	24.1
	MVP	53.1	46.6	34.8	44.3	30.2	21.6	72.7	61.7	51.1
	DCMEM	**83.6**	**72.4**	**66.3**	**62.6**	**43.9**	**35.0**	**84.1**	**73.5**	**67.0**
0.5	MVAE	27.0	14.3	7.7	17.7	2.4	0.9	36.4	22.1	13.6
	MoPoE	32.6	20.2	12.4	46.2	26.6	17.0	42.1	32.6	23.4
	MEME	20.4	4.8	2.4	16.9	1.4	0.2	19.8	4.2	1.8
	MVP	58.7	45.4	36.4	31.0	18.5	13.0	52.4	44.2	36.4
	DCMEM	**84.3**	**75.7**	**68.0**	**51.3**	**37.2**	**28.0**	**79.9**	**68.5**	**60.3**
0.25	MVAE	28.2	13.1	7.9	27.0	12.0	7.3	30.5	17.4	10.5
	MoPoE	19.4	5.1	2.4	25.4	11.6	6.1	26.9	10.2	5.8
	MEME	16.9	1.7	0.2	16.1	1.0	0.1	17.2	1.8	0.3
	MVP	41.7	30.1	20.4	20.6	6.5	4.2	31.4	20.9	17.7
	DCMEM	**73.6**	**62.0**	**51.3**	**53.1**	**39.8**	**32.6**	**82.7**	**70.9**	**64.3**

Table 6: Quantitative comparison of clustering performance based on latent representations under different missing rates in the Caption modality on the CUBICC dataset. Fraction indicates the proportion of paired samples relative to the full training set. Each model is trained on a dataset consisting of paired data at a proportion of Fraction and unimodal Image data at a proportion of 1-Fraction, and evaluated on the complete test set.

Fraction	Methods	Image Representation			Caption Representation			Joint Representation		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
0.75	MVAE	29.0	12.1	7.5	22.6	11.3	3.8	33.7	26.9	15.1
	MoPoE	27.9	15.8	7.8	38.2	23.0	15.2	42.0	27.9	17.0
	MEME	24.5	9.1	4.7	24.8	15.7	6.5	24.5	16.2	7.9
	MVP	58.8	48.2	37.2	42.9	30.1	21.6	52.3	50.5	36.7
	DCMEM	**82.0**	**70.5**	**63.2**	**60.8**	**44.7**	**33.9**	**83.9**	**74.5**	**67.2**
0.5	MVAE	28.0	11.3	6.4	20.4	6.3	3.1	46.2	33.1	23.0
	MoPoE	27.8	13.1	8.0	33.0	19.3	12.0	31.0	18.9	11.6
	MEME	17.0	1.3	0.3	16.7	1.4	0.2	17.1	1.6	0.3
	MVP	50.0	36.0	27.1	45.8	27.4	19.0	64.1	50.9	41.8
	DCMEM	**81.0**	**68.4**	**61.9**	**57.2**	**38.8**	**29.5**	**82.0**	**70.1**	**63.4**
0.25	MVAE	29.5	17.3	8.9	21.2	6.0	2.9	35.9	24.6	15.7
	MoPoE	24.5	7.6	4.5	25.0	10.3	5.4	27.6	11.7	7.0
	MEME	15.9	0.8	0.1	16.4	1.1	0.1	16.3	1.0	0.1
	MVP	51.1	33.4	25.8	40.1	24.1	14.7	54.5	41.8	31.1
	DCMEM	**65.6**	**55.2**	**42.2**	**54.7**	**40.4**	**29.3**	**67.4**	**62.9**	**50.4**

undesirably low distances in off-diagonal entries, which implies confusion between unrelated classes. Notably, only our model consistently achieves a strong diagonal with low intra-class distances and high inter-class distances across both datasets, highlighting its superior capability in capturing and preserving cross-modal semantic structure.

C.5 Ablation Study

The explicit mathematical definitions of the objective functions used in the ablation study are as follows:

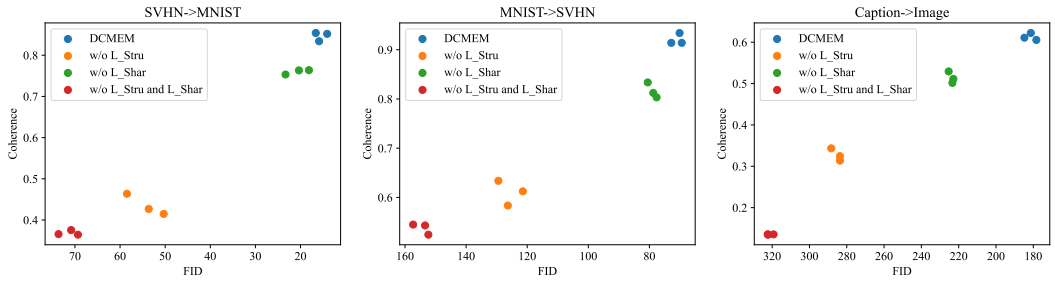


Figure 31: Generation performance with different modules ablated. $\mathcal{L}_{Stru}$ represents the structured representation learning module and $\mathcal{L}_{Shar}$ represents the shared representations alignment module.

Table 7: Clustering performance of joint latent representations with different modules ablated. $\mathcal{L}_{Stru}$ represents the structured representation learning module and $\mathcal{L}_{Shar}$ represents the shared representations alignment module.

Datasets			MNIST_SVHN			CUBICC		
$\mathcal{L}_{ELBO}$	$\mathcal{L}_{Stru}$	$\mathcal{L}_{Shar}$	ACC	NMI	ARI	ACC	NMI	ARI
✓			50.6±3.1	32.4±2.1	27.4±2.2	26.4±0.7	13.7±0.2	10.3±0.1
✓	✓		91.5±2.4	80.6±1.8	82.0±1.2	84.6±1.8	74.1±0.9	68.6±2.1
✓		✓	72.7±1.9	63.5±1.3	57.7±1.2	49.3±1.3	50.7±1.9	35.0±0.9
✓	✓	✓	**99.5±0.1**	**98.4±0.2**	**98.9±0.1**	**86.3±1.8**	**76.8±2.8**	**71.5±3.1**

(1) $\mathcal{L}_{\text{ELBO}}$ denotes the variational lower bound derived in Section 3.1. Under our mutual supervision setup, it includes both $s \to z \to t$ and $t \to z \to s$ directions. For the $s \to z \to t$ direction, the ELBO term is given by: $\mathcal{L}^{s\to t}_{\text{ELBO}} = \mathbb{E}_{q_\phi(z_s,w_s|s)}\left[\log \frac{p_{\psi_z}(z_s|t)p(w_s)p_\theta(s|z_s,w_s)}{q_{\phi_z}(z_s|s)q_{\phi_w}(w_s|s)}\right] + \log q_{\phi_z,\phi}(t \mid s) + \log p(t)$. A symmetric term is used for the $t \to z \to s$ direction. Together, they form the total $\mathcal{L}_{\text{ELBO}}$ used in training.

(2) $\mathcal{L}_{\text{Stru}}$ corresponds to the Structured Representation Learning term introduced in Section 3.1. It also includes bidirectional modeling. For example, the $s \to z \to t$ direction includes a reconstruction term and a latent distribution alignment term: $\mathcal{L}^{s\to t}_{\text{Stru}} = \mathbb{E}_{q_\phi(\boldsymbol{z}_s,\boldsymbol{w}_s|\boldsymbol{s})p_D(\boldsymbol{s})} \log p_\theta\left(\boldsymbol{s} \mid \boldsymbol{z}_s, \boldsymbol{w}_s\right) - \mathbb{E}_{p_D(\boldsymbol{s},\boldsymbol{t})}\left[D_{KL}\left(q_{\phi_z}\left(\boldsymbol{z}_s = \boldsymbol{z} \mid \boldsymbol{s}\right) || q_{\psi_z}\left(\boldsymbol{z}_t = \boldsymbol{z} \mid \boldsymbol{t}\right)\right)\right]$ and vice versa for the $t \to z \to s$ direction.

(3) $\mathcal{L}_{\text{Shar}}$ corresponds to the Shared Representations Alignment term introduced in Section 3.1. It captures the mutual information between z_s and z_t, defined as: $\mathcal{L}_{\text{Shar}} = \alpha I(z_s; z_t)$, where $I(\cdot;\cdot)$ is estimated via contrastive learning.

To evaluate the contribution of each component in our model, we perform an ablation study by selectively removing the structured representation learning ($\mathcal{L}_{Stru}$) and the shared representations alignment ($\mathcal{L}_{Shar}$). The results are summarized in Figure 31 and Table 7. As shown in Figure 31, we evaluate the contribution of each module to generation performance using FID and coherence scores across three tasks. The full model consistently achieves the lowest FID and highest coherence scores, indicating superior visual fidelity and semantic consistency. Removing either $\mathcal{L}_{Stru}$ or $\mathcal{L}_{Shar}$ leads to a clear decline in performance, confirming the necessity of both components. Table 7 reports the clustering performance of different ablation settings on MNIST-SVHN and CUBICC datasets. We observe that omitting $\mathcal{L}_{Shar}$ results in a modest performance drop as the model loses the alignment constraint for shared latent features, leading to suboptimal cross-modal representation. In contrast, removing $\mathcal{L}_{Stru}$ causes a significant decline in all metrics. This suggests that without proper disentanglement of shared and modality-specific information, the model fails to preserve meaningful semantic structure in the shared space. Overall, the ablation results demonstrate that both structured representation learning and shared representations alignment are indispensable for achieving strong performance in both generation and clustering tasks.

To further isolate our architectural contribution, we conduct experiments on the CUBICC dataset by enhancing MVP with the VampPrior mechanism. MVP is selected as the strongest non-VampPrior baseline in terms of cross-modal generation and its competitive performance in clustering and classification under various pairing rates (Figures 12, 23; Tables 5, 6). The resulting variant, MVP_VP,

Table 8: Generation performance of MVP_VP (The subscript of each metric indicates the observed modality. For example, Coherence_Image (Fraction = 0.25) denotes that the training data consists of 25% paired samples and 75% unimodal Image samples).

Fraction	Methods	Coherence_Image	FID_Image	Coherence_Caption	FID_Caption
0.75	MVP	0.242	253.243	0.239	250.773
	MVP_VP	0.241	265.464	0.234	251.384
	DCMEM	**0.497**	**203.981**	**0.515**	**207.708**
0.5	MVP	0.237	256.383	0.224	255.310
	MVP_VP	0.215	260.604	0.223	259.447
	DCMEM	**0.517**	**204.286**	**0.471**	**212.218**
0.25	MVP	0.231	259.064	0.195	256.498
	MVP_VP	0.217	250.562	0.187	271.769
	DCMEM	**0.313**	**214.203**	**0.294**	**221.351**

Table 9: Classification accuracy of MVP_VP (The subscript of each metric indicates the observed modality. For example, Image Representation_Image (Fraction = 0.25) denotes that the training data consists of 25% paired samples and 75% unimodal Image samples).

Fraction	Methods	Image Repres entation_Image	Caption Repres entation_Image	Image Repres entation_Caption	Caption Repres entation_Caption
0.75	MVP	0.864	0.561	0.877	0.566
	MVP_VP	0.824	0.512	0.804	0.532
	DCMEM	**0.866**	**0.631**	**0.873**	**0.645**
0.5	MVP	0.865	0.557	0.825	0.537
	MVP_VP	0.754	0.520	0.755	0.485
	DCMEM	**0.871**	**0.618**	**0.879**	**0.620**
0.25	MVP	0.702	0.518	0.676	0.458
	MVP_VP	0.635	0.453	0.647	0.386
	DCMEM	**0.712**	**0.526**	**0.717**	**0.576**

uses pseudo-points from the missing modality to construct a Gaussian mixture prior, which guides latent learning from the observed modality. As shown in the Tables 8, 9 and 10, MVP_VP does not yield consistent improvements over the original MVP baseline. On the contrary, it often leads to a degradation in generation metrics, classification accuracy and clustering performance, particularly at lower pairing rates. We hypothesize that this is due to a mismatch in modeling assumptions: MVP relies on cycle-consistency alignment, which degenerates to a trivial alignment (i.e., with itself) when only one modality is present, yielding zero loss for such cases. The introduction of VampPrior forces these unpaired samples to align with a prior constructed from the missing modality, introducing a non-trivial loss term that may disrupt the overall optimization, especially since the alignment does not follow the same cyclic mechanism as MVP's original design. Apart from these results, it is worth noting that two baseline models, MMVM and MEME , which use a similar VampPrior strategy, also underperform compared to our model. This further indicates that our performance gains stem not only from the use of VampPrior, but from the integration of disentangled representation learning and mutual information alignment within a unified mutual supervision framework, which ensures consistent robustness under both paired and missing data scenarios.

C.6 Parameter Analysis

To evaluate the impact of the shared representations alignment component, we conduct a parameter analysis on its weighting factor α. As illustrated in Figure 32, we plot the FID and Coherence scores under different values of α across three cross-modal generation tasks. The results reveal that the model achieves optimal performance when α is set to 0.5 or 1. In this range, the alignment module effectively bridges modality gaps by aligning latent representations, thereby preserving semantic consistency and enhancing generation quality. In contrast, when α is too small, the alignment term contributes minimally and results in suboptimal cross-modal coherence. On the other hand, excessively large α values may lead to overfitting or over-alignment which adversely affects performance. In addition, Table 11 reports the clustering performance of joint latent representations under different α values. We observe that the model maintains strong performance when α ranges from 0.1 to 1. However,

Table 10: Clustering accuracy of MVP_VP (The subscript of each metric indicates the observed modality. For example, Image Representation_Image (Fraction = 0.25) denotes that the training data consists of 25% paired samples and 75% unimodal Image samples).

Fraction	Methods	Image Repres entation_Image	Caption Repres entation_Image	Joint Repres entation_Image
0.75	MVP	58.8	42.9	52.3
	MVP_VP	29.4	32.2	34.5
	DCMEM	**82.0**	**60.8**	**83.9**
0.5	MVP	50.0	45.8	64.1
	MVP_VP	40.2	36.7	44.2
	DCMEM	**81.0**	**57.2**	**82.0**
0.25	MVP	51.1	40.1	54.5
	MVP_VP	25.0	25.4	27.8
	DCMEM	**65.6**	**54.7**	**67.4**

Fraction	Methods	Image Repres entation_Caption	Caption Repres entation_Caption	Joint Repres entation_Caption
0.75	MVP	53.1	44.3	72.7
	MVP_VP	41.9	32.3	39.6
	DCMEM	**83.6**	**62.6**	**84.1**
0.5	MVP	58.7	31.0	52.4
	MVP_VP	37.8	34.0	42.9
	DCMEM	**84.3**	**51.3**	**79.9**
0.25	MVP	41.7	20.6	31.4
	MVP_VP	24.8	19.9	22.7
	DCMEM	**73.6**	**53.1**	**82.7**

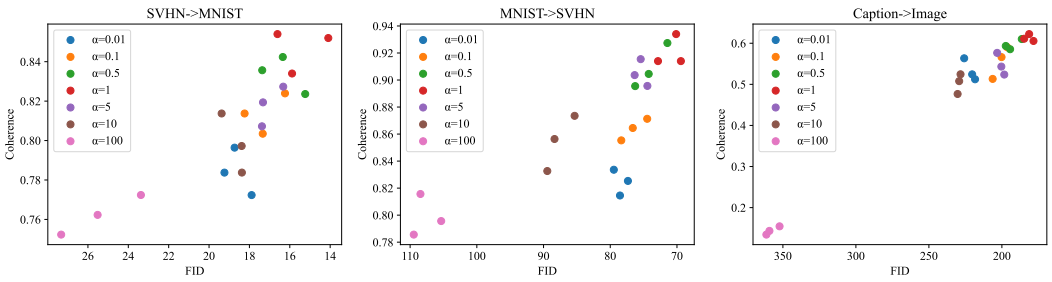


Figure 32: Generation performance under different α values.

Table 11: Clustering performance of joint latent representations under different α values.

Datasets	MNIST_SVHN			CUBICC		
α	ACC	NMI	ARI	ACC	NMI	ARI
0.01	94.4±1.2	90.0±1.0	91.4±1.5	84.5±2.1	74.1±1.7	68.6±1.7
0.1	97.9±0.3	95.8±0.2	96.7±0.3	**86.7±1.1**	75.7±1.4	69.8±1.9
0.5	99.1±0.2	**98.6±0.4**	97.9±0.3	85.6±1.6	76.4±1.3	71.2±1.5
1	**99.5±0.1**	98.4±0.2	**98.9±0.1**	86.3±1.8	**76.8±2.8**	**71.5±3.1**
5	96.5±1.1	91.1±1.3	92.4±1.2	83.4±1.4	72.3±0.8	66.6±1.6
10	92.3±1.9	82.1±1.4	83.8±1.7	75.0±1.3	69.7±0.7	59.3±0.9
100	85.9±1.7	72.6±1.4	71.8±1.8	67.0±1.1	49.3±1.4	40.4±1.3

when α exceeds this range, the clustering performance degrade significantly, further confirming the importance of a well-balanced alignment strength. Based on both generation and clustering results, we recommend setting α between 0.5 and 1 in practice for robust and consistent performance.

C.7 Computational Resources

All experiments are conducted on a machine equipped with an NVIDIA GeForce RTX 2080 Ti GPU and 64 GB of RAM. For the MNIST-SVHN dataset, each run uses 4 CPU workers and approximately 10 GB of GPU memory, with an average training time of around 8 hours per run. We evaluate 9 different methods, each with 3 random seeds, resulting in a total compute time of approximately 216 GPU hours ($8 \times 9 \times 3$). For the CUBICC dataset, each run uses 2 CPU workers and approximately 9 GB of GPU memory. Each training run takes about 16 hours on average. Evaluating 9 methods over 3 seeds results in a total compute time of roughly 432 GPU hours ($16 \times 9 \times 3$). For the Human Breast Cancer dataset, each run uses 2 CPU workers and around 4 GB of GPU memory. The average runtime is approximately 2 hours. With 14 methods and 3 seeds, the total compute time amounts to about 84 GPU hours ($2 \times 14 \times 3$). In total, the experiments require approximately 732 GPU hours. Additional GPU time is used during model development and hyperparameter tuning, which is not included in the above calculation.