FLOW-EMBEDDED MULTIMODAL VARIATIONAL AUTOENCODERS FOR INCOMPLETE MULTI-VIEW CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Incomplete multi-view clustering aims to uncover the underlying cluster structure from data observed across multiple views with missing modalities. Recently, generative modeling has emerged as a promising paradigm for this task, as it provides a principled framework for capturing complex data distributions and inferring unobserved information from the available views. In this paper, we propose **F**low-**E**mbedded **M**ultimodal **VA**riational auto**E**ncoders (FEMVAE) with Gaussian mixture priors to learn expressive latent distributions and a coherent shared clustering structure from incomplete multi-view data. Specifically, we integrate normalizing flows into variational autoencoders to construct flexible posterior distributions within a unified framework, while leveraging a Gaussian mixture prior to impose a cluster-aware structure on the learned latent space. Moreover, to fully align the view-specific latent representations, we employ a pairwise optimal transport strategy to promote distributional consistency across views. This enables the complementary information from different views to be effectively integrated despite missing modalities. Once trained, our model can naturally recover missing data modalities. Extensive experiments on several benchmark incomplete multi-view datasets demonstrate the effectiveness of the proposed method, which consistently achieves competitive clustering performance against state-of-the-art approaches under different missing-view settings.

1 INTRODUCTION

Multi-view clustering exploits cross-view dependencies and complementary information to uncover latent cluster structure, allowing ambiguities within individual views to be resolved through evidence from other views (Dong et al., 2023; Fang et al., 2023; Liu et al., 2026). However, missing views limit the availability of cross-view correspondences, hindering the joint modeling of heterogeneous observations. Moreover, variations in view availability across samples introduce discrepancies in the information available for representation learning, complicating the construction of a unified latent space that preserves cluster structure. Incomplete multi-view clustering (IMVC) seeks to address these challenges by integrating the available observations to infer a consistent clustering structure under partial view availability (Zhong et al., 2025; Yin et al., 2025; Xu et al., 2026).

Existing IMVC methods can be broadly categorized into non-imputation-based and imputation-based approaches. Non-imputation-based methods learn clustering structures directly from available observations. Matrix factorization and subspace learning methods capture shared latent structures through observation masks or cross-view correspondences (Hu & Chen, 2018; Li et al., 2014). Graph-based methods integrate observed sample relationships into consensus structures (Qin et al., 2025; Wang et al., 2022), while deep representation learning methods employ adaptive fusion, semantic consistency, or knowledge distillation to accommodate varying view availability (Yu et al., 2022; Xu et al., 2023; 2024; Wu et al., 2025; Dai et al., 2025). Although these methods avoid imputation errors, they do not explicitly recover the complementary information carried by missing views. When the remaining views provide insufficient evidence to distinguish clusters, consistency learning and fusion alone may leave ambiguities in sample assignments unresolved.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

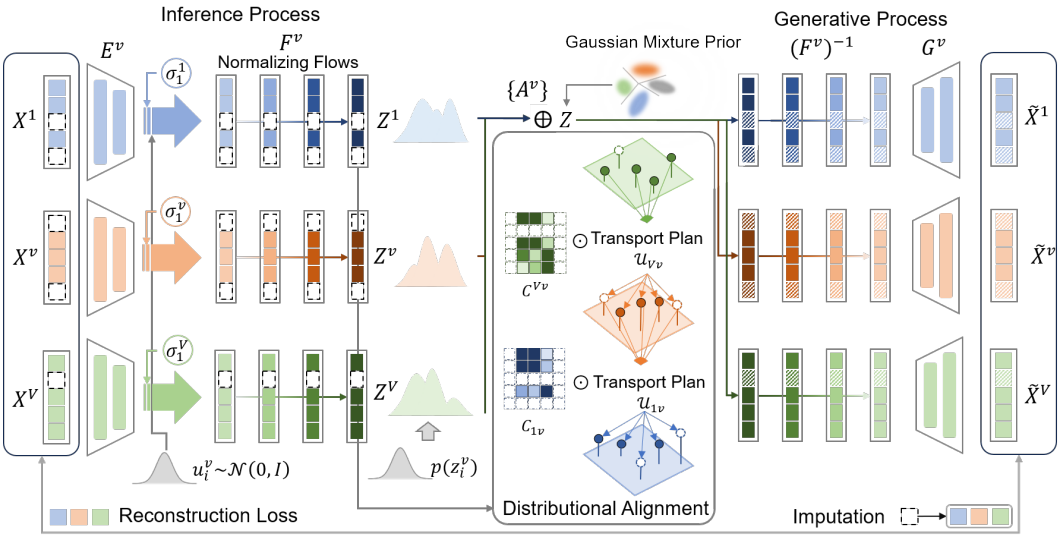

Figure 1: Overview of the proposed framework. Given incomplete multi-view data $X^v * v = 1^V$, each encoder E^v maps the available observations to low-dimensional embeddings, which are transformed into view-specific latent representations as $Z^v = F^v(E^v(X^v) + \sigma_1^v U^v)$, where $\boldsymbol{u}_i^v \sim \mathcal{N}(0, \mathbf{I})$ and F^v acts row-wise. The resulting posteriors are regularized toward $p(\boldsymbol{z}_i^v) = \mathcal{N}(\boldsymbol{z}_i^v \mid 0, \mathbf{I})$, with missing rows in Z^v set to zero. The view-specific representations are adaptively fused as $Z = \sum * v \in \mathcal{V} A^v Z^v$, where $A^v = \mathrm{diag}(\alpha_1^v, \ldots, \alpha_N^v)$, $\alpha_i^v = 0$ for missing views, and $\sum_{v \in \mathcal{V}} \alpha_i^v = 1$. A Gaussian mixture prior and pairwise distributional consistency constraint respectively promote cluster structure and cross-view coherence. Finally, $\widetilde{X}^v = G^v((F^v)^{-1}(Z))$ reconstructs observed views and imputes missing views, shown as solid and mosaic-patterned blocks, respectively.

Imputation-based methods recover missing information to support clustering. Matrix and kernel completion methods exploit low-rank structure and cross-view dependencies (Liu et al., 2017; Xing et al., 2025), but restrictive structural assumptions can limit their ability to capture complex nonlinear relationships. Cross-view prediction methods estimate absent features or representations through learned mappings (Lin et al., 2021; Jiang et al., 2025; Wang et al., 2025b); however, deterministic predictions do not characterize the multiple plausible outcomes that may arise when observed views cannot uniquely determine missing ones. Relation- and prototype-guided methods use neighborhoods or cluster representatives to guide completion (Zhuang et al., 2025; Wang et al., 2025a; Du et al., 2026; Yuan et al., 2024; Zhang et al., 2026b), making their estimates sensitive to unreliable neighborhoods or inaccurate preliminary cluster assignments. Generative methods formulate missing-view recovery as a distribution modeling problem. Existing approaches employ adversarial learning (Xu et al., 2019; Wang et al., 2021), variational inference (Cai et al., 2024; Gao & Pu, 2025), or diffusion models (Wen et al., 2024; Zhang et al., 2025; 2026a) to generate missing views. However, these methods do not explicitly address how to model and propagate the latent uncertainty induced by incomplete observations. Adversarial methods focus on distribution matching, variational methods may be limited by posterior parameterization, while diffusion methods rely on iterative generation. Normalizing flows (NFs) enable expressive distribution modeling with tractable density evaluation through invertible transformations (Rezende & Mohamed, 2015; Dinh et al., 2017; Su & Wu, 2018; Selvan et al., 2020; Li et al., 2025). DiCMoR (Wang et al., 2023) and BURG (Jin et al., 2025) employ NFs for cross-view distribution alignment and missing-view recovery. However, their focus on feature alignment and transfer leaves the potential of NFs to enhance view-specific, observation-conditioned posteriors for latent inference underexplored.

To address these challenges, we propose a unified flow-based variational framework FEMVAE for incomplete multi-view clustering. The overall architecture is illustrated in Figure 1. First, for each available view, we embed a view-specific normalizing flow into the VAE posterior inference process, allowing the posterior to capture more expressive view-specific latent distributions while retaining

a tractable density evaluation. Second, we adaptively aggregate the available view-level posterior samples into a unified latent representation. We impose a cluster-oriented Gaussian mixture prior on this representation to encourage cluster structure and introduce a pairwise distributional consistency constraint between each pair of view-specific posteriors to promote cross-view semantic agreement. These components are jointly optimized by maximizing the ELBO together with the pairwise distributional alignment term. Finally, we pass the unified representation through the inverse flows corresponding to the missing views to reconstruct the unobserved data. Our work makes the following key contributions:

- We propose a unified framework that integrates normalizing flows with variational autoencoders to learn more expressive posterior distributions.
- A pairwise optimal transport strategy is introduced to align the latent distributions of different views, promoting cross-view distributional consistency and enabling effective integration of complementary information even when certain modalities are missing.
- Extensive experimental results demonstrates that our model achieves competitive clustering performance across multiple benchmark datasets and different missing view settings.

2 METHODOLOGY

2.1 NOTATIONS

An incomplete multi-view dataset is denoted as $\mathcal{X} = \{\mathbf{X}^v\}_{v=1}^V$, where $\mathbf{X}^v \in \mathbb{R}^{N_v \times D_v}$ denotes the v-th view with N_v samples, each of which is represented in a D_v-dimensional space. The dataset contains a total of N samples across all views. The i-th sample in view v is denoted as $\boldsymbol{x}_i^v \in \mathbb{R}^{D_v}$ and $\boldsymbol{x}^v$ is used to denote a generic sample when the sample index is clear from the context. Accordingly, the multi-view representation of a sample is denoted by $\{\boldsymbol{x}^v\}_{v=1}^V$, where some views may be unavailable due to missing observations. The availability of each view is represented by a binary view indicator matrix $\mathbf{M} \in \{0,1\}^{N \times V}$, where M_{iv} denotes whether the v-th view of the i-th sample is observed, with $M_{iv} = 1$ for an observed view and $M_{iv} = 0$ for a missing view. We assume that each sample has at least one observed view, i.e., $\sum_{v=1}^V M_{iv} \geq 1$ for all $i \in \{1, \ldots, N\}$. The identity matrix is denoted by $\mathbf{I}$. The goal of incomplete multi-view clustering (IMVC) is to partition all N samples into K clusters by leveraging the information from the available views despite the presence of missing views.

2.2 FLOW-BASED INFERENCE WITH A GAUSSIAN MIXTURE PRIOR

To learn a unified latent representation from the available views of incomplete multi-view data, we formulate a variational framework that jointly models view-level representations and their integration into a unified latent space. Let $\boldsymbol{z}^v$ denote the latent representation associated with the v-th available view $\boldsymbol{x}^v$ and $\boldsymbol{z}$ denote the unified representation obtained by integrating the available view-level representations.

Inference Process. Given the aforementioned assumption, we define the corresponding inference processes as follows:

$$q(\boldsymbol{z}, \{\boldsymbol{z}^v\} \mid \{\boldsymbol{x}^v\}) = q(\boldsymbol{z} \mid \{\boldsymbol{z}^v\}) \prod_{v \in \mathcal{V}} q(\boldsymbol{z}^v \mid \boldsymbol{x}^v), \tag{1}$$

where $q(\boldsymbol{z}^v \mid \boldsymbol{x}^v)$ denotes the posterior distribution of the latent representation inferred from the v-th available view, while $q(\boldsymbol{z} \mid \{\boldsymbol{z}^v\})$ denotes the integration process that maps the available view-level representations into a unified latent space. $M_{iv} \in \{0,1\}$ indicates whether $\boldsymbol{x}_i^v$ is observed and $\mathcal{V}_i = \{v \in \{1, \ldots, V\} : M_{iv} = 1\}$ denotes the set of available views for sample i. For a generic sample, we use $\mathcal{V}$ to denote its set of available views. Traditional variational autoencoders typically approximate latent-variable posteriors with simple Gaussian distributions, which may limit their ability to capture complex latent distributions. In contrast, normalizing flows employ a sequence of invertible transformations to flexibly transform a simple base distribution and explicitly evaluate the resulting probability density via the change-of-variables formula. However, most existing methods treat normalizing flows as an additional enhancement stage separate from the variational autoencoder, without exploiting their full potential to improve the variational inference process. Therefore,

we seek to directly incorporate normalizing flows into the variational inference process to enable more expressive posterior distributions while retaining the variational inference framework. To this end, we introduce additional random variables $\boldsymbol{u}^v$ into the normalizing flows to construct a flexible probabilistic modeling in the latent space. Specifically, for each available view, we model the latent-variable posterior $q(\boldsymbol{z}^v \mid \boldsymbol{x}^v)$ using a view-conditioned normalizing flow.

$$q(\boldsymbol{z}^v \mid \boldsymbol{x}^v) = \int \delta(\boldsymbol{z}^v - F^v_{\boldsymbol{x}^v}(\boldsymbol{u}^v))p(\boldsymbol{u}^v)d\boldsymbol{u} = \int \delta(\boldsymbol{z}^v - F^v(\sigma_1^v \boldsymbol{u}^v + E(\boldsymbol{x}^v)))p(\boldsymbol{u}^v)d\boldsymbol{u}, \quad (2)$$

where $p(\boldsymbol{u}^v)$ denotes a simple base distribution, typically chosen as a standard Gaussian distribution $\mathcal{N}(0, \mathbf{I})$, while $\delta(\cdot)$ denotes the Dirac delta function that enforces the deterministic relationship $\boldsymbol{z}^v = F^v_{\boldsymbol{x}^v}(\boldsymbol{u}^v)$. $E(\boldsymbol{x}^v)$ denotes the feature representation extracted from the v-th available view and $\sigma_1^v > 0$ is a scalar scale parameter controlling the magnitude of the base noise for view v. The transformation $F^v(\sigma_1^v \boldsymbol{u}^v + E(\boldsymbol{x}^v))$, equivalently denoted as $F^v_{\boldsymbol{x}^v}(\boldsymbol{u}^v)$, is invertible with respect to $\boldsymbol{u}^v$ and is conditioned on the observed view $\boldsymbol{x}^v$. Thus, $F^v_{\boldsymbol{x}^v}$ can be regarded as a view-conditioned normalizing flow that transforms the base variable $\boldsymbol{u}^v$ into the view-level latent representation $\boldsymbol{z}^v$. With the obtained latent representations $\{\boldsymbol{z}^v\}_{v=1}^{\mathcal{V}}$, we obtain $q(\boldsymbol{z} \mid \{\boldsymbol{z}^v\})$ by aggregating the observed views through an adaptive integration mechanism:

$$\begin{aligned} \boldsymbol{z} &= \sum_{v \in \mathcal{V}} \alpha_v \boldsymbol{z}^v, \\ \text{s.t.} \quad \alpha_v &\geq 0, \quad \sum_{v \in \mathcal{V}} \alpha_v = 1, \end{aligned} \quad (3)$$

Here, α_v denotes the integration weight of the v-th available view. This weighting strategy allows the integration to account for the different contributions of the available views. The resulting $\boldsymbol{z}$ is used as the unified latent representation for clustering and view generation.

Generative Process. Based on the above inference process, we formulate the corresponding generative process by specifying priors for the latent variables and generating each view from the unified representation. The joint distribution is given by:

$$p(\{\boldsymbol{x}^v\}, \boldsymbol{z}, \{\boldsymbol{z}^v\}) = p(\{\boldsymbol{z}^v\})p(\boldsymbol{z})p(\{\boldsymbol{x}^v\} \mid \boldsymbol{z}) = p(\boldsymbol{z}) \prod_{v \in \mathcal{V}} p(\boldsymbol{z}^v)p(\boldsymbol{x}^v \mid \boldsymbol{z}). \quad (4)$$

Here, $p(\boldsymbol{z}^v) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is used as the prior for each view-level latent representation. For the unified latent variable $\boldsymbol{z}$, however, an isotropic Gaussian prior may be overly restrictive. As described in the inference process, its aggregated posterior is effectively a mixture of sample-dependent distributions jointly induced by heterogeneous observations, adaptive integration weights and varying available-view patterns. Moreover, since each view-level posterior is obtained through a nonlinear flow transformation, the resulting aggregated posterior is generally not expected to follow a unimodal isotropic Gaussian distribution and may instead exhibit multimodality, asymmetric density and complex dependencies among latent dimensions. Consequently, a standard Gaussian prior may impose an excessively restrictive geometry on the unified latent space, potentially leading to a substantial mismatch with its aggregated posterior. To better accommodate the potentially multimodal structure of the aggregated posterior and encourage a cluster-aware organization of the unified latent space, we model

$$p(\boldsymbol{z}) = \sum_{o=1}^{O} \pi_o \mathcal{N}\left(\boldsymbol{z} \mid \boldsymbol{\mu}_o, \boldsymbol{\sigma}_o^2 \mathbf{I}\right), \quad (5)$$

where O denotes the number of mixture components and o indexes the corresponding latent clusters. The parameters π_o, $\boldsymbol{\mu}_o$ and $\boldsymbol{\sigma}_o^2$ denote the mixture weight, mean and diagonal variance of the o-th Gaussian component, respectively. The mixture weights satisfy $\pi_o \geq 0$ and $\sum_{o=1}^{O} \pi_o = 1$. Unlike a unimodal Gaussian prior that assumes a single compact latent distribution, the Gaussian mixture prior explicitly models the multimodal structure of the latent space by assigning each component to a cluster-specific region. This formulation enables the prior distribution to capture the inherent semantic diversity of incomplete multi-view data and provides a probabilistic reference for learning unified latent representations from different available-view patterns. Finally, each view is generated from the unified latent representation through its corresponding inverse flow and view-specific decoder. Specifically, the inverse transformation $(F^v)^{-1}$ maps $\boldsymbol{z}$ from the shared latent space to the

pre-flow latent space associated with the v-th view. For continuous observations, the conditional likelihood is defined as

$$p(\boldsymbol{x}^v \mid \boldsymbol{z}) = \mathcal{N}\big(\boldsymbol{x}^v \mid G^v\big((F^v)^{-1}(\boldsymbol{z})\big), (\sigma_2^v)^2\mathbf{I}\big), \tag{6}$$

where G^v denotes the decoder for the v-th view, $G^v((F^v)^{-1}(\boldsymbol{z}))$ specifies the conditional mean, and σ_2^v is a learnable scalar standard deviation shared across all dimensions of the v-th view. The inverse flows translate the shared representation into view-specific latent coordinates, while the decoders map these coordinates to their respective observation spaces.

Missing-View Recovery. For the i-th sample, only the observed views $\{\boldsymbol{x}^v : v \in \mathcal{V}\}$ are used to infer the view-level latent representations and the unified representation $\boldsymbol{z}$. For a missing view $v \notin \mathcal{V}$, the corresponding view-level latent representation cannot be inferred because $\boldsymbol{x}^v$ is unavailable. Once $\boldsymbol{z}$ has been obtained, the generative model provides a conditional distribution for every view, including the missing ones. For each $v \notin \mathcal{V}$, we use its conditional mean as the completed view:

$$\widehat{\boldsymbol{x}}^v = \mathbb{E}_{p(\boldsymbol{x}^v|\boldsymbol{z})}[\boldsymbol{x}^v] = G^v\left((F^v)^{-1}(\boldsymbol{z})\right). \tag{7}$$

During training, reconstruction likelihoods are evaluated only for observed views with $M_{iv} = 1$. Genuinely missing observations are not used as reconstruction targets.

2.3 Variational Learning Objective

Given the generative model and the variational inference procedure defined above, we aim to learn the model parameters by maximizing the marginal log-likelihood $\max \log p(\{\boldsymbol{x}^v\})$. However, directly optimizing the marginal likelihood is generally intractable because it requires integrating over the shared and view-level latent variables. We therefore introduce the variational posterior defined above to approximate the intractable true posterior and obtain a tractable evidence lower bound (ELBO) on the marginal log-likelihood. Applying Jensen's inequality yields

$$\log p(\{\boldsymbol{x}^v\}) \geq \mathbb{E}_{q(\boldsymbol{z},\{\boldsymbol{z}^v\}|\{\boldsymbol{x}^v\})} \log\left[\frac{p(\{\boldsymbol{x}^v\},\boldsymbol{z},\{\boldsymbol{z}^v\})}{q(\boldsymbol{z},\{\boldsymbol{z}^v\} \mid \{\boldsymbol{x}^v\})}\right] = -\mathcal{L}_z - \mathcal{L}_{z^v} - \mathcal{L}_{rec} = -\mathcal{L}_{ELBO}, \tag{8}$$

where

$$\begin{aligned}
\mathcal{L}_z &= \mathbb{E}_{\prod_{v\in\mathcal{V}} q(\boldsymbol{z}^v|\boldsymbol{x}^v)} D_{KL}\left[q(\boldsymbol{z} \mid \{\boldsymbol{z}^v\}) \| p(\boldsymbol{z})\right], \\
\mathcal{L}_{z^v} &= \sum_{v\in\mathcal{V}} D_{KL}\left[q(\boldsymbol{z}^v \mid \boldsymbol{x}^v) \| p(\boldsymbol{z}^v))\right], \\
\mathcal{L}_{rec} &= \sum_{v\in\mathcal{V}} \mathbb{E}_{q(\boldsymbol{z}|\{\boldsymbol{z}^v\})\prod_{v=1}^{V} q(\boldsymbol{z}^v|\boldsymbol{x}^v)} \log p(\boldsymbol{x}^v \mid \boldsymbol{z}).
\end{aligned}$$

The resulting objective consists of three components. First, $\mathcal{L}_z$ regularizes the integrated latent representation by encouraging its conditional distribution $q(\boldsymbol{z} \mid \{\boldsymbol{z}^v\})$ to approach the shared prior $p(\boldsymbol{z})$. Since the integrated posterior depends on the stochastic view-level representations, this divergence is further averaged over their joint variational distribution. Second, $\mathcal{L}_{z^v}$ regularizes each view-specific posterior $q(\boldsymbol{z}^v \mid \boldsymbol{x}^v)$ toward its corresponding prior $p(\boldsymbol{z}^v)$, which helps maintain a structured view-level latent space. Finally, $\mathcal{L}_{\text{rec}}$ measures the negative expected log-likelihood of all observed views conditioned on the integrated representation. Minimizing this term encourages $\boldsymbol{z}$ to preserve the information shared by the observed views and to provide sufficient information for their reconstruction. Using the distributional assumptions and flow transformations defined in Subsection 2.2, we further derive tractable expressions for the three ELBO components. The resulting expressions are given below.

For the shared latent variable $\mathcal{L}_z$, the conditional KL divergence can be written by substituting the flow-based prior as:

$$\mathcal{L}_z = -H(q(\boldsymbol{z} \mid \{\boldsymbol{z}^v\})) - \mathbb{E}_{q(\boldsymbol{z}|\{\boldsymbol{z}^v\})}\left[\log \sum_{o=1}^{O} \pi_o \frac{1}{(2\pi\sigma_o^2)^{d/2}} \exp\left(-\frac{\|\boldsymbol{z}-\boldsymbol{\mu}_o\|_2^2}{2\sigma_o^2}\right)\right] \tag{9}$$

Here, σ_o is a scalar standard deviation shared across all dimensions, such that each dimension has variance σ_o^2 and the covariance matrix is given by $\sigma_o^2\mathbf{I}$. This formulation consists of two terms: the conditional entropy of the integration distribution and the expected log-density of the shared latent representation under the Gaussian mixture prior. The second term encourages the learned latent

representations to match the cluster-oriented prior distribution, thereby introducing cluster-specific structural information into the latent space. The expectation can be efficiently estimated using Monte Carlo sampling. The complete derivation is presented in the Appendix A.3.

For the view-level regularization term $\mathcal{L}_{z^v}$, the invertibility of the view-conditioned flow allows each KL divergence to be evaluated as an expectation over the base distribution. With standard Gaussian distributions for both $p(\boldsymbol{u}^v)$ and $p(\boldsymbol{z}^v)$, the resulting expression is

$$\mathcal{L}_{z^v} = \sum_{v\in\mathcal{V}} \mathbb{E}_{p(\boldsymbol{u}^v)} \left[\frac{1}{2}\|F^v(\sigma_1^v \boldsymbol{u}^v + E(\boldsymbol{x}^v))\|^2 - \frac{1}{2}\|\boldsymbol{u}^v\|^2 - \log \left| \det \frac{\partial F^v_{\boldsymbol{x}^v}(\boldsymbol{u}^v)}{\partial \boldsymbol{u}^v} \right| \right]. \tag{10}$$

The difference between the two squared norms follows from the Gaussian log-densities of $\boldsymbol{z}^v$ and $\boldsymbol{u}^v$, while the log-determinant accounts for the density change introduced by $F^v_{\boldsymbol{x}^v}$. Thus, the expectation can be estimated using samples from $p(\boldsymbol{u}^v)$. The detailed derivation is provided in the Appendix A.4.

For the reconstruction term, we model each observed view using an isotropic Gaussian likelihood whose mean is given by the view-specific decoder $G^v((F^v)^{-1}(\boldsymbol{z}))$. Taking the negative expected log-likelihood yields

$$\mathcal{L}_{rec} = -\sum_{v\in\mathcal{V}} \mathbb{E}_{q(\boldsymbol{z}|\{\boldsymbol{z}^v\})\prod_{v=1}^V q(\boldsymbol{z}^v|\boldsymbol{x}^v)} \left[\frac{1}{2(\sigma_2^v)^2} \|G^v((F^v)^{-1}(\boldsymbol{z})) - \boldsymbol{x}^v\|^2 \right]. \tag{11}$$

The resulting term measures the expected reconstruction negative log-likelihood across the available views, with the squared reconstruction error of each view scaled by its inverse observation variance $1/(\sigma_2^v)^2$. The expectation is estimated using samples of the view-level latent variables and the integrated latent variable drawn according to the inference process.

2.4 Pairwise Distributional Alignment

The ELBO governs inference and generation but does not explicitly align the latent distributions produced by different view-specific flows. The representations being fused may therefore follow different distributions in the shared latent space. To encourage distributional coherence, we compare each view distribution with a common reference formed as a weighted mixture of all view distributions, using their expected fusion weights to reflect their average contributions. When the discrepancy is convex in its second argument, this construction yields an upper bound expressed in terms of pairwise view discrepancies. The following theorem establishes this bound and motivates our distributional coherence regularizer.

Theorem 2.1 (Pairwise upper bound for individual modality incoherence) *Let $\mu_1, \ldots, \mu_V$ be probability distributions, where $V \geq 2$. Let $w_1, \ldots, w_V$ be random fusion weights such that $w_j > 0$ and $\sum_{j=1}^V w_j = 1$. Define $\bar{w}_j = \mathbb{E}[w_j]$, $\mu = \sum_{j=1}^V \bar{w}_j \mu_j$. Assume that the distribution discrepancy $\mathcal{D}$ satisfies the following properties:*

1. *$\mathcal{D}(P, Q) \geq 0$ for any distributions P and Q;*
2. *$\mathcal{D}(P, P) = 0$ for any distribution P;*
3. *for any fixed P, the map $Q \mapsto \mathcal{D}(P, Q)$ is convex.*

Then, for every $v \in \{1, \ldots, V\}$,

$$\mathcal{D}(\mu_v, \mu) \leq \sum_{\substack{v_1, v_2 = 1 \\ v_1 \neq v_2}}^{V} \mathcal{D}(\mu_{v_1}, \mu_{v_2}). \tag{12}$$

The proof of Theorem 2.1 is provided in Appendix A.5. Theorem 2.1 provides a common pairwise upper bound for the discrepancy of every view-specific distribution from the mixture reference distribution. Hence, minimizing the pairwise discrepancies simultaneously constrains the distributional incoherence of all views. The result is not tied to a particular discrepancy: it applies whenever $\mathcal{D}$

is nonnegative, vanishes on identical distributions, and is convex in its second argument. Examples include the 1-Wasserstein distance $\mathcal{W}_1$, the p-th power of the p-Wasserstein distance $\mathcal{W}_p^p$ for $p \geq 1$, and the squared maximum mean discrepancy MMD^2, which measures the squared distance between kernel mean embeddings of two distributions. We adopt the 1-Wasserstein distance (see Appendix A.6 for its definition) to measure distributional discrepancies between view-specific latent distributions. An empirical estimate can be obtained from samples drawn from the corresponding flow posteriors.

Consider a minibatch of B incomplete multi-view samples, and let $M_{iv} \in \{0, 1\}$ denote the view availability indicator, where $M_{iv} = 1$ indicates that the v-th view of the i-th sample is observed, and $M_{iv} = 0$ otherwise. Here, $i \in \{1, \ldots, B\}$ indexes the samples and $v \in \{1, \ldots, V\}$ indexes the views. For each view v, we define the set of observed sample indices as $\mathcal{I}_v = \{i \in \{1, \ldots, B\} : M_{iv} = 1\}$, and denote the number of available samples in this view by $B_v = |\mathcal{I}_v|$. For every observed sample $i \in \mathcal{I}_v$, a latent sample $\boldsymbol{z}_i^v \in \mathbb{R}^d$ is obtained from $q(\boldsymbol{z}^v|\boldsymbol{x}_i^v)$, where d represents the shared latent dimension. For two views v_1 and v_2 that contain observed samples in the current minibatch ($B_{v_1}, B_{v_2} > 0$), we define the transport cost matrix $C^{v_1,v_2} \in \mathbb{R}_+^{B_{v_1} \times B_{v_2}}$ as

$$C_{ij}^{v_1,v_2} = \|\boldsymbol{z}_i^{v_1} - \boldsymbol{z}_j^{v_2}\|_2, \qquad i \in \mathcal{I}_{v_1}, \quad j \in \mathcal{I}_{v_2}. \tag{13}$$

The admissible transport plans have uniform marginals over the observed samples:

$$\mathcal{U}_{v_1,v_2} = \left\{ T \in \mathbb{R}_+^{B_{v_1} \times B_{v_2}} : T\mathbf{1}_{B_{v_2}} = \frac{\mathbf{1}_{B_{v_1}}}{B_{v_1}}, T^\top \mathbf{1}_{B_{v_1}} = \frac{\mathbf{1}_{B_{v_2}}}{B_{v_2}} \right\}. \tag{14}$$

where T_{ij} represents the amount of probability mass transported from the i-th latent sample of view v_1 to the j-th latent sample of view v_2. The row and column constraints ensure that the empirical distributions of the two views preserve their original uniform masses, with each observed sample contributing probability mass $1/B_{v_1}$ and $1/B_{v_2}$, respectively. To obtain a differentiable and computationally efficient approximation of the optimal transport problem, we adopt an entropy-regularized optimal transport formulation. Specifically, the empirical Wasserstein discrepancy between two view-specific latent distributions is computed as

$$\widehat{\mathcal{W}}_{1,\varepsilon}^{v_1,v_2} = \min_{T \in \mathcal{U}_{v_1,v_2}} \sum_{i \in \mathcal{I}_{v_1}} \sum_{j \in \mathcal{I}_{v_2}} T_{ij} C_{ij}^{v_1,v_2} + \varepsilon \sum_{i \in \mathcal{I}_{v_1}} \sum_{j \in \mathcal{I}_{v_2}} T_{ij}(\log T_{ij} - 1). \tag{15}$$

where $\varepsilon > 0$ controls the strength of the entropy regularization. The formulation uses the negative entropy form $\sum_{i,j} T_{ij}(\log T_{ij} - 1)$, whose derivative with respect to T is $\log T$, facilitating efficient optimization with the Sinkhorn algorithm. The additional entropy term encourages a smoother transport plan and enables efficient optimization using the Sinkhorn algorithm. The empirical 1-Wasserstein distance provides a minibatch approximation of the distributional discrepancy between the two observed latent distributions. Importantly, the optimal transport coupling does not require every sample to have observations in both views. Instead, it establishes a probabilistic correspondence between the available latent samples and measures the discrepancy between their empirical distributions. Let $\mathcal{P}_B = \{(v_1, v_2) : v_1 < v_2,\ B_{v_1} > 0,\ B_{v_2} > 0\}$ denote the set of valid view pairs in the current minibatch, where only view pairs with available latent samples are considered. We defined

$$\mathcal{L}_{\text{align}} = \sum_{(v_1,v_2) \in \mathcal{P}_B} \widehat{\mathcal{W}}_1^{v_1,v_2}. \tag{16}$$

The alignment loss minimizes the Wasserstein discrepancy among the marginal latent distributions inferred from different observed views. Therefore, it encourages view-specific flows to produce semantically consistent latent representations despite incomplete observations.

2.5 Overall Loss

Combining the variational objective with the distributional coherence regularizer, we minimize the following overall training loss:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \beta\mathcal{L}_z + \gamma\mathcal{L}_{z^v} + \lambda\mathcal{L}_{align}. \tag{17}$$

Here, $\mathcal{L}_{rec}$ encourages faithful reconstruction of the observed views, while $\mathcal{L}_z$ and $\mathcal{L}_{z^v}$ regularize the fused and view-specific latent representations, respectively. The alignment term $\mathcal{L}_{align}$ encourages agreement among the view-specific latent distributions. The nonnegative hyperparameters β, γ and λ balance these regularization terms against the reconstruction objective. The overall training procedure is summarized in Algorithm 1.

Table 1: Clustering performance comparison across five datasets under four different missing rates.

	Missing Rates	0.1			0.3			0.5			0.7		
	Metrics	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
CUB	APADC	61.93	63.79	43.55	64.23	63.28	46.24	59.70	62.17	44.11	50.33	54.49	35.33
	RPCIC	71.37	70.50	57.24	67.50	67.60	54.28	59.43	60.98	44.52	58.53	61.14	45.14
	COMPLETER	52.97	65.47	45.98	60.73	68.88	52.78	51.90	61.84	45.11	19.43	17.37	0.73
	HSACC	60.13	70.43	51.60	60.67	66.57	46.61	50.30	58.28	35.08	39.40	46.67	22.70
	RGCL	54.83	57.45	43.86	46.33	48.24	32.43	48.00	42.83	25.78	46.67	40.66	24.05
	DGIMVCM	52.10	55.25	39.42	50.70	55.03	39.51	41.77	48.61	30.17	41.90	48.62	27.65
	DCG	74.57	72.11	59.77	71.38	67.82	55.70	68.98	66.30	53.33	62.87	60.97	45.83
	BRIDGE	46.17	61.61	37.19	48.50	50.17	32.20	42.83	51.29	31.31	44.67	54.54	34.18
	BURG	71.50	71.60	60.48	58.00	61.85	45.86	57.33	58.89	45.12	53.83	56.58	45.12
	FEMVAE	**75.08**	**73.55**	**62.10**	**76.00**	**71.11**	**59.75**	**72.37**	**66.46**	**54.36**	**66.83**	**63.02**	**50.08**
Caltech7-5V	APADC	61.20	61.40	46.41	60.90	61.00	45.70	55.01	54.01	35.81	53.90	50.10	32.81
	RPCIC	79.81	72.46	66.67	76.47	68.81	61.65	49.74	49.40	33.52	64.79	58.53	48.91
	COMPLETER	64.30	63.60	45.60	54.20	50.31	33.10	45.21	37.91	20.80	34.61	25.01	14.40
	HSACC	54.40	62.72	38.73	51.54	58.54	32.55	50.71	56.92	31.59	52.10	58.11	33.16
	RGCL	53.21	41.98	33.58	53.21	46.20	37.39	49.93	44.49	34.85	49.36	38.93	33.18
	DGIMVCM	52.83	49.81	36.37	53.56	50.33	36.76	53.87	48.98	38.71	52.03	47.86	35.89
	DCG	56.98	48.14	38.45	57.56	49.10	39.68	54.34	43.46	35.04	55.19	44.14	36.44
	BRIDGE	57.07	47.95	37.00	43.79	39.20	27.50	43.50	30.86	21.14	31.64	16.96	12.21
	BURG	56.07	59.52	46.89	59.14	58.23	45.99	58.21	53.71	39.31	60.36	55.55	43.93
	FEMVAE	**81.00**	**74.54**	**69.18**	**82.99**	**74.76**	**70.98**	**80.76**	**70.72**	**66.70**	**78.96**	**68.58**	**65.04**
Handwritten	APADC	79.45	79.19	70.46	66.35	73.01	50.73	79.35	81.55	72.60	57.03	65.26	51.90
	RPCIC	90.86	83.15	81.15	87.25	80.91	76.87	82.36	76.93	71.17	77.59	74.74	67.01
	COMPLETER	82.18	77.73	68.92	78.70	69.08	58.14	74.73	67.49	50.21	68.86	62.41	41.06
	HSACC	82.47	83.61	73.48	82.86	**85.96**	76.95	82.79	84.60	75.65	75.54	76.51	60.22
	RGCL	92.30	85.71	83.88	90.95	79.41	75.38	86.50	79.10	74.46	83.15	74.23	68.50
	DGIMVCM	85.74	88.60	82.58	85.42	84.18	79.19	**93.90**	**87.90**	**87.00**	78.34	77.82	70.57
	DCG	72.69	74.66	63.21	77.10	75.55	66.37	69.65	70.78	59.62	75.93	70.96	63.22
	BRIDGE	80.21	82.13	74.89	73.20	80.29	68.80	67.84	60.85	47.04	56.46	51.41	36.30
	BURG	86.12	84.24	79.16	76.75	75.69	64.72	72.82	72.75	60.16	68.70	68.03	49.84
	FEMVAE	**94.15**	**88.61**	**83.92**	**92.04**	84.87	**84.66**	91.46	83.96	82.39	**90.68**	**82.22**	**80.68**
Scene-15	APADC	40.90	42.00	24.60	39.90	40.50	23.00	39.00	38.90	22.20	38.60	37.90	21.80
	RPCIC	37.25	35.83	21.08	39.89	40.76	23.85	37.07	40.52	22.06	35.00	38.22	20.81
	COMPLETER	37.00	41.89	23.60	40.04	42.41	24.22	36.64	38.99	19.70	35.37	37.05	17.58
	HSACC	37.69	41.95	21.91	37.04	40.10	19.93	39.59	41.41	21.99	38.64	39.37	21.26
	RGCL	33.25	32.23	16.49	35.59	33.23	18.50	32.48	29.22	14.59	23.59	20.31	10.42
	DGIMVCM	50.50	**49.70**	**32.60**	46.90	46.40	**29.80**	45.20	**43.00**	27.30	42.60	39.70	24.40
	DCG	37.34	41.04	22.10	38.23	40.97	22.43	35.92	40.86	21.61	35.83	39.05	20.46
	BRIDGE	29.56	33.42	17.12	31.36	31.95	17.35	30.18	29.90	16.07	25.09	24.19	12.46
	BURG	37.32	36.50	19.39	31.24	31.74	13.75	27.63	27.47	10.10	23.28	24.34	7.31
	FEMVAE	**50.90**	44.20	30.80	**47.92**	**47.97**	29.67	**46.62**	42.52	**27.61**	**45.00**	**40.08**	**25.90**
CCV	APADC	15.19	14.31	4.71	14.59	13.15	4.45	12.31	8.43	2.53	10.66	5.24	1.28
	RPCIC	19.47	18.60	6.93	14.51	10.78	3.58	13.21	9.26	2.97	14.44	10.02	3.49
	COMPLETER	20.32	19.73	5.96	20.43	17.32	3.71	19.64	16.41	4.89	18.40	14.77	6.17
	HSACC	19.25	16.48	5.23	19.72	16.21	5.18	18.49	14.89	3.76	17.34	14.62	3.83
	RGCL	17.55	15.68	5.80	17.78	13.45	5.35	17.85	13.61	5.64	17.55	12.94	5.27
	DGIMVCM	**21.74**	19.09	8.36	21.42	17.61	**7.77**	19.76	16.04	6.82	18.59	14.10	5.91
	DCG	19.52	17.13	7.10	17.50	15.11	5.75	18.25	14.10	6.13	17.53	12.89	5.44
	BRIDGE	19.18	18.54	7.74	18.63	17.07	7.52	17.79	14.41	6.25	11.48	8.06	2.38
	BURG	14.66	11.12	3.57	13.85	10.03	2.80	13.02	8.98	2.19	12.83	8.19	2.16
	FEMVAE	21.46	**20.89**	**8.39**	**22.17**	**19.41**	7.53	**21.04**	**17.77**	**6.86**	**20.68**	**17.86**	**6.41**

3 RESULTS

3.1 EXPERIMENT SETUP

Datasets, Baselines and Metrics. We conduct experiments on five commonly used benchmark datasets for multi-view learning: CUB (Wah et al., 2011), Handwritten (van Breukelen et al., 1998), CCV (Jiang et al., 2011), Scene-15 (Xu et al., 2020) and Caltech7-5V (Li et al., 2015). Further details of these datasets are reported in Appendix A.8. For comparison, we select ten representative

Table 2: Results of ablation experiments on the CUB and Handwritten datasets.

Components			CUB				Handwritten			
Align	Flow	GMM	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7
	✓	✓	0.74/0.71/0.60	0.73/0.70/0.59	0.71/0.62/0.53	0.62/0.61/0.49	0.86/0.82/0.81	0.83/0.79/0.77	0.82/0.76/0.75	0.80/0.75/0.72
✓		✓	0.67/0.69/0.54	0.68/0.69/0.54	0.68/0.65/0.53	0.66/0.62/0.48	0.85/0.80/0.79	0.85/0.79/0.78	0.84/0.79/0.79	0.82/0.76/0.74
✓	✓		0.74/0.72/0.54	0.75/0.70/0.55	0.70/0.62/0.51	0.64/0.61/0.43	0.81/0.80/0.77	0.72/0.77/0.70	0.73/0.77/0.80	0.72/0.75/0.70
✓	✓	✓	**0.75/0.73/0.62**	**0.76/0.71/0.60**	**0.72/0.66/0.54**	**0.67/0.63/0.50**	**0.94/0.87/0.84**	**0.92/0.85/0.85**	**0.91/0.84/0.82**	**0.91/0.82/0.81**

state-of-the-art methods for incomplete multi-view clustering, namely APADC (Xu et al., 2023), RPCIC (Yuan et al., 2024), COMPLETER (Lin et al., 2021), HSACC (Ding et al., 2025), RGCL (Zhuang et al., 2025), DGIMVCM (Zhang et al., 2026b), DCG (Zhang et al., 2025), BRIDGE (Jiang et al., 2025) and BURG (Jin et al., 2025). We assess the clustering quality using three standard metrics: Clustering Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI). The implementation and experimental settings are detailed in Appendix A.9.

3.2 Experimental Results and Analysis

To comprehensively evaluate the clustering performance of FEMVAE, we compare it with nine baselines on five benchmark datasets under four missing rates. Table 1 reports the results in terms of ACC, NMI, and ARI. Overall, FEMVAE achieves the best results in 50 of the 60 metric comparisons and ranks among the top two in 59, demonstrating strong overall performance. In particular, it consistently outperforms all baselines across all metrics and missing rates on CUB and Caltech7-5V. On Caltech7-5V, its ACC exceeds the strongest baseline by 22.55 and 14.17 percentage points at missing rates of 0.5 and 0.7, respectively. At the highest missing rate of 0.7, FEMVAE achieves the best results across all three metrics on all five datasets, with ACC improvements ranging from 2.09 to 14.17 percentage points over the strongest baseline for each dataset. These results support the effectiveness of FEMVAE in maintaining discriminative clustering representations under substantial view missingness. To illustrate the learned cluster structure, we visualize the integrated latent representations on CUB at a missing rate of 0.3. As shown in Figure 2, the representations form clear clusters, suggesting effective integration of the available views. A detailed analysis is provided in Appendix A.10. We also analyze the sensitivity to the loss weights β, γ and λ to examine their effects on clustering performance. The results show stable performance over reasonable parameter ranges, with detailed results and analysis provided in Appendix A.11.

3.3 Ablation Analysis

To assess component contributions, we ablate the flow-based posterior, cross-view alignment loss, and GMM prior, which respectively promote posterior flexibility, cross-view consistency, and latent cluster structure. At a missing rate of 0.7, replacing the GMM prior with a standard Gaussian prior reduces ACC from 0.67 to 0.64 on CUB and from 0.91 to 0.72 on Handwritten (Table 2), highlighting its importance under severe view missingness. Removing the posterior flow or alignment loss also reduces Handwritten ACC to 0.82 and 0.80, respectively, further supporting their complementary contributions to clustering under incomplete observations. Detailed results and analysis are provided in Appendix A.12.

4 Conclusion

In this paper, we propose a flow-embedded multimodal variational autoencoder for incomplete multi-view clustering. Our model unifies the normalizing flow and the variational autoencoders into an integrated framework, which elevates the expressive power of variational inference by learning more flexible posterior distributions. Furthermore, we introduce a Gaussian mixture prior to impose a cluster-aware structure on the latent space and employ an optimal transport strategy to align view-specific latent representations, effectively facilitating information integration across incomplete views. Extensive experiments on benchmark incomplete multi-view datasets demonstrate that the proposed method achieves competitive clustering performance under various missing-view settings. These results validate the effectiveness of combining expressive probabilistic modeling with cross-view distribution alignment for incomplete multi-view clustering.

REFERENCES

Hongmin Cai, Weitian Huang, Sirui Yang, Siqi Ding, Yue Zhang, Bin Hu, Fa Zhang, and Yiu-Ming Cheung. Realize generative yet complete latent representation for incomplete multi-view learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3637–3652, 2024. doi: 10.1109/TPAMI.2023.3346869.

Yuzhuo Dai, Jiaqi Jin, Zhibin Dong, Siwei Wang, Xinwang Liu, En Zhu, Xihong Yang, Xinbiao Gan, and Yu Feng. Imputation-free and alignment-free: Incomplete multi-view clustering driven by consensus semantic learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5071–5081, 2025.

Xiaojian Ding, Lin Zhao, Xian Li, and Xiaoying Zhu. Incomplete multi-view clustering via hierarchical semantic alignment and cooperative completion. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 52175–52195. Curran Associates, Inc., 2025. doi: 10.52202/085713-1739. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/4b2a917e30e1bb1aff055b4d8c6c081c-Paper-Conference.pdf.

Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-Linear Independent Components Estimation. In *International Conference on Learning Representations*, 2015. URL https://mlanthology.org/iclr/2015/dinh2015iclr-nice/.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HkpbnH9lx.

Zhibin Dong, Siwei Wang, Jiaqi Jin, Xinwang Liu, and En Zhu. Cross-view topology based consistent and complementary information for deep multi-view clustering. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19383–19394, 2023. doi: 10.1109/ICCV51070.2023.01781.

Yiming Du, Ziyu Wang, Jian Li, Rui Ning, and Lusi Li. Deep incomplete multi-view clustering via hierarchical imputation and alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20941–20949, 2026. doi: 10.1609/aaai.v40i25.39235.

Uno Fang, Man Li, Jianxin Li, Longxiang Gao, Tao Jia, and Yanchun Zhang. A comprehensive survey on multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35 (12):12350–12368, 2023. doi: 10.1109/TKDE.2023.3270311.

Xin Gao and Jian Pu. Deep incomplete multi-view learning via cyclic permutation of vaes. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 99609–99643, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/f74054328beeb0c21a9b8e99da557f5a-Paper-Conference.pdf.

Menglei Hu and Songcan Chen. Doubly aligned incomplete multi-view clustering. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 2262–2268. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/313. URL https://doi.org/10.24963/ijcai.2018/313.

Xiaorui Jiang, Buyun He, Peng Yuan Zhou, Xinyue Chen, Jingcai Guo, Jie Xu, and Yong Liao. A unified framework to bridge complete and incomplete deep multi-view clustering under non-iid missing patterns. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 594–603, 2025. doi: 10.1109/ICCV51701.2025.00063.

Yu-Gang Jiang, Guangnan Ye, Shih-Fu Chang, Daniel Ellis, and Alexander C. Loui. Consumer video understanding: a benchmark database and an evaluation of human and machine performance. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450303361. doi: 10.1145/1991996.1992025. URL https://doi.org/10.1145/1991996.1992025.

Jiaqi Jin, Siwei Wang, Zhibin Dong, Xihong Yang, Xinwang Liu, En Zhu, and Kunlun He. Deep incomplete multi-view clustering with distribution dual-consistency recovery guidance. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1016–1026, 2025. doi: 10.1109/ICCV51701.2025.00102.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/d139db6a236200b21cc7f752979132d0-Paper.pdf.

Conghui Li, Chern Hong Lim, and Xin Wang. Gmm-based vae model with normalising flow for effective stochastic segmentation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 84457–84480. Curran Associates, Inc., 2025. doi: 10.52202/085713-2831. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/79ee64e62fd3ee763961ce3d6aed4971-Paper-Conference.pdf.

Shao-Yuan Li, Yuan Jiang, and Zhi-Hua Zhou. Partial multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014. doi: 10.1609/aaai.v28i1.8973. URL https://doi.org/10.1609/aaai.v28i1.8973.

Yeqing Li, Feiping Nie, Heng Huang, and Junzhou Huang. Large-scale multi-view spectral clustering via bipartite graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, Feb. 2015. doi: 10.1609/aaai.v29i1.9598. URL https://ojs.aaai.org/index.php/AAAI/article/view/9598.

Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. Completer: Incomplete multi-view clustering via contrastive prediction. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11169–11178, 2021. doi: 10.1109/CVPR46437.2021.01102.

Xinwang Liu, Miaomiao Li, Lei Wang, Yong Dou, Jianping Yin, and En Zhu. Multiple kernel k-means with incomplete kernels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pp. 2259–2265, 2017. doi: 10.1609/aaai.v31i1.10893. URL https://ojs.aaai.org/index.php/AAAI/article/view/10893.

Xinwang Liu, Ke Liang, Jun Wang, Suyuan Liu, Xiangke Wang, and Huaimin Wang. Two decades of multi-view clustering: Taxonomy, application, and challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3744–3764, 2026. doi: 10.1109/TPAMI.2025.3640109.

Yalan Qin, Guorui Feng, and Xinpeng Zhang. Fast incomplete multi-view clustering by flexible anchor learning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 50235–50245. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/qin25b.html.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1530–1538, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/rezende15.html.

Raghavendra Selvan, Frederik Faye, Jon Middleton, and Akshay Pai. Uncertainty quantification in medical image segmentation with normalizing flows. In *Machine Learning in Medical Imaging: 11th International Workshop, MLMI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings*, pp. 80–90, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-59860-0. doi: 10.1007/978-3-030-59861-7_9. URL https://doi.org/10.1007/978-3-030-59861-7_9.

Jianlin Su and Guang Wu. f-vaes: Improve vaes with conditional flows, 2018. URL https://arxiv.org/abs/1809.05861.

Martijn van Breukelen, Robert PW Duin, David MJ Tax, and JE Den Hartog. Handwritten digit recognition by combined classifiers. *Kybernetika*, 34(4):381–386, 1998. URL http://www.kybernetika.cz/content/1998/4/381.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. *The caltech-ucsd birds-200-2011 dataset*, Jul 2011.

Haiyue Wang, Wensheng Zhang, Quan Wang, and Xiaoke Ma. Adaptive structural-guided multi-level representation learning with graph contrastive for incomplete multi-view clustering. *Information Fusion*, 119:103035, 2025a. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2025.103035. URL https://www.sciencedirect.com/science/article/pii/S1566253525001083.

Qianqian Wang, Zhengming Ding, Zhiqiang Tao, Quanxue Gao, and Yun Fu. Generative partial multi-view clustering with adaptive fusion and cycle consistency. *IEEE Transactions on Image Processing*, 30:1771–1783, 2021. doi: 10.1109/TIP.2020.3048626.

Siwei Wang, Xinwang Liu, Li Liu, Wenxuan Tu, Xinzhong Zhu, Jiyuan Liu, Sihang Zhou, and En Zhu. Highly-efficient incomplete largescale multiview clustering with consensus bipartite graph. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9766–9775, 2022. doi: 10.1109/CVPR52688.2022.00955.

Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21968–21977, 2023. doi: 10.1109/ICCV51070.2023.02013.

Ziyu Wang, Yiming Du, Yao Wang, Rui Ning, and Lusi Li. Deep incomplete multi-view clustering via multi-level imputation and contrastive alignment. *Neural Networks*, 181:106851, 2025b. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2024.106851. URL https://www.sciencedirect.com/science/article/pii/S0893608024007755.

Jie Wen, Shijie Deng, Waikeung Wong, Guoqing Chao, Chao Huang, Lunke Fei, and Yong Xu. Diffusion-based missing-view generation with the application on incomplete multi-view clustering. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52762–52778. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wen24c.html.

Benyu Wu, Wei Du, Jun Wang, and Guoxian Yu. Imputation-free incomplete multi-view clustering via knowledge distillation. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 6570–6578. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/731. URL https://doi.org/10.24963/ijcai.2025/731. Main Track.

Lei Xing, Xinhu Zheng, Yao Lu, and Badong Chen. Incomplete multi-view subspace clustering based on robust matrix completion. *Neurocomputing*, 621:129240, 2025. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2024.129240. URL https://www.sciencedirect.com/science/article/pii/S0925231224020113.

Cai Xu, Ziyu Guan, Wei Zhao, Hongchang Wu, Yunfei Niu, and Beilei Ling. Adversarial incomplete multi-view clustering. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 3933–3939. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/546. URL https://doi.org/10.24963/ijcai.2019/546.

Gehui Xu, Jie Wen, Chengliang Liu, Bing Hu, Yicheng Liu, Lunke Fei, and Wei Wang. Deep variational incomplete multi-view clustering: Exploring shared clustering structures. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16147–16155, 2024. doi: 10.1609/aaai.v38i14.29548. URL https://doi.org/10.1609/aaai.v38i14.29548.

Gehui Xu, Chengliang Liu, Yabo Liu, Yong Xu, and Jie Wen. Partial contrastive learning for partially view-aligned multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(10):12068–12081, 2026. doi: 10.1109/TPAMI.2026.3694999.

Jie Xu, Chao Li, Liang Peng, Yazhou Ren, Xiaoshuang Shi, Heng Tao Shen, and Xiaofeng Zhu. Adaptive feature projection with distribution alignment for deep incomplete multi-view clustering. *IEEE Transactions on Image Processing*, 32:1354–1366, 2023. doi: 10.1109/TIP.2023.3243521.

Ye Xu, Xiaodong Yu, Tian Wang, and Zezhong Xu. Scene-15 image dataset. Figshare, CC BY 4.0, 2020. https://figshare.com/articles/dataset/Scene-15/12103434.

Jun Yin, Pei Wang, Shiliang Sun, and Zhonglong Zheng. Incomplete multi-view clustering via multi-level contrastive learning. *IEEE Transactions on Knowledge and Data Engineering*, 37(8):4716–4727, 2025. doi: 10.1109/TKDE.2025.3568795.

Xiao Yu, Hui Liu, Yuxiu Lin, Yan Wu, and Caiming Zhang. Auto-weighted sample-level fusion with anchors for incomplete multi-view clustering. *Pattern Recognition*, 130:108772, 2022. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2022.108772. URL https://www.sciencedirect.com/science/article/pii/S0031320322002539.

Honglin Yuan, Shiyun Lai, Xingfeng Li, Jian Dai, Yuan Sun, and Zhenwen Ren. Robust prototype completion for incomplete multi-view clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 10402–10411, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681397. URL https://doi.org/10.1145/3664647.3681397.

Yuanyang Zhang, Yijie Lin, Weiqing Yan, Li Yao, Xinhang Wan, Guangyuan Li, Chao Zhang, Guanzhou Ke, and Jie Xu. Incomplete multi-view clustering via diffusion contrastive generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22650–22658, 2025.

Yuanyang Zhang, Yijie Lin, Xinhang Wan, Jie Xu, Li Yao, Weiqing Yan, and Chang Tang. Structure-aware conditional diffusion generation for incomplete multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 38(5):3168–3181, 2026a. doi: 10.1109/TKDE.2026.3673320.

Zhenghao Zhang, Jun Xie, Xingchen Chen, Tao Yu, Hongzhu Yi, Kaixin Xu, Yuanxiang Wang, Tianyu Zong, Xinming Wang, Jiahuan Chen, Guoqing Chao, Feng Chen, Zhepeng Wang, and Jungang Xu. Dynamic deep graph learning for incomplete multi-view clustering with masked graph reconstruction loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28600–28608, 2026b. doi: 10.1609/aaai.v40i34.40091.

Zhenhao Zhong, Zhibin Gu, Pengpeng Yang, Yaqian zhou, and Ruiqiang Guo. Gaussian regression-driven tensorized incomplete multi-view clustering with dual manifold regularization. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 72102–72131. Curran Associates, Inc., 2025. doi: 10.52202/085713-2420. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/684c59d614fe6ae74a3be8c3ef07e061-Paper-Conference.pdf.

Deyin Zhuang, Jian Dai, Xingfeng Li, Xi Wu, Yuan Sun, and Zhenwen Ren. Robust graph contrastive learning for incomplete multi-view clustering. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 7282–7290. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/810. URL https://doi.org/10.24963/ijcai.2025/810. Main Track.

A APPENDIX

A.1 AI USE STATEMENT

For this manuscript, large language models were used exclusively for language polishing and improving the clarity and readability of the writing. Beyond these linguistic assistance tasks, large language models made no substantive contributions to the conception, methodology, analysis, experiments, or completion of this work. All AI-assisted text was carefully reviewed and revised by the authors, who take full responsibility for the final content of this manuscript.

A.2 RELATED WORK

A.2.1 IMPUTATION-FREE INCOMPLETE MULTI-VIEW CLUSTERING

Non-imputation-based IMVC methods seek to identify a shared clustering structure from available observations without explicitly reconstructing missing views. Early approaches mainly rely on matrix factorization and subspace learning to connect samples with different view availability. PVC (Li et al., 2014) employs nonnegative matrix factorization to learn a common latent space, using paired samples to establish cross-view correspondence while incorporating singly observed samples into representation learning. DAIMC (Hu & Chen, 2018) introduces weighted semi-nonnegative matrix factorization with instance-level and basis-level alignment, allowing available observations to contribute to a consistent representation despite missing views. Graph-based approaches instead organize observed relationships into consensus structures. IMVC-CBG (Wang et al., 2022) learns a consensus bipartite graph connecting samples to shared anchors, providing an efficient alternative to constructing full sample similarity matrices. FIML (Qin et al., 2025) jointly optimizes anchor learning, graph construction, and graph partitioning, allowing the learned anchors and sample relationships to adapt to the clustering objective. ASA-IC (Yu et al., 2022) further introduces sample-level adaptive weighting to fuse the available sample-to-anchor relationships, accounting for differences in view reliability across individual samples. Deep approaches extend these ideas through nonlinear representation learning and cross-view information integration. APADC (Xu et al., 2023) projects available observations into a common feature space through adaptive feature projection, maximizes mutual information to extract shared cluster information, and reduces the distribution discrepancy between complete and incomplete samples. DVIMC (Xu et al., 2024) learns view-specific variational posteriors and aggregates available evidence through a product-of-experts formulation. It combines a Gaussian mixture prior with a coherence objective to encourage cluster structure and strengthen shared information. I2MVC (Wu et al., 2025) transfers clustering knowledge from a teacher trained on complete samples to a student capable of processing incomplete observations, thereby extending the learned clustering structure through knowledge distillation. FreeCSL (Dai et al., 2025) constructs a shared semantic space using consensus prototypes and incorporates within-view structural information to enhance semantic discrimination without explicit imputation or instance-level alignment. These approaches accommodate missing views through latent alignment, relational aggregation, or semantic transfer, avoiding direct reliance on reconstructed values. Nevertheless, their clustering decisions remain constrained by the evidence available in observed views and learned cross-sample relationships. When this evidence is insufficient to distinguish neighboring clusters, consistency learning and fusion may leave ambiguous assignments unresolved.

A.2.2 IMPUTATION-BASED INCOMPLETE MULTI-VIEW CLUSTERING

Imputation-based IMVC methods explicitly estimate unavailable data, representations, or similarities and incorporate the recovered information into clustering. Matrix and kernel completion approaches couple structural recovery with partition learning. MKKM-IK (Liu et al., 2017) optimizes missing kernel entries together with kernel combination weights and clustering variables, enabling the clustering objective to guide similarity completion. IMVC-RMC (Xing et al., 2025) integrates correntropy-based robust matrix completion with subspace clustering and adaptive graph fusion to accommodate missing observations and unreliable reconstruction residuals. Although these formulations exploit cross-view dependencies, their completion quality depends on the selected kernels, reconstruction models, and structural assumptions. Cross-view prediction approaches learn mappings from observed representations to missing ones. COMPLETER (Lin et al., 2021) combines con-

trastive learning with dual prediction, encouraging shared information across views while learning predictive dependencies for missing-representation recovery. BRIDGE (Jiang et al., 2025) addresses distribution shifts induced by non-IID missing patterns through domain-adversarial adaptation and combines learned cross-view predictions with target-view cluster means, balancing sample-specific estimates with cluster-level information. Other approaches make structural or semantic guidance central to completion. MICA (Wang et al., 2025b) constructs an adaptive cross-view graph to select reliable views for feature-, data-, and reconstruction-level imputation, while contrastive alignment promotes agreement among cluster assignments. RGCL (Zhuang et al., 2025) recovers missing data using a multi-view consistency transfer relationship graph, then introduces noise-robust graph contrastive learning and cross-view graph-level alignment to reduce the influence of erroneous contrastive pairs. ASMLGC (Wang et al., 2025a) combines multilevel representation learning with global and local structural guidance, using graph contrastive learning to strengthen consistency in the recovered representations. DIMVC-HIA (Du et al., 2026) first estimates missing cluster assignments through cross-view contrastive similarity and subsequently recovers missing features from intra-view, intra-cluster statistics, with semantic and assignment alignment refining the clustering structure. RPCIC (Yuan et al., 2024) addresses incorrect correspondence among view-specific prototypes through robust prototype discriminative learning and uses prototype fusion to complete missing information. DGIMVCM (Zhang et al., 2026b) guides imputation with a global graph while learning dynamic view-specific graphs, incorporating structural contrastive learning and masked graph reconstruction to improve representation quality. These methods provide informative completion constraints, but prediction errors, unreliable neighborhoods, or inaccurate preliminary assignments can affect the recovered features; moreover, point estimates do not explicitly describe alternative missing-view realizations compatible with the same observations. Generative approaches introduce distributional modeling into the recovery process. AIMC (Xu et al., 2019) combines reconstruction, adversarial learning, and an aligned clustering objective to jointly infer missing data and learn a common latent space. GP-MVC (Wang et al., 2021) uses view-specific adversarial generators conditioned on shared representations, with adaptive fusion and cycle consistency supporting cross-view recovery. In related variational multimodal clustering research, CMVAE (Cai et al., 2024) incorporates latent cluster structure into a multimodal variational model and further explores diffusion-based generation, whereas MVP (Gao & Pu, 2025) uses cyclic permutations of variational autoencoders to establish cross-view generative relationships and encourage consistency among posterior distributions. Diffusion-based methods offer another route to missing-view recovery: DMVG (Wen et al., 2024) combines diffusion generation with data augmentation to support learning under high missing rates; DCG (Zhang et al., 2025) integrates diffusion generation with instance- and category-level contrastive learning; and SACDG (Zhang et al., 2026a) introduces structural conditions into the diffusion process to guide representation recovery. These generative studies already incorporate probabilistic modeling, but generating plausible missing views does not by itself establish that sample-specific uncertainty is faithfully represented and retained in clustering decisions. Posterior expressiveness, cross-view distributional consistency, and the interaction between latent inference and cluster structure therefore remain relevant considerations for uncertainty-aware IMVC. Normalizing flows (NFs) provide a flexible framework for modeling complex probability distributions through a sequence of invertible transformations (Rezende & Mohamed, 2015; Dinh et al., 2015; 2017; Kingma & Dhariwal, 2018; Su & Wu, 2018; Selvan et al., 2020; Li et al., 2025). Starting from a simple base distribution, NFs can construct expressive target distributions while retaining tractable density evaluation through the change-of-variables formula. Their invertibility also enables bidirectional mapping between latent and data spaces, making NFs suitable for probabilistic generation and missing-view reconstruction. Recent NF-based recovery methods use invertible flows primarily to align or transfer feature distributions across modalities or views, with inverse flows used to recover missing representations . DiCMoR (Wang et al., 2023) aligns modality-specific latent features with class-conditioned Gaussian distributions and transfers representations across modalities, whereas BURG (Jin et al., 2025) maps view-specific features to a shared Gaussian reference space and uses fused representations to recover missing views through inverse flows. However, existing methods mainly use NFs to model and transfer feature distributions for recovery, rather than to increase the expressiveness of a view-specific, observation-conditioned posterior. Consequently, the expressive transformations of NFs are not directly applied to the posterior used for latent inference, limiting its capacity to represent complex, observation-dependent latent distributions.

A.3 Expansion of the KL Regularization Term $\mathcal{L}_z$

Given the cluster-oriented Gaussian mixture prior, the latent variable distribution is defined as Eq. (5)

For the o-th Gaussian component, its probability density can be written as

$$\mathcal{N}(\boldsymbol{z} \mid \boldsymbol{\mu}_o, \boldsymbol{\sigma}_o^2\mathbf{I}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\sigma}_o^2\mathbf{I}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_o)^T(\boldsymbol{\sigma}_o^2\mathbf{I})^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_o)\right).$$

Since the covariance matrix is isotropic, we have $|\boldsymbol{\sigma}_o^2\mathbf{I}| = (\sigma_o^2)^d$ and $(\boldsymbol{\sigma}_o^2\mathbf{I})^{-1} = \frac{1}{\sigma_o^2}\mathbf{I}$. Therefore, the Gaussian density can be simplified as

$$\mathcal{N}(\boldsymbol{z} \mid \boldsymbol{\mu}_o, \boldsymbol{\sigma}_o^2\mathbf{I}) = \frac{1}{(2\pi\sigma_o^2)^{d/2}} \exp\left(-\frac{\|\boldsymbol{z}-\boldsymbol{\mu}_o\|_2^2}{2\sigma_o^2}\right).$$

Substituting it into the GMM prior gives

$$p(\boldsymbol{z}) = \sum_{o=1}^{O} \pi_o \frac{1}{(2\pi\sigma_o^2)^{d/2}} \exp\left(-\frac{\|\boldsymbol{z}-\boldsymbol{\mu}_o\|_2^2}{2\sigma_o^2}\right).$$

Taking the logarithm, the log prior probability is obtained as

$$\log p(\boldsymbol{z}) = \log \sum_{o=1}^{O} \pi_o \frac{1}{(2\pi\sigma_o^2)^{d/2}} \exp\left(-\frac{\|\boldsymbol{z}-\boldsymbol{\mu}_o\|_2^2}{2\sigma_o^2}\right)$$

The KL divergence $\mathcal{L}_z$ can be decomposed as

$$\begin{aligned} &D_{KL}\left[q(\boldsymbol{z} \mid \{\boldsymbol{z}^v\}) \| p(\boldsymbol{z})\right] = \mathbb{E}_{q(\boldsymbol{z}|\{\boldsymbol{z}^v\})} \log \frac{q(\boldsymbol{z} \mid \{\boldsymbol{z}^v\})}{p(\boldsymbol{z})} = -H(q(\boldsymbol{z} \mid \{\boldsymbol{z}^v\})) - \mathbb{E}_{q(\boldsymbol{z}|\{\boldsymbol{z}^v\})} \log p(\boldsymbol{z}) \\ &= -H(q(\boldsymbol{z} \mid \{\boldsymbol{z}^v\})) - \mathbb{E}_{q(\boldsymbol{z}|\{\boldsymbol{z}^v\})} \left[\log \sum_{o=1}^{O} \pi_o \frac{1}{(2\pi\sigma_o^2)^{d/2}} \exp\left(-\frac{\|\boldsymbol{z}-\boldsymbol{\mu}_o\|_2^2}{2\sigma_o^2}\right)\right] \end{aligned} \tag{18}$$

A.4 Derivation of the View-Specific KL Term $\mathcal{L}_{z^v}$

Using the view-conditioned normalizing flow defined in Eq. (2), we express the posterior density in terms of the base variable $\boldsymbol{u}^v$ and rewrite the KL divergence as follows.

$$\begin{aligned} &D_{KL}\left[q(\boldsymbol{z}^v \mid \boldsymbol{x}^v) \| p(\boldsymbol{z}^v)\right] = \mathbb{E}_{q(\boldsymbol{z}^v|\boldsymbol{x}^v)} \log \frac{q(\boldsymbol{z}^v \mid \boldsymbol{x}^v)}{p(\boldsymbol{z}^v)} = \int q(\boldsymbol{z}^v \mid \boldsymbol{x}^v) \log \frac{q(\boldsymbol{z}^v \mid \boldsymbol{x}^v)}{p(\boldsymbol{z}^v)} d\boldsymbol{z}^v \\ &= \int\int \delta(\boldsymbol{z}^v - F_{\boldsymbol{x}^v}^v(\boldsymbol{u}^v)) p(\boldsymbol{u}^v) \log \frac{\int \delta(\boldsymbol{z}^v - F_{\boldsymbol{x}^v}^v(\boldsymbol{u}_*^v)) p(\boldsymbol{u}_*^v) d\boldsymbol{u}_*^v}{p(\boldsymbol{z}^v)} d\boldsymbol{z}^v d\boldsymbol{u}^v \\ &= \int p(\boldsymbol{u}^v) \log \frac{\int \delta(F_{\boldsymbol{x}^v}^v(\boldsymbol{u}^v) - F_{\boldsymbol{x}^v}^v(\boldsymbol{u}_*^v)) p(\boldsymbol{u}_*^v) d\boldsymbol{u}_*^v}{p(F_{\boldsymbol{x}^v}^v(\boldsymbol{u}^v))} d\boldsymbol{u}^v \end{aligned} \tag{19}$$

Let $\boldsymbol{m}^v = F_{\boldsymbol{x}^v}^v(\boldsymbol{u}_*^v), \boldsymbol{u}_*^v = H_{x^v}^v(\boldsymbol{m}^v)$ and $\det\left[\frac{\partial \boldsymbol{u}_*^v}{\partial \boldsymbol{m}^v}\right] = 1 \Big/ \det\left[\frac{\partial \boldsymbol{m}^v}{\partial \boldsymbol{u}_*^v}\right] = 1 \Big/ \det\left[\frac{\partial F_{\boldsymbol{x}^v}^v(\boldsymbol{u}_*^v)}{\partial \boldsymbol{u}_*^v}\right]$.

$$\begin{aligned} &\int p(\boldsymbol{u}^v) \log \frac{\int \delta(F_{\boldsymbol{x}^v}^v(\boldsymbol{u}^v) - F_{\boldsymbol{x}^v}^v(\boldsymbol{u}_*^v)) p(\boldsymbol{u}_*^v) d\boldsymbol{u}_*^v}{p(F_{\boldsymbol{x}^v}^v(\boldsymbol{u}^v))} d\boldsymbol{u}^v \\ &= \int p(\boldsymbol{u}^v) \log \frac{\int \delta(F_{\boldsymbol{x}^v}^v(\boldsymbol{u}^v) - \boldsymbol{m}^v) p(H_{\boldsymbol{x}^v}^v(\boldsymbol{m}^v)) \left|\det \frac{\partial H_{\boldsymbol{x}^v}^v(\boldsymbol{m}^v)}{\partial \boldsymbol{m}^v}\right| d\boldsymbol{m}^v}{p(F_{\boldsymbol{x}^v}^v(\boldsymbol{u}^v))} d\boldsymbol{u}^v \\ &= \int p(\boldsymbol{u}^v) \log \frac{\int \delta(F_{\boldsymbol{x}^v}^v(\boldsymbol{u}^v) - \boldsymbol{m}^v) p(H_{\boldsymbol{x}^v}^v(\boldsymbol{m}^v)) \left|\det \frac{\partial F_{\boldsymbol{x}^v}^v(\boldsymbol{u}_*^v)}{\partial \boldsymbol{u}_*^v}\right|^{-1} d\boldsymbol{m}^v}{p(F_{\boldsymbol{x}^v}^v(\boldsymbol{u}^v))} d\boldsymbol{u}^v \\ &= \int p(\boldsymbol{u}^v) \log \frac{p(\boldsymbol{u}^v)}{p(F_{\boldsymbol{x}^v}^v(\boldsymbol{u}^v)) \left|\det \frac{\partial F_{\boldsymbol{x}^v}^v(\boldsymbol{u}^v)}{\partial \boldsymbol{u}^v}\right|} d\boldsymbol{u}^v \\ &= \mathbb{E}_{p(\boldsymbol{u}^v)} \left[\frac{1}{2}\|F^v(\sigma_1^v \boldsymbol{u}^v + E(\boldsymbol{x}^v))\|^2 - \frac{1}{2}\|\boldsymbol{u}\|^2 - \log \left|\det \frac{\partial F_{\boldsymbol{x}^v}^v(\boldsymbol{u}^v)}{\partial \boldsymbol{u}^v}\right|\right] \end{aligned} \tag{20}$$

A.5 Pairwise Upper Bound on Distributional Incoherence

Theorem A.1 (Pairwise upper bound for individual modality incoherence) *Let $\mu_1, \ldots, \mu_V$ be probability distributions, where $V \geq 2$. Let $w_1, \ldots, w_V$ be random fusion weights such that $w_j > 0$ and $\sum_{j=1}^{V} w_j = 1$. Define $\bar{w}_j = \mathbb{E}[w_j]$, $\mu = \sum_{j=1}^{V} \bar{w}_j \mu_j$. Assume that the distribution discrepancy $\mathcal{D}$ satisfies the following properties:*

1. *$\mathcal{D}(P, Q) \geq 0$ for any distributions P and Q;*
2. *$\mathcal{D}(P, P) = 0$ for any distribution P;*
3. *for any fixed P, the map $Q \mapsto \mathcal{D}(P, Q)$ is convex.*

Then, for every $v \in \{1, \ldots, V\}$,

$$\mathcal{D}(\mu_v, \mu) \leq \sum_{\substack{v_1, v_2 = 1 \\ v_1 \neq v_2}}^{V} \mathcal{D}(\mu_{v_1}, \mu_{v_2}).$$

Proof A.1 *First, by the assumptions on the random fusion weights,*

$$\bar{w}_j = \mathbb{E}[w_j] > 0, \qquad \sum_{j=1}^{M} \bar{w}_j = \mathbb{E}\left[\sum_{j=1}^{V} w_j\right] = 1.$$

Thus, $\mu = \sum_{j=1}^{V} \bar{w}_j \mu_j$ is a convex combination of the modality distributions. Fix any $v \in \{1, \ldots, V\}$. Because $\mathcal{D}(\mu_v, \cdot)$ is convex, applying the definition of convexity to the above convex combination yields

$$\mathcal{D}(\mu_v, \mu) = \mathcal{D}\left(\mu_v, \sum_{j=1}^{V} \bar{w}_j \mu_j\right) \leq \sum_{j=1}^{V} \bar{w}_j \mathcal{D}(\mu_v, \mu_j).$$

The term corresponding to $j = v$ is zero, since $\mathcal{D}(\mu_v, \mu_v) = 0$. Hence,

$$\mathcal{D}(\mu_v, \mu) \leq \sum_{\substack{j=1 \\ j \neq m}}^{V} \bar{w}_j \mathcal{D}(\mu_v, \mu_j).$$

Moreover, $\bar{w}_j \leq 1$ because the $\bar{w}_j$ are nonnegative and sum to one. Together with the nonnegativity of $\mathcal{D}$, this implies

$$\sum_{\substack{j=1 \\ j \neq v}}^{V} \bar{w}_j \mathcal{D}(\mu_v, \mu_j) \leq \sum_{\substack{j=1 \\ j \neq v}}^{V} \mathcal{D}(\mu_v, \mu_j).$$

Finally, the sum on the right contains the ordered pairs (v, j) with $j \neq v$. All of these pairs are included in the set of ordered pairs (v_1, v_2) with $v_1 \neq v_2$. Every additional term is nonnegative; therefore,

$$\sum_{\substack{j=1 \\ j \neq v}}^{V} \mathcal{D}(\mu_v, \mu_j) \leq \sum_{\substack{v_1, v_2 = 1 \\ v_1 \neq v_2}}^{V} \mathcal{D}(\mu_{v_1}, \mu_{v_2}).$$

Combining the above inequalities, we obtain, for every $v \in \{1, \ldots, V\}$,

$$\mathcal{D}(\mu_v, \mu) \leq \sum_{\substack{j=1 \\ j \neq v}}^{V} \bar{w}_j \mathcal{D}(\mu_v, \mu_j) \leq \sum_{\substack{j=1 \\ j \neq v}}^{V} \mathcal{D}(\mu_v, \mu_j) \leq \sum_{\substack{v_1, v_2 = 1 \\ v_1 \neq v_2}}^{V} \mathcal{D}(\mu_{v_1}, \mu_{v_2}).$$

Thus, the divergence between any individual modality distribution and the aggregate distribution is bounded above by the sum of all pairwise modality divergences, as claimed.

A.6 Definition of the 1-Wasserstein distance

Let P and Q be probability distributions on $\mathbb{R}^d$ with finite first moments. Their 1-Wasserstein distance is defined as

$$\mathcal{W}_1(P,Q) = \inf_{\gamma \in \Pi(P,Q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\boldsymbol{z} - \boldsymbol{z}'\|_2 \, \gamma(\mathrm{d}\boldsymbol{z}, \mathrm{d}\boldsymbol{z}'). \tag{21}$$

Here, $\mathcal{W}_1$ denotes 1-Wasserstein distance. $\boldsymbol{z}$ and $\boldsymbol{z}'$ denote points in the latent space and $\Pi(P,Q)$ is the set of joint probability measures whose first and second marginals are P and Q, respectively. Each γ specifies a transport plan between the two distributions and the integral is its total transport cost.

For empirical distributions represented by B equally weighted latent samples, this definition reduces to the finite-dimensional optimization problem in Eq. (15).

A.7 Algorithm

Algorithm 1 Training with incomplete multiview observations

Require: Incomplete dataset $\mathcal{X} = \{\mathbf{X}^v\}_{v=1}^V$; view indicator matrix $\mathbf{M}$; batch size B; pretraining epochs N_{pre}; training epochs N_{train}; loss weights β, γ, λ; learning rate η; number of clusters K.

Ensure: Cluster labels $\{y_i\}_{i=1}^N$.

Stage 1: Pretraining

for $e = 1, \ldots, N_{\text{pre}}$ **do**
  for each minibatch $\mathcal{B}$ of size at most B **do**
    for each sample $i \in \mathcal{B}$ **do**
      Determine its observed-view set $\mathcal{V}_i = \{v : M_{iv} = 1\}$.
      Sample view-level representations $\{\boldsymbol{z}_i^v\}_{v \in \mathcal{V}_i}$ using the view-conditioned flows in Eq. (2).
      Integrate $\{\boldsymbol{z}_i^v\}_{v \in \mathcal{V}_i}$ into $\boldsymbol{z}_i$ using Eq. (3).
    end for
    Compute $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{align}}$ using Eqs. (11) and (16).
    Update the model parameters by minimizing $\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_{\text{align}}$.
  end for
end for

GMM initialization

Compute an integrated representation $\boldsymbol{z}_i$ for every training sample using its observed views.
Apply K-means with K clusters to $\{\boldsymbol{z}_i\}_{i=1}^N$.
Initialize the GMM parameters $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k\}_{k=1}^K$ from the K-means assignments.

Stage 2: Joint training

for $e = 1, \ldots, N_{\text{train}}$ **do**
  for each minibatch $\mathcal{B}$ of size at most B **do**
    Obtain $\{\boldsymbol{z}_i^v\}_{v \in \mathcal{V}_i}$ and $\boldsymbol{z}_i$ for each $i \in \mathcal{B}$ as in Stage 1.
    Compute $\mathcal{L}_{\text{rec}}$, $\mathcal{L}_z$, $\mathcal{L}_{z^v}$, and $\mathcal{L}_{\text{align}}$ using Eqs. (11), (9), (10) and (16), respectively.
    Update the model parameters by minimizing $\mathcal{L}_{\text{total}}$ in Eq. (17).
  end for
end for

Stage 3: Missing-view imputation and clustering

for $i = 1, \ldots, N$ **do**
  Compute the integrated representation $\boldsymbol{z}_i$ from the observed views $\{v : M_{iv} = 1\}$.
  for each missing view v such that $M_{iv} = 0$ **do**
    Impute $\hat{\boldsymbol{x}}_i^v$ using Eq. (7) .
  end for
  Construct the completed multiview sample $\tilde{\boldsymbol{x}}_i^v = M_{iv}\boldsymbol{x}_i^v + (1 - M_{iv})\hat{\boldsymbol{x}}_i^v$, for $v = 1, \ldots, V$.
  Compute the fused representation $\tilde{\boldsymbol{z}}_i$ from $\{\tilde{\boldsymbol{x}}_i^v\}_{v=1}^V$.
end for

Apply K-means with K clusters to $\{\tilde{\boldsymbol{z}}_i\}_{i=1}^N$.
return the resulting cluster labels $\{y_i\}_{i=1}^N$.

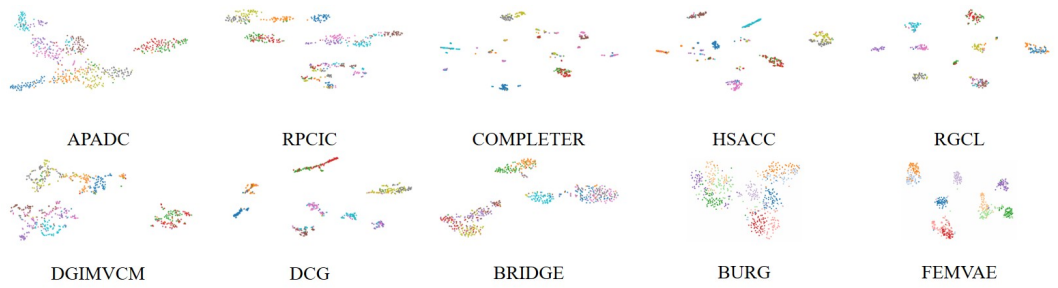

Figure 2: Visualization of the representations learned by different methods on CUB at a missing rate of 0.3. Different colors indicate different ground-truth classes.

A.8 Dataset Details

We conduct experiments on five widely used multi-view benchmark datasets, namely CUB, Handwritten, CCV, Scene-15, and Caltech7-5V. CUB contains 11,788 bird images from 200 categories, and we use the visual features provided for multi-view learning. Handwritten consists of 2,000 handwritten digit samples from 10 classes, represented by six views with feature dimensions of 216, 76, 64, 6, 240, and 47. CCV contains 9,317 consumer videos belonging to 20 semantic categories, with three released audio/visual feature representations. Scene-15 comprises 4,485 images from 15 indoor and outdoor scene categories, represented by three views with dimensions of 20, 59, and 40. Caltech7-5V is a seven-class subset of Caltech101 containing 1,400 samples and five views, whose feature dimensions are 40, 254, 928, 512, and 1,984.

A.9 Implementation Details

All experiments are conducted on a local server equipped with two NVIDIA GeForce RTX 3090 GPUs, each with 24 GB of memory, and 128 GB of RAM. For all datasets, the batch size is set to 256. The encoder follows a fully connected architecture with the structure d_v-512-1024-512-d, where d_v denotes the input dimension of the v-th view and d is set to the number of clusters K. ReLU activation is applied to all hidden layers, and the decoder mirrors the encoder structure in reverse. Each view is equipped with a four-layer conditional affine coupling flow with hidden dimensions of 128-128, a scale clipping factor of 1.0, and identity initialization.Both the pretraining and joint training stages use the Adam optimizer, with initial learning rates of 0.001 and 0.0003, respectively.A StepLR scheduler reduces the learning rate by a factor of 0.5 every 20 epochs. The pretraining and joint training stages run for 50 and 150 epochs, respectively. After pretraining, K-means is applied to the shared latent representations to initialize the K-component diagonal GMM. The Sinkhorn regularization coefficient is set to 0.1 with 20 iterations. The trade-off coefficients β, γ, and λ are selected from $\{0.01, 0.1, 1, 10, 100\}$.

A.10 Visualization Analysis

To qualitatively assess the learned representations, we compare FEMVAE with the competing methods on CUB at a missing rate of 0.3. Figure 2 presents the two-dimensional visualizations, where each point represents a sample and different colors denote different ground-truth classes. We examine both the compactness of samples belonging to the same class and the separation between samples from different classes.

Several competing methods, including APADC, BRIDGE, and BURG, exhibit noticeable class mixing or dispersed sample groups. In these visualizations, samples from the same class may spread across relatively broad regions, while samples from different classes occupy overlapping regions. Some competing methods also produce compact groups, but multiple classes remain mixed within individual groups, suggesting that compactness does not necessarily imply clear class separation. In comparison, FEMVAE produces relatively compact groups with clearer separation between many

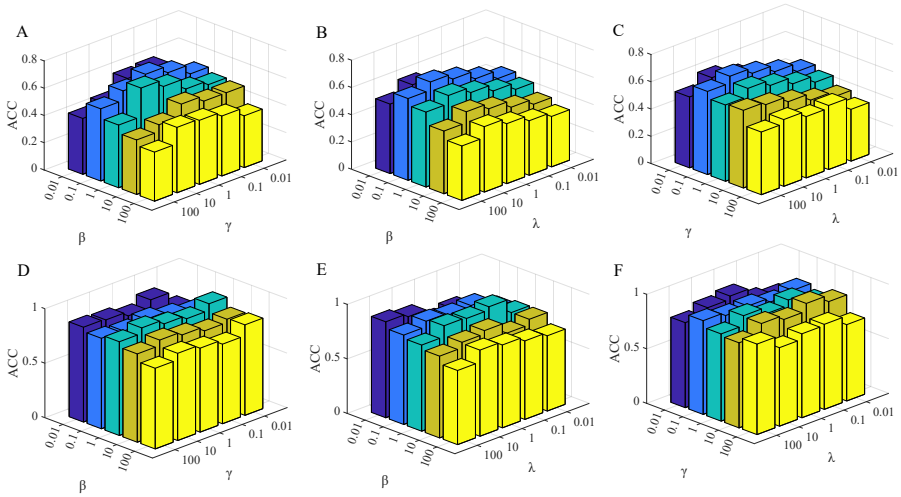

Figure 3: Parameter analysis on the CUB and Handwritten datasets. A Performance on the CUB dataset with $\lambda = 10$, while β and γ are varied within the predefined range. B Performance on the CUB dataset with $\gamma = 1$, while β and λ are varied. C Performance on the CUB dataset with $\beta = 0.1$, while γ and λ are varied. D Performance on the Handwritten dataset with $\lambda = 0.1$, while β and γ are varied. E Performance on the Handwritten dataset with $\gamma = 10$, while β and λ are varied. F Performance on the Handwritten dataset with $\beta = 1$, while γ and λ are varied.

classes. Although partial overlap remains between some classes, the overall arrangement shows a more distinguishable cluster structure in the projected space.

These observations suggest that FEMVAE can retain useful class-discriminative information from incomplete observations. The observed structure is consistent with the intended roles of its flexible view-specific posterior modeling, cluster-oriented Gaussian mixture prior, and cross-view distributional alignment, which jointly aim to preserve informative latent representations while encouraging cluster structure and semantic agreement. This qualitative evidence complements the quantitative results in Table 1, where FEMVAE achieves the highest ACC, NMI, and ARI on CUB under the same missing rate.

A.11 Parameter Analysis

To examine how the different learning objectives influence clustering performance, we investigate the sensitivity of the three loss weights β, γ, and λ, which control the regularization of the unified latent representation, the regularization of view-specific latent representations, and cross-view distributional alignment, respectively. The reconstruction coefficient is fixed to one throughout the experiments. Each parameter is varied over $0.01, 0.1, 1, 10, 100$ while the remaining parameters are fixed to representative values. To reduce the influence of randomness, the same missing-view masks and random seeds are used across different parameter settings. We evaluate the resulting clustering performance using ACC, NMI, and ARI. We further visualize the parameter sensitivity on the CUB and HandWritten datasets under a missing rate of 0.7, as shown in Figure 3.

As shown in Figure 3(A)-(C), the performance on CUB varies smoothly across a broad range of parameter combinations, although the sensitivity differs among the three loss weights. When λ is fixed, varying β and γ changes the balance between the unified representation and view-specific representations, resulting in a relatively smooth performance surface rather than an abrupt change. Similarly, when γ or β is fixed, the performance changes gradually with λ, indicating that the distributional alignment objective can contribute effectively when its weight is appropriately balanced with the representation regularization terms. Importantly, the performance does not exhibit a simple monotonic relationship with any individual parameter. Instead, the higher-performing regions generally arise from a suitable combination of the three weights, suggesting that the different objectives are complementary and should be jointly balanced during optimization.

A similar trend can be observed on the Handwritten dataset in Figure 3(D)-(F). The clustering performance remains within a relatively stable range over a substantial portion of the parameter space, while more noticeable changes occur when one objective is assigned an excessively small or large weight. In particular, increasing a single coefficient to an extreme value does not consistently improve the performance, since over-emphasizing one objective can weaken the contribution of the other learning objectives. This observation is consistent with the role of β, γ, and λ in coordinating the unified representation, view-specific information, and cross-view distributional consistency. The two datasets also exhibit different response patterns to the same parameters, indicating that the relative importance of the individual objectives depends on the characteristics of the underlying multi-view data.

Overall, the sensitivity analysis demonstrates that the proposed method does not rely on a narrowly tuned parameter configuration. Instead, competitive clustering performance can be obtained across a reasonably broad range of β, γ, and λ, while extreme parameter values may lead to performance degradation due to an imbalance among the learning objectives. Based on this analysis, we use a fixed set of parameter values for the remaining experiments, which provides a consistent configuration across different datasets and missing-view conditions.

A.12 Ablation Analysis

We conduct ablation experiments to evaluate the contributions of three key components: the view-conditioned normalizing flow posterior, the cross-view alignment loss, and the GMM prior. Specifically, we construct three variants. Gaussian Posterior replaces the flow-based posterior with a conventional diagonal Gaussian posterior. w/o Alignment removes the cross-view alignment loss $\mathcal{L}_{\text{align}}$. Gaussian Prior replaces the GMM prior with a standard Gaussian prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$. All other components and experimental settings remain unchanged. We report ACC, NMI, and ARI under four missing rates $\eta \in \{0.1, 0.3, 0.5, 0.7\}$ on CUB and Handwritten, with the results summarized in Table 2. The Gaussian Posterior variant obtains lower ACC than the full model across all evaluated settings. For example, on CUB at $\eta = 0.1$, its ACC and ARI are 0.67 and 0.54, respectively, compared with 0.75 and 0.62 for the full model. On Handwritten at $\eta = 0.7$, its ACC/NMI/ARI values are $0.82/0.76/0.74$, whereas the full model achieves $0.91/0.82/0.81$. These comparisons support the benefit of a more expressive flow-based posterior for learning clustering-relevant representations from incomplete observations. Removing the alignment loss leads to a larger decline in ACC on CUB as the missing rate increases, compared with the full model. The ACC/NMI/ARI values of w/o Alignment decrease from $0.74/0.71/0.60$ at $\eta = 0.1$ to $0.62/0.61/0.49$ at $\eta = 0.7$. On Handwritten at $\eta = 0.7$, this variant achieves $0.80/0.75/0.72$, falling below the full model on all three metrics. Without $\mathcal{L}_{\text{align}}$, the model lacks an explicit constraint encouraging distributional consistency across views. These results support the usefulness of cross-view alignment for integrating the available information under incomplete observations. The Gaussian Prior variant also exhibits substantial performance degradation. On CUB, its ACC decreases from 0.74 to 0.64, while its ARI decreases from 0.54 to 0.43 as η increases from 0.1 to 0.7. The contribution of the GMM prior is particularly evident on Handwritten: at $\eta = 0.7$, replacing it with a standard Gaussian prior reduces ACC from 0.91 to 0.72 and ARI from 0.81 to 0.70. These results suggest that the multimodal structure provided by the GMM prior offers useful guidance for organizing the unified latent representations into clusters. The full model maintains relatively stable performance on Handwritten, with ACC/NMI/ARI changing from $0.94/0.87/0.84$ at $\eta = 0.1$ to $0.91/0.82/0.81$ at $\eta = 0.7$. It achieves the highest ACC, NMI, and ARI across all evaluated missing rates on both CUB and Handwritten. The results support the complementary contributions of flexible posterior modeling, cross-view distributional alignment, and a cluster-oriented GMM prior to incomplete multi-view clustering.