HARNESSING CAUSALITY-DRIVEN META-LEARNING FOR IMPUTATION IN INCOMPLETE MULTI-VIEW CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years, incomplete multi-view clustering algorithms (IMVC) have attracted extensive attention. In IMVC, the quality of missing-view imputation is a key factor determining clustering quality. However, existing IMVC approaches predominantly exploit statistical correlations and cross-view contrastive information for view completion. They fail to account for two intrinsic data characteristics: 1) the inherent causal relationships that define the logical connections between samples; 2) the superfluous features that lack discriminative power for downstream clustering tasks. This joint deficiency ultimately renders clustering predictions heavily dependent on extrinsic observed data patterns, resulting in suboptimal clustering performance. To address these issues, we propose a novel IMVC method based on Causal Meta-Learning Imputation (CMLI), which embeds causal knowledge into the imputation module orchestrated by meta-learning. Specifically, CMLI adopts meta-learning to distill critical causal information from features, thereby improving the discrimination of task-irrelevant information in the imputation process. Leveraging the Information Bottleneck (IB) principle, CMLI further optimizes the trade-off between compressing local features and preserving global features by simultaneously minimizing one mutual information term while maximizing another. Finally, the consistency learning between the label features effectively exploits cross-view clustering consistency and complementarity, resulting in a unified framework that couples clustering assignment with multi-view representation learning. Comprehensive experimental results validate that our proposed method surpasses state-of-the-art approaches.

1 INTRODUCTION

Driven by the need to handle complex and heterogeneous data, Multi-View Clustering (MVC) has been widely adopted for multimodal data processing (Cao et al., 2025; Wang et al., 2025; Sattari & Yazici, 2025). However, real-world multi-view data often suffers from missing views due to hardware faults, storage corruption, or network transmission errors. Consequently, existing MVC approaches struggle to handle such incomplete multi-view data. As a practical extension of MVC that dispenses with the completeness assumption, Incomplete Multi-View Clustering (IMVC) seeks to reveal common clustering patterns from partially observed multi-view data (Wen et al., 2022).

Based on whether the missing views are recovered prior to clustering, existing IMVC methods are broadly divided into imputation-free and imputation-based methods. Imputation-free IMVC methods posit that partial views contain sufficient information for downstream tasks, aiming to fully leverage the clustering structures inherent in the available views (Tan et al., 2018; Bai et al., 2024). In contrast, imputation-based IMVC methods aim to restore information diversity. They impute the original views through techniques including tensor factorization, low-rank matrix completion, and deep generative models before executing the standard MVC pipeline. For example, Yin & Sun (2021) learn a shared latent representation through matrix factorization, which is then used to reconstruct the missing views. Li et al. (2023) recover missing views by leveraging the sample-prototype relationships. Yao et al. (2024) design a dual tensor constraint module for missing-view reconstruction.

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.

These imputation-based methods undoubtedly strengthen the data foundation for subsequent clustering, yet they mainly reconstruct missing views based on statistical feature-wise or sample-wise dependency (Liu et al., 2026; Zhang et al., 2026). Modeling only statistical correlations is intrinsically limited to undirected associations, rendering it vulnerable to data incompleteness and incapable of confounder discrimination. Additionally, imputed data often harbor irrelevant information beyond mere noise, including spurious entries or even view-private attributes. Such vulnerabilities significantly hinder the efficacy of existing imputation-based methods in scenarios with limited resources.

To address the above issues, we introduce a novel incomplete multi-view clustering framework named CMLI, which imputes missing views from a causal perspective using meta-learning. The proposed framework is illustrated in Figure 1. Firstly, we compress task-irrelevant details in raw view-specific data to obtain the compressed representation $\mathbf{Z}^v$, thus providing a more reliable data foundation. Notably, we fill missing values in missing views by mean imputation before applying dual embeddings, which incorporate cross-view information for subsequent causal imputation. Then, we utilize dual embeddings to construct view-specific causal graphs $\mathbf{CG}^i$, which are then refined into convolution kernels via meta-learning for imputation of missing views. The recovered representations $\mathbf{Z}^v_{rev}$ are then aligned with the observed views via contrastive learning over a limited set of paired samples. Additionally, we adopt an attention layer to fuse each $\mathbf{Z}^v$, and design two feature extractors to produce high-level features and label features, respectively. To simultaneously capture consistent and complementary information, we perform information preservation by maximizing the mutual information between high-level features $\mathbf{H}^i$ and fused feature $\mathbf{Z}$. Building on this, we further introduce a cluster consistency learning over the label features $\mathbf{Q}^i$ and $\mathbf{Q}$, which not only enhances accuracy but also enables end-to-end clustering by fully leveraging the implicit clustering consistency. Our main contributions can be summarized as follows:

- We propose a novel IMVC framework that incorporates causal meta-learning for imputation, tackling the challenge of incomplete views by leveraging causality rather than relying solely on traditional statistical correlations.
- To achieve reliable representation learning and end-to-end clustering, we innovatively integrate IB principle with cluster consistency learning within CMLI, thereby alleviating the encoding of task-irrelevant information while enhancing the exploration of task-relevant information.
- Extensive experiments are conducted to evaluate CMLI, demonstrating its superiority over state-of-the-art IMVC methods across multiple datasets under varying missing rates.

2 Related work

2.1 Deep Incomplete Multi-View Clustering with View Imputation

Benefiting from the strong feature extraction ability of deep learning, deep incomplete multi-view clustering (DIMVC) has witnessed substantial advancements. According to the techniques used for missing-view completion, current DIMVC approaches fall into four main categories: 1) Graph-based approaches (Zhang et al., 2026; Chao et al., 2025) utilize information from neighboring complete views via constructed graph structures to infer missing views. 2) Generative network-based approaches (Wen et al., 2024; Wang et al., 2025) learn representations from the observed views to drive generative models such as GANs or diffusion models for reconstructing the missing views. 3) Entropy-based approaches (Lin et al., 2022; Yin et al., 2025) aim to minimize the conditional entropy to bidirectionally predict the paired missing views. 4) Prototype-based approaches (Li et al., 2023; Yuan et al., 2025) recover missing data from instance-specific information by selecting prototypes and building sample-prototype relations with the observed samples. Although existing DIMVC approaches achieve effective view imputation through various strategies, they overlook the inherent causal relations and feature redundancy in samples.

2.2 Causal discovery with missing data

Missing data poses a ubiquitous and serious challenge in real-world applications, as incomplete observations can undermine causal conclusions. The mechanisms of missingness can be categorized

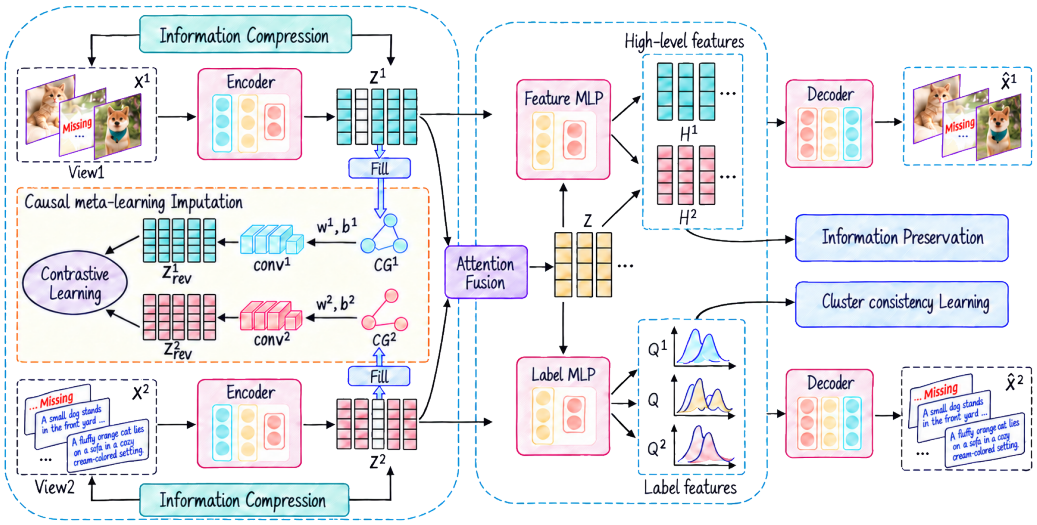


Figure 1: The architecture of the proposed CMLI framework. As shown, our CMLI comprises four core modules: (a) View-specific autoencoders with information compression for extracting initial deep features. These deep features are then fused by the attention layer. (b) Causal meta-learning imputation modules for constructing causal graphs over deep features and imputing the missing views through meta-convolution kernel. (c) The feature MLP that learns high-level features for information preservation. (d) The label MLP that learns label features for cluster consistency learning.

into three types: 1) Missing Completely At Random (MCAR): the missingness is purely random and independent of both the observed and unobserved variables. 2) Missing At Random (MAR): the missingness depends on other observed variables. 3) Missing Not At Random (MNAR): the missingness depends on the unobserved data itself. In both MCAR and MAR cases, the underlying data distribution can be identified from incomplete observations without additional assumptions (Mohan & Pearl, 2021). Accordingly, most causal discovery methods for missing data are mainly built upon the MCAR and MAR assumption (Gao et al., 2022; Wang et al., 2020). Traditional list-wise deletion methods perform poorly in small samples or samples with high missing rates, motivating a shift toward imputation-based approaches. Recent studies have emphasized the intrinsic coupling of causal discovery and data imputation, including directly integrating causal structure learning into imputation (Woldemariam et al., 2026), adopting the EM algorithm to jointly optimize causal discovery with imputation (Gao et al., 2022), and building bidirectional interactions linking causal discovery to imputation (Zhu & Zhao, 2025).

3 METHODOLOGY

3.1 PRELIMINARIES

Let $\mathcal{X} = \{\mathbf{X}^v\}_{v=1}^V$ denote a incomplete multi-view dataset with N samples across V views, where $\mathbf{X}^v = [x_1^v, x_2^v, \ldots, x_N^v]^\top \in \mathbb{R}^{N \times D_v}$ represents the instance set of the v-th view in a D_v-dimensional space. The matrix $\mathbf{M} \in \{0,1\}^{N \times V}$ indicates the positions of missing views, where $\mathbf{M}_{ij} = 1$ if the j-th view of the i-th sample is missing and $\mathbf{M}_{ij} = 0$ otherwise. We assume that each sample is available in at least one view, i.e. $\sum_{j=1}^V \mathbf{M}_{ij} < V, \forall i \in \{1, \cdots, N\}$. Our objective is to cluster these N samples that may have missing views into K clusters.

The Information Bottleneck (IB) principle (Tishby & Zaslavsky, 2015) characterizes the optimal trade-off between compressing input data and preserving predictive information about the output data within the latent representation. It suggests that discarding useless and redundant information helps learn a robust compressed representation $\mathbf{Z}$. The essence of IB lies in compressing irrelevant information by minimizing the mutual information (MI) $I(\mathbf{X}; \mathbf{Z})$ between the source variable $\mathbf{X}$

and $\mathbf{Z}$, while preserving relevant information by maximizing the MI $I(\mathbf{Z}; y)$ between $\mathbf{Z}$ and the target variable y. The objective function of IB principle can be written as:

$$\max_{\mathbf{Z}} I(\mathbf{Z}, y) - \lambda I(\mathbf{X}, \mathbf{Z}), \tag{1}$$

where λ is a balance parameter. Recent studies have increasingly leveraged the IB for multi-view learning (Yan et al., 2024; Wan et al., 2021).

3.2 Information compression and preservation

As an effective approach to mitigating the instability of manifold structure in high-dimensional space, deep autoencoders are widely adopted for unsupervised representation learning (Lin et al., 2021; Guo et al., 2017). Since each view carries its own distinctive feature information, we design a view-specific autoencoder to learn the latent representation $\mathbf{Z}^v \in \mathbb{R}^{N \times D}$ by reducing the reconstruction loss:

$$\mathcal{L}_{\mathbf{R}} = \sum_{v=1}^{V} \sum_{i=1}^{N} \left\| \mathbf{x}_i^v - D_{\phi^v}^{(v)} \left(\mathbf{z}_i^v \right) \right\|_2^2, \tag{2}$$

where $\mathbf{x}_i^v$ and $D_{\phi^v}^{(v)}$ denote the i-th sample of $\mathbf{X}^v$ and the decoder of the v-th view with the parameters ϕ^v, respectively. The latent representation $\mathbf{z}_i^v$ is produced as:

$$\mathbf{z}_i^v = E_{\theta^v}^{(v)}(\mathbf{x}_i^v) \tag{3}$$

where $E_{\theta^v}^{(v)}$ is the encoder of v-th view with the parameters θ^v.

To further compress redundant information and encourage robust feature representations, we integrate IB principle into the representation learning process. Following (Yu et al.; Tian et al., 2026; Yin et al., 2024), we use Cauchy-Schwarz (CS) divergence as a non-parametric MI estimator to implement the objective function described in Eq.(1). Formally, the CS divergence is defined as:

$$\mathrm{D}_{\mathrm{cs}}(p; q) = -\log \left(\frac{\left(\int p(x) q(x)\, dx \right)^2}{\int p(x)^2\, dx \int q(x)^2\, dx} \right), \tag{4}$$

The CS-based MI estimation is then derived by substituting the joint distribution $p(x, z)$ for $p(x)$ and the product of marginals $p(x)p(z)$ for $q(x)$ in Eq.(4):

$$\begin{aligned} I_{\mathrm{CS}} = & \log \left(\int p^2(x, z)\, dx\, dz \right) + \log \left(\int p^2(x)\, p^2(z)\, dx\, dz \right) \\ & - 2 \log \left(\int p(x, z)\, p(x)\, p(z)\, dx\, dz \right). \end{aligned} \tag{5}$$

Note that I_{CS} is formulated in terms of quadratic density functionals and interactions among probability densities. Thus, we can adopt multi-kernel Gram matrices to approximate functionals involved in the I_{CS} so as to better fit the complex distributions. The multi-kernel matrix $\mathbf{K}$ derived from the Gaussian kernel over the sample set $\{x_i\}_{i=1}^N$ is defined as:

$$\begin{aligned} \mathbf{K} &= f\left(\mathbf{K}^{(m)} \right) = f\left(\left[k^{(m)}(x_i, x_j) \right]_{i,j} \right), \\ k^{(m)}(x_i, x_j) &= \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_m^2} \right), \end{aligned} \tag{6}$$

where $f(\cdot)$ denotes the multi-kernel aggregate function, $\mathbf{K}^{(m)}$ denotes the kernel matrix computed using the m-th kernel, and σ_m denotes the kernel bandwidth. With this, we can approximate the quadratic density functionals in Eq.(5) using the kernel matrices $\mathbf{K}^x$ and $\mathbf{K}^z$, constructed from $\mathbf{X}^v$ and $\mathbf{Z}^v$, respectively. The approximation is presented as follows, and its derivation is provided in the Appendix A.

$$\begin{aligned} I_{\mathrm{CS}}(X; Z) \approx & \log \left(\frac{1}{N^2} \mathrm{tr}(\mathbf{K}^x \mathbf{K}^z) \right) + \log \left(\frac{1}{N^4} \mathbf{1}^\top \mathbf{K}^x \mathbf{1} \cdot \mathbf{1}^\top \mathbf{K}^z \mathbf{1} \right) \\ & - 2 \log \left(\frac{1}{N^3} \mathbf{1}^\top \mathbf{K}^x \mathbf{K}^z \mathbf{1} \right), \end{aligned} \tag{7}$$

where $\mathbf{1}$ denotes an N-dimensional vector of all ones.

Accordingly, we can achieve information compression by directly minimizing $I_{CS}(\mathbf{X}^v, \mathbf{Z}^v)$, which corresponds to the second term in Eq.(1). Regarding the first term $I(\mathbf{Z}, y)$, the target variable y represents the latent cluster assignment, which is not accessible during training. To this end, we adopt $I(\mathbf{Z}^v, \mathbf{Z})$ as an approximation to achieve information preservation. The unified representation $\mathbf{Z}$ is obtained by fusing the view-specific representations $\{\mathbf{Z}^v\}_{v=1}^V$ via a self-attention layer. Details of this approximation are provided in Appendix B. Since the bottom-level feature sets $\{\mathbf{Z}^v\}_{v=1}^V$ often contain significant noise and redundancy, we apply a feature MLP to learn high-level feature representations $\{\mathbf{H}^v\}_{v=1}^V$. This process aids in preserving high-level semantics and ensuring cross-view consistency. Finally, the loss function for information compression and preservation is formulated as follows:

$$\mathcal{L}_{\mathbf{IB}} = \alpha \sum_{v=1}^{V} I_{CS}\left(\mathbf{X}^v, \mathbf{Z}^v\right) - \beta \sum_{v=1}^{V} I_{CS}\left(\mathbf{H}^v, \mathbf{Z}\right), \tag{8}$$

where α and β are fixed balancing coefficients set manually before training.

3.3 Causal meta-learning Imputation

Assume that $\{\mathbf{Z}^v\}_{v=1}^V$ satisfy causal identifiability. For each view, we initially impute the missing values in $\mathbf{Z}^v$ using the mean of the corresponding values from the available views to construct $\mathbf{S}^v$. Then, we quantify the underlying causal dependencies by projecting features in $\mathbf{S}^v$ into the t-dimensional structural embedding space and observational embedding space to yield the corresponding dual embedding representations $\mathbf{SR} \in \mathbb{R}^{D\times t}$ and $\mathbf{OR} \in \mathbb{R}^{D\times t}$. Furthermore, inspired by the concept of multi-kernel learning (Gönen & Alpaydın, 2011; Lu et al., 2022), we construct corresponding structural representations $\left\{\mathbf{SR}^{(m)}\right\}_{m=1}^M$ from M embedding spaces so as to improve stability of the causal graph. The dual embedding representations are defined as follows:

$$\begin{aligned} \mathbf{SR}^{(m)} &= f_1^{(m)}\left(s_{:1}^v, s_{:2}^v, \cdots, s_{:D}^v\right) = \left\{sr_1^{(m)}, sr_2^{(m)}, \cdots, sr_D^{(m)}\right\}^\top, \\ \mathbf{OR} &= f_2\left(s_{:1}^v, s_{:2}^v, \cdots, s_{:D}^v\right) = \left\{or_1, or_2, \cdots, or_D\right\}^\top, \end{aligned} \tag{9}$$

where $f_1^{(m)}(\cdot)$ and $f_2(\cdot)$ denote the mapping operation. By leveraging dual embedding representations, causal relationships are captured to build a causal graph $\mathbf{CG} \in \mathbb{R}^{D\times D}$ as follows:

$$\mathbf{CG} = \frac{1}{M}\sum_{m=1}^{M} \mathrm{Nor}\left[\tanh(\mathbf{SR}^{(m)} \times \mathbf{OR}^\top) + \mathbf{I}\right], \tag{10}$$

where $\mathrm{Nor}\left[\cdot\right]$ is the normalization operation, and $\tanh$ is an activation function. Next, we aim to refine crucial information from causal graph to generate weights that effectively guide the prediction of imputation values. To this end, we first employ a two-layer fully connected extractor to capture the information as parameters of meta-convolution kernel, which is formulated as:

$$\begin{aligned} \mathbf{CI}^{(1)} &= \sigma(\mathbf{CGW}^{(11)} + b^{(11)})\mathbf{W}^{(12)} + b^{(12)}, \\ \mathbf{CI}^{(2)} &= \sigma(\mathbf{CGW}^{(21)} + b^{(21)})\mathbf{W}^{(22)} + b^{(22)}, \end{aligned} \tag{11}$$

where $\mathbf{W}^{(11)}$,$\mathbf{W}^{(12)}$,$\mathbf{W}^{(21)}$, and $\mathbf{W}^{(22)}$ are learnable weight parameters, $b^{(11)}$, $b^{(12)}$, $b^{(21)}$, and $b^{(22)}$ are the corresponding biases. The refined information $\mathbf{CI}^{(1)} \in \mathbb{R}^{D\times 2a\gamma}$ and $\mathbf{CI}^{(2)} \in \mathbb{R}^{D\times 2a}$ are subsequently used for weight generation, where a and γ denote the number and size of the meta convolution kernels, respectively. By performing row-wise slicing on $\mathbf{CI}^{(1)}$ and $\mathbf{CI}^{(2)}$, we can transform them into the weight and bias of meta-kernel convolution for each feature d $(1 \le d \le D)$. Specifically, these are represented as the weight tensor $\varphi_{meta}^d \in \mathbb{R}^{2a\times 1\times\gamma\times 1}$ and the bias vector $b_{meta}^d \in \mathbb{R}^{2a}$, respectively. Then, we can obtain the imputed values using these meta convolution kernels, where the process is enhanced by a residual connection with a gating mechanism, as shown below:

$$\begin{aligned} [\mathbf{R}^d, \mathbf{T}^d] &= Padding(\varphi_{meta}^d * \mathbf{z}_{:d}^v + b_{meta}^d), \\ \mathbf{U}^d &= (\mathbf{R}^d + Res(\mathbf{z}_{:d}^v)) \odot \sigma(\mathbf{T}^d), \\ \hat{\mathbf{z}}_{:d}^v &= Sum(\mathbf{U}^d), \\ \hat{\mathbf{Z}}^v &= concat\left(\hat{\mathbf{z}}_{:1}^v, \hat{\mathbf{z}}_{:2}^v, \cdots, \hat{\mathbf{z}}_{:D}^v\right), \end{aligned} \tag{12}$$

where $\hat{\mathbf{Z}}$ and $\hat{\mathbf{z}}_{:d}$ are the predicted imputation values, $\mathbf{R}^d, \mathbf{T}^d, \mathbf{U}^d \in \mathbb{R}^{2a\times N\times 1}$ are nonlinear features, $\odot$ is the Hadamard product. Finally, the recovered complete representation of each view $\mathbf{Z}^v_{rev}$ can be imputed by the corresponding value in $\hat{\mathbf{Z}}$, which is formulated as:

$$\mathbf{Z}^v_{rev} = \hat{\mathbf{Z}}^v \odot \mathbf{M}_{:v} + \mathbf{Z}^v \odot (1 - \mathbf{M}_{:v}) \tag{13}$$

For remaining components of $\mathbf{Z}^v$ that are not used for imputation, we use them as supervision for evaluating reconstruction quality, deriving the following reconstruction loss:

$$\mathcal{L}_{mr} = \sum_{v=1}^{V} \frac{\left\| \left(\hat{\mathbf{Z}}^v - \mathbf{Z}^v \right) \odot (1 - \mathbf{M}_{:v}) \right\|_2^2}{\sum_{i=1}^{N} (1 - \mathbf{M}_{iv})} \tag{14}$$

Additionally, to improve the cross-view consistency, we apply contrastive learning (Chen et al., 2020) on a limited set of paired samples between recovered view $\mathbf{Z}^m_{rev}$ and the other observed views, thereby aligning the recovered representation with the available observations:

$$\mathcal{L}_{rcl} = -\frac{1}{2N} \sum_{m=1}^{V} \sum_{n\neq m} \sum_{i=1}^{N} \log \frac{e^{sim[(\mathbf{Z}^m_{rev})_i, \mathbf{Z}^n_i]/\tau_R}}{\sum_{j=1}^{N} \sum_{v=m,n} e^{sim[(\mathbf{Z}^m_{rev})_i, \mathbf{Z}^v_j]/\tau_R}}, \tag{15}$$

where $sim(\cdot,\cdot)$ denotes the cosine similarity, and τ_R is the temperature parameter to modulate the smoothness. In general, the overall loss of the causal meta-learning imputation module is formulated as:

$$\mathcal{L}_{\mathbf{I}} = \mathcal{L}_{mr} + \mathcal{L}_{rcl}. \tag{16}$$

3.4 CLUSTER CONSISTENCY LEARNING

Unlike conventional approaches that necessitate post-processing steps such as k-means or spectral clustering, we seek to jointly optimize the clustering objective alongside the representation learning objective. To this end, we introduce a label MLP that shares weights across different views to directly generate label features $\{\mathbf{Q}^v\}_{v=1}^V \in \mathbb{R}^{N\times K}$ for end-to-end clustering. Motivated by the analogy between the clustering objective and the contrastive learning objective (Chen et al., 2020), we adopt a category-level contrastive loss that enhances model robustness and promotes cluster consistency:

$$\begin{aligned} \mathcal{L}_{ccl} &= \frac{1}{2} \sum_{m=1}^{V} \sum_{n\neq m} \ell^{(mn)} + \sum_{v=1}^{V} \sum_{j=1}^{K} s^v_j \log s^v_j, \\ \ell^{(mn)} &= -\frac{1}{K} \sum_{j=1}^{K} \log \frac{e^{sim(\mathbf{Q}^m_{:j}, \mathbf{Q}^n_{:j})/\tau_C}}{\sum_{k=1}^{K} \sum_{v=m,n} e^{sim(\mathbf{Q}^m_{:j}, \mathbf{Q}^v_{:k})/\tau_C}}, \end{aligned} \tag{17}$$

where $s^v_j = \frac{1}{N} \sum_{i=1}^{N} q^v_{ij}$ is the regularization term preventing all samples from being assigned to the same cluster (Huang et al., 2020; Chao et al., 2024), and τ_C denotes the temperature parameter to modulate the smoothness. Furthermore, to learn complementary information, we leverage a self-supervised confidence enhancement approach formulated via KL divergence. Specifically, we first feed the unified representation $\mathbf{Z}$ into the feature MLP to obtain the corresponding label feature $\mathbf{Q}$ as soft cluster assignment. Subsequently, the high-confidence target is derived by assigning each instance to its maximum soft cluster assignment:

$$\mathbf{q}_{ij} = \max\{\mathbf{Q}_{ij}, \mathbf{Q}^v_{ij}\}, \tag{18}$$

which is used as the target for auxiliary generation of the final soft cluster assignment. The following normalization operation further enhance the confidence and blur instances at cluster boundaries:

$$\mathbf{p}_{ij} = \frac{\mathbf{q}^2_{ij}}{\sum_{j=1}^{k} \mathbf{q}^2_{ij}}, \tag{19}$$

Based on KL divergence, the optimization objective of the clustering results is derived as follows:

$$\mathcal{L}_{hc} = KL(\mathbf{P} \| \mathbf{Q}) = \sum_{i=1}^{N} \sum_{j=1}^{K} \mathbf{p}_{ij} \log \frac{\mathbf{p}_{ij}}{\mathbf{q}_{ij}}, \tag{20}$$

Accordingly, we formulate the overall loss for the clustering consistency learning module as:

$$\mathcal{L}_{\mathbf{C}} = \mathcal{L}_{ccl} + \mathcal{L}_{hc}. \tag{21}$$

3.5 THE OBJECTIVE FUNCTION

In summary, the total loss of our model comprises four terms and is formulated as:

$$\mathcal{L} = \mathcal{L}_{\mathbf{R}} + \mathcal{L}_{\mathbf{IB}} + \lambda_1 \mathcal{L}_{\mathbf{I}} + \lambda_2 \mathcal{L}_{\mathbf{C}} \tag{22}$$

where λ_1 and λ_2 are trade-off coefficients. Since our model constitutes an end-to-end clustering framework free of post-processing, we can optimize all components jointly in a unified manner.

4 EXPERIMENTS

To validate the effectiveness of our CMLI, we conduct extensive experiments to answer the following questions: (Q1) Does CMLI outperform state-of-the-art IMVC methods? (Q2) Does each component of CMLI positively contribute to performance? (Q3) How sensitive is the performance of CMLI to variations in its key hyperparameters? (Q4) What clustering structures does CMLI identify?

4.1 EXPERIMENTAL SETUP

We evaluated our method on six representative datasets: Synthetic3d (Kumar et al., 2011), CUB (Wah et al., 2011), Fashion (Xiao et al., 2017), LandUse-21 (Yang & Newsam, 2010), Handwritten (Asuncion et al., 2007) and Scene-15 (Fei-Fei & Perona, 2005). Due to space limitations, additional details regarding and experimental results are provided in Appendix C to F . Following (Lin et al., 2024), we simulate incomplete multi-view data by randomly selecting m instances and removing one view from each. Consequently, the missing rate is defined as the proportion of incomplete instances as m/n, where n is the total number of instances. Three widely used clustering metrics are employed to facilitate a comprehensive evaluation, namely: Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI). Greater values on these metrics correspond to better clustering results. Our model is implemented based on PyTorch 2.10.0 and trained on a desktop computer configured with NVIDIA GeForce RTX 5090 and 32G RAM.

4.2 COMPARATIVE RESULTS ANALYSIS (Q1)

We compare CMLI with eight state-of-the-art IMVC methods, including DCP (Lin et al., 2022), APADC (Xu et al., 2023), ICMVC (Chao et al., 2024), RPCIC (Yuan et al., 2024), IMDC-DIA (Liu et al., 2026), IMC-MCL (Yin et al., 2025), DCG (Zhang et al., 2025), and DIMVC-HIA (Du et al., 2026). To ensure a fair comparison, all baselines adopted the experimental settings from their original implementations, with consistent pre-processing and missing-pattern generation.

We present the comparison results of all methods under different missing rates in Table 1. It can be observed that: (1) Our method achieves the best or second-best results in comparisons across four datasets, three missing rates, and three evaluation metrics. For example, at a missing rate of 0.1 on LandUse-21, our method surpasses the second-best method by 3.27%, 4.91%, and 5.47% on ACC, NMI, and ARI, respectively. This demonstrates the superior effectiveness of our proposed method over the baselines. Moreover, we observe that our method achieves larger overall improvements in NMI and ARI than in ACC. This indicates that our CMLI not only enhances the clustering accuracy, but also improved the clustering structure and category consistency. (2) As the missing rate increases, our method exhibits relatively small overall declines in ACC, NMI, and ARI, which demonstrates its strong robustness to missing data. We also observe fluctuating results on CUB, whereas the results on the other datasets decrease monotonically as the missing rate increases. We hypothesize that this behavior may result from variations in view quality and the regularizing effect of moderate missingness of the CUB dataset.

4.3 ABLATION STUDIES (Q2)

To assess the contribution of each component, we performed an ablation study on three datasets with a missing rate of 0.3. As shown in Table 2, the best results are achieved when all strategies are applied together. Notably, removing the causal meta-learning imputation module $\mathcal{L}_{\mathcal{I}}$ causes a significant performance degradation on all three datasets, which demonstrates that the designed

Table 1: Clustering performance of all methods on the four datasets. The best and second-best results are marked in boldface and by underlining, respectively.

Datasets	Missing rate	0.1			0.3			0.5		
	Metrics	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Synthetic3d	DCP(2022)	89.83	68.22	72.27	89.00	66.38	70.75	84.00	57.93	60.03
	APADC(2023)	89.83	67.68	72.31	84.67	63.41	61.79	84.50	56.06	60.01
	ICMVC(2024)	84.67	58.49	61.13	81.83	53.52	55.82	83.67	53.79	58.48
	RPCIC(2024)	66.17	50.52	47.11	65.67	42.60	40.28	65.67	41.51	38.90
	IMDC-DIA(2025)	86.33	62.00	64.68	85.17	58.41	61.26	80.67	51.80	52.62
	IMC-MCL(2025)	86.67	61.22	65.51	84.67	55.97	60.58	82.00	53.00	55.67
	DCG(2025)	91.17	71.14	75.95	88.50	64.20	69.34	87.33	61.81	66.89
	DIMVC-HIA(2026)	75.17	38.94	41.17	72.67	33.27	36.50	67.17	24.26	26.54
	OURS	**91.50**	**72.54**	**76.86**	**90.33**	**68.64**	**73.67**	**88.67**	**63.94**	**69.62**
CUB	DCP(2022)	65.00	69.97	52.52	24.67	37.21	13.04	28.67	40.91	15.63
	APADC(2023)	54.00	58.30	44.77	42.67	50.83	32.53	35.00	44.99	23.18
	ICMVC(2024)	64.33	64.95	51.89	56.67	53.85	37.09	45.17	43.09	25.77
	RPCIC(2024)	66.17	67.44	53.02	54.33	58.70	40.50	58.83	62.95	44.26
	IMDC-DIA(2025)	43.83	42.11	25.74	40.50	44.64	30.10	36.83	39.33	23.39
	IMC-MCL(2025)	62.67	68.79	53.89	62.17	66.32	52.70	64.33	63.54	49.31
	DCG(2025)	73.50	69.76	56.95	66.50	66.67	50.60	65.33	**65.59**	50.64
	DIMVC-HIA(2026)	62.83	66.02	50.43	48.83	54.78	37.33	52.67	58.51	43.79
	OURS	**75.67**	**72.16**	**60.95**	**77.17**	**71.90**	**61.70**	**67.83**	63.55	**51.26**
Fashion	DCP(2022)	73.78	83.61	71.33	72.10	80.63	68.97	81.26	80.89	72.65
	APADC(2023)	57.48	72.98	44.86	46.36	61.68	25.55	52.63	64.07	31.79
	ICMVC(2024)	81.42	83.70	75.66	80.84	81.42	73.42	74.85	72.36	64.14
	RPCIC(2024)	73.14	70.67	59.50	53.25	60.99	43.31	47.26	58.00	38.47
	IMDC-DIA(2025)	61.19	58.99	47.90	59.10	56.05	44.67	57.99	58.19	45.39
	IMC-MCL(2025)	76.16	84.79	72.67	73.27	76.91	63.60	73.19	74.06	62.68
	DCG(2025)	95.08	89.95	89.72	92.09	85.38	83.99	89.71	81.68	79.68
	DIMVC-HIA(2026)	93.91	88.96	87.67	86.07	84.97	80.84	80.54	78.32	74.85
	OURS	**95.31**	**90.71**	**90.27**	**93.51**	**87.71**	**86.80**	**90.62**	**83.34**	**81.45**
LandUse-21	DCP(2022)	26.33	29.92	13.54	25.86	28.71	12.15	25.48	27.99	11.77
	APADC(2023)	23.62	28.61	9.23	20.48	26.45	6.98	18.43	19.81	4.85
	ICMVC(2024)	24.05	29.74	11.58	25.14	30.26	12.34	25.10	27.72	11.76
	RPCIC(2024)	19.95	24.55	7.14	17.14	20.80	5.27	15.14	17.46	3.53
	IMDC-DIA(2025)	25.48	29.22	11.00	24.43	28.17	10.40	22.43	25.00	8.16
	IMC-MCL(2025)	22.43	26.66	9.58	22.24	27.27	9.46	21.33	24.78	8.45
	DCG(2025)	25.43	29.76	12.77	25.52	28.60	12.36	25.81	27.74	11.73
	DIMVC-HIA(2026)	23.86	27.13	11.15	24.62	28.54	11.04	25.14	26.58	11.39
	OURS	**27.19**	**31.39**	**14.28**	**26.62**	**31.08**	**14.15**	**26.76**	**29.78**	**14.03**

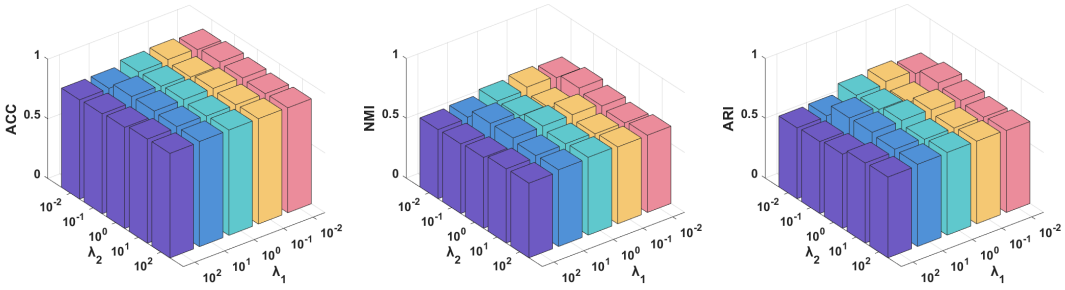
Table 2: Ablation results on Synthetic3d, Fashion, and LandUse-21 at 0.3 missing rate

Datasets			Synthetic3d			Fashion			LandUse-21		
$\mathcal{L}_I$	$\mathcal{L}_{IB}$	$\mathcal{L}_C$	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
×	✓	✓	57.17	25.19	26.68	49.73	46.90	20.33	17.38	18.49	1.97
✓	×	✓	88.00	63.58	68.47	82.35	82.69	75.96	24.29	29.54	13.52
✓	✓	×	88.33	64.98	69.33	48.40	62.11	26.96	12.81	16.28	2.60
✓	✓	✓	**90.33**	**68.64**	**73.67**	**93.51**	**87.71**	**86.80**	**26.62**	**31.08**	**14.15**

imputation module can effectively recover the latent representations for facilitating the downstream clustering.

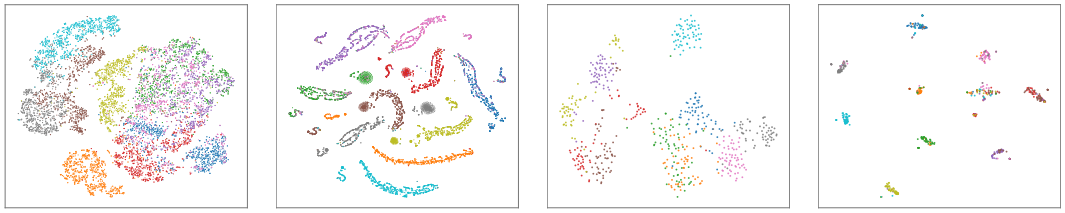
4.4 Parameter sensitivity analysis (Q3)

Our objective loss function primarily involves two trade-off parameters, λ_1 and λ_2. To evaluate their impact, we conducted a parameter sensitivity analysis by varying both parameters within the range of 10^{-2} to 10^2 on Synthetic3d dataset at a missing rate of 0.3. As illustrated in Fig. 2, the clustering performance of our CMLI remains stable despite variations in λ_1 and λ_2, thereby validating the robustness of our method.



(a) ACC (b) NMI (c) ARI

Figure 2: Parameter sensitivity analysis on Synthetic3d at a missing rate of 0.3.



(a) Raw Features (b) Latent Features (c) Raw Features (d) Latent Features

Figure 3: t-SNE visualization of Fashion and CUB with the missing rate of 0.3. (a) and (b) show the raw and latent features on Fashion; (c) and (d) show the raw and latent features on CUB.

4.5 Visualization Analysis (Q4)

To intuitively demonstrate the superiority of CMLI, we present the t-SNE visualization of both the raw features and the latent representations learned by CMLI on two datasets. It can be observed from Fig. 3 that the latent representations exhibit considerably stronger category discriminability, suggesting that our CMLI effectively captures useful discriminative structure in the embedding space.

5 Conclusion

This work presents a novel causality-driven meta-learning approach for imputation in incomplete multi-view clustering. CMLI innovatively integrates causal discovery and meta-learning for imputing incomplete view data. The IB-based information compression and preservation module maximally compresses task-irrelevant features while preserving task-relevant features, facilitating high-quality missing view recovery. By fully exploiting consistency and complementary information, the cluster consistency learning module enables end-to-end clustering. Experiments demonstrate the superior performance of CMLI in data imputation and clustering tasks.

AI use statement

In this work, we only used generative AI tools for language polishing. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Arthur Asuncion, David Newman, et al. Uci machine learning repository, 2007.

Shunshun Bai, Qinghai Zheng, Xiaojin Ren, and Jihua Zhu. Graph-guided imputation-free incomplete multi-view clustering. *Expert Systems with Applications*, 258:125165, 2024.

Jianwei Cao, Yanli Hu, Zhen Tan, and Xiang Zhao. Cross-modal multi-task learning for multimedia event extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11454–11462, 2025.

Guoqing Chao, Yi Jiang, and Dianhui Chu. Incomplete contrastive multi-view clustering with high-confidence guiding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 11221–11229, 2024.

Guoqing Chao, Kaixin Xu, Xijiong Xie, and Yongyong Chen. Global graph propagation with hierarchical information transfer for incomplete contrastive multi-view clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 15713–15721, 2025.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

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.

Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 2, pp. 524–531. IEEE, 2005.

Erdun Gao, Ignavier Ng, Mingming Gong, Li Shen, Wei Huang, Tongliang Liu, Kun Zhang, and Howard Bondell. Missdag: Causal discovery in the presence of missing data with continuous additive noise models. *Advances in Neural Information Processing Systems*, 35:5024–5038, 2022.

Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268, 2011.

Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. Improved deep embedded clustering with local structure preservation. In *Ijcai*, volume 17, pp. 1753–1759, 2017.

Jiabo Huang, Shaogang Gong, and Xiatian Zhu. Deep semantic clustering by partition confidence maximisation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8846–8855. IEEE, 2020.

Abhishek Kumar, Piyush Rai, and Hal Daume. Co-regularized multi-view spectral clustering. *Advances in neural information processing systems*, 24, 2011.

Haobin Li, Yunfan Li, Mouxing Yang, Peng Hu, Dezhong Peng, and Xi Peng. Incomplete multi-view clustering via prototype-based imputation. *arXiv preprint arXiv:2301.11045*, 2023.

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. IEEE, 2021.

Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4447–4461, 2022.

Yijie Lin, Jie Zhang, Zhenyu Huang, Jia Liu, Xi Peng, et al. Multi-granularity correspondence learning from long-term noisy videos. In *International conference on learning representations*, volume 2024, pp. 44188–44207, 2024.

Jiyuan Liu, Xinwang Liu, Xinhang Wan, Ke Liang, Weixuan Liang, Huijun Wu, Kehua Guo, et al. Incomplete multi-view deep clustering with data imputation and alignment. *Advances in Neural Information Processing Systems*, 38:89119–89142, 2026.

Yihang Lu, Haonan Xin, Rong Wang, Feiping Nie, and Xuelong Li. Scalable multiple kernel k-means clustering. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 4279–4283, 2022.

Karthika Mohan and Judea Pearl. Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534):1023–1037, 2021.

Saeid Sattari and Adnan Yazici. Semantic deep learning and adaptive clustering for handling multimodal multimedia information retrieval. *Multimedia Tools and Applications*, 84(13):11795–11831, 2025.

Qiaoyu Tan, Guoxian Yu, Carlotta Domeniconi, Jun Wang, and Zili Zhang. Incomplete multi-view weak-label learning. In *Ijcai*, volume 18, pp. 2703–2709, 2018.

Zhen Tian, Enlai Ouyang, Quan Zou, and Xiaoqiang Yan. Multi-view clustering with cauchy-schwarz mutual information maximin. *IEEE Transactions on Image Processing*, 2026.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pp. 1–5. Ieee, 2015.

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

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, pp. 10085–10092, 2021.

Xinxin Wang, Yongshan Zhang, and Yicong Zhou. Multimodal remote sensing image clustering with multiscale spectral-spatial anchor graphs. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–12, 2025.

Yuhao Wang, Vlado Menkovski, Hao Wang, Xin Du, and Mykola Pechenizkiy. Causal discovery from incomplete data: a deep learning approach. *arXiv preprint arXiv:2001.05343*, 2020.

Jie Wen, Zheng Zhang, Lunke Fei, Bob Zhang, Yong Xu, Zhao Zhang, and Jinxing Li. A survey on incomplete multiview clustering. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(2):1136–1149, 2022.

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 *Forty-first international conference on machine learning*, 2024.

Asrat Ayele Woldemariam, Yingchun Zhou, and Yuqi Qiu. Causal discovery with incomplete data integrated from multiple sources. *Journal of Survey Statistics and Methodology*, pp. smag016, 2026.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

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.

Wenbiao Yan, Yiyang Zhou, Yifei Wang, Qinghai Zheng, and Jihua Zhu. Multi-view semantic consistency based information bottleneck for clustering. *Knowledge-Based Systems*, 288:111448, 2024.

Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pp. 270–279, 2010.

Mingze Yao, Huibing Wang, Yawei Chen, and Xianping Fu. Between/within view information completing for tensorial incomplete multi-view clustering. *IEEE transactions on multimedia*, 27: 1538–1550, 2024.

Jun Yin and Shiliang Sun. Incomplete multi-view clustering with reconstructed views. *IEEE Transactions on Knowledge and Data Engineering*, 35(3):2671–2682, 2021.

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.

Ming Yin, Xin Liu, Junli Gao, Haoliang Yuan, Taisong Jin, Shengwei Zhang, and Lingling Li. Mvaibnet: Multiview disentangled representation learning with information bottleneck. *IEEE Transactions on Industrial Informatics*, 20(10):11511–11520, 2024. doi: 10.1109/TII.2024.3397357.

S Yu, X Yu, S Løkse, R Jenssen, and JC Principe. Cauchy-schwarz divergence information bottleneck for regression. arxiv 2024. *arXiv preprint arXiv:2404.17951*.

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*, pp. 10402–10411, 2024.

Honglin Yuan, Yuan Sun, Fei Zhou, Jing Wen, Shihua Yuan, Xiaojian You, and Zhenwen Ren. Prototype matching learning for incomplete multi-view clustering. *IEEE Transactions on Image Processing*, 34:828–841, 2025.

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.

Zhenghao Zhang, Jun Xie, Xingchen Chen, Tao Yu, Hongzhu Yi, Kaixin Xu, Yuanxiang Wang, Tianyu Zong, Xinming Wang, Jiahuan Chen, et al. 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, 2026.

Kun Zhu and Chunhui Zhao. When causality meets missing data: Fusing key information to bridge causal discovery and imputation in time series via bidirectional meta-learning. *Information Fusion*, 117:102811, 2025.

APPENDIX

A DERIVATION OF EQUATION (7)

Given the formulation of CS-based estimation expressed by $p(x, z)$ and $p(x)p(z)$ as follows:

$$\begin{aligned} I_{\mathrm{CS}} = \log\left(\int p^2(x,z)\,dx\,dz\right) + \log\left(\int p^2(x)\,p^2(z)\,dx\,dz\right) \\ - 2\log\left(\int p(x,z)\,p(x)\,p(z)\,dx\,dz\right). \end{aligned} \tag{23}$$

For the first term, the quadratic density $\int p^2(x, z)dx$ can be expressed as:

$$\int p^2(x,z)\,dx\,dz = \mathbb{E}_{p(x,z)}[p(x,z)]. \tag{24}$$

As the true joint distribution $p(x, z)$ is unknown, we approximate $p(x_i, z_i)$ using kernel density estimation as:

$$\hat{p}(x_i, z_i) = \frac{1}{N}\sum_{j=1}^{N} k(x_i, x_j) \cdot k(z_i, z_j), \tag{25}$$

where $k(\cdot, \cdot)$ denote the similarity between samples based on Gaussian kernel. The symmetry, positive definiteness, boundedness, and integrability of the Gaussian kernel make it suitable for CS-based estimation. Thus, Eq.(24) can be further formulated as:

$$\begin{aligned} \int p^2(x,z)\,dx\,dz &\approx \frac{1}{N}\sum_{i=1}^{N}\hat{p}(x_i, z_i) \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k(x_i, x_j)\cdot k(z_i, z_j) \\ &= \frac{1}{N^2}\sum_{i,j}^{N} K^x_{ij} K^z_{ij}. \end{aligned} \tag{26}$$

For the term $\int p^2(x)\,p^2(z)\,dx\,dz$, we can express it as:

$$\begin{aligned} \int p^2(x)\,p^2(z)\,dx\,dz &= \mathbb{E}_{p(x)}[p(x)] \cdot \mathbb{E}_{p(z)}[p(z)] \\ &\approx \frac{1}{N}\sum_{i=1}^{N}\hat{p}(x_i) \cdot \frac{1}{N}\sum_{q=1}^{N}\hat{p}(z_q) \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k(x_i, x_j) \cdot \frac{1}{N^2}\sum_{q=1}^{N}\sum_{r=1}^{N} k(z_q, z_r) \\ &= \frac{1}{N^4}\sum_{i,j,q,r}^{N} K^x_{ij} K^z_{qr}. \end{aligned} \tag{27}$$

Similarly, the term $\int p(x,z)\,p(x)\,p(z)\,dx\,dz$ can be expressed as:

$$\begin{aligned}\int p(x,z)\,p(x)\,p(z)\,dx\,dz &= \mathbb{E}_{p(x,z)}[p(x)p(z)] \\ &\approx \frac{1}{N}\sum_{i=1}^{N}\hat{p}(x_i)\hat{p}(z_i) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{1}{N}\sum_{j=1}^{N}k(x_i,x_j)\right)\left(\frac{1}{N}\sum_{q=1}^{N}k(z_i,z_q)\right)\right] \\ &= \frac{1}{N^3}\sum_{i,j,q}^{N}K^x_{ij}K^z_{iq}.\end{aligned} \tag{28}$$

Now, we approximate all quadratic density functionals via kernel density estimation, which avoids explicit parametric density estimation. Therefore, after further simplification, Eq.(23) can be formulated as:

$$\begin{aligned}I_{\mathrm{CS}}(X;Z) \approx\ &\log\left(\frac{1}{N^2}\mathrm{tr}(\mathbf{K}^x\mathbf{K}^z)\right)+\log\left(\frac{1}{N^4}\mathbf{1}^\top\mathbf{K}^x\mathbf{1}\cdot\mathbf{1}^\top\mathbf{K}^z\mathbf{1}\right) \\ &-2\log\left(\frac{1}{N^3}\mathbf{1}^\top\mathbf{K}^x\mathbf{K}^z\mathbf{1}\right),\end{aligned} \tag{29}$$

B Approximation of the Maximization of $I(\mathbf{Z}, y)$ and $I(\mathbf{Z}^v, \mathbf{Z})$

To begin with, we can apply the chain rule of MI to expand $I(\mathbf{Z}, y)$ as:

$$\begin{aligned}I(\mathbf{Z};y) &= I(y;\mathbf{Z}^v,\mathbf{Z}) - I(\mathbf{Z}^v;y|\mathbf{Z}) \\ &= I(\mathbf{Z}^v;y) + I(\mathbf{Z};y|\mathbf{Z}^v) - I(\mathbf{Z}^v;y|\mathbf{Z}) \\ &\geq I(\mathbf{Z}^v;y) - I(\mathbf{Z}^v;y|\mathbf{Z}).\end{aligned} \tag{30}$$

Then, we can derive the following MI inequality using conditional entropy.

$$\begin{aligned}I(\mathbf{Z}^v;y|\mathbf{Z}) &= H(\mathbf{Z}^v|\mathbf{Z}) - H(\mathbf{Z}^v|\mathbf{Z},y) \\ &\leq H(\mathbf{Z}^v|\mathbf{Z}) \\ &= H(\mathbf{Z}^v) - I(\mathbf{Z}^v,\mathbf{Z}).\end{aligned} \tag{31}$$

From the above two equations, we obtain a lower bound on $I(\mathbf{Z}, y)$ as:

$$I(\mathbf{Z};y) \geq I(\mathbf{Z}^v;y) + I(\mathbf{Z}^v;\mathbf{Z}) - H(\mathbf{Z}^v), \tag{32}$$

The above equation demonstrates that increasing $I(\mathbf{Z}^v, \mathbf{Z})$ enlarges the lower bound on $I(\mathbf{Z}; y)$, as $H(Z_v)$ is regularized by the objective of feature representation learning with compression when $I(\mathbf{Z}^v; y)$ is maximal. From a different angle, this can be analyzed through information decomposition. Formally, we can derive the following inequality:

$$\begin{aligned}I(\mathbf{Z},y) &= I(\mathbf{Z};\mathbf{Z}_v,y) - I(\mathbf{Z};\mathbf{Z}_v|y) \\ &= I(\mathbf{Z};\mathbf{Z}_v) + I(\mathbf{Z};y|\mathbf{Z}_v) - I(\mathbf{Z}_v;\mathbf{Z}|y) \\ &\geq I(\mathbf{Z};\mathbf{Z}_v) - I(\mathbf{Z};\mathbf{Z}_v|y)\end{aligned} \tag{33}$$

Since $I(\mathbf{Z}; \mathbf{Z}_v|y)$ is indirectly regularized by the compression objective in feature representation learning, increasing $I(\mathbf{Z}; \mathbf{Z}_v)$ helps enhance $I(\mathbf{Z}, y)$.

C DATASETS

We provide the detailed information about the datasets used in our experiments in Table 3. These datasets are described in detail in the following paragraph.

Table 3: The detailed information of incomplete multi-view datasets in experiments.

Datasets	Samples	Clusters	Views	Dimensionality
Synthetic3d	600	3	3	3/3/3
CUB	600	10	2	1024/300
Handwritten	2000	10	6	6/47/64/67/76/240
LandUse-21	2100	21	3	20/40/59
Scene-15	4485	15	3	20/40/59
Fashion	10000	10	3	784/784/784

Synthetic3d consists of 600 samples generated from view-specific Gaussian mixture models, with three three-dimensional views and three clusters. **CUB** contains 600 bird samples from 10 categories. Each sample is represented by a 1024-dimensional visual feature and a 300-dimensional textual feature. **Handwritten** is a widely used handwritten digit dataset, which consists of 2,000 images from 10 classes. Each sample is characterized by six types of visual features. **LandUse-21** comprises 2,100 satellite images spanning 21 land use categories. From these images, PHOG and LBP features are extracted to serve as two complementary views. **Scene-15** is a dataset of 4,485 images covering 15 indoor and outdoor scene categories, each represented by multiple complementary visual descriptors. **Fashion** is derived from Fashion-MNIST and contains 10,000 clothing images from 10 categories. Multiple views are constructed by applying different transformations to the original images.

D COMPARATIVE RESULTS ANALYSIS

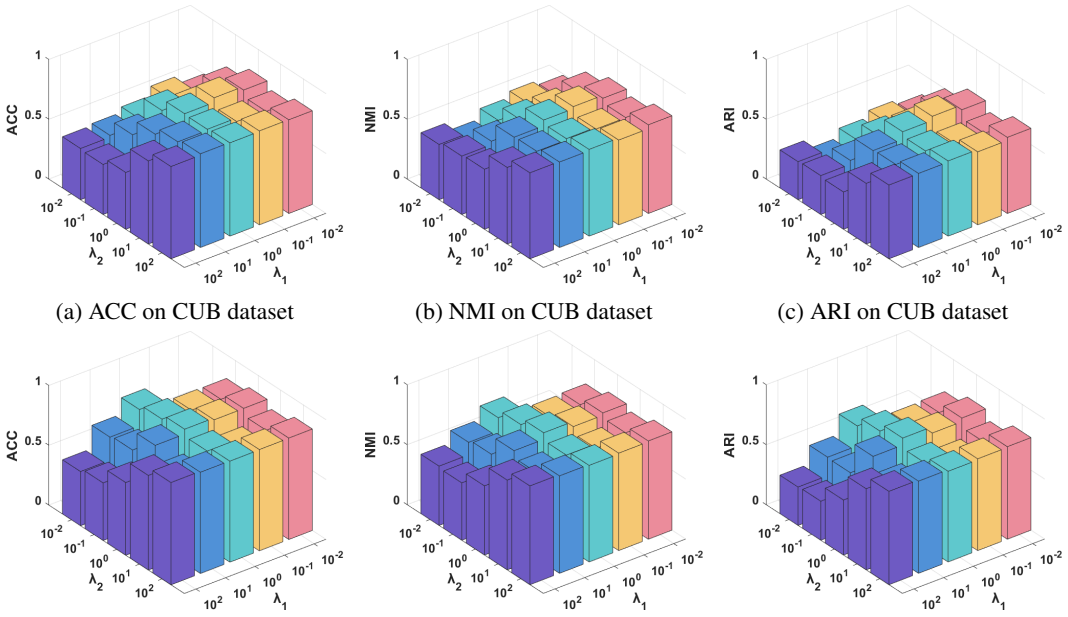
We show the supplementary comparative results of Handwritten and Scene-15 datasets in Table 4. For each dataset, we have the following observations. On Handwritten, although DIMVC-HIA obtains higher ACC and ARI at a missing rate of 0.1, our method achieves the highest results across metrics when missing rate is 0.3 and 0.5. On Scene-15, our method achieves the highest ACC at every missing rate and the highest ARI at a missing rate 0.3 and 0.5, while its NMI remains close to the best results. These findings indicate that the method maintains competitive clustering performance as the proportion of missing data increases.

Table 4: Clustering performance of all methods on Handwritten and Scene-15 datasets. The best and second-best results are marked in boldface and by underlining, respectively.

Datasets	Missing rate	0.1			0.3			0.5		
	Metrics	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Handwritten	DCP(2022)	74.55	81.73	66.58	68.80	76.94	59.85	73.90	76.86	56.88
	APADC(2023)	71.55	78.72	66.13	82.10	80.85	74.03	57.90	63.62	44.77
	ICMVC(2024)	81.30	79.26	72.49	78.55	75.98	67.05	81.90	77.07	71.31
	RPCIC(2024)	71.30	65.71	56.84	62.45	54.32	28.43	40.80	38.04	8.36
	IMDC-DIA(2025)	86.15	77.19	72.69	82.35	71.11	65.71	69.30	61.23	53.07
	IMC-MCL(2025)	84.70	83.69	76.51	78.70	77.02	66.91	77.35	73.42	67.57
	DCG(2025)	81.45	80.49	72.88	79.10	76.16	69.11	78.25	77.05	70.53
	DIMVC-HIA(2026)	**89.38**	82.02	**78.98**	79.25	75.78	69.14	73.35	67.74	60.81
	OURS	85.30	**84.33**	78.60	**85.70**	**82.15**	**77.63**	**82.50**	**77.09**	**71.34**
Scene-15	DCP(2022)	39.02	43.67	24.39	39.44	42.47	23.04	40.65	43.33	23.72
	APADC(2023)	40.16	**44.34**	**24.75**	38.84	42.37	24.36	41.29	42.89	24.36
	ICMVC(2024)	36.99	40.97	22.22	39.06	40.83	22.54	36.01	37.91	20.36
	RPCIC(2024)	29.81	27.45	12.68	26.04	25.86	9.83	21.29	24.18	9.44
	IMDC-DIA(2025)	25.48	29.22	11.00	21.90	26.20	7.94	20.57	23.76	7.73
	IMC-MCL(2025)	30.50	37.78	18.33	34.45	40.04	21.02	35.50	41.27	21.29
	DCG(2025)	37.70	42.94	23.02	40.96	**44.63**	25.08	40.82	**43.83**	25.79
	DIMVC-HIA(2026)	40.00	40.60	22.80	41.75	39.52	22.72	38.47	38.00	20.33
	OURS	**42.85**	43.85	24.72	**42.36**	43.18	**25.38**	**42.54**	43.45	**25.87**

E PARAMETER SENSITIVITY ANALYSIS

To further evaluate the sensitivity of the model, we conducted extensive experiments on the CUB and Handwritten datasets under various parameter combinations. Fig. 4 presents the experimental results. As shown in the figure, λ_1 has a greater influence in model performance than λ_2, while λ_2 exhibits some variation across different datasets. This validates the effectiveness of our causal meta-learning imputation method and suggests that the clustering consistency learning module is partly dependent on the dataset structure. In summary, our model maintains robustness over a relatively wide parameter range without exhibiting drastic performance fluctuations.



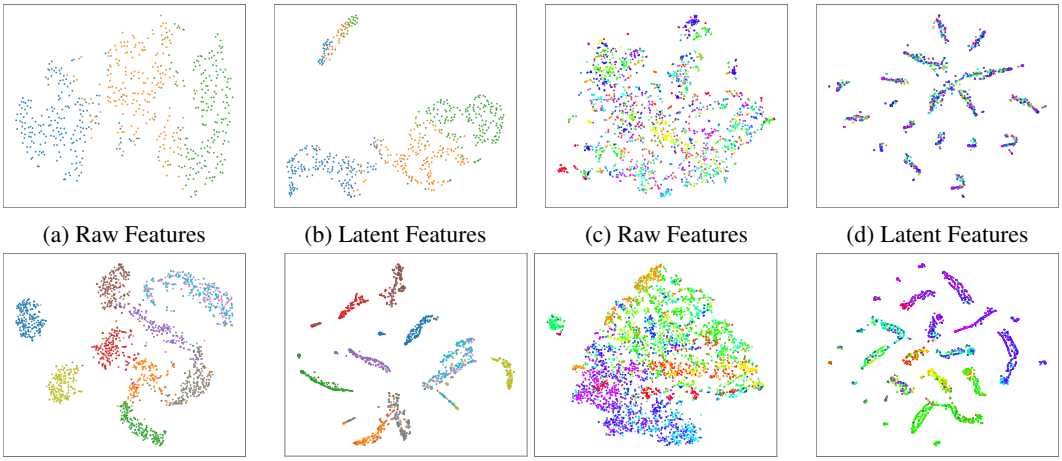
(a) ACC on CUB dataset (b) NMI on CUB dataset (c) ARI on CUB dataset

(d) ACC on Handwritten dataset (e) NMI on Handwritten dataset (f) ARI on Handwritten dataset

Figure 4: Parameter sensitivity analysis on CUB and Handwritten datasets at a missing rate of 0.3.

F VISUALIZATION ANALYSIS

To further examine the learned representations, we visualize the raw and latent features of four datasets using t-SNE. Compared with the raw features, the latent features generally form more compact and structured groups. Class separation is especially apparent when comparing subplots 5a and 5b, as well as 5e and 5f. In the other two datasets, the initially diffuse distributions become more organized, although some classes remain mixed. These observations suggest that the learned representations capture useful discriminative structure, while the degree of class separation varies across datasets.



(a) Raw Features (b) Latent Features (c) Raw Features (d) Latent Features

(e) Raw Features (f) Latent Features (g) Raw Features (h) Latent Features

Figure 5: t-SNE visualizations of raw and latent features at a missing rate of 0.3. Subplots (a) and (b) correspond to Synthetic3D, (c) and (d) to LandUse-21, (e) and (f) to Handwritten, and (g) and (h) to Scene-15. In each pair, the first subplot shows raw features and the second shows latent features.