Instance-Aligned Semantic Reconstruction for Incomplete Multi-View Clustering

Weiqing Yan[1] , **Yongteng Du**[1] , **Peng Song**[1] , **Chang Tang**[2*]
[1]School of Computer and Control Engineering, Yantai University, Yantai, China
[2]School of Software Engineering, Huazhong University of Science and Technology, Wuhan, China
wqyan@tju.edu.cn, duyt5216@qq.com, pengsong@ytu.edu.cn, tangchang@hust.edu.cn

Abstract

Incomplete multi-view clustering (IMVC) aims to exploit complementary information from multiple views with missing observations. Recent diffusion-based approaches have shown promise for view completion; however, they often fail to capture instance-aligned global correlations across views and suffer from inefficient inference and loosely coupled optimization. In this paper, we propose IASR, an Instance-Aligned Semantic Reconstruction framework for IMVC. IASR formulates missing-view recovery as a cross-view token alignment generator, in which noisy targets, observed views, and timestep embeddings are jointly represented as tokens and interact to capture long-range cross-view dependencies throughout the denoising trajectory. To stabilize representation learning under generative noise, we further introduce a stable–active dual-encoder representation architecture with a generative-adaptive contrastive learning strategy that tightly couples view completion and clustering. Extensive experiments on eight benchmark datasets demonstrate that IASR consistently outperforms state-of-the-art IMVC methods, especially under high missing-rate settings, while achieving improved inference efficiency.

1 Introduction

Real-world objects can often be described from multiple sources or through diverse feature extractors, resulting in multi-view data[Yan *et al.*, 2023b; Han *et al.*, 2023; Jin *et al.*, 2023]. For example, autonomous driving systems commonly rely on heterogeneous sensors such as LiDAR, cameras, and millimeter-wave radar to perceive and navigate the surrounding environment[Cui *et al.*, 2024; Tang and Liu, 2022]; in image understanding tasks, raw RGB images are frequently accompanied by handcrafted features such as SIFT and HOG[Yan *et al.*, 2024b; Wen *et al.*, 2020], or deep features extracted from different convolutional layers. Multi-view clustering (MVC) aims to exploit cross-view consistency and complementary information to partition multi-view data into semantically coherent clusters[Yan *et al.*, 2023a; Fang *et al.*, 2023; Li *et al.*, 2024]. However, most existing MVC methods implicitly assume that all views are complete and strictly aligned, an assumption that rarely holds in real-world scenarios. Due to sensor failures, occlusions, asynchronous acquisition, or privacy constraints, data are often partially missing. Consequently, incomplete multi-view clustering (IMVC), which explicitly addresses view-missing scenarios, has emerged as a critical and challenging problem[Yan *et al.*, 2024a].

The core objective of IMVC is to achieve consistent clustering under view-missing conditions[Chao *et al.*, 2024]. Existing IMVC methods can be broadly categorized into non-imputation-based methods and imputation-based methods.

Non-imputation-based methods aim to learn clustering representations directly from observed views through alignment strategies, attention mechanisms, or shared embeddings, thereby avoiding noise introduced by erroneous imputation. For example, APADC[Xu *et al.*, 2023] proposes an adaptive feature projection strategy that maps different views into a common space via distribution alignment, eliminating the need for explicit imputation. DIMVC[Xu *et al.*, 2022] constructs a complementarity mining mechanism based on high-dimensional nonlinear mappings, transforming cross-view complementary information into high-confidence supervisory signals to achieve consistent clustering. DVIMC[Xu *et al.*, 2024], built upon a variational autoencoder framework, adopts a Product-of-Experts mechanism to aggregate information from visible views and explore shared structures. However, due to the lack of explicit modeling of missing data, these methods often exhibit limited robustness when confronted with severe missingness or high-noise scenarios.

In contrast, imputation-based methods explicitly generate missing views or their latent representations to recover complete multi-view structures. Representative deep learning approaches such as COMPLETER[Lin *et al.*, 2021] and IMC-MCL[Yin *et al.*, 2025]leverage contrastive learning and dual-prediction mechanisms to minimize conditional entropy in the latent space for view recovery. DIVIDE[Lu *et al.*, 2024] follows a similar paradigm by designing cross-view decoders to reconstruct missing representations. MICA[Wang *et al.*, 2025] introduces a multi-level imputation strategy to preserve topological consistency across views. Furthermore, with the advancement of generative models, BURG[Jin *et al.*, 2025] employs flow-based models to perform cross-view distribution transfer. Despite these progresses, existing imputation-

based methods often struggle to accurately capture the complex distributions of missing views, resulting in limited generation quality and degraded clustering performance under high missing-rate settings.

Motivated by the strong capability of diffusion models in modeling complex data distributions, recent studies have begun to explore diffusion-based frameworks for incomplete multi-view clustering. However, existing diffusion-based IMVC methods still suffer from notable limitations. First, most approaches rely on local convolutional architectures or static conditioning schemes, which are inadequate for capturing long-range, instance-aligned cross-view correlations. Second, the iterative denoising process introduces substantial computational overhead, severely restricting their practical applicability, particularly under high missing-rate scenarios.

To address these challenges, we propose IASR, a cross-view token alignment generator for incomplete multi-view clustering. Unlike previous methods, IASR formulates missing-view recovery as an instance-specific and globally conditioned generation process, in which noisy targets, observed views, and timestep embeddings are treated as equally important tokens and jointly processed through a global token-level correlation operator. This design enables explicit modeling of long-range cross-view dependencies throughout the entire denoising trajectory. The main contributions of this work are summarized as follows:

- We design a stable-active dual-encoder architecture that integrates a momentum-updated stable branch with an active branch to jointly learn robust and view-specific representations. This architecture prevents representation collapse and ensures stable learning under severe view missingness.
- Propose a novel cross-view token alignment generator framework for incomplete multi-view clustering, which jointly represents noisy targets, observed views, and timestep embeddings as tokens and models their interactions to capture long-range cross-view dependencies throughout the denoising trajectory, enabling efficient and high-quality imputation.
- Propose a generative-adaptive contrastive learning framework based on a tri-level contrastive paradigm with an adaptive affinity matrix, which explicitly enforces intra-view stability, cross-view complementarity, and semantic consistency of generated representations, thereby yielding substantial performance improvements under high missing-rate scenarios.

2 Related Work

2.1 Incomplete Multi-View Clustering

Early traditional methods predominantly relied on techniques such as matrix factorization[Hu and Chen, 2018], kernel learning[Liu *et al.*, 2021], and graph learning[Zhang *et al.*, 2020] to replenish missing information. However, these shallow models often struggle to capture the nonlinear structures inherent in high-dimensional data. Recently, Deep Incomplete Multi-View Clustering (DIMVC) has emerged as the mainstream approach due to its robust feature representation capabilities. Existing works can be primarily categorized into four classes: (1) Cross-view prediction-based methods aim to learn consistent representations in the latent space by maximizing mutual information. For instance, DCP[Lin *et al.*, 2022] maximizes mutual information via contrastive learning and minimizes conditional entropy via dual prediction, thereby recovering missing views; (2) Generative Adversarial Network (GAN)-based methods leverage adversarial training to fit the distribution of missing data. Representative works such as GP-MVC[Wang *et al.*, 2021] generate missing views from shared representations using view-specific generators and cycle consistency constraints. (3) Graph Neural Network (GNN)-based methods focus on guiding data recovery utilizing topological structures. CRTC[Wang *et al.*, 2023] and ICMVC[Chao *et al.*, 2024] transfer sample neighbor relationships from observed views to missing ones and employ GCNs for feature completion; (4) Direct clustering learning aim to circumvent noise introduced by explicit imputation by enhancing feature discriminability. ProImp[Li *et al.*, 2023] utilize sample-prototype relationships and robust contrastive losses to recover data, effectively preserving instance commonality and view diversity.

2.2 Generative-Assisted Missing-view Learning

Generative models have gradually evolved from pure data synthesizers into powerful assistants for representation learning. The central idea is to introduce reconstruction or generation objectives as auxiliary regularization signals, encouraging models to capture underlying semantic distributions and latent manifold structures. This paradigm has demonstrated remarkable success across various domains, including image classification[Azizi *et al.*, 2023], object detection[Lin *et al.*, 2023], and image generation[Tian *et al.*, 2023].

In early studies on IMVC, reconstruction- or prediction-based objectives were commonly employed to complement discriminative learning. Recently, probabilistic diffusion models[Ho *et al.*, 2020] have shown superior capability in modeling complex data distributions, providing rich semantic priors for downstream tasks. Wen et al.[2024] pioneered the application of conditional diffusion models to IMVC by conditioning the generation of missing views on observed ones, thereby facilitating clustering while completing data. Subsequently, DCG [2025] combines diffusion processes with contrastive learning to reduce the reliance on massive paired data and leverages diffusion priors to enhance clustering compactness.

3 Method

We propose IASR, a unified framework for IMVC. As illustrated in Fig. 1, IASR consists of three core components: S-ADE, CTAG, and G-ACL. S-ADE learns robust and view-specific representations via a momentum-updated stable branch and an active branch. CTAG formulates missing-view recovery as an instance-aligned token generation process to capture long-range cross-view dependencies. Meanwhile, G-ACL enforces a tri-level contrastive objective to ensure intra-view stability, cross-view complementarity, and semantic consistency. We detail each module below.

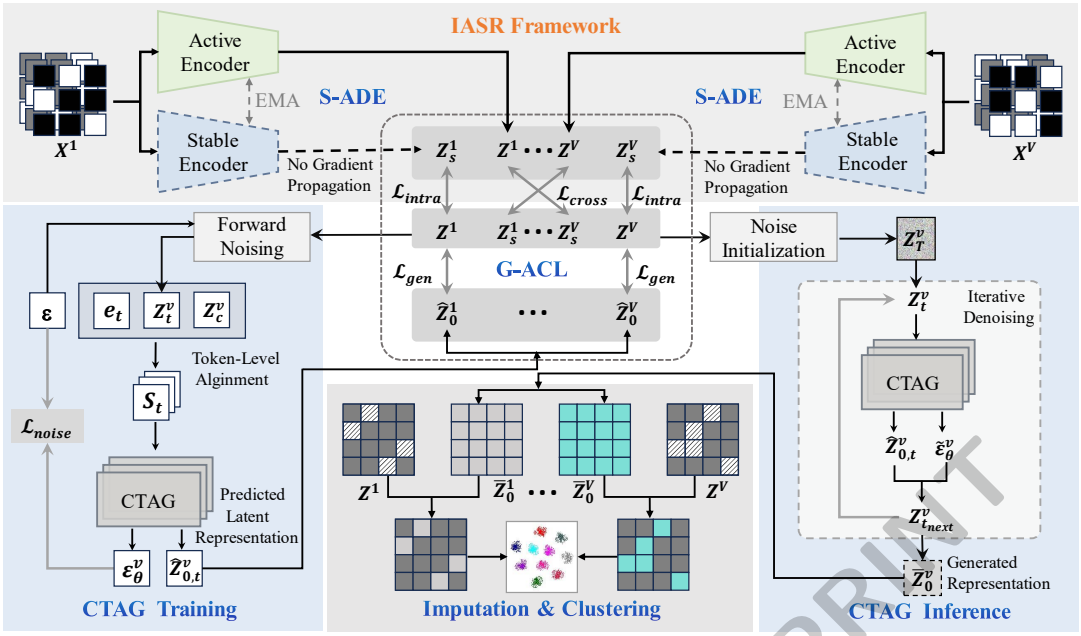


Figure 1: Overview of the IASR framework. IASR consists of three core components: Stable–Active Dual Encoders (S-ADE), Cross-view Token Alignment Generator (CTAG), and Generative-Adaptive Contrastive Learning (G-ACL). S-ADE employs a momentum-updated stable branch and an active branch to learn robust and view-specific representations under severe view missingness. CTAG reformulates missing-view recovery as an instance-aligned token generation process, modeling noisy targets, observed views, and timestep embeddings as unified tokens to capture long-range cross-view dependencies along the denoising trajectory. Meanwhile, G-ACL enforces a tri-level contrastive objective to ensure intra-view stability, cross-view complementarity, and semantic consistency of generated representations.

3.1 Representation Learning with Stable-Active Dual Encoders(S-ADE)

To alleviate the adverse impact of early-stage generative noise on representation learning and to prevent degenerate solutions during joint generation and clustering optimization, we introduce an asymmetric dual-encoder architecture consisting of an active encoder and a stable encoder. Both encoders share the same network architecture but follow different parameter update strategies during training.

Active Encoder. The active encoder is responsible for modeling the views and is optimized end-to-end via back-propagation. Specifically, for the v-th view, the active encoder is defined as:

$$\mathbf{z}^{(v)} = f^{(v)}(\mathbf{x}^{(v)}; \boldsymbol{\theta}^{(v)}), \tag{1}$$

where $\boldsymbol{\theta}^{(v)}$ denotes the learnable parameters. This encoder enables flexible feature adaptation as training progresses.

Stable Encoder. The stable encoder shares the same architecture as the active encoder but is updated without gradient back-propagation. For the v-th view, it produces a stable representation:

$$\mathbf{z}_s^{(v)} = f_s^{(v)}(\mathbf{x}^{(v)}; \boldsymbol{\theta}_s^{(v)}), \quad \boldsymbol{\theta}_s^{(v)} \leftarrow m\boldsymbol{\theta}_s^{(v)} + (1-m)\boldsymbol{\theta}^{(v)}, \tag{2}$$

where $\boldsymbol{\theta}_s^{(v)}$ is updated via an exponential moving average (EMA) of the active encoder parameters $\boldsymbol{\theta}^{(v)}$, with $m \in [0, 1)$ denoting the momentum coefficient. The stable encoder evolves smoothly across training iterations, acting as a regularization anchor that constrains the optimization of the active encoder, thereby enabling the latter to focus on learning noise-robust and task-relevant representations.

3.2 Cross-view Token Alignment Generator (CTAG)

We formulate missing-view recovery as a globally conditioned generation process that explicitly enforces instance-level consistency. Unlike conventional conditional generation approaches that rely on static feature concatenation, our method injects conditioning information via token-level global correlation modeling.

Token-Level Global Correlation Modeling. Given an instance with target view representation $\mathbf{z}^{(v)}$ and a set of observed conditional view representations $\{\mathbf{z}_c^{(k)}\}$, we first apply forward diffusion to the target view:

$$\mathbf{z}_t^{(v)} = \sqrt{\bar{\alpha}_t}\,\mathbf{z}_0^{(v)} + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{3}$$

where t denotes the diffusion timestep and $\bar{\alpha}_t$ follows a predefined noise schedule.

At each timestep t, we project the timestep embedding, the noisy target view, and the observed conditional views into token representations along the token dimension:

$$\mathbf{S}_t = \left[\mathbf{e}_t,\ \mathbf{z}_t^{(v)},\ \mathcal{Z}_c\right], \mathcal{Z}_c = \left\{\mathbf{z}_c^{(1)}, \ldots, \mathbf{z}_c^{(K)} \mid k \neq v\right\} \tag{4}$$

where $\mathcal{Z}_c$ denotes the set of available conditional view tokens, and $\mathbf{e}_t$ is the positional embedding at timestep t. Unlike conventional conditional diffusion models that impose an

asymmetric conditioning hierarchy, we adopt a symmetry-preserving tokenization scheme, in which the noisy target, conditional views, and timestep embedding are treated as homogeneous tokens to avoid predefined conditioning directions. In the CTAG module, the noise prediction network $\boldsymbol{\epsilon}_\theta$ models the input token sequence $\mathbf{S}_t$ through a global token-level correlation operator, which captures instance-specific dependencies among temporal information, target view, and conditional views within a unified latent space. Formally, the noise prediction is defined as:

$$\boldsymbol{\epsilon}_\theta(\mathbf{S}_t) = \mathbf{W}_o \left(\phi\Big(\mathbf{S}_t \mathbf{W}_{Q_1} \left(\mathbf{S}_t \mathbf{W}_{Q_2} \right)^\top \Big) \mathbf{S}_t \mathbf{W}_{Q_3} \right), \quad (5)$$

where $\mathbf{W}_{Q_1}$, $\mathbf{W}_{Q_2}$, $\mathbf{W}_{Q_3}$, and $\mathbf{W}_o$ are learnable projection matrices, and $\phi(\cdot)$ normalizes the values to the range $[0, 1]$. The model is trained by minimizing the noise prediction error:

$$\mathcal{L}_{\text{noise}} = \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{S}_t)\|_2^2 . \quad (6)$$

Fast Sampling for View Completion. At inference, we utilize the trained CTAG network to recover missing views via a fast reverse denoising process. To balance generation fidelity and efficiency, we employ a deterministic DDIM-style sampling strategy with stride-based timestep skipping, where the inference timesteps follow a sparsified schedule $\mathcal{T} = t_1, \ldots, t_L$ with stride S, resulting in $L \ll T$. We employ a Classifier-Free Guidance (CFG) strategy to explicitly align the generated target with observed views. Specifically, within the input sequence $\mathbf{S}_t$, the target representation is initialized as pure Gaussian noise at $t = T$, denoted as $\mathbf{z}_T^{(v)}$. The adjusted noise prediction is:

$$\tilde{\boldsymbol{\epsilon}}_t = (1 + w) \cdot \boldsymbol{\epsilon}_\theta(\mathbf{S}_t) - w \cdot \boldsymbol{\epsilon}_\theta(\mathbf{S}_t^{\emptyset}), \quad (7)$$

where $\mathbf{S}_t^{\emptyset}$ is the sequence with conditional view tokens removed, and w is the guidance scale. Crucially, to facilitate intermediate supervision in the subsequent contrastive learning stage, we explicitly derive the predicted latent representation $\hat{\mathbf{z}}_{0,t}^{(v)}$ from the current noisy state $\mathbf{z}_t^{(v)}$ at each step:

$$\hat{\mathbf{z}}_{0,t}^{(v)} = \frac{\mathbf{z}_t^{(v)} - \sqrt{1 - \bar{\alpha}_t}\tilde{\boldsymbol{\epsilon}}_t}{\sqrt{\bar{\alpha}_t}}. \quad (8)$$

Based on this representation, the reverse update from t to t_{next} follows the deterministic transition:

$$\mathbf{z}_{t_{next}}^{(v)} = \sqrt{\bar{\alpha}_{t_{next}}}\hat{\mathbf{z}}_{0,t}^{(v)} + \sqrt{1 - \bar{\alpha}_{t_{next}}}\tilde{\boldsymbol{\epsilon}}_t, \quad (9)$$

where $\bar{\alpha}_{t_{next}}$ denotes the cumulative noise coefficient at timestep t_{next}. After iterating through all timesteps, the recovered target representations $\bar{\mathbf{z}}_0^{(v)}$ are integrated with the incomplete multi-view representations to form a complete feature set for clustering.

3.3 Generative-Adaptive Contrastive Learning (G-ACL)

To learn discriminative and completion-aware representations under incomplete views, we propose a Generative-Adaptive Contrastive Learning framework. The proposed objective jointly enforces instance-level stability, cross-view semantic alignment, and generation-consistent regularization within a unified contrastive formulation.

At the core of G-ACL is a temperature-scaled InfoNCE-style contrastive loss equipped with an adaptive affinity matrix M, which allows each sample to align with multiple semantically related neighbors:

$$\mathcal{L}_{\text{con}}(Z, Z', M) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{M_{ij}}{\sum_{l=1}^{N} M_{il}} \cdot \log\frac{\exp(\text{sim}(z_i, z'_j)/\tau)}{\sum_{k=1}^{N}\exp(\text{sim}(z_i, z'_k)/\tau)}, \quad (10)$$

where $\text{sim}(\cdot)$ denotes cosine similarity, τ is the temperature parameter and N denotes the number of samples in the current mini-batch. The affinity matrix $M \in [0, 1]^{N\times N}$ encodes soft positive relationships and is constructed via a multi-step random-walk kernel affinity over normalized features, capturing local neighborhood semantics beyond strict one-to-one correspondence.

Intra-View Contrastive Loss. This loss enforces representation stability by aligning features extracted by the active encoder with those from the Stable Encoder:

$$\mathcal{L}_{\text{intra}} = \sum_{v=1}^{V}\mathcal{L}_{\text{con}}(Z^v, Z_s^v, M^v). \quad (11)$$

Cross-View Contrastive Loss. To capture complementary semantics across views, we impose bidirectional alignment between different views:

$$\mathcal{L}_{\text{cross}} = \sum_{v=1}^{V}\sum_{u\neq v}\frac{1}{2}\left[\mathcal{L}_{\text{con}}(Z^v, Z_s^u, M^v) + \mathcal{L}_{\text{con}}(Z^v, Z_s^u, M^u)\right]. \quad (12)$$

Generative Consistency Loss. Since the CTAG network explicitly predicts the latent representation $\hat{\mathbf{z}}_{0,t}^{(v)}$, we enforce generative consistency by aligning this prediction with the corresponding view features through a contrastive objective:

$$\mathcal{L}_{\text{gen}} = \sum_{v=1}^{V}\mathcal{L}_{\text{con}}\big(Z^v, \hat{Z}_{0,t}^v, M^v\big). \quad (13)$$

The overall training objective is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{gen}} + \lambda_1\mathcal{L}_{\text{intra}} + \lambda_2\mathcal{L}_{\text{cross}} + \lambda_3\mathcal{L}_{\text{noise}}, \quad (14)$$

where λ_1, λ_2, and λ_3 are the weighting coefficients.

4 Experiments

We evaluate the proposed IASR on eight widely used multi-view datasets, including *HandWritten*, *synthetic3d*, *MNIST-USPS*, *CIFAR-10*, *ORL*, *OutdoorScene*, *COIL20*, and *100Leaves*. These datasets cover diverse modalities such as images and handcrafted features, with the number of samples ranging from 400 to 50,000. Clustering performance is evaluated using Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI). Higher values of these metrics indicate better clustering performance. Among

Dataset	Method	0.1			0.3			0.5			0.7		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
HandWritten	Completer(CVPR'21)	67.31	70.95	50.15	55.74	60.55	29.38	57.82	61.66	35.61	59.60	58.60	35.42
	DIMVC(AAAI'22)	68.26	63.54	53.48	74.78	64.21	56.91	72.02	63.78	53.81	67.37	61.56	51.88
	APADC(TIP'23)	71.19	77.52	62.11	68.91	75.03	58.65	68.22	71.65	55.01	76.78	72.40	60.50
	DVIMC(AAAI'24)	25.73	16.37	9.03	25.78	15.31	8.80	25.23	15.85	8.45	23.87	14.51	7.78
	Divide(AAAI'24)	82.06	76.82	70.56	84.24	75.92	70.92	85.14	76.32	71.66	82.67	72.29	66.99
	MICA(Neural Netw'25)	65.85	69.02	57.61	66.30	61.52	49.98	59.25	52.03	41.41	35.85	39.48	22.61
	DCG(AAAI'25)	29.95	51.22	29.26	29.80	51.15	28.20	29.40	48.75	27.98	29.35	48.28	28.34
	BURG(ICCV'25)	89.15	80.88	77.76	84.15	74.18	68.03	68.40	64.43	51.61	50.45	53.07	31.53
	IMC-MCL(TKDE'25)	48.98	58.65	37.18	56.25	61.95	42.35	58.82	65.09	48.27	70.88	67.92	55.52
	IASR (Ours)	**93.35**	**87.50**	**86.00**	**92.21**	**85.60**	**83.58**	**88.36**	**80.45**	**77.58**	**84.23**	**73.46**	**69.63**
synthetic3d	Completer(CVPR'21)	91.30	73.07	76.31	89.53	66.94	71.48	86.67	61.37	64.66	83.33	54.45	56.80
	DIMVC(AAAI'22)	88.61	65.33	69.60	88.93	65.31	70.30	84.95	55.01	60.85	60.00	31.11	31.19
	APADC(TIP'23)	73.03	56.50	50.68	73.27	50.29	47.44	78.23	50.52	52.32	79.63	50.44	51.80
	DVIMC(AAAI'24)	80.77	52.48	55.30	77.73	44.70	46.57	73.70	41.16	40.14	69.40	34.66	33.50
	Divide(AAAI'24)	85.63	58.75	62.34	81.70	50.53	53.62	78.30	44.96	47.02	73.17	37.72	37.61
	MICA(Neural Netw'25)	64.83	44.32	37.98	63.00	37.08	34.06	60.33	34.34	29.89	61.67	29.91	29.44
	DCG(AAAI'25)	92.50	75.00	79.31	87.50	62.69	67.25	83.50	54.65	58.44	79.33	47.31	50.07
	BURG(ICCV'25)	90.33	70.28	73.52	88.00	64.76	67.00	58.33	42.51	34.66	58.33	40.60	35.25
	IMC-MCL(TKDE'25)	68.87	33.67	34.93	77.77	43.60	46.93	72.70	38.44	39.68	73.43	36.85	38.71
	IASR (Ours)	**96.97**	**87.02**	**91.13**	**95.03**	**80.77**	**85.71**	**92.73**	**74.54**	**79.53**	**91.67**	**71.71**	**76.68**
MNIST_USPS	Completer(CVPR'21)	88.68	93.27	87.68	82.23	89.15	78.72	85.20	88.20	81.28	84.97	82.99	76.98
	DIMVC(AAAI'22)	79.87	81.62	73.54	79.41	82.39	75.30	80.52	78.29	72.18	83.15	77.29	73.16
	APADC(TIP'23)	97.96	94.98	92.93	96.77	92.27	92.93	95.39	89.33	90.00	91.89	83.23	82.77
	DVIMC(AAAI'24)	93.16	91.65	89.00	86.70	86.69	81.16	80.99	81.64	74.36	79.02	76.28	68.40
	Divide(AAAI'24)	93.26	90.06	88.97	91.64	86.49	85.48	91.84	82.81	82.89	87.85	75.86	75.05
	MICA(Neural Netw'25)	98.80	96.61	97.35	68.26	77.83	66.35	57.58	71.71	58.66	39.44	57.85	40.70
	DCG(AAAI'25)	97.76	94.31	95.10	88.20	85.33	80.18	94.02	86.71	87.38	92.48	83.60	84.21
	BURG(ICCV'25)	83.42	79.93	73.95	69.36	67.68	55.17	55.14	54.79	37.08	52.42	49.41	33.99
	IMC-MCL(TKDE'25)	90.59	91.37	87.82	86.02	87.30	82.47	88.26	85.86	83.07	92.90	84.87	85.10
	IASR (Ours)	**98.96**	**97.14**	**97.71**	**97.85**	**94.33**	**95.30**	**96.72**	**91.72**	**92.87**	**94.30**	**86.76**	**87.81**
cifar10	Completer(CVPR'21)	80.85	85.61	75.33	88.56	86.91	83.79	92.34	86.38	86.04	92.61	83.96	84.60
	DIMVC(AAAI'22)	97.10	92.84	93.76	95.76	89.94	90.93	94.19	86.75	87.67	92.95	84.20	85.21
	APADC(TIP'23)	96.68	91.93	92.83	95.56	89.38	90.48	90.14	85.06	83.71	84.15	80.35	75.99
	DVIMC(AAAI'24)	57.94	72.09	52.62	38.58	59.04	35.29	42.26	60.24	35.40	35.32	52.40	28.42
	Divide(AAAI'24)	97.07	92.63	93.68	95.74	89.74	90.90	93.96	86.23	87.19	92.17	83.21	83.57
	MICA(Neural Netw'25)	96.67	91.64	92.83	87.28	88.37	83.55	85.56	84.64	79.76	73.11	75.08	63.32
	DCG(AAAI'25)	97.02	92.48	93.57	95.94	90.04	91.30	94.65	87.38	88.63	93.13	84.62	85.67
	BURG(ICCV'25)	61.41	67.60	52.23	57.21	58.28	37.40	47.83	50.05	28.64	34.97	39.28	17.61
	IMC-MCL(TKDE'25)	62.76	63.87	45.94	60.63	65.55	51.51	60.36	67.91	50.85	52.64	61.99	47.72
	IASR (Ours)	**97.77**	**94.19**	**95.16**	**96.53**	**91.32**	**92.53**	**95.32**	**88.75**	**90.01**	**93.92**	**85.98**	**87.14**

Table 1: Clustering performance comparison on four dual-view datasets under different missing rates. The best results are highlighted in bold, and the second-best results are underlined.

them, Table 1 presents the comparison results on the first four dual-view datasets, while Table 2 reports experiments on the remaining four multi-view datasets using baseline methods that support an arbitrary number of views.

All experiments are implemented in PyTorch and conducted on a single NVIDIA RTX 3090 GPU. We adopt the Adam optimizer. The batch size is set to 512 for the large-scale dataset CIFAR-10 and 256 for the remaining datasets. The learning rate is individually tuned for each dataset to ensure optimal convergence. To simulate missing-view scenarios, we randomly discard partial views of each sample with a missing rate $\eta \in \{0.1, 0.3, 0.5, 0.7\}$, while ensuring that at least one view is retained for each sample. All experiments are independently repeated with five different random seeds, and the reported results are averaged over these five runs.

4.1 Comparison with SOTA Methods

We compare the proposed **IASR** with nine SOTA IMVC methods, including *Completer*, *DIMVC*, *APADC*, *DVIMC*, *Divide*, *MICA*, *DCG*, *BURG*, and *IMC-MCL*. All baseline methods are evaluated under different missing rates using the same evaluation metrics.

Table 1 summarizes the average clustering performance on four representative dual-view datasets. On the *synthetic3d* dataset, our method significantly outperforms all competing baselines. In particular, under a high missing rate of $\eta = 0.7$, IASR achieves improvements of 12.04%, 21.27%, and 24.88% over the second-best method in terms of ACC, NMI, and ARI, respectively, demonstrating its superiority and effectiveness for incomplete multi-view clustering. Moreover, on the *HandWritten* dataset with a missing rate of 0.1, IASR surpasses the second-best method *BURG* by 4.2% in ACC.

Dataset	Methods	$\eta = 0.1$			$\eta = 0.3$			$\eta = 0.5$			$\eta = 0.7$		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
ORL	DCP(TPAMI'23)	39.30	65.24	12.66	40.45	60.33	6.89	31.55	51.38	9.96	25.40	45.01	4.18
	DIMVC(AAAI'22)	53.10	75.15	38.93	50.68	68.73	22.88	47.15	65.96	29.79	30.43	51.93	13.30
	DVIMC(AAAI'24)	49.90	70.44	33.38	48.05	66.01	27.03	38.30	57.26	15.42	27.20	47.46	29.85
	MICA(Neural Netw'25)	6.75	16.51	0.61	13.00	28.37	2.57	11.25	27.36	2.07	9.07	17.22	0.99
	BURG(ICCV'25)	62.00	80.76	50.38	59.25	76.99	42.39	47.00	66.77	26.03	43.75	65.81	25.26
	IASR	**73.80**	**86.11**	**62.97**	**73.20**	**84.43**	**60.55**	**66.10**	**79.69**	**50.70**	**58.05**	**73.81**	**41.00**
OutdoorScene	DCP(TPAMI'23)	45.08	38.23	21.63	54.26	44.65	28.86	48.25	39.99	21.89	47.34	40.09	20.28
	DIMVC(AAAI'22)	37.34	23.84	16.97	36.28	24.94	18.72	33.73	21.29	14.64	34.85	20.56	13.62
	DVIMC(AAAI'24)	48.90	34.38	24.74	43.58	27.11	19.04	41.50	25.66	18.06	36.05	19.16	12.48
	MICA(Neural Netw'25)	51.13	56.05	39.54	49.92	51.97	36.04	50.91	52.05	37.39	53.96	52.79	39.27
	BURG(ICCV'25)	55.80	51.28	38.03	47.66	44.53	28.33	43.97	41.01	24.93	42.56	34.82	17.66
	IASR	**74.55**	**63.11**	**54.65**	**73.94**	**62.50**	**53.87**	**73.57**	**61.56**	**53.56**	**71.78**	**58.77**	**50.61**
COIL20	DCP(TPAMI'23)	63.65	78.55	54.60	64.53	77.15	54.70	60.56	75.65	50.18	60.19	73.36	40.55
	DIMVC(AAAI'22)	57.00	73.84	51.28	52.55	73.04	49.31	58.27	71.61	51.72	53.73	70.91	48.90
	DVIMC(AAAI'24)	51.64	63.25	40.91	48.39	57.59	35.56	45.47	51.92	29.07	38.71	45.85	22.47
	MICA(Neural Netw'25)	68.00	78.66	62.05	74.71	82.84	69.63	75.00	79.36	64.29	64.79	73.28	55.74
	BURG(ICCV'25)	59.86	78.16	57.53	61.94	73.01	52.85	56.67	70.68	49.88	55.14	74.63	48.05
	IASR	**78.68**	**86.84**	**76.04**	**77.74**	**85.85**	**73.83**	**78.22**	**85.70**	**74.06**	**73.33**	**81.56**	**67.52**
100Leaves	DCP(TPAMI'23)	57.68	81.99	46.90	50.29	75.15	33.85	45.51	69.23	20.36	39.19	63.96	14.01
	DIMVC(AAAI'22)	66.97	83.52	49.87	72.52	83.31	57.32	58.82	73.96	39.13	51.34	68.06	31.15
	DVIMC(AAAI'24)	40.04	67.82	26.01	36.01	63.61	20.22	32.86	60.64	16.27	26.57	56.25	11.07
	MICA(Neural Netw'25)	12.94	42.70	4.25	12.81	38.49	2.97	11.50	36.04	2.03	10.87	34.34	1.52
	BURG(ICCV'25)	69.12	87.86	62.52	57.06	77.77	37.93	45.38	70.08	25.23	36.03	63.09	16.22
	IASR	**86.04**	**93.51**	**80.31**	**76.35**	**86.63**	**64.35**	**66.08**	**80.31**	**50.63**	**54.91**	**74.23**	**37.67**

Table 2: Performance evaluation extended to multi-view datasets under varying missing rates.

As the missing rate increases to 0.7, the ACC of *MICA* drops sharply by 38.7%, whereas IASR exhibits a much smaller degradation of 9.12% and still maintains an ACC of 84.23% .

We further extend the evaluation to scenarios with more than two views, and the results are reported in Table 2. On *OutdoorScene* (4 views) and *100Leaves* (3 views), IASR consistently demonstrates overwhelming advantages over competing methods. For instance, on *100Leaves* with $\eta = 0.1$, IASR achieves an ACC of 86.04%, outperforming the second-best method *BURG* by more than 16%, which clearly validates the effectiveness of modeling global correlations for capturing long-range cross-view dependencies.

4.2 Ablation Studies

We conduct comprehensive ablation studies to evaluate both the architectural components and the contrastive objectives. Table 3 analyzes the impact of key modules, while Table 4 examines the contribution of each loss term. In addition, qualitative visualization is provided to further analyze the effect of conditional generation guidance.

Impact of the Key Modules. Table 3 summarizes the ablation results of the proposed S-ADE, CTAG, and G-ACL modules. As shown, replacing S-ADE with a standard autoencoder causes a significant performance drop, highlighting the importance of stability-enhanced representation learning under severe view-missing scenarios. Similarly, when the missing-view representations are generated by replacing CTAG with a UNet-based prediction architecture, the clustering performance degrades noticeably, indicating that architectures with local receptive fields struggle to capture long-range cross-view dependencies. Moreover, removing the generative

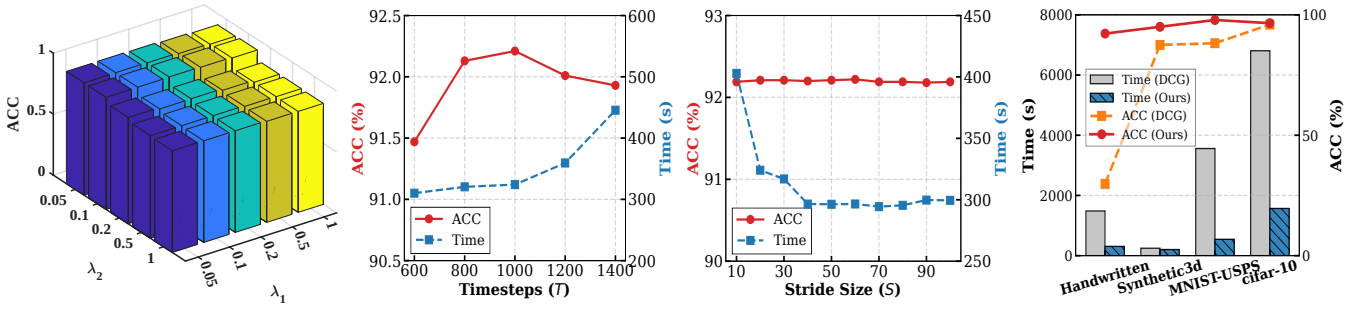
Modules			Performance		
S-ADE	**CTAG**	**G-ACL**	**ACC**	**NMI**	**ARI**
×	✓	✓	72.95	67.46	58.74
✓	×	✓	83.56	80.89	74.86
✓	✓	×	79.47	72.09	66.62
✓	✓	✓	**92.21**	**85.60**	**83.85**

Table 3: Ablation study on different model components.

consistency loss $\mathcal{L}_{\text{gen}}$ while retaining $\mathcal{L}_{\text{intra}}$ and $\mathcal{L}_{\text{cross}}$ also results in inferior performance, suggesting that decoupling view completion from clustering hinders unified optimization. Overall, the full model consistently achieves the best results, validating the complementary and synergistic contributions of all components.

Analysis of the G-ACL Module. Table 4 shows that removing any single loss in G-ACL module leads to clear degradation, confirming their complementarity. In particular, excluding $\mathcal{L}_{\text{gen}}$ causes the largest drop, highlighting its role in enforcing generation–semantic consistency. When optimized alone, neither $\mathcal{L}_{\text{intra}}$ nor $\mathcal{L}_{\text{cross}}$ suffices for effective clustering, as the former mainly stabilizes intra-view representations while the latter lacks generation-aware supervision. Although $\mathcal{L}_{\text{gen}}$ alone performs relatively better, the best results are achieved only when all three losses are jointly optimized, validating the necessity of the tri-level contrastive objective.

Analysis of the CTAG Module. To further investigate the role of conditional information in the CTAG component, we visualize the learned representations using t-SNE under two ablation settings: unconditioned generation (i.e., the w is set as -1 in Eq.(7)) and global correlation-conditioned genera-

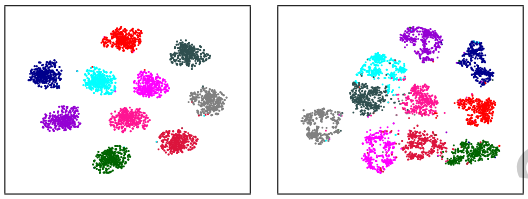


(a) Impact of λ_1 and λ_2 (b) Impact of Timesteps T (c) Impact of Stride Size S (d) Efficiency vs. Accuracy

Figure 2: Experimental analysis on hyperparameters and performance comparison. **(a)** Impact of parameters in total loss ; **(b)** Impact of sampling timesteps T on accuracy and time; **(c)** Analysis of stride size S (with fixed $T = 1000$); **(d)** Running time comparison with the baseline (DCG).

Loss Configuration	Loss Terms			Performance		
	$\mathcal{L}_{intra}$	$\mathcal{L}_{cross}$	$\mathcal{L}_{gen}$	ACC	NMI	ARI
w/o $\mathcal{L}_{intra}$	×	✓	✓	90.41	82.58	81.77
w/o $\mathcal{L}_{cross}$	✓	×	✓	91.60	83.67	82.47
w/o $\mathcal{L}_{gen}$	✓	✓	×	79.47	72.09	66.62
Only $\mathcal{L}_{intra}$	✓	×	×	60.45	49.58	40.86
Only $\mathcal{L}_{cross}$	×	✓	×	42.15	46.12	33.72
Only $\mathcal{L}_{gen}$	×	×	✓	91.15	83.26	81.76
Ours (Full)	✓	✓	✓	**92.21**	**85.60**	**83.85**

Table 4: Ablation study on different loss configurations.



(a) Conditional (b) Unconditional

Figure 3: t-SNE visualization on the MNIST-USPS dataset with a missing rate of 0.3. (a) Clustering results guided by conditional view features; (b) Distribution of clustering results under the unconditioned setting.

tion, as shown in Fig. 3. As illustrated in Fig. 3(a), the unconditioned setting produces relatively scattered clusters with ambiguous inter-class boundaries, indicating insufficient semantic alignment in the latent space. In contrast, Fig. 3(b) shows that incorporating conditional guidance leads to more compact intra-class distributions and clearer inter-class separation. This observation suggests that global correlation-aware conditioning effectively guides the diffusion process to generate semantically consistent representations, thereby enhancing discriminative structure in the learned space.

4.3 Parameter and Efficiency Analysis

We study the effects of the loss-weight parameters λ in Eq. (14), the sampling timesteps T, and the stride size S on both performance and inference efficiency in the Fast Sampling for View Completion section, and report the runtime comparison in Fig. 2. Fig. 2(a) presents the clustering accuracy on the Handwritten dataset with a missing rate of 0.3 under different combinations of λ_1 and λ_2 ranging from 0.05 to 1.0. As observed, when $\lambda_1 \in [0.05, 1.0]$ and $\lambda_2 \in [0.05, 0.2]$, the ACC consistently exceeds 0.90. This suggests that these hyperparameters lie within a stable region, highlighting the robustness. In addition, the sensitivity of λ_3 is analyzed in a similar manner. Accordingly, we set $\lambda_1 = 0.2$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.01$ in all experiments.

In this paper, we use the sampling timesteps T to control the number of denoising iterations. As shown in Fig. 2(b), increasing T from 600 to 1000 steadily improves clustering accuracy, which peaks at $T = 1000$, while further increases bring marginal gains with higher cost. Thus, we set $T = 1000$ by default. With $T = 1000$ fixed, Fig. 2(c) shows that the stride $S = 40$ significantly reduces running time while maintaining stable accuracy, demonstrating robustness to stride variations. Fig. 2(d) compares the running time of our method with DCG [2025]. From the efficiency perspective (left axis), our method consistently requires less running time than DCG on all datasets, achieving over a $4\times$ speedup on the large-scale *CIFAR-10* dataset. From the performance perspective (right axis), our method consistently attains higher clustering accuracy. In particular, on the challenging *Handwritten* dataset, DCG fails to converge effectively, whereas our method maintains accuracy above 90%. These results demonstrate a superior balance between effectiveness and efficiency.

5 Conclusion

In this paper, we proposed IASR, a unified framework that reformulates missing-view recovery as an instance-aligned token generation process for incomplete multi-view clustering. By introducing a stable–active dual-encoder architecture(S-ADE), IASR learns robust and view-adaptive representations while preventing representation collapse under severe view missingness. The proposed cross-view token alignment generator (CTAG) explicitly captures long-range cross-view dependencies along the denoising trajectory, and the generative-adaptive contrastive learning (G-ACL) strategy tightly couples view completion with discriminative representation learning. Extensive experiments on eight benchmarks demonstrate that IASR achieves superior clustering performance, especially under high missing rates, while significantly improving inference efficiency.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62471426), the Shandong Provincial Natural Science Foundation (No.ZR2024MF060). The authors would also like to thank the anonymous reviewers for their valuable comments and suggestions.

References

[Azizi *et al.*, 2023] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023.

[Chao *et al.*, 2024] Guoqing Chao, Yi Jiang, and Dianhui Chu. Incomplete Contrastive Multi-View Clustering with High-Confidence Guiding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10):11221–11229, March 2024.

[Cui *et al.*, 2024] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. Drive As You Speak: Enabling Human-Like Interaction With Large Language Models in Autonomous Vehicles. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 902–909, 2024.

[Fang *et al.*, 2023] 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, December 2023.

[Han *et al.*, 2023] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted Multi-View Classification With Dynamic Evidential Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2551–2566, February 2023.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[Hu and Chen, 2018] Menglei Hu and Songcan Chen. Doubly Aligned Incomplete Multi-view Clustering. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 2262–2268, Stockholm, Sweden, July 2018. International Joint Conferences on Artificial Intelligence Organization.

[Jin *et al.*, 2023] Jiaqi Jin, Siwei Wang, Zhibin Dong, Xinwang Liu, and En Zhu. Deep Incomplete Multi-View Clustering with Cross-View Partial Sample and Prototype Alignment. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11600–11609, June 2023.

[Jin *et al.*, 2025] 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, March 2025.

[Li *et al.*, 2023] 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.

[Li *et al.*, 2024] Liang Li, Yuangang Pan, Jie Liu, Yue Liu, Xinwang Liu, Kenli Li, Ivor W. Tsang, and Keqin Li. BGAE: Auto-Encoding Multi-View Bipartite Graph Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 36(8):3682–3696, August 2024.

[Lin *et al.*, 2021] 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)*, pages 11169–11178, June 2021.

[Lin *et al.*, 2022] 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*, pages 1–14, 2022.

[Lin *et al.*, 2023] Shaobo Lin, Kun Wang, Xingyu Zeng, and Rui Zhao. Explore the power of synthetic data on few-shot object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 638–647, 2023.

[Liu *et al.*, 2021] Xinwang Liu, Miaomiao Li, Chang Tang, Jingyuan Xia, Jian Xiong, Li Liu, Marius Kloft, and En Zhu. Efficient and Effective Regularized Incomplete Multi-View Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2634–2646, August 2021.

[Lu *et al.*, 2024] Yiding Lu, Yijie Lin, Mouxing Yang, Dezhong Peng, Peng Hu, and Xi Peng. Decoupled Contrastive Multi-View Clustering with High-Order Random Walks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(13):14193–14201, March 2024.

[Tang and Liu, 2022] Huayi Tang and Yong Liu. Deep Safe Incomplete Multi-view Clustering: Theorem and Algorithm. In *Proceedings of the 39th International Conference on Machine Learning*, pages 21090–21110. PMLR, June 2022.

[Tian *et al.*, 2023] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, 36:48382–48402, 2023.

[Wang *et al.*, 2021] 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.

[Wang *et al.*, 2023] Yiming Wang, Dongxia Chang, Zhiqiang Fu, Jie Wen, and Yao Zhao. Incomplete Multiview Clustering via Cross-View Relation Transfer. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(1):367–378, January 2023.

[Wang *et al.*, 2025] Ziyu Wang, Yiming Du, Yao Wang, Rui Ning, and Lusi Li. Deep Incomplete Multi-view Cluster-

ing via Multi-level Imputation and Contrastive Alignment. *Neural Networks*, 181:106851, January 2025.

[Wen *et al.*, 2020] Jie Wen, Yong Xu, and Hong Liu. Incomplete Multiview Spectral Clustering With Adaptive Graph Learning. *IEEE Transactions on Cybernetics*, 50(4):1418–1429, April 2020.

[Wen *et al.*, 2024] 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*, June 2024.

[Xu *et al.*, 2022] Jie Xu, Chao Li, Yazhou Ren, Liang Peng, Yujie Mo, Xiaoshuang Shi, and Xiaofeng Zhu. Deep incomplete multi-view clustering via mining cluster complementarity. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 8761–8769, 2022.

[Xu *et al.*, 2023] 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.

[Xu *et al.*, 2024] 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, pages 16147–16155, 2024.

[Yan *et al.*, 2023a] Weiqing Yan, Meiqi Gu, Jinlai Ren, Guanghui Yue, Zhaowei Liu, Jindong Xu, and Weisi Lin. Collaborative structure and feature learning for multi-view clustering. *Information Fusion*, 98:101832, October 2023.

[Yan *et al.*, 2023b] Weiqing Yan, Yuanyang Zhang, Chenlei Lv, Chang Tang, Guanghui Yue, Liang Liao, and Weisi Lin. GCFAgg: Global and Cross-View Feature Aggregation for Multi-View Clustering. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19863–19872, Vancouver, BC, Canada, June 2023. IEEE.

[Yan *et al.*, 2024a] Weiqing Yan, Kanglong Liu, Wujie Zhou, and Chang Tang. Deep incomplete multi-view clustering via dynamic imputation and triple alignment with dual optimization. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[Yan *et al.*, 2024b] Weiqing Yan, Yuanyang Zhang, Chang Tang, Wujie Zhou, and Weisi Lin. Anchor-sharing and cluster-wise contrastive network for multiview representation learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2):3797–3807, 2024.

[Yin *et al.*, 2025] 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, August 2025.

[Zhang *et al.*, 2020] Pei Zhang, Siwei Wang, Jingtao Hu, Zhen Cheng, Xifeng Guo, En Zhu, and Zhiping Cai. Adaptive Weighted Graph Fusion Incomplete Multi-View Subspace Clustering. *Sensors*, 20:5755, October 2020.

[Zhang *et al.*, 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(21):22650–22658, April 2025.