TOPOLOGY-DISTILLED PROMPT PROPAGATION FOR INCOMPLETE MULTI-VIEW CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Incomplete multi-view clustering (IMVC) aims to learn clusters from samples with partially observed views. Existing methods usually follow two directions, completing missing features or aligning the available views without imputation. Feature completion may inject biased synthetic signals because the missing view is not uniquely determined by the observed ones. Imputation-free methods avoid raw reconstruction, but they often rely on static consistency or pairwise alignment among observed views, which provides limited sample-specific guidance for refining missing-view representations. These limitations motivate a routing perspective in which the model learns how observed-view cues are encoded, exchanged, and fused across samples and missing patterns, rather than prescribing this process through a completion target or fixed alignment rule. We propose Topology-Distilled Prompt Propagation (TDPP), a framework that formulates missing-view representation learning as context-conditioned latent refinement. Its core module, cross-view prompt propagation, encodes observed-view cues into instance-specific prompts and iteratively exchanges them across views and samples under observation masks. This process enables missing-view representations to absorb reliable context while blocking messages from unobserved sources. To keep the propagated context consistent with the clustering objective, a self-supervised teacher trained on fully observed samples provides reference representations for the student. The teacher further guides the student with quality-aware feature distillation to account for unequal view reliability and sparse multi-hop topology distillation to preserve the neighborhood structure learned from complete data. Experiments on benchmark datasets show that TDPP outperforms recent state-of-the-art IMVC methods across missing ratios.

1 INTRODUCTION

Multi-view clustering (MVC) aims to group unlabeled samples by exploiting heterogeneous observations from multiple sensors, modalities, or feature extractors (Chen et al. (2026b); Liu et al. (2026a); Dong et al. (2026b)). Its advantage over single-view clustering comes from the fact that different views often provide complementary information about the same latent semantics. When all views are available, a model can compare these observations and learn a consensus structure that is more stable than any individual view (Cui et al. (2026); Dong et al. (2025)). This property has made MVC useful in multimedia analysis, remote sensing, bioinformatics, and other multi-source learning scenarios (Liu et al. (2026a); Dong et al. (2025)).

The fully observed setting is rarely guaranteed in practice (Jiang et al. (2025); Wang et al. (2026)). Views may be missing due to sensor failures, transmission errors, annotation cost, privacy restrictions, or feature extraction failures (Yu et al. (2025)). Incomplete multi-view clustering (IMVC) therefore relaxes the full-view assumption and seeks to partition all samples when each sample is observed by only a subset of views (Liu et al. (2025b); Qin et al. (2025)). What makes IMVC difficult is not only that some feature vectors disappear. Missingness also breaks the paths through which complementary information is compared. Two samples from the same class may be observed through different view subsets, while two available views may have unequal semantic reliability for the same sample. The model has to cluster with uneven observations and uncertain neighborhoods, rather than merely fill empty features.

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

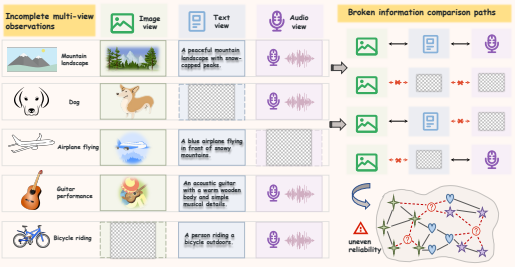


Figure 1: Background and challenge of incomplete multi-view clustering. Missing views break cross-view comparison paths and make view reliability vary across samples. Useful context should be routed through valid observations, while unavailable views should not act as information sources.

Figure 1 depicts the structure of this incomplete setting. Complete samples allow heterogeneous views to verify and complement one another. When some views are absent, only part of the comparison graph remains valid, and the available views may support different semantic relations for different samples. Directly completing missing views can create artificial links, while aligning only the observed views leaves missing-view representations weakly conditioned. The useful signal therefore lies in valid observed-view context and in the neighborhood structure that can still be estimated from complete samples. Recent IMVC methods can be viewed as attempts to repair these broken information paths. Similarity-level imputation and cross-view affinity learning infer missing cross-view structures from observed relations (Yu et al. (2025); Liu et al. (2025b)). Anchor, graph, and structural optimization methods introduce compact or interpretable supports for organizing partial observations (Qin et al. (2025); Li et al. (2025); Wu et al. (2026a); Zhang et al. (2026b); Zhao et al. (2026)). Tensor, subspace, and unaligned multi-view models study related structural constraints when correspondences are incomplete or unreliable (Ji & Feng (2025); Jiang et al. (2026); Gu et al. (2026); Zhou et al. (2026)). Diffusion or generative methods synthesize missing structures, while contrastive and unified frameworks strengthen agreement under non-IID, disorderly, or severe missing patterns (Li et al. (2026); Zhang et al. (2026a); Ding et al. (2025); Zhao et al. (2025a); Jiang et al. (2025)). These advances are valuable, but they lead to a natural question. *How can an IMVC model route reliable context from observed views to missing representations without committing to ambiguous imputations or being confined to the observed-view subspace?*

This question reflects the tension between two common designs. Imputed features, graphs, or assignments may reshape local neighborhoods and cluster centroids when the missing view is not uniquely determined by the available views (Yu et al. (2025); Li et al. (2026); Zhang et al. (2026a); Jiang et al. (2025)). Imputation-free alignment avoids raw-view hallucination, but it often provides limited instance-specific guidance for missing-view placeholders (Dai et al. (2025); Wu et al. (2025)). The issue is further amplified by view imbalance, since view reliability can vary across samples and should not be treated as a fixed view-level property (Liu et al. (2025a)).

Motivated by this question, we propose *Topology-Distilled Prompt Propagation* (TDPP), which treats incomplete representation learning as mask-aware latent information routing. TDPP compresses observed-view cues into compact prompts and propagates them across views and samples under observation masks. In this way, missing-view representations can be conditioned on reliable cross-view context, while unobserved sources are prevented from sending messages. To keep this routing process aligned with the clustering objective, a self-supervised teacher learned from fully observed samples provides view-quality signals and a neighborhood topology for the student. The resulting framework avoids committing to a single imputed view and instead refines the latent representation through mask-aware communication and structure-aware distillation.

Our main contributions are summarized as follows.

- We propose TDPP, which formulates IMVC as mask-aware latent information routing. Rather than prescribing the fusion process through a predefined feature-completion target or fixed cross-view alignment constraint, TDPP learns how observed-view cues should be encoded, exchanged, and integrated under observation masks to produce clusterable representations.

- We design a cross-view prompt propagation (CVPP) module that compresses observed-view information into compact prompts and iteratively routes them across views and samples under observation masks. This enables missing-view representations to be refined by reliable observations while suppressing messages from unobserved sources.

- We introduce a topology-distilled student learning strategy, where a self-supervised teacher trained on complete samples provides quality-aware feature guidance and neighborhood-level topology supervision. Extensive experiments on benchmark datasets, together with ablation and visualization studies, validate the effectiveness and robustness of TDPP under different missing ratios.

2 Inspiration and Motivation

Most IMVC methods either impute missing views, risking biased signals when cross-view correlations are weak, or align only the observed subspace, which offers limited instance-specific guidance for refining missing representations. Prompt-based conditioning provides an alternative direction, carrying compact context without committing to a reconstructed view.

Our design is inspired by the role of prompts as compact context carriers. Recent studies view prompt engineering as a parameter-efficient way to condition large pretrained models through input context or a small set of learnable tokens, instead of updating the full backbone (Liu et al., 2026b). In large language models, this idea is often associated with arranging task descriptions, retrieved knowledge, or reasoning context around the input. In multimodal models, prompts further serve as compact context carriers that adapt vision-language or multimodal Transformers to downstream tasks with limited trainable parameters. A prompt exposes task-relevant context in a controllable form, rather than acting as a full replacement for the missing information.

Recent missing-modality studies make this principle more concrete. A recent survey on multimodal learning with missing modality separates data-level imputation from representation-focused and architecture-focused strategies, and notes prompt or adapter tuning as a practical route for incomplete inputs (Wu et al., 2026b). RAGPT retrieves related instances and generates context-aware prompts to compensate for incomplete multimodal inputs, while MemPrompt stores modality-specific semantic information in a prompt memory and uses shared prompts to generalize across missing patterns (Lang et al., 2025; Zhao et al., 2025b). ComP propagates cross-modal prompts to relieve modality imbalance in incomplete emotion recognition (He et al., 2026). AOEPT further argues that conditioning only on observed modalities can confine inference to a modality-reduced subspace, and introduces modality-contextualized prompts to recover missing-modality priors (Lang et al., 2026). These methods show that prompts can carry cross-modal context effectively. However, many of them are developed under supervised objectives that guide prompt learning and settings where modalities are treated as potential information sources. IMVC lacks these conditions. Without labels, there is no external signal to guide what prompts encode. Without the completeness assumption, unobserved views become noise sources if allowed to transmit. This shifts the problem from cross-modal compensation toward conditioned routing under a clustering objective.

TDPP is designed around this shift from compensation to routing. Unlike existing prompt-based missing-modality methods, it restricts information propagation to valid observation paths under an explicit mask and couples this routing process with structural supervision from complete samples. The teacher further provides view-quality guidance and a sparse multi-hop topology prior to preserve clustering-relevant relationships. In this way, mask-conditioned routing and topology-distilled supervision jointly constrain the propagated representations.

3 Proposed Method

3.1 Problem Definition

Consider a multi-view dataset consisting of N unlabeled samples across V distinct views. Denote the entire dataset as $\mathcal{X} = \{\mathcal{X}_i\}_{i=1}^{N}$, where the i-th sample is represented by a collection of view-specific feature vectors $\mathcal{X}_i = \{\mathbf{x}_i^v \in \mathbb{R}^{d_v}\}_{v=1}^{V}$, and d_v denotes the dimensionality of the v-th view.

To describe the incomplete setting, let $\mathcal{O}_i \subseteq \{1, \ldots, V\}$ denote the set of observed-view indices for sample i. We assume that each sample has at least one observed view. The sample indices are then divided into a complete set $\mathcal{D}_C = \{i \mid |\mathcal{O}_i| = V\}$, where all views are available, and an incomplete set $\mathcal{D}_I = \{i \mid 1 \leq |\mathcal{O}_i| < V\}$, where at least one view is missing.

The objective of IMVC is to partition the entire dataset $\mathcal{X}$ into K distinct clusters by learning a consensus representation that leverages both complete and partial observations. The overall pipeline of our proposed method addresses this challenge and is illustrated in Fig. 2.

As shown in Fig. 2, TDPP follows a teacher-student pipeline. The teacher is trained on complete samples to provide complete-view representations and prototypes, from which view-quality estimates and a sparse multi-hop topology prior are derived. The student then processes complete and incomplete samples together, where observed views are encoded normally and missing views start from learnable placeholders. CVPP transfers compact context from observed views to missing-view representations, while feature, quality-aware, and topology distillation keep the propagated representations aligned with the cluster structure learned from complete data.

3.2 Teacher: Learning Structural Priors from Complete Views

The objective of the teacher phase is to learn a structural prior from the complete subset $\mathcal{D}_C$, serving as a geometric target for subsequent distillation. This requires constructing a latent space that captures both intra-sample cross-view consistency and inter-sample semantic separability.

While the framework is agnostic to the feature extraction architecture, we employ view-specific Auto-Encoders (AEs) as the backbone. The teacher combines self-reconstruction, cross-view translation, and semantic alignment objectives through the aggregate objective $\mathcal{L}_{\text{teacher}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{cross}}(\mathcal{L}_{\text{cross}} + \mathcal{L}_{\text{mask}}) + \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}$.

In particular, K-means is periodically applied to the concatenated complete-view features $\mathbf{z}^{\text{tch}}_{\text{cat},i} = [\mathbf{z}^1_{\text{tch},i}, \ldots, \mathbf{z}^V_{\text{tch},i}] \in \mathbb{R}^{Vd}$, yielding global centroids $\mathbf{C} = [\mathbf{C}^1, \ldots, \mathbf{C}^V] \in \mathbb{R}^{K\times Vd}$. Here, $\mathbf{C}^v \in \mathbb{R}^{K\times d}$ is the view-v block of the global centroids. Learnable per-view prototypes $\mathbf{A}^v \in \mathbb{R}^{K\times d}$ are aligned with the corresponding block $\mathbf{C}^v$. This retains the class-level structure subsequently queried by CVPP. Upon convergence, the teacher is frozen and provides view-aligned feature targets, the prototypes $\mathbf{A}^v$, from which view-quality cues and a sparse multi-hop topology prior are derived for student distillation. Complete definitions of the teacher objectives and the EMA target encoder are provided in Appendix A.

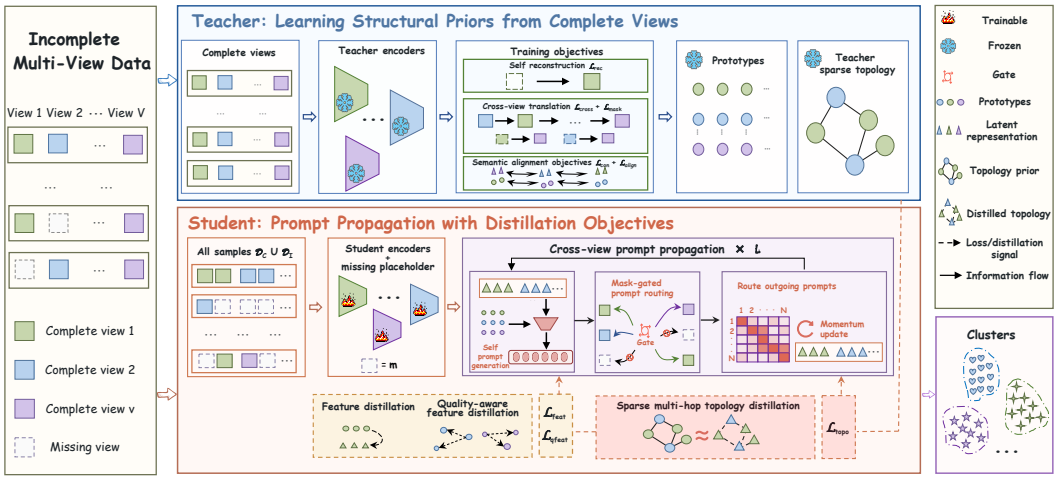


Figure 2: Overall framework of TDPP. The teacher phase learns complete-view representations and prototypes, from which view-quality cues and a sparse multi-hop topology prior are derived. The student phase handles all samples by encoding observed views and assigning learnable placeholders to missing views. CVPP routes compact prompts through observation masks and refines representations with cross-sample attention. Feature, quality-aware feature, and sparse multi-hop topology distillation jointly guide the student toward clusterable representations without reconstructing raw missing views.

3.3 STUDENT: PROMPT PROPAGATION WITH DISTILLATION OBJECTIVES

The objective of the student phase is to learn representations for all samples by propagating contextual cues from observed views to missing views, rather than reconstructing raw features.

The student encoders f^v_{stu} share architecture with the teacher and are initialized from frozen teacher weights. A learnable missing placeholder $\mathbf{m} \in \mathbb{R}^d$ replaces missing-view encodings.

$$\mathbf{z}_i^{u,(0)} = \begin{cases} f^u_{\text{stu}}(\mathbf{x}_i^u) & u \in \mathcal{O}_i \\ \mathbf{m} & u \notin \mathcal{O}_i \end{cases}. \tag{1}$$

3.3.1 CROSS-VIEW PROMPT PROPAGATION

When a view is absent, the placeholder at that view contains no sample-specific content. Treating it as an ordinary token would allow missing sources to participate in cross-view fusion, while explicit feature reconstruction can introduce imputation errors. CVPP therefore uses compact prompts as carriers of contextual cues. Observed views send prompts to other views under the observation mask, and missing sources are blocked from contributing.

Within each iteration, a view first summarizes its current state with respect to the frozen teacher class structure. The resulting self-prompt is fused with incoming prompts and the current feature, then refined by masked cross-sample attention within the same view. The refined token is split into an updated feature and outgoing prompts, both updated with momentum for stable propagation. Starting from zero prompts and learnable missing placeholders, repeated CVPP blocks yield multi-view representations whose missing views are informed by observed-view context. We detail each stage below.

(a) Self prompt generation. All cross-view prompts are initialized to $\mathbf{0}$ and will be updated in step (c). At the l-th CVPP iteration, view v first measures how its current feature $\mathbf{z}_i^{v,(l-1)}$ relates to the frozen teacher prototypes $\mathbf{A}^v$:

$$\mathbf{s}_i^v = \text{softmax}\left(\frac{\mathbf{z}_i^{v,(l-1)}\mathbf{A}^{v\top}}{\tau_s}\right), \qquad \mathbf{a}_i^v = \mathbf{s}_i^v\mathbf{A}^v. \tag{2}$$

The resulting summary $\mathbf{a}_i^v$ captures class-level context, while the bottleneck MLP_P compresses it with the current feature into a compact structural summary.

$$\mathbf{p}^v_{\text{self},i} = \text{MLP}_P\big([\mathbf{z}_i^{v,(l-1)}, \mathbf{a}_i^v]\big) \in \mathbb{R}^p, \quad p \ll d. \tag{3}$$

This self-prompt serves as a local descriptor that will be fused with incoming cross-view information in the next step.

(b) Mask-gated prompt routing. View v then gathers prompts that other views have sent to it in the previous iteration, but retains only those from actually observed views:

$$\mathbf{p}^v_{\text{in,i}} = \bigoplus_{u \neq v} \Big(\mathbf{p}_i^{u\to v,(l-1)} \cdot \mathbb{I}[u \in \mathcal{O}_i]\Big) \in \mathbb{R}^{(V-1)p}. \tag{4}$$

The observation gating prevents missing source views from injecting noise. The collected prompts are concatenated with the current feature and the self-prompt, then linearly projected into a unified token:

$$\mathbf{g}_i^v = \mathbf{W}_{\text{proj}}\,[\mathbf{z}_i^{v,(l-1)}, \mathbf{p}^v_{\text{self},i}, \mathbf{p}^v_{\text{in,i}}]. \tag{5}$$

Sample relations within each view are modeled by applying masked multi-head self-attention (MSA) to the batch tokens $\{\mathbf{g}_i^v\}_{i=1}^B$. The attention mask is derived from $\mathcal{O}^v = \{i \mid v \in \mathcal{O}_i\}$ and is applied to key/value positions, preventing samples missing view v from serving as keys or values while retaining them as queries

$$\mathbf{G}^v = \text{MSA}\big(\{\mathbf{g}_i^v\},\ \text{mask} = \neg\mathcal{O}^v\big). \tag{6}$$

Thus, samples missing view v can still be updated as queries, but they cannot provide context to other samples in that view.

(c) Route outgoing prompts and momentum update. From the refined token $\mathbf{G}^v[i]$, a linear recovery head extracts an updated feature and a set of candidate outgoing prompts:

$$[\tilde{\mathbf{z}}_i^v,\ \{\tilde{\mathbf{p}}_i^{v\to u}\}_{u\neq v}] = \mathbf{W}_{\text{rec}}\,\mathbf{G}^v[i]. \tag{7}$$

Both are merged with their previous values via exponential moving averages to stabilise the propagation. The feature update applies to all samples:

$$\mathbf{z}_i^{v,(l)} = \lambda_f\,\mathbf{z}_i^{v,(l-1)} + (1-\lambda_f)\,\tilde{\mathbf{z}}_i^v. \tag{8}$$

For the prompts, only observed views are allowed to send information, missing views output zeros:

$$\mathbf{p}_i^{v\to u,(l)} = \begin{cases} \lambda_p\,\mathbf{p}_i^{v\to u,(l-1)} + (1-\lambda_p)\,\tilde{\mathbf{p}}_i^{v\to u}, & v\in\mathcal{O}_i \\ \mathbf{0}, & v\notin\mathcal{O}_i. \end{cases} \tag{9}$$

This gating separates receiving information from providing information. If view v is missing for sample i, its feature $\mathbf{z}_i^v$ can still be updated through attention, but it cannot send prompts to other views. The missing view is therefore refined using observed-view context without letting placeholders become unreliable sources. During student training, some complete samples are additionally processed by the student CVPP with one view temporarily masked, while the frozen teacher targets are extracted from their original complete inputs, providing partial-observation supervision without an explicit reconstruction loss.

3.3.2 Distillation Objectives

While CVPP determines how observed-view information is propagated, the following objectives constrain the propagated representations using the complete-view teacher. The propagated features are supervised on complete samples, where the frozen teacher provides reliable view-wise targets. Let $\hat{\mathbf{z}}_{\text{stu},i}^v = \text{norm}(\mathbf{z}_i^{v,(L)})$ and $\hat{\mathbf{z}}_{\text{tch},i}^v = \text{norm}(\mathbf{z}_{\text{tch},i}^v)$ denote the normalized student and teacher features. Feature distillation directly anchors each student representation to its teacher counterpart via mean squared error:

$$\mathcal{L}_{\text{feat}} = \frac{1}{N_C V}\sum_{i\in\mathcal{D}_C}\sum_{v=1}^{V}\left\|\hat{\mathbf{z}}_{\text{stu},i}^v - \hat{\mathbf{z}}_{\text{tch},i}^v\right\|_2^2. \tag{10}$$

Feature distillation weights all views equally, yet a view can be observed but still less discriminative for a particular neighborhood. Quality-aware feature distillation addresses this by estimating sample-specific view reliability from the frozen teacher. Specifically, TDPP asks whether a single view preserves the local neighborhood defined by the fused teacher representation.

Let $\mathbf{h}_{\text{tch},i} = [\hat{\mathbf{z}}_{\text{tch},i}^1, \cdots, \hat{\mathbf{z}}_{\text{tch},i}^V]$ and let $\mathcal{N}_k(i)$ be the k nearest neighbors of sample i in this fused space. For view v, the local contrastive score is

$$Q_i^v = -\frac{1}{k}\sum_{j\in\mathcal{N}_k(i)}\log\frac{\exp\left(\hat{\mathbf{z}}_{\text{tch},i}^{v\top}\hat{\mathbf{z}}_{\text{tch},j}^v/\tau_q\right)}{\sum_{j'\neq i}\exp\left(\hat{\mathbf{z}}_{\text{tch},i}^{v\top}\hat{\mathbf{z}}_{\text{tch},j'}^v/\tau_q\right)}. \tag{11}$$

Smaller values indicate that view v keeps the teacher neighbors close while separating them from other samples. The scores are converted into confidence weights

$$\omega_i^v = \frac{\exp(-Q_i^v/\tau_\omega)}{\sum_{u=1}^{V}\exp(-Q_i^u/\tau_\omega)}. \tag{12}$$

The quality-aware feature distillation objective is then defined as:

$$\mathcal{L}_{\text{qfeat}} = \frac{1}{N_C}\sum_{i\in\mathcal{D}_C}\sum_{v=1}^{V}\omega_i^v\left\|\hat{\mathbf{z}}_{\text{stu},i}^v - \hat{\mathbf{z}}_{\text{tch},i}^v\right\|_2^2. \tag{13}$$

This weighting makes the student follow the teacher more strongly on views that preserve local structure, while reducing the influence of less informative view-specific targets. We retain $\mathcal{L}_{\text{feat}}$ as a view-uniform anchor, while $\mathcal{L}_{\text{qfeat}}$ adaptively emphasizes locally reliable view-specific targets.

Feature-level alignment, whether uniform or quality-weighted, does not constrain how samples are arranged relative to one another. For clustering, this relational structure is crucial because nearby samples should remain close after prompt propagation. Sparse multi-hop topology distillation addresses this by transferring the teacher's neighborhood structure to the student.

A sparse k'-NN graph is built in the fused teacher space $\{\mathbf{h}_{\text{tch},i}\}_{i\in\mathcal{D}_C}$, with Gaussian edge weights

$$S_{ij}^{\text{tch}} = \begin{cases} \exp\left(-\left\|\mathbf{h}_{\text{tch},i} - \mathbf{h}_{\text{tch},j}\right\|_2^2 / (2\sigma^2)\right), & j \in \mathcal{N}_{k'}(i), \\ 0, & \text{otherwise}, \end{cases} \tag{14}$$

where σ is estimated from the neighbor distances. Row-normalizing $\mathbf{S}^{\text{tch}}$ gives the one-step transition matrix $\mathbf{T}$. Short random walks on this graph define the teacher topology

$$\tilde{p}_{\text{tch}}(j|i) = \frac{\sum_{h=0}^{H}(\mathbf{T}^h)_{ij}}{\sum_{j'}\sum_{h=0}^{H}(\mathbf{T}^h)_{ij'} + \epsilon}. \tag{15}$$

which preserves immediate neighbors while allowing higher-order neighborhood cues to contribute.

The student is encouraged to reproduce this relational structure from its final representation. Let $\mathbf{h}_{\text{stu},i} = [\hat{\mathbf{z}}_{\text{stu},i}^1, \cdots, \hat{\mathbf{z}}_{\text{stu},i}^V]$ be the fused student feature. Its soft neighborhood distribution over complete samples is

$$q_{\text{stu}}(j|i) = \frac{\exp\left(-\left\|\mathbf{h}_{\text{stu},i} - \mathbf{h}_{\text{stu},j}\right\|_2^2 / \tau_t\right)}{\sum_{j'}\exp\left(-\left\|\mathbf{h}_{\text{stu},i} - \mathbf{h}_{\text{stu},j'}\right\|_2^2 / \tau_t\right)}. \tag{16}$$

The sparse multi-hop topology distillation objective minimizes the cross-entropy from the teacher distribution to the student distribution:

$$\mathcal{L}_{\text{topo}} = -\frac{1}{N_C}\sum_{i\in\mathcal{D}_C}\sum_{j}\tilde{p}_{\text{tch}}(j|i)\log q_{\text{stu}}(j|i). \tag{17}$$

The neighborhood-based distillation terms are used only when the teacher structure is sufficiently informative. If the complete subset is too small or the graph becomes overly dense, neighborhood supervision can become unstable. With the validity indicator $\eta \in \{0, 1\}$, the student objective is

$$\mathcal{L}_{\text{student}} = \lambda_{\text{feat}}\mathcal{L}_{\text{feat}} + \eta\left(\lambda_{\text{qfeat}}\mathcal{L}_{\text{qfeat}} + \lambda_{\text{topo}}\mathcal{L}_{\text{topo}}\right). \tag{18}$$

When $\eta = 0$, the student falls back to feature distillation, avoiding unreliable structural constraints while retaining teacher-guided alignment. After training, clustering is performed by applying K-means to the concatenated normalized student features $\mathbf{h}_{\text{stu},i}$ for all samples.

4 Experiments

Experimental setup and implementation details are provided in Appendix B.

4.1 Main Results

Table 1 compares TDPP with recent IMVC methods under four missing ratios. Across the shown settings, the gains are most evident on NUS-WIDE and Scene15. Under the 50% missing ratio, TDPP improves the best competing ACC on NUS-WIDE from 62.77% to 77.38% and ARI from 42.42% to 64.84%. On Scene15, ACC and ARI are higher than the strongest baseline by 5.00% and 8.17%. These gains suggest that routing compact latent prompts through observed views is more reliable than directly filling missing raw features when cross-view information is uneven.

The advantage remains under severe incompleteness. At the 80% missing ratio, TDPP achieves 64.29% ACC on NUS-WIDE, 46.53% on Scene15, and 29.78% on LandUse-21, exceeding the strongest baselines by 3.94%, 4.21%, and 0.53%, respectively. The smaller margin on LandUse-21 indicates that the benefit varies with the degree of view complementarity, but the overall trend is stable. By propagating information only from observed sources and distilling teacher quality and topology cues, TDPP preserves cluster structure without relying on potentially noisy raw-view reconstruction.

Table 1: Clustering performance on four incomplete multi-view datasets at different missing ratios. Results are reported in percentage.

Ratio	Methods	CUB			NUS-WIDE			Scene15			LandUse-21		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
0%	DIVIDE (AAAI 2024)	71.61	73.53	61.05	61.86	48.68	40.25	43.87	44.70	27.18	30.44	35.68	15.79
	CANDY (NeurIPS 2024)	76.83	75.24	64.61	63.19	48.76	42.84	42.45	43.29	25.65	28.32	32.26	14.29
	I2MVC (IJCAI 2025)	53.28	55.61	35.81	47.27	36.79	27.07	21.81	21.58	7.79	18.10	22.47	6.12
	CAMEL (ICLR 2026)	62.03	55.70	45.12	56.56	39.69	33.44	35.20	32.18	18.64	27.08	32.11	12.33
	RAC-DMVC (AAAI 2026)	74.33	72.30	61.92	50.61	41.83	30.71	47.52	47.19	30.09	30.16	35.16	15.60
	GAVIM (AAAI 2026)	78.61	79.20	67.09	52.90	43.32	33.50	39.09	41.46	24.15	28.49	36.11	14.55
	DIMVC-HIA (AAAI 2026)	79.72	72.66	63.18	58.58	46.83	38.19	32.25	34.31	16.37	14.42	17.65	3.49
	Ours	**80.72**	**79.34**	**70.22**	**77.70**	**67.43**	**65.10**	**59.06**	**60.80**	**44.05**	**32.37**	**39.57**	**18.12**
20%	DIVIDE (AAAI 2024)	71.28	73.17	60.74	61.64	47.83	40.42	43.41	44.54	26.70	28.84	33.72	14.40
	CANDY (NeurIPS 2024)	73.61	74.86	64.03	61.63	48.34	41.20	43.50	43.73	26.15	28.14	31.95	13.95
	I2MVC (IJCAI 2025)	41.94	47.64	30.33	38.64	27.65	17.58	21.55	19.27	6.72	17.40	20.17	5.50
	CAMEL (ICLR 2026)	51.03	45.47	34.17	53.70	35.68	29.27	28.44	27.25	13.70	25.29	29.88	10.06
	RAC-DMVC (AAAI 2026)	69.83	69.21	57.64	57.82	46.16	37.15	47.36	47.01	29.95	28.68	32.14	14.59
	GAVIM (AAAI 2026)	77.11	**77.40**	66.45	48.07	38.55	27.11	41.50	41.73	24.56	26.75	33.70	13.15
	DIMVC-HIA (AAAI 2026)	62.28	60.86	45.78	54.41	41.43	33.20	35.95	37.91	18.90	15.70	18.25	4.31
	Ours	**81.78**	76.94	**68.07**	**77.40**	**67.23**	**64.23**	**57.88**	**60.94**	**44.77**	**30.57**	**37.54**	**16.28**
50%	DIVIDE (AAAI 2024)	70.11	72.63	60.56	62.66	48.99	41.68	42.27	43.09	25.68	29.71	33.90	15.00
	CANDY (NeurIPS 2024)	72.44	71.98	60.31	62.77	48.82	42.42	42.32	42.58	25.12	28.29	32.26	14.20
	I2MVC (IJCAI 2025)	35.94	37.61	19.89	24.20	17.71	7.16	20.85	18.99	6.50	15.17	18.21	3.97
	CAMEL (ICLR 2026)	42.50	36.69	23.21	49.22	30.67	23.42	27.46	25.42	11.44	21.97	25.56	7.22
	RAC-DMVC (AAAI 2026)	64.06	65.65	52.68	60.11	46.68	39.58	45.68	45.12	28.06	27.94	31.25	14.16
	GAVIM (AAAI 2026)	73.61	**74.48**	**61.13**	46.89	36.94	23.91	41.20	39.04	23.53	25.79	32.48	12.73
	DIMVC-HIA (AAAI 2026)	58.72	54.04	40.51	51.41	37.57	28.77	36.80	38.92	19.80	19.63	22.00	6.83
	Ours	**74.00**	67.42	55.98	**77.38**	**67.65**	**64.84**	**50.68**	**55.71**	**36.23**	**29.83**	**35.61**	**15.76**
80%	DIVIDE (AAAI 2024)	64.94	67.27	52.98	55.05	45.02	36.32	42.32	42.53	25.12	29.25	34.47	15.18
	CANDY (NeurIPS 2024)	69.11	**71.56**	**59.38**	57.87	46.19	37.80	40.90	41.35	23.73	28.21	31.86	13.95
	I2MVC (IJCAI 2025)	34.06	34.26	16.42	23.29	15.47	5.03	20.79	19.77	6.90	14.79	18.46	3.79
	CAMEL (ICLR 2026)	41.10	38.47	22.39	37.13	23.16	13.78	27.11	25.26	11.06	18.73	23.15	6.02
	RAC-DMVC (AAAI 2026)	65.78	63.33	48.96	60.35	44.83	39.04	39.26	40.78	23.06	26.59	29.68	12.58
	GAVIM (AAAI 2026)	61.17	64.74	47.61	47.18	35.70	25.56	35.44	35.22	19.72	23.95	28.98	8.86
	DIMVC-HIA (AAAI 2026)	49.17	50.71	35.85	30.00	19.76	11.73	38.66	37.55	19.69	18.35	19.39	7.16
	Ours	**70.23**	65.93	52.99	**64.29**	**52.29**	**46.97**	**46.53**	**47.14**	**29.38**	**29.78**	**34.96**	**15.64**

Table 2: Ablation study at 50% missing rate. Results are reported in percentage.

Variant	Components			NUS-WIDE			Scene15		
	CVPP	Quality-aware	Topology	ACC	NMI	ARI	ACC	NMI	ARI
Full	✓	✓	✓	**77.38**	**67.65**	**64.84**	**50.68**	**55.71**	**36.23**
No CVPP	–	✓	✓	32.05	24.17	10.60	28.67	34.81	15.70
No Topology	✓	✓	–	62.33	49.38	42.80	45.22	48.24	29.74
No Quality-aware	✓	–	✓	72.68	61.07	57.14	49.90	55.31	34.62
Feature Distillation Only	✓	–	–	61.90	48.13	41.70	45.80	46.33	29.47

4.2 Ablation Study

To examine the contribution of each component, we conduct ablation experiments on NUS-WIDE and Scene15 under the 50% missing ratio, as reported in Table 2. The most pronounced degradation comes from removing CVPP. On NUS-WIDE, ACC drops from 77.38% to 32.05% and ARI from 64.84% to 10.60%, on Scene15, ACC also decreases by 22.01%. This large drop shows that mask-aware cross-view routing is the main source of the gains, while distillation alone is not sufficient when missing views provide no sample-specific input.

The topology and quality terms provide complementary gains on top of CVPP. Removing sparse multi-hop topology distillation leads to a clear decline on both datasets, especially on NUS-WIDE where ARI decreases by 22.04%.

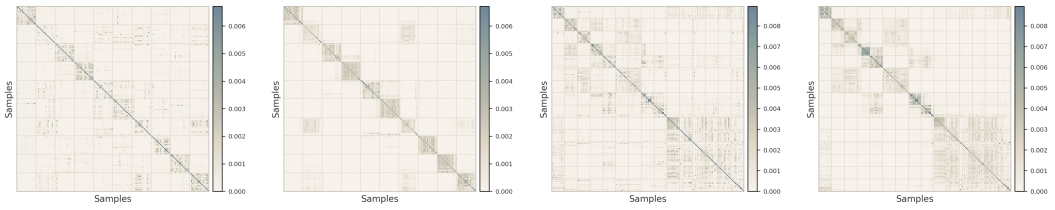
This supports the motivation that pointwise feature matching is insufficient for preserving the clustering structure learned by the teacher. The quality-aware weighting brings a smaller but consistent improvement, with larger gains on NUS-WIDE than on Scene15, indicating that view reliability is more important when the two views differ substantially in semantic strength. Finally, the feature-distillation-only variant remains below the full model on all metrics, showing that TDPP benefits from combining local feature alignment with view-aware and topology-aware structural supervision rather than relying on a single distillation signal.

TDPP remains stable once CVPP reaches several propagation iterations and across a broad range of distillation weights. Detailed sensitivity analyses are provided in Appendix C.1.

4.3 Topology Preservation Analysis

The ablation results above show that topology distillation contributes substantially to clustering performance. We further visualize the teacher and student topology patterns to examine how the neighborhood structure learned from complete views is preserved after student training.

For a direct comparison, the teacher and student fused representations are rendered using the same soft-affinity kernel on the same class-balanced complete samples. Therefore, corresponding entries in the two maps represent the same sample pairs. As shown in Fig. 3, the student preserves many of the high-affinity neighborhood blocks learned by the teacher on both NUS-WIDE and Scene15, rather than spreading affinities uniformly across samples.



(a) NUS-WIDE teacher (b) NUS-WIDE student (c) Scene15 teacher (d) Scene15 student

Figure 3: Teacher-student soft topology patterns on matched complete samples.

Additional visualizations are provided in Appendix C.2.

5 Conclusion

In this paper, we introduced TDPP, a topology-distilled prompt propagation framework for IMVC. TDPP learns a structural prior from fully observed samples with a self-supervised teacher, then guides a student that operates on all samples through CVPP. Rather than reconstructing missing raw views, the student exchanges compact semantic prompts among observed views and refines missing-view representations in the latent space. A quality-aware feature distillation term emphasizes reliable view-specific signals, while sparse multi-hop topology distillation transfers the teacher's global neighborhood structure to stabilize clustering under severe incompleteness. Experiments on four benchmark datasets show that TDPP consistently outperforms recent IMVC methods across missing ratios. Ablation, sensitivity, and visualization results further confirm the roles of prompt propagation, view reliability estimation, and topology preservation in learning coherent and discriminative representations.

Use of Large Language Models

Large language models (LLMs) were used for language polishing, literature retrieval and discovery, and assistance in drafting parts of the manuscript. All LLM-assisted content was reviewed and verified by the authors, who take full responsibility for the final manuscript.

REFERENCES

Wenlan Chen, Lu Gao, Daoyuan Wang, Fei Guo, and Cheng Liang. Geometry-aware variational information maximization for deep incomplete multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20289–20297, 2026a.

Zhaoliang Chen, William K Cheung, Hong-Ning Dai, Byron Choi, and Jiming Liu. Gated variational graph autoencoders as experts with competition and consensus for multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20445–20453, 2026b.

Jinrong Cui, Xiaohuang Wu, Wai Keung Wong, Linlin Tang, Sen Xu, and Jie Wen. Deep multi-view clustering via cluster-semantic guidance. *IEEE Transactions on Image Processing*, 35:4659–4672, 2026.

Dengxin Dai and Luc Van Gool. Ensemble projection for semi-supervised image classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2072–2079, 2013.

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

Yu Ding, Katsuya Hotta, Chunzhi Gu, Ao Li, Jun Yu, and Chao Zhang. Learning to discriminate while contrasting: Combating false negative pairs with coupled contrastive learning for incomplete multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 37(10): 6046–6060, 2025.

Shihao Dong, Yue Liu, Xiaotong Zhou, Yuhui Zheng, Huiying Xu, and Xinzhong Zhu. RAC-DMVC: Reliability-aware contrastive deep multi-view clustering under multi-source noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20835–20843, 2026a.

Zhibin Dong, Meng Liu, Siwei Wang, Ke Liang, Yi Zhang, Suyuan Liu, Jiaqi Jin, Xinwang Liu, and En Zhu. Enhanced then progressive fusion with view graph for multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15518–15527, 2025.

Zhibin Dong, Shengju Yu, Siwei Wang, Xinwang Liu, and En Zhu. Cross-view graph matching for unsupervised learning with unaligned multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 38(4):2221–2233, 2026b.

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

Zhibin Gu, Zhenhao Zhong, Xi Zhang, and Bing Li. Constant degree matrix-driven incomplete multi-view clustering via connectivity-structure and embedding tensor learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Ruiming Guo, Mouxing Yang, Yijie Lin, Xi Peng, and Peng Hu. Robust contrastive multi-view clustering against dual noisy correspondence. *Advances in Neural Information Processing Systems*, 37:121401–121421, 2024.

Wen-Jue He, Xiaofeng Zhu, and Zheng Zhang. Cross-modal prompting for balanced incomplete multi-modal emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17463–17471, 2026. doi: 10.1609/aaai.v40i21.38800.

Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. Scalable deep multimodal learning for cross-modal retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 635–644, 2019.

Jintian Ji and Songhe Feng. Anchors crash tensor: Efficient and scalable tensorial multi-view subspace clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2660–2675, 2025.

Hongyu Jiang, Hong Tao, Zhangqi Jiang, and Chenping Hou. Unaligned multi-view clustering via diversified anchor graph fusion. *Pattern Recognition*, 170:111977, 2026.

Xiaorui Jiang, Buyun He, Peng Yuan Zhou, Xinyue Chen, Jingcai Guo, Jie Xu, and Yong Liao. A unified framework to BRIDGE complete and incomplete deep multi-view clustering under Non-IID missing patterns. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 594–603, 2025.

Jian Lang, Zhangtao Cheng, Ting Zhong, and Fan Zhou. Retrieval-augmented dynamic prompt tuning for incomplete multimodal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18035–18043, 2025. doi: 10.1609/aaai.v39i17.33984.

Jian Lang, Rongpei Hong, Ting Zhong, and Fan Zhou. AOEPT: Breaking the implicit modality-reduction bottleneck in modality-missing prompt tuning. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

Ao Li, Sanlin Mei, Dehua Miao, Fengwei Gu, Tianyu Gao, and Jiahui Pan. Diffusion-imputation one-step incomplete multi-view clustering with anchor graph regularization. *Neurocomputing*, pp. 133291, 2026.

Mengyao Li, Xu Zhou, Jiapeng Zhang, Zhibang Yang, Cen Chen, and Kenli Li. TOTF: Missing-aware encoders for clustering on multi-view incomplete attributed graphs. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 3054–3062, 2025.

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, 2023.

Chengliang Liu, Jie Wen, Yong Xu, Bob Zhang, Liqiang Nie, and Min Zhang. Reliable representation learning for incomplete multi-view missing multi-label classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4940–4956, 2025a.

Qian Liu, Huibing Wang, Jinjia Peng, Yawei Chen, Mingze Yao, Xianping Fu, and Yang Wang. Consensus-guided incomplete multi-view clustering via cross-view affinities learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 5761–5769, 2025b.

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

Yao-Yang Liu, Zhen Zheng, Feng Zhang, Jin-Cheng Feng, Yi-Yang Fu, Ji-Dong Zhai, Bing-Sheng He, Xiao Zhang, Xiao-Yong Du, et al. A comprehensive taxonomy of prompt engineering techniques for large language models. *Frontiers of Computer Science*, 20:2003601, 2026b. doi: 10.1007/s11704-025-50058-z.

Yiding Lu, Yijie Lin, Mouxing Yang, Dezhong Peng, Peng Hu, and Xi Peng. Decoupled contrastive multi-view clustering with high-order random walks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14193–14201, 2024.

Yalan Qin, Guorui Feng, and Xinpeng Zhang. Fast incomplete multi-view clustering by flexible anchor learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 50235–50245. PMLR, 2025.

Youqing Wang, Yu Cao, Jinlu Wang, Xiang Xu, Jiapu Wang, Tengfei Liu, Junbin Gao, and Jipeng Guo. Aligning collaborative view recovery and tensorial subspace learning via latent representation for incomplete multi-view clustering. In *The Fourteenth International Conference on Learning Representations*, 2026.

Benyu Wu, Wei Du, Jun Wang, and Guoxian Yu. Imputation-free incomplete multi-view clustering via knowledge distillation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 6570–6578, 2025.

Chunming Wu, Zihan Fang, Zhiling Cai, Shide Du, Wei Huang, Jinbo Li, and Shiping Wang. SOI-Net: Structural optimization-inspired interpretable network for incomplete multi-view clustering. *IEEE Transactions on Multimedia*, 28:196–208, 2026a.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *Transactions on Machine Learning Research*, 2026b. arXiv:2409.07825.

Shengju Yu, Zhibin Dong, Siwei Wang, Pei Zhang, Yi Zhang, Xinwang Liu, Naiyang Guan, Tiejun Li, and Yiu ming Cheung. Simple yet effective incomplete multi-view clustering: Similarity-level imputation and intra-view hybrid-group prototype construction. In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025.

Changqing Zhang, Zongbo Han, Huazhu Fu, Joey Tianyi Zhou, Qinghua Hu, et al. CPM-Nets: Cross partial multi-view networks. *Advances in Neural Information Processing Systems*, 32, 2019.

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

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, 2026b.

Liang Zhao, Shubin Ma, Bo Xu, and Qingchen Zhang. Dual-learning based penalized multi-align clustering for multi-view incomplete and disorderly data. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1318–1326, 2025a.

Shuping Zhao, Yulong Chen, Jie Wen, Lunke Fei, Jinrong Cui, and Tingting Chai. Confident block diagonal structure-aware invariable graph completion for incomplete multi-view clustering. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yihan Zhao, Wei Xi, Xiao Fu, and Jizhong Zhao. Enhancing multimodal model robustness under missing modalities via memory-driven prompt learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 2458–2466, 2025b. doi: 10.24963/ijcai.2025/274.

Haochen Zhou, Guofeng Ding, Mouxing Yang, Peng Hu, Yijie Lin, and Xi Peng. Uncover underlying correspondence for robust multi-view clustering. In *The Fourteenth International Conference on Learning Representations*, 2026.

A Teacher Objective Definitions

While the framework is agnostic to the specific feature extraction architecture, we employ view-specific Auto-Encoders (AEs) as the backbone network. The AE framework is selected for its established capability in unsupervised representation learning, where minimizing view-specific reconstruction errors ensures the preservation of view-exclusive information within the latent embeddings.

(a) Self reconstruction. For each view $v \in \{1, \ldots, V\}$, an encoder $f^v_{\text{tch}} : \mathbb{R}^{d_v} \to \mathbb{R}^d$ maps the input $\mathbf{x}^v_i$ to a latent representation $\mathbf{z}^v_i$, and a corresponding decoder $g^v_{\text{tch}} : \mathbb{R}^d \to \mathbb{R}^{d_v}$ reconstructs the input. The self-reconstruction loss over the $N_C = |\mathcal{D}_C|$ complete samples is

$$\mathcal{L}_{\text{rec}} = \frac{1}{N_C} \sum_{i=1}^{N_C} \sum_{v=1}^{V} \left\| g^v_{\text{tch}}(\mathbf{z}^v_i) - \mathbf{x}^v_i \right\|_2^2. \tag{A.1}$$

(b) Cross-view translation. To promote the extraction of view-invariant features, we introduce cross-view reconstruction capabilities. A predictor $\psi^{u\to v}(\cdot)$ translates the latent representation of view u to reconstruct the input of view v, yielding

$$\mathcal{L}_{\text{cross}} = \frac{1}{N_C}\sum_{i=1}^{N_C}\sum_{u=1}^{V}\sum_{v\neq u}\left\|\psi^{u\to v}(\mathbf{z}_i^u)-\mathbf{x}_i^v\right\|_2^2. \tag{A.2}$$

To make this translation more robust to missing data, we randomly mask one view of selected complete samples and reconstruct it from the remaining view via the corresponding predictor:

$$\mathcal{L}_{\text{mask}} = \frac{1}{N_C}\sum_{i=1}^{N_C}\sum_{u\ \text{masked}}\left\|\psi^{v\to u}(\mathbf{z}_i^v)-\mathbf{x}_i^u\right\|_2^2, \qquad v\neq u. \tag{A.3}$$

(c) Semantic alignment objectives. We employ contrastive learning to maximize agreement between views of the same complete sample. During teacher training, stochastic view-dropout is used as an augmentation, and the augmented input is denoted by $\tilde{\mathbf{x}}_i^u$. The online projection is $\mathbf{r}_i^u = \phi^u(f_{\text{tch}}^u(\tilde{\mathbf{x}}_i^u))$, while $\bar{\mathbf{r}}_i^v$ denotes the target projection produced by the EMA target encoder of view v. The InfoNCE objective is

$$\mathcal{L}_{\text{con}} = -\frac{1}{N_C}\sum_{i=1}^{N_C}\frac{1}{V(V-1)}\sum_{u=1}^{V}\sum_{v\neq u}\log\frac{\exp(\mathbf{r}_i^u\cdot\bar{\mathbf{r}}_i^v/\tau_c)}{\sum_{j=1}^{N_C}\exp(\mathbf{r}_i^u\cdot\bar{\mathbf{r}}_j^v/\tau_c)}, \tag{A.4}$$

where τ_c is a temperature parameter.

To introduce a clustering-friendly structure without hard assignments, we concatenate the complete-view teacher features as $\mathbf{z}_{\text{cat},i}^{\text{tch}} = [\mathbf{z}_{\text{tch},i}^1,\ldots,\mathbf{z}_{\text{tch},i}^V]\in\mathbb{R}^{Vd}$ and periodically apply K-means to obtain global centroids $\mathbf{C}=[\mathbf{C}^1,\ldots,\mathbf{C}^V]\in\mathbb{R}^{K\times Vd}$. Each block $\mathbf{C}^v\in\mathbb{R}^{K\times d}$ is the centroid component associated with view v. The learnable per-view prototypes $\mathbf{A}^v\in\mathbb{R}^{K\times d}$ are pulled towards their corresponding centroid blocks. The alignment loss is

$$\mathcal{L}_{\text{align}} = \sum_{v=1}^{V}\sum_{k=1}^{K}\left\|\mathbf{A}_k^v-\mathtt{sg}[\mathbf{C}_k^v]\right\|_2^2, \tag{A.5}$$

where $\mathbf{A}_k^v,\mathbf{C}_k^v\in\mathbb{R}^d$ and $\mathtt{sg}$ stops gradients through the centroid targets. During back-propagation, only the view-specific prototypes $\mathbf{A}^v$ are updated, while $\mathbf{C}$ remains fixed. Because all blocks $\{\mathbf{C}^v\}_{v=1}^V$ come from the same centroid in the concatenated space, they share a consistent cluster index k across views. This encourages different views to follow a shared clustering structure. After teacher training, the learned class-level structure is kept fixed when the student propagates information among incomplete views.

B Experimental Setup and Implementation Details

B.1 Datasets and Evaluation Metrics

CUB (Zhang et al. (2019)) contains bird images from multiple categories. We use the first ten categories for evaluation. Two views are constructed from GoogLeNet-based visual features and doc2vec-based textual features.

Scene15 (Dai & Van Gool (2013)) includes 4,485 images from 15 indoor and outdoor scene categories. Each image is represented by two hand-crafted feature views, namely 20-dimensional GIST features and 59-dimensional PHOG features.

NUS-WIDE (Hu et al. (2019)) consists of 9,000 image-text pairs from 10 categories. We use 4,096-dimensional visual features extracted from the fc7 layer of a 19-layer VGGNet, together with 300-dimensional textual features produced by a sentence CNN.

LandUse-21 (Lin et al. (2023)) contains 2,100 satellite images covering 21 land-use categories. We adopt PHOG and LBP descriptors as two views.

Three widely used metrics in MVC are adopted for evaluation: accuracy (ACC), normalized mutual information (NMI), and adjusted rand index (ARI). Higher values of these metrics indicate better clustering performance. For a missing ratio of r, we randomly select $r\%$ of the samples as incomplete and randomly remove views from these samples while ensuring that at least one view remains observed for each sample. The same missing-view masks are used for all compared methods under each setting.

B.2 Baselines and Implementations

We compare TDPP with seven recent state-of-the-art incomplete multi-view clustering methods, including DIVIDE (Lu et al. (2024)), CANDY (Guo et al. (2024)), I2MVC (Wu et al. (2025)), RAC-DMVC (Dong et al. (2026a)), GAVIM (Chen et al. (2026a)), DIMVC-HIA (Du et al. (2026)), and CAMEL (Gu et al. (2026)).

For all comparison methods, we use the authors' publicly released implementations whenever available; when only official MATLAB code is provided, we implement a Python version following the released code. Hyperparameters follow the settings recommended in the original papers or official implementations, with limited dataset-specific adjustments when required by our evaluation protocol. All reported results are obtained from our own runs under the same experimental protocol, without modifying the core algorithms of the compared methods. Reproduced results may therefore differ slightly from those originally reported due to differences in implementation environments, datasets, missing-view protocols, or parameter configurations.

B.3 Training and Implementation Details

All input views are standardized before training. The teacher is trained on complete samples, and the student is initialized from the teacher and then trained on all samples. The main implementation settings are summarized in Table 3. For all datasets, the teacher and student are trained for 180 and 150 epochs using Adam with learning rate 5×10^{-4}, weight decay 10^{-4}, batch size 256, warm-up and cosine decay. We use $(\lambda_{\text{rec}}, \lambda_{\text{cross}}, \lambda_{\text{con}}, \lambda_{\text{align}}, \lambda_{\text{qfeat}}) = (0.1, 0.5, 0.5, 1.0, 1.0)$ and perform K-means on the concatenated normalized student features. All experiments are implemented in PyTorch 2.1.2 with CUDA 11.8 on an NVIDIA GeForce RTX 4060 GPU. For each experimental setting, training is performed with a fixed random seed, and the reported ACC, NMI, and ARI are obtained by averaging the corresponding results over the final three training epochs.

Table 3: Implementation settings of TDPP. Encoder denotes the view-specific encoder, the decoder is symmetric.

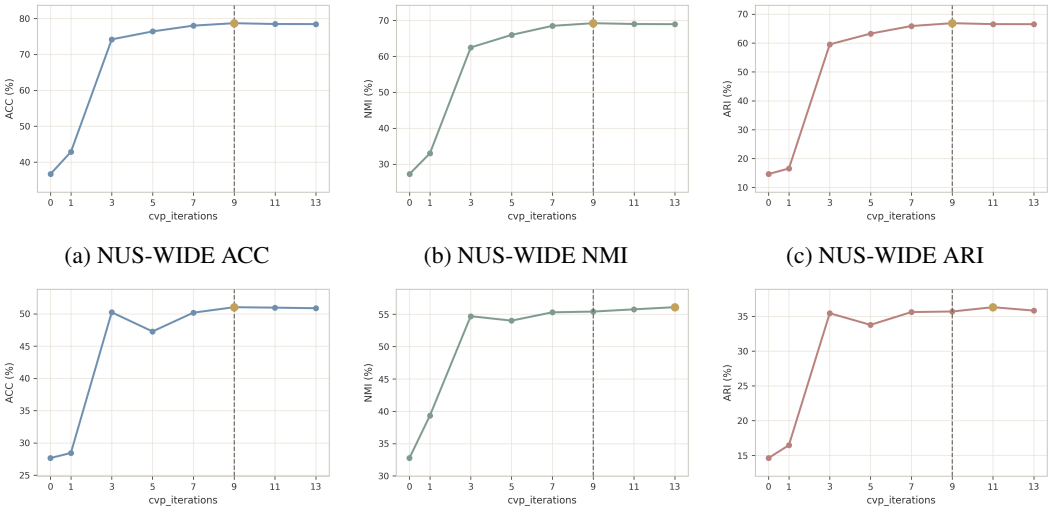
Dataset	**Encoder**	**Prompt dim.**	**MSA**	**CVPP iter.**	$\boldsymbol{\lambda}_{\text{feat}}$	$\boldsymbol{\lambda}_{\text{topo}}$
CUB	$[d_v, 2048, 2048, 2048, 128]$	64	2 layers, 8 heads	9	10	0.001
NUS-WIDE	$[d_v, 1024, 1024, 1024, 128]$	64	2 layers, 8 heads	9	10	5
Scene15	$[d_v, 1024, 1024, 1024, 128]$	64	2 layers, 8 heads	9	10	5
LandUse-21	$[d_v, 1024, 1024, 1024, 128]$	64	2 layers, 8 heads	9	10	0.001

C Additional Experimental Analyses

C.1 Parameter Sensitivity Analysis

Figures 4–6 show the sensitivity of TDPP on NUS-WIDE and Scene15 under the 50% missing ratio. These sweeps are used to examine trends rather than to replace the fixed settings used in the main comparison. When CVPP is disabled, all metrics drop clearly, confirming that prompt propagation is essential. Adding several iterations brings a substantial improvement, after which the curves become smoother as most cross-view information has already been exchanged. The trend is consistent across ACC, NMI, and ARI, indicating that CVPP improves both clustering accuracy and cluster agreement.

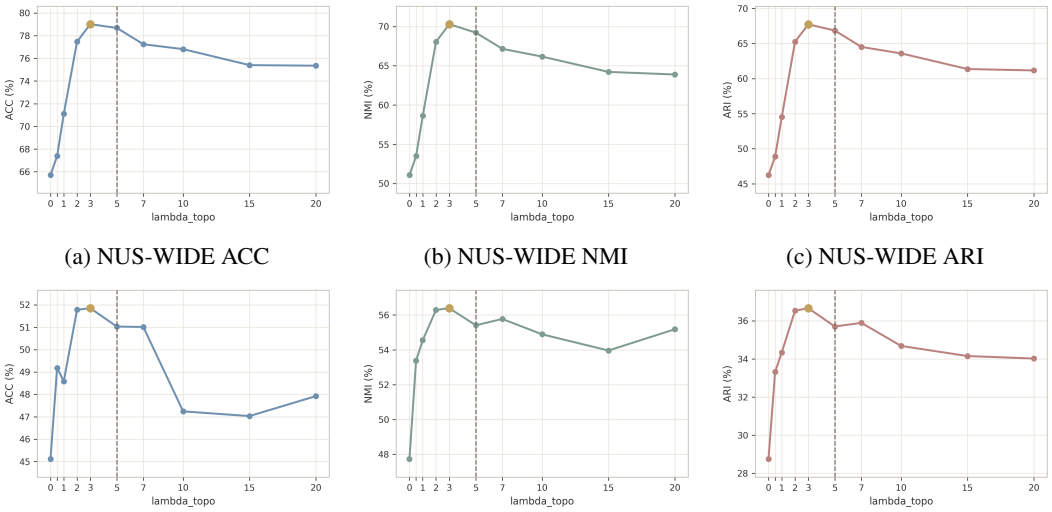
We then study the effect of the topology distillation weight λ_{topo}. Removing this term consistently weakens TDPP, while a non-zero weight improves the three metrics over a broad range. The curves



(a) NUS-WIDE ACC (b) NUS-WIDE NMI (c) NUS-WIDE ARI

(d) Scene15 ACC (e) Scene15 NMI (f) Scene15 ARI

Figure 4: Sensitivity to the number of CVPP iterations on NUS-WIDE and Scene15.



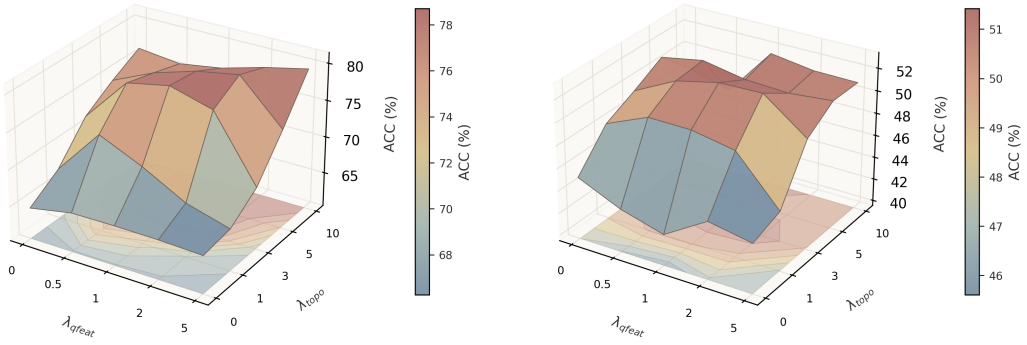
(a) NUS-WIDE ACC (b) NUS-WIDE NMI (c) NUS-WIDE ARI

(d) Scene15 ACC (e) Scene15 NMI (f) Scene15 ARI

Figure 5: Sensitivity to the topology distillation weight λ_{topo} on NUS-WIDE and Scene15.

do not collapse around a single narrow value, suggesting that sparse multi-hop topology distillation acts as a stable structural regularizer rather than a finely tuned correction.

The joint analysis of λ_{qfeat} and λ_{topo} further clarifies their interaction. Activating topology brings a clear gain, and the best regions appear when quality-aware feature distillation is used together with topology preservation. This supports the design choice of combining view reliability with global neighborhood structure, rather than relying on either signal in isolation.



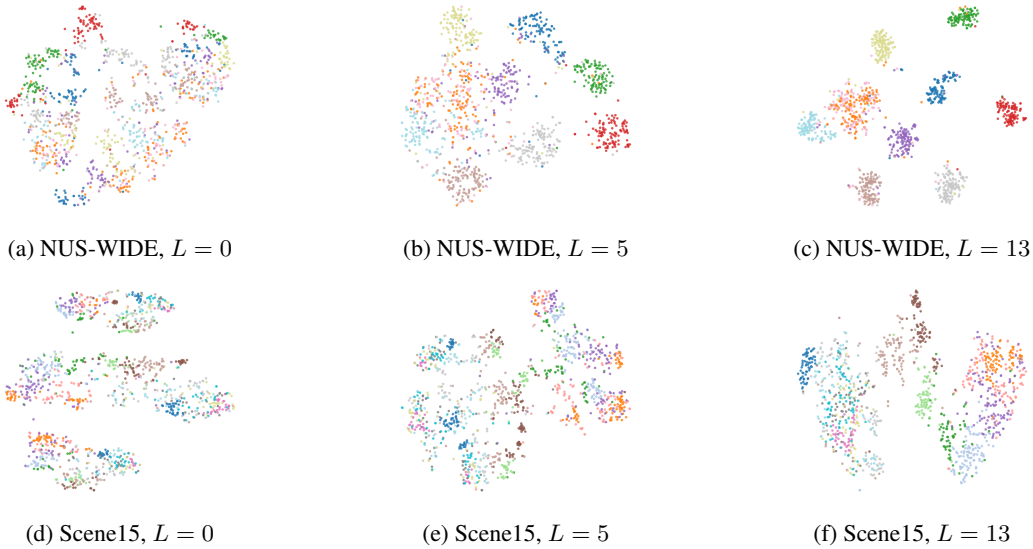
(a) NUS-WIDE (b) Scene15

Figure 6: Joint sensitivity of λ_{qfeat} and λ_{topo} in terms of ACC.

C.2 Additional Visualizations

C.2.1 Representation Evolution across CVPP Iterations

To further inspect the learned representations, we visualize TDPP on NUS-WIDE and Scene15 under the 50% missing ratio. Figure 7 shows the t-SNE embeddings of the student representations with different numbers of CVPP iterations. When CVPP is disabled, samples from the same class are often split across several regions, and different classes are mixed in local neighborhoods.



(a) NUS-WIDE, $L = 0$ (b) NUS-WIDE, $L = 5$ (c) NUS-WIDE, $L = 13$

(d) Scene15, $L = 0$ (e) Scene15, $L = 5$ (f) Scene15, $L = 13$

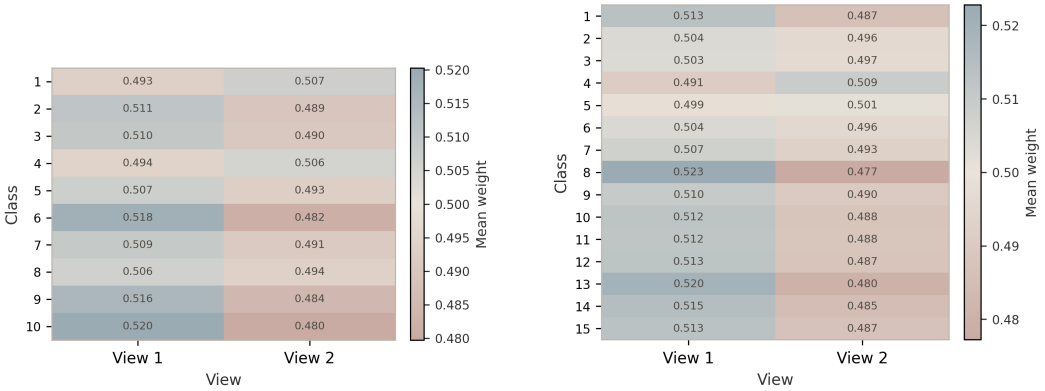
Figure 7: Latent representation evolution under CVPP propagation. Points are colored by ground-truth labels.

As prompt propagation proceeds, the embeddings become more organized. At $L = 13$, many classes form more compact regions with clearer separation, especially on NUS-WIDE, while Scene15 also

shows less fragmented semantic groups despite its more heterogeneous scene categories. This qualitative behavior is consistent with the quantitative results, where increasing CVPP iterations improves ACC, NMI, and ARI.

C.2.2 CLASS-WISE VIEW RELIABILITY

The quality maps in Fig. 8 provide complementary evidence for view-aware distillation. TDPP does not treat every view as equally reliable. The weights change across semantic classes, allowing feature supervision to emphasize views whose teacher representations better preserve local structure.



(a) NUS-WIDE class-view weights

(b) Scene15 class-view weights

Figure 8: Teacher-estimated class-wise view reliability.