UNCOVER UNDERLYING CORRESPONDENCE FOR ROBUST MULTI-VIEW CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-view clustering (MVC) aims to group unlabeled data into semantically meaningful clusters by leveraging cross-view consistency. However, real-world datasets collected from the web often suffer from noisy correspondence (NC), which breaks the consistency prior and results in unreliable alignments. In this paper, we identify two critical forms of NC that particularly harm clustering: i) category-level mismatch, where semantically consistent samples from the same class are mistakenly treated as negatives; and ii) sample-level mismatch, where collected cross-view pairs are misaligned and some samples may even lack any valid counterpart. To address these challenges, we propose a generative framework that formulates noisy correspondence learning in MVC as maximum likelihood estimation over underlying cross-view correspondences. The objective is elegantly solved via an Expectation–Maximization algorithm: in the E-step, soft correspondence distributions are inferred across views, capturing class-level relations while adaptively down-weighting noisy or unalignable samples through GMM-guided marginals; in the M-step, the embedding network is updated to maximize the expected log-likelihood. Extensive experiments on both synthetic and real-world noisy datasets demonstrate that our method significantly improves clustering robustness. The code will be released upon acceptance.

1 INTRODUCTION

Describing the same object from multiple perspectives (Yan et al., 2021) or modalities (Sharma et al., 2018), multi-view data have become increasingly prevalent in real-world applications. To exploit such data, contrastive multi-view clustering (MVC) has emerged as a powerful unsupervised paradigm. Relying on the consistency prior that views from the same instance should be semantically aligned, contrastive MVC pulls positive pairs (*i.e.*, views of the same instance) closer while pushing negative pairs (*i.e.*, views from different instances) apart in the embedding space. Through this process, it could learn a shared embedding space across views and group unlabeled samples into semantically meaningful clusters.

However, this prior is often difficult to satisfy. In practice, multi-view datasets are commonly constructed by crawling paired data from web, such as images with their associated alt text (Wang et al., 2015). This automatic process inevitably introduces the noisy correspondence (NC) problem (Huang et al., 2021) , where cross-view pairs are incorrectly matched. Such noise undermines the cross-view consistency prior and severely distorts the semantic structure of the learned embedding space.

In this paper, we identify two major types of NC that are particularly harmful to clustering: i) *Category-level mismatch*, where views from different modalities but belonging to the same class are mistakenly treated as negatives by contrastive MVC methods, despite their underlying semantic consistency; ii) *Sample-level mismatch*, which manifests in two scenarios: alignable mispairs, where a sample is wrongly paired with an mismatched view despite having a correct counterpart elsewhere; and unalignable samples, where no valid counterpart exists due to corruption, noise, or poor data quality. Such issues are especially prevalent in webly collected data, where the pairwise noise rate can exceed 20% (Sharma et al., 2018; Wang et al., 2015). Critically, manually verifying or cleaning these correspondence is prohibitively expensive, underscoring the need for robust multi-view clustering methods. To address NC, recent works mainly adopt either pairwise reweighting or

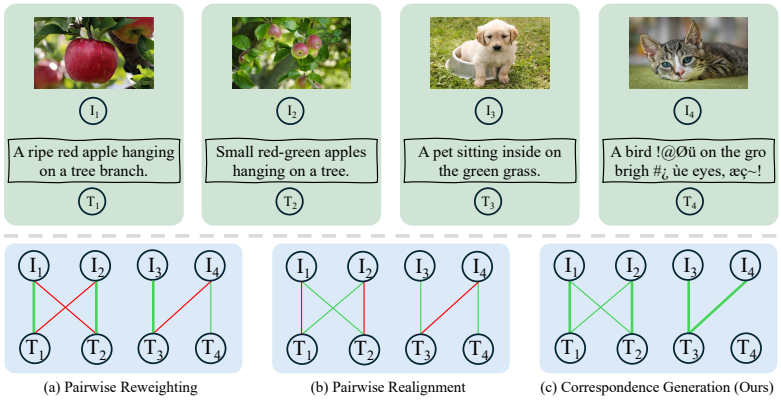


Figure 1: (*Top*) Examples of multi-view data, including noisy pairs I_4–T_4. (*Bottom*) Illustration of three paradigms for handling noisy correspondence, where green edges denote discovered correspondences and red edges indicate potential but undiscovered ones. (a) Pairwise reweighting, which applies robust contrastive losses to down-weight potentially noisy pairs during training but retains the original correspondences unchanged; (b) Pairwise realignment, which reassigns samples to more plausible cross-view counterpart; (c) Correspondence generation (Ours), which directly uncovers latent correspondences and filters out noise.

realignment strategies, as illustrated in Fig. 2. However, both approaches overlook category-level semantics and unalignable samples, leading to suboptimal results in clustering.

In this paper, we shift from existing discriminative contrastive objective to a generative one. Specifically, we formulate noisy correspondence learning in MVC as a maximum likelihood estimation problem for the underlying correspondence distribution, where the correspondences across views are modeled as unobserved latent variables. Unlike previous methods that focus on verifying whether given positive or negative pairs are correctly aligned, our formulation uncovers the underlying correspondences without heavily relying on pre-defined (potentially noisy) pairs. By maximizing the overall log-likelihood, we capture the semantic structure in a principled and probabilistic manner.

To effectively optimize the proposed objective, we develop an Expectation-Maximization (EM) algorithm CorreGen. In the E-step, the goal is to infer a latent correspondence distribution across views. We first estimate the marginal likelihood of each sample by fitting a Gaussian Mixture Model in the embedding space. Intuitively, this estimation assigns higher probabilities to samples that lie in large and coherent clusters, while noisy or unalignable samples receive lower probabilities. These marginals serve as constraints to solve an optimal transport formulation, yielding a soft many-to-many assignment that captures category-level relationships across views. In the M-step, the estimated correspondences are used to maximize the expected log-likelihood, updating the embedding network such that semantically consistent pairs are assigned higher probabilities. Iterating between the two steps gradually uncovers reliable correspondences and refines robust cluster representations. In summary, the contribution of our work can be summarized as follows:

- We propose a novel generative perspective for noisy correspondence learning in multi-view clustering, formulating latent cross-view alignments as a maximum likelihood problem solved via the EM algorithm.
- We introduce a principled E-step solution that jointly models category-level correspondences and suppresses sample-level noise by leveraging GMM-guided marginals. Extensive experiments on both synthetic and real-world noisy datasets validate the effectiveness of our approach. Notably, our method achieves 10% accuracy improvements on the challenging UMPC-Food101 dataset.

2 Related Work

Robust Multi-view Clustering aims to handle imperfections that commonly occur in real-world datasets. These imperfections can be broadly categorized into two types: i) *Incomplete Multi-view*

Problem (IMP) arises when some views are missing, resulting in incomplete cross-view information. Representative solutions adopt completion-based strategies, such as predictive learning (Lin et al., 2021), adversarial generation (Li et al., 2019), or diffusion models (Zhang et al., 2025), which aim to impute the missing views and recover complete multi-view representations; ii) *Partially-view aligned Problem* (PVP) occurs when the correspondences across views are misaligned. For example, in multi-camera surveillance, images of the same person from different cameras may be temporally asynchronous (Huang et al., 2020). To address this, recent works (Yang et al., 2022b; 2021) design robust realignment objectives that leverage manually curated correspondence information to realign the misaligned pairs.

Although both PVP and NC address erroneous cross-view correspondences, the NC problem studied in this paper differs in two significant aspects. First, misalignments in NC are unobserved, with no manually verified labels or alignment indicators available. Second, NC encompasses not only instance-level mismatches, but also category-level misalignments and even unalignable samples that lack valid counterparts across views.

Noisy Correspondence learning was first introduced in cross-modal retrieval (Huang et al., 2021), where mismatched image-text pairs are mistakenly treated as true positives. Recently, this problem has garnered increasing attention across a range of domains, including video reasoning (Lin et al., 2024), graph matching (Lin et al., 2023), person re-identification (Yang et al., 2022a) and multi-view clustering (Sun et al., 2024; 2025). Existing solutions can be broadly categorized into two groups: i) *Reweighting-based methods* (Yang et al., 2024) aim to reduce the impact of mismatched pairs by assigning them lower weights during training. For example, Huang et al. (2021) adjust the margins in triplet contrastive loss to account for false positives; ii) *Realignment-based methods* (Lin et al., 2024) attempt to reassign each sample to a more plausible counterpart across views, thereby mitigating alignment errors.

Although existing methods achieve promising results, they mainly refine given positive pairs while overlooking potential category-level correspondences, leading to suboptimal clustering performance. Different from these discriminative approaches, we propose a generative objective for noisy correspondence learning in MVC, which assigns higher likelihoods to semantically consistent samples and uncoverst latent correspondences. Notably, our optimization does not rely heavily on off-the-shelf pairs, thereby mitigating the noisy correspondence problem from a new perspective.

3 METHOD

In this section, we first introduce the problem setting and formalize correspondence learning in multi-view clustering (MVC) as a generative maximum likelihood estimation problem in Sec. 3.1. To optimize this objective, we propose **CorreGen**, an EM-based framework in Sec. 3.2, and detail its two steps in Sec. 3.2.1 and Sec. 3.2.2.

3.1 PROBLEM DEFINITION

Given a multi-view dataset $\{(\boldsymbol{x}_i^{(1)}, \ldots, \boldsymbol{x}_i^{(V)})\}_{i=1}^N$ with N instances observed from V views, the goal of MVC is to learn an encoder f_θ that maps each view $\boldsymbol{x}_i^{(v)}$ into a shared embedding space, *i.e.*, $\boldsymbol{z}_i^{(v)} = f_\theta(\boldsymbol{x}_i^{(v)})$. Ideally, the distribution of these embeddings should form C well-separated semantic clusters, such that traditional clustering algorithms (*e.g.*, K-means (McQueen, 1967)) can easily distinguish them.

To achieve this goal, recent contrastive MVC methods pull positive pairs (*i.e.*, views of the same instance) closer while pushing negative pairs (*i.e.*, views from different instances) apart in the embedding space. Formally, for any pair of views (v_1, v_2) with $v_1 \neq v_2$, the positive and negative sets are defined as

$$\mathcal{P}^+_{v_1,v_2} = \bigcup_{i=1}^{N} \{\, (\boldsymbol{x}_i^{(v_1)}, \boldsymbol{x}_i^{(v_2)}, t_{ii}^{12} = 1)\}, \quad \mathcal{P}^-_{v_1,v_2} = \bigcup_{i=1}^{N} \bigcup_{j=1, j\neq i}^{N} \{\, (\boldsymbol{x}_i^{(v_1)}, \boldsymbol{x}_j^{(v_2)}, t_{ij}^{12} = 0)\}, \quad (1)$$

where $t_{ij}^{12} \in \{0, 1\}$ is an indicator variable that equals 1 if $\boldsymbol{x}_i^{(v_1)}$ and $\boldsymbol{x}_j^{(v_2)}$ belong to the same instance, and 0 otherwise. Nevertheless, contrastive MVC essentially formulates an instance-level

discriminative task, which overlooks the intrinsic cluster structure of data. As a result, real-world multi-view datasets are particularly vulnerable to the *noisy correspondence* problem, where the assumed cross-view alignment fails to hold. For clarity, we formalize its two manifestations, namely *category-level mismatch* and *sample-level mismatch*, as defined below.

Definition 1 (Category-level mismatch)**.** *Consider a cross-view pair* $(\boldsymbol{x}_i^{(v_1)}, \boldsymbol{x}_j^{(v_2)}, t_{ij}^{12})$*, where* $t_{ij}^{12} \in \{0, 1\}$ *denotes whether the pair is treated as positive or negative. Let* $c_i^{(v_1)}$ *and* $c_j^{(v_2)}$ *be the oracle class labels of* $\boldsymbol{x}_i^{(v_1)}$ *and* $\boldsymbol{x}_j^{(v_2)}$*, respectively. A category-level mismatch occurs if* $c_i^{(v_1)} = c_j^{(v_2)}$ *but* $t_{ij}^{12} = 0$*, i.e., samples from the same semantic class are incorrectly assigned as a negative pair.*

In other words, category-level mismatch occurs when semantically related instances are mistakenly treated as negatives. Ideally, all cross-view pairs of samples from the same class should be regarded as positives with $t_{ij}^{12} = 1$, rather than only those from the same instance.

Definition 2 (Sample-level mismatch)**.** *Consider a cross-view pair* $(\boldsymbol{x}_i^{(v_1)}, \boldsymbol{x}_i^{(v_2)}, t_{ii}^{12})$*, where* $c_i^{(v_1)}$ *and* $c_i^{(v_2)}$ *denote the oracle class labels of* $\boldsymbol{x}_i^{(v_1)}$ *and* $\boldsymbol{x}_i^{(v_2)}$*, respectively. A sample-level mismatch occurs if either i)* $c_i^{(v_1)} \neq c_i^{(v_2)}$*, or ii) at least one of* $c_i^{(v_1)}$ *or* $c_i^{(v_2)}$ *does not correspond to any valid class. In both cases, the pair cannot be regarded as a valid positive correspondence.*

Specifically, sample-level mismatch admits two scenarios: i) *alignable mispaired*: although the constructed pair is incorrect, the sample $\boldsymbol{x}_i^{(v_1)}$ still has a valid counterpart $\boldsymbol{x}_k^{(v_2)}$ in the other view. This case often co-occurs with category-level mismatch; ii) *unalignable mispaired*: there is no valid counterpart exists, *e.g.*, the sample $\boldsymbol{x}_i^{(v_1)}$ might be corrupted or purely noisy data.

These two types of complex noisy correspondence motivate a more fundamental question: can we reduce the reliance on pre-defined pairs and instead directly model the intrinsic relationships that couple different views? Building on this intuition, we adopt a generative formulation that maximizes the marginal log-likelihood of the observed multi-view data.

$$\theta^* = \arg\max_{\theta} \sum_{v=1}^{V} \sum_{i=1}^{N} \log p(\boldsymbol{x}_i^{(v)}; \theta), \tag{2}$$

In multi-view clustering, each sample in one view may be associated with multiple counterparts in another view. Since these associations are unknown a prior, we treat them as latent variables. By aggregating over all unordered view pairs (v_i, v_j), the objective can be reformulated as:

$$\theta^* = \arg\max_{\theta} \sum_{v_1}^{V} \sum_{i}^{N} \sum_{v_2}^{V} \log \sum_{j}^{N} p(\boldsymbol{x}_i^{(v_1)}, \boldsymbol{x}_j^{(v_2)}; \theta). \tag{3}$$

Maximizing this marginal likelihood implicitly encourages the model to learn a meaningful joint distribution $p(\boldsymbol{x}_i^{(v_1)}, \boldsymbol{x}_j^{(v_2)}; \theta)$. In particular, to maximize the inner summation over j, the parameters θ must assign higher joint probability to semantically consistent pairs, thereby revealing the underlying cross-view correspondences in a probabilistic sense.

Compared with discriminative objectives, this generative formulation offers two key advantages: i) it alleviates the heavy reliance on pre-defined positive and negative pairs, making it naturally robust to sample-level unmatchable cases; ii) it captures many-to-many probabilistic correspondences across views, which better reflects the complex coupling of real-world multi-view data and mitigates category-level mismatch. However, the nested summation in Eq. (3) makes direct optimization intractable. To address this, we cast the objective into the Expectation–Maximization (EM) framework and present the theoretical derivation in the next section.

3.2 Correspondence Generation via Expectation–Maximization

To simplify the derivation of the joint log-likelihood defined in Eq. (3), we first consider a subset of the objective involving only two views:

$$\theta^* = \arg\max_{\theta} \sum_{i=1}^{N} \log \sum_{j=1}^{N} p(\boldsymbol{x}_i^{(v_1)}, \boldsymbol{x}_j^{(v_2)}; \theta). \tag{4}$$

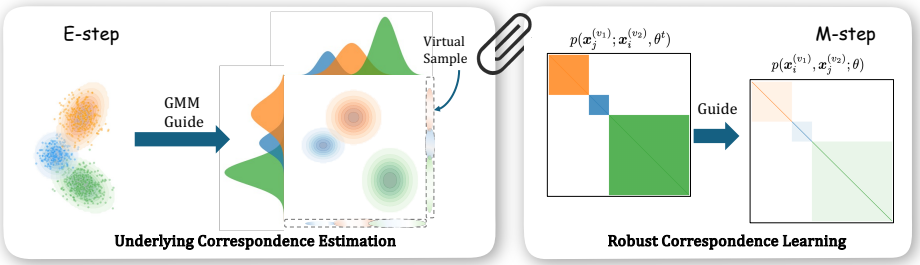


Figure 2: The framework of CorreGen.

Directly optimizing Eq. (4) is intractable due to the nested log-sum over latent variables. To address this, we introduce an auxiliary distribution $Q(\boldsymbol{x}_j^{(v_2)})$ over $\boldsymbol{x}_j^{(v_2)}$ such that $\sum_{j=1}^{N} Q(\boldsymbol{x}_j^{(v_2)}) = 1$. This allows us to derive a lower bound:

$$\sum_{i=1}^{N} \log \sum_{j=1}^{N} p(\boldsymbol{x}_i^{(v_1)}, \boldsymbol{x}_j^{(v_2)}; \theta) = \sum_{i=1}^{N} \log \sum_{j=1}^{N} Q(\boldsymbol{x}_j^{(v_2)}) \frac{p(\boldsymbol{x}_i^{(v_1)}, \boldsymbol{x}_j^{(v_2)}; \theta)}{Q(\boldsymbol{x}_j^{(v_2)})}, \tag{5}$$

$$\geq \sum_{i=1}^{N} \sum_{j=1}^{N} Q(\boldsymbol{x}_j^{(v_2)}) \log \frac{p(\boldsymbol{x}_i^{(v_1)}, \boldsymbol{x}_j^{(v_2)}; \theta)}{Q(\boldsymbol{x}_j^{(v_2)})}, \tag{6}$$

where the inequality follows from Jensen's inequality. The bound becomes tight when $Q(\boldsymbol{x}_j^{(v_2)}) = p(\boldsymbol{x}_j^{(v_2)}; \boldsymbol{x}_i^{(v_1)}, \theta)$, *i.e.*, when the auxiliary distribution matches the posterior under the current parameters θ^t. Substituting this choice of Q into the bound gives:

$$\theta^* = \arg\max_{\theta} \sum_{i}^{N} \sum_{j}^{N} Q(\boldsymbol{x}_j^{(v_2)}) \log p(\boldsymbol{x}_i^{(v_1)}, \boldsymbol{x}_j^{(v_2)}; \theta) - \sum_{i}^{N} \sum_{j}^{N} Q(\boldsymbol{x}_j^{(v_2)}) \log Q(\boldsymbol{x}_j^{(v_2)}) \tag{7}$$

$$= \arg\max_{\theta} \sum_{i}^{N} \sum_{j}^{N} p(\boldsymbol{x}_j^{(v_1)}; \boldsymbol{x}_i^{(v_2)}, \theta^{(t)}) \log p(\boldsymbol{x}_i^{(v_1)}, \boldsymbol{x}_j^{(v_2)}; \theta), \tag{8}$$

where the entropy term $-\sum_i^N \sum_j^N Q(\boldsymbol{x}_j^{(v_2)}) \log Q(\boldsymbol{x}_j^{(v_2)})$ is omitted since it is independent of θ. In the **E-step**, we estimate the posterior distribution $p(\boldsymbol{x}_j^{(v_2)}; \boldsymbol{x}_i^{(v_1)}, \theta^{(t)})$, which provides a soft assignment of correspondences between samples across views. In the **M-step**, we maximize the weighted log-likelihood in Eq. (8), updating the parameters θ guided by the correspondences inferred in the E-step. By aggregating over all views, above derivation naturally generalizes to multiple views.

3.2.1 E-STEP: ESTIMATING UNDERLYING CORRESPONDENCES

In the E-step, we estimate the posterior distribution of latent correspondences $p(\boldsymbol{x}_j^{(v_1)}; \boldsymbol{x}_i^{(v_2)}, \theta^{(t)})$ under the current parameters $\theta^{(t)}$:

$$p(\boldsymbol{x}_j^{(v_1)}; \boldsymbol{x}_i^{(v_2)}, \theta^{(t)}) = \frac{p(\boldsymbol{x}_i^{(v_1)}, \boldsymbol{x}_j^{(v_2)}; \theta^{(t)})}{p(\boldsymbol{x}_i^{(v_1)}; \theta^{(t)})}, \tag{9}$$

which naturally decomposes the estimation into two parts, namely, the marginal distribution of individual views and the joint distribution across views.

First, we estimate the joint distribution between views v_1 and v_2, represented as a matrix $\boldsymbol{P} \in \mathbb{R}_+^{N \times N}$ where each entry $\boldsymbol{P}_{ij} = p(\boldsymbol{x}_i^{(v_1)}, \boldsymbol{x}_j^{(v_2)}; \theta^{(t)})$. A good estimate of $\boldsymbol{P}$ should not only satisfy the marginal constraints but also capture the semantic dependency between the two views. To this end, we introduce a correlation function $s(\boldsymbol{z}_i^{(v_1)}, \boldsymbol{z}_j^{(v_2)})$ (e.g., cosine similarity) to measure the semantic correlations of a sample pair under the current parameters $\theta^{(t)}$, with $\boldsymbol{z}_i^{(v)} = f_{\theta^{(t)}}(\boldsymbol{x}_i^{(v)})$, and the

expected correlation is defined as

$$\mathbb{E}_{\boldsymbol{P}}[s] = \sum_{i=1}^{N}\sum_{j=1}^{N} \boldsymbol{P}_{ij}\, s(\boldsymbol{z}_i^{(v_1)}, \boldsymbol{z}_j^{(v_2)}). \tag{10}$$

We then seek the optimal joint distribution by maximizing this expectation:

$$\begin{aligned} \boldsymbol{P}^* &= \underset{\boldsymbol{P}\in\Pi(\boldsymbol{p}^{(v_1)},\boldsymbol{p}^{(v_2)})}{\arg\max}\ \mathbb{E}_{\boldsymbol{P}}[s] \\ \text{s.t}\quad \Pi(\boldsymbol{p}^{(v_1)},\boldsymbol{p}^{(v_2)}) &= \left\{ \boldsymbol{P}\in\mathbb{R}_+^{N\times N} \,\middle|\, \boldsymbol{P}\mathbf{1}_N = \boldsymbol{p}^{(v_1)}, \boldsymbol{P}^\top\mathbf{1}_N = \boldsymbol{p}^{(v_2)} \right\} \end{aligned} \tag{11}$$

This formulation ensures that the estimated joint distribution preserves the marginal constraints while assigning higher probability mass to semantically correlated pairs. However, due to the sample-level unalignable problem, there may exist outliers whose joint probabilities with all other samples should ideally be close to zero. To handle these outliers and obtain a more realistic joint distribution, we first introduce a virtual sample for each view to represent the outliers.

Virtual Sample for Partial Alignment. Let ρ denote the potential noise rate of the virtual samples, which corresponds to the marginal probability of the virtual sample. We then augment the joint distribution to $\tilde{\boldsymbol{P}} \in \mathbb{R}_+^{(N+1)\times(N+1)}$, ensuring that the total probability mass assigned to outliers equals ρ. Formally, $\tilde{\boldsymbol{P}}$ satisfies

$$\tilde{\boldsymbol{P}}\mathbf{1}_{N+1} = [\boldsymbol{p}^{(v_1)}; \rho], \quad \tilde{\boldsymbol{P}}^\top\mathbf{1}_{N+1} = [\boldsymbol{p}^{(v_2)}; \rho]. \tag{12}$$

This construction enables the model to absorb unalignable or noisy samples into the virtual mass. The final joint distribution is then obtained by dropping the last row and column of $\tilde{\boldsymbol{P}}$, *i.e*, $\boldsymbol{P}^* = \tilde{\boldsymbol{P}}_{1:N,1:N}$.

Recall from Eq. (9) that estimating the posterior requires both the joint distribution $p(\boldsymbol{x}_i^{(v_1)}, \boldsymbol{x}_j^{(v_2)}; \theta^{(t)})$ and the marginal distribution $p(\boldsymbol{x}_i^{(v_1)}; \theta^{(t)})$. In the expectation formulation Eq. (11), these marginals act as constraints on the feasible set of couplings $\Pi(\boldsymbol{p}^{(v_1)}, \boldsymbol{p}^{(v_2)})$, which essentially determines how many valid counterparts each sample can align with. Under category-level mismatch, the number of valid counterparts is not uniform but depends on the size and structure of its semantic class. Therefore, the marginal distribution should naturally reflect this variability: *samples from larger clusters or closer to cluster centers are assigned higher alignment mass, while outliers receive lower probabilities.*

GMM-guided Marginal Estimation. We assume that each sample is generated from a latent semantic cluster, which can be approximated by an anisotropic Gaussian distribution $\boldsymbol{x}_i^{(v)} \sim \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$. Accordingly, we fit the embedding space of each view with a Gaussian Mixture Model (GMM) and compute the posterior responsibility of each cluster for every sample. The marginal probability is then estimated as

$$p(\boldsymbol{x}_i^{(v)}; \theta^{(t)}) = \frac{m^{d_i} - 1}{m - 1} \cdot \frac{N_c}{N}, \tag{13}$$

$$d_i = \exp\left(-\epsilon \sqrt{(\boldsymbol{z}_i^{(v)} - \boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{z}_i^{(v)} - \boldsymbol{\mu}_c)} \right), \tag{14}$$

where N_c is the number of samples assigned to cluster c by GMM, ϵ and m are shaping parameters. Concretely, we first compute the Mahalanobis distance (Eq. (14) between each sample and its cluster center, and map the result through an exponential kernel to obtain an assignment confidence d_i. This confidence is further passed through a curve-shaping function $\frac{m^{d_i}-1}{m-1}$, which amplifies the contrast between high- and low-confidence samples: samples closer to the cluster center receive disproportionately higher weights, while distant ones are smoothly down-weighted rather than suppressed abruptly. Finally, the re-scaled confidence is combined with the cluster proportion N_c/N to yield the final probability to fill the marginal distribution in Eq. (11). In practice, we set $\epsilon = 0.1$ and $m = 10$, and apply a momentum update to stabilize training.

Proposition 1. *We can solve Eq. (11) by effecient scaling algorithm if adding an entropy regularization $-\lambda\mathcal{H}(\boldsymbol{P})$ to it, where λ is a regularization factor. Specifically, the optimal $\boldsymbol{P}^*$ can derived*

through the following iterations:

$$\begin{aligned} \boldsymbol{P}^* &= \mathit{Diag}(\boldsymbol{u}) \exp(\boldsymbol{S}/\lambda) \mathit{Diag}(\boldsymbol{v}), \\ \textit{with iteration update} \quad \boldsymbol{u} &\leftarrow \boldsymbol{p}^{(v_1)}/(\exp(\boldsymbol{S}/\lambda)\boldsymbol{v}), \quad \boldsymbol{v} \leftarrow \boldsymbol{p}^{(v_1)}/(\exp(\boldsymbol{S}^\top/\lambda)\boldsymbol{u}). \end{aligned} \tag{15}$$

where $\boldsymbol{u} \in \mathbb{R}_+^N$, $\boldsymbol{v} \in \mathbb{R}_+^N$ are two scaling vectors, and $\boldsymbol{S}$ denotes the correlation matrix where $\boldsymbol{S}_{ij} = s(\boldsymbol{z}_i^{(v_1)}, \boldsymbol{z}_j^{(v_2)})$. The proof is in Appendix A.

3.2.2 M-STEP: ROBUST CORRESPONDENCE LEARNING

In the M-step, we maximize the overall log-likelihood of the observed data based on the estimated posterior distribution. To make Eq. (8) tractable, we approximate the joint distribution $p(\boldsymbol{x}_i^{(v_1)}, \boldsymbol{x}_j^{(v_2)}; \theta)$ by normalizing the similarity scores of embeddings in the latent space

$$p(\boldsymbol{x}_i^{(v_1)}, \boldsymbol{x}_j^{(v_2)}; \theta) = \frac{\exp(s(\boldsymbol{z}_i^{(v_1)}, \boldsymbol{z}_j^{(v_2)})/\tau)}{\sum_{m=1}^N \sum_{n=1}^N \exp(s(\boldsymbol{z}_m^{(v_1)}, \boldsymbol{z}_n^{(v_2)})/\tau)} \tag{16}$$

where $\boldsymbol{z}_i^{(v)} = f_\theta(\boldsymbol{x}_i^{(v)})$ denotes the embedding of $\boldsymbol{x}_i^{(v)}$ and τ is a temperature parameter. Substituting this parameterization into Eq. (8), the M-step objective becomes

$$\theta^* = \arg\max_\theta \sum_{i=1}^N \sum_{j=1}^N \boldsymbol{Q}_{ij} \log \frac{\exp(s(\boldsymbol{z}_i^{(v_1)}, \boldsymbol{z}_j^{(v_2)})/\tau)}{\sum_{m=1}^N \sum_{n=1}^N \exp(s(\boldsymbol{z}_m^{(v_1)}, \boldsymbol{z}_n^{(v_2)})/\tau)}, \text{ where } \boldsymbol{Q}_{ij} = \frac{\boldsymbol{P}_{ij}^*}{\boldsymbol{p}_i^{(v_1)}}, \tag{17}$$

where $s(\cdot, \cdot)$ denotes a similarity function and $\boldsymbol{P}^*$ is the estimated joint distribution from the E-step. Unlike contrastive objectives that rely on manually defined positive/negative pairs, this formulation leverages the soft correspondences $\boldsymbol{P}^*$ inferred in the E-step, thereby mitigating the negative affects of noisy correspondence and enabling more robust representation learning. Importantly, we find that the widely used InfoNCE loss can be unified into our framework as a special case as stated below.

Proposition 2. *If the marginal distribution $p(\boldsymbol{x}_i^{(v)}; \theta)$ is uniform and the posterior probability degenerates to $p(\boldsymbol{x}_i^{(v_2)}; \boldsymbol{x}_i^{(v_1)}, \theta) = 1$ (i.e., only paired cross-view samples are treated as valid positives), then Eq. (8) reduces to the standard InfoNCE contrastive objective:*

$$\theta^* = \arg\max_\theta \sum_i^N \log \frac{\exp(s(\boldsymbol{z}_i^{(v_1)}, \boldsymbol{z}_j^{(v_2)})/\tau)}{\sum_{n=1}^N \exp(s(\boldsymbol{z}_i^{(v_1)}, \boldsymbol{z}_n^{(v_2)})/\tau)} \tag{18}$$

The proof is in Appendix B,

4 EXPERIMENTS

In this section, we conduct extensive experiments to evaluate the effectiveness of our method in addressing both category-level and sample-level noisy correspondence. Our study is guided by the following research questions: **Q1**: Does our method outperform existing robust MVC approaches under noisy correspondence (Section 4.2)? **Q2**: Can our method reliably uncover underlying category-level correspondences across views (Section 4.3)? **Q3**: How does performance vary under different levels of mismatch (Appendix D)? **Q4**: How sensitive is our method to hyperparameter choices (Appendix E)? **Q5**: Are the proposed components crucial for the improvements (Appendix F)?

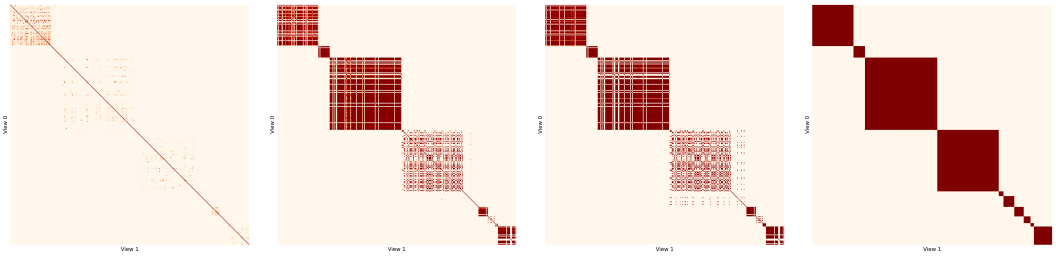
4.1 EXPERIMENTAL SETUP

Datasets. We evaluate our method on four widely used datasets: Scene15 (Fei-Fei & Perona, 2005), Caltech101 (Li et al., 2015), LandUse21 (Yang & Newsam, 2010), and UMPC-Food101 (Wang et al., 2015). Notably, UMPC-Food101 contains images from 101 food categories paired with recipes crawled from the web, which inevitably introduces substantial irrelevant or noisy information. Representative examples of such noisy image–text pairs are provided in Appendix G.

Baselines. We compare CorreGen against seven state-of-the-art MVC methods, including DCP (Lin et al., 2022), SURE (Yang et al., 2022b), GCFAgg (Yan et al., 2023), CGCN (Wang et al., 2024),

Table 1: The clustering performance with different mismatch ratio (MR). The best results and second best result are marked in **bold** and underline. All the results are the mean of five individual runs.

MR Ratio	Method	Scene15			LandUse21			Caltech101			UMPC-Food101		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
0%	DCP	40.16	42.71	23.00	24.20	30.88	11.70	51.91	74.91	47.57	16.33	36.56	7.50
	SURE	43.41	44.33	25.71	23.14	29.20	10.62	38.94	65.64	27.28	29.86	46.37	19.22
	GCFAgg	33.58	32.91	16.76	23.48	26.75	10.80	34.27	55.57	19.98	16.12	30.03	6.55
	CGCN	41.34	40.09	24.64	23.57	26.88	10.40	36.40	66.72	24.72	29.58	39.57	14.69
	DIVIDE	44.57	45.98	28.43	32.50	39.44	18.16	62.20	83.30	50.50	36.20	57.92	27.72
	CANDY	42.55	41.67	25.41	30.94	36.33	16.20	67.64	84.06	60.02	33.10	53.06	22.39
	ROLL	47.61	48.71	30.86	29.43	33.78	15.24	17.83	42.75	13.43	23.65	47.22	16.43
	Ours	**50.25**	**48.92**	**32.87**	**32.87**	**39.52**	**18.54**	**68.52**	**84.45**	**63.45**	**49.77**	**58.36**	**35.73**
20%	DCP	35.88	37.63	16.51	24.20	28.46	10.10	43.99	70.83	35.43	17.83	35.63	8.45
	SURE	37.26	35.56	19.94	24.67	27.45	10.91	35.91	60.06	24.56	20.30	32.89	8.99
	GCFAgg	33.11	27.64	15.29	23.86	23.30	9.11	28.90	47.47	13.81	11.28	19.48	2.94
	CGCN	35.96	35.73	20.10	24.52	26.38	10.36	33.01	64.17	24.41	28.01	38.36	13.63
	DIVIDE	41.91	40.16	24.84	30.89	35.93	16.21	55.65	70.72	50.92	31.41	51.21	22.70
	CANDY	41.05	40.41	24.44	30.54	35.45	15.99	65.79	82.29	60.03	30.41	50.36	20.36
	ROLL	44.86	46.96	28.71	29.33	33.11	15.16	20.39	46.44	15.03	21.26	43.05	13.73
	Ours	**48.04**	**47.36**	**30.75**	**32.26**	**38.76**	**17.83**	**68.01**	**84.23**	**62.78**	**46.76**	**55.22**	**32.46**
50%	DCP	25.28	25.24	5.78	24.01	26.95	8.37	41.52	69.35	29.59	13.36	24.04	4.60
	SURE	28.16	26.52	13.16	22.67	24.91	9.94	26.89	52.51	18.73	11.06	21.51	3.20
	GCFAgg	21.07	11.26	5.14	24.48	22.56	8.92	22.16	36.65	8.89	6.70	11.02	0.80
	CGCN	35.99	33.07	19.47	20.62	23.35	7.83	37.74	65.66	28.20	20.71	31.44	8.51
	DIVIDE	39.67	36.47	22.69	29.75	33.17	15.23	38.81	59.18	33.03	25.21	44.47	16.00
	CANDY	41.25	39.02	23.93	29.09	32.56	14.77	60.30	78.60	55.16	28.80	48.69	19.03
	ROLL	42.41	44.49	26.43	28.65	32.81	15.01	18.57	43.50	13.68	20.97	38.54	11.89
	Ours	**45.07**	**44.97**	**27.87**	**32.03**	**37.98**	**17.84**	**66.60**	**83.61**	**62.38**	**42.57**	**51.79**	**27.29**
80%	DCP	21.46	21.15	2.87	21.17	22.59	7.17	32.13	58.16	20.78	32.17	46.60	21.90
	SURE	24.57	23.68	9.90	17.57	19.61	5.94	23.61	49.01	15.97	8.81	18.32	2.19
	GCFAgg	11.53	3.08	0.90	17.38	15.17	4.44	16.61	32.57	5.78	3.58	6.90	0.14
	CGCN	28.81	25.42	12.89	20.29	20.70	7.32	35.32	63.83	25.77	18.13	29.48	6.92
	DIVIDE	35.90	32.95	19.63	28.56	31.74	14.32	27.42	53.68	21.56	24.78	42.98	15.63
	CANDY	38.27	36.08	20.74	28.44	31.39	14.01	54.17	77.30	53.79	27.59	48.10	17.62
	ROLL	37.62	38.27	21.19	25.67	28.42	11.96	20.83	45.58	13.97	31.39	40.96	15.69
	Ours	**40.96**	**41.74**	**24.74**	**31.52**	**37.21**	**17.75**	**64.74**	**82.77**	**61.78**	**43.00**	**53.03**	**27.12**



(a) Warmup (10 epoch) (b) 100 epoch (c) 200 epoch (d) Ground Truth

Figure 3: Estimated posterior distributions over the course of training on the Caltech101 dataset.

DIVIDE (Lu et al., 2024), CANDY (Guo et al., 2024), and ROLL (Sun et al., 2025). For fair comparison, we apply a view realignment strategy to the learned representations following prior studies (Guo et al., 2024; Sun et al., 2025), where realignment is consistently performed within batches of 512 to ensure fair evaluation.

Implementation Details. CorreGen introduces a generative objective for MVC that can be seamlessly integrated into existing contrastive frameworks. We implement it on top of DIVIDE (Lu et al., 2024) as the base model. More details are provided in Appendix C.

4.2 Performance Comparision (Q1)

Since MVC is an unsupervised task, category-level correspondences depend on the underlying class sizes and distributions, making category-level mismatch an intrinsic challenge rather than one that can be explicitly specified. Therefore, in this section, we focus on evaluating model performance under different *sample-level mismatch* setting, which includes two cases: i) *alignable mispairs*, caused by instance-level permutations across views; and ii) *unalignable mispairs*, caused by noisy or corrupted samples. We control these two factors using the Mismatch Rate (MR) and Corruption Rate (CR), with detailed construction described in Appendix C.

Table 2: The clustering performance on four multi-view datasets with different Mismatch Rate (MR) and Corruption Rate (CR).

Setting	Method	Scene15			LandUse21			Caltech101			UMPC-Food101		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
MR 0.2 CR 0.2	DCP	36.50	40.52	21.55	24.62	29.19	11.37	43.03	69.34	37.81	12.97	28.99	4.71
	SURE	37.93	38.53	21.23	24.48	28.32	11.02	33.71	58.99	20.69	13.14	25.66	4.95
	GCFAgg	29.59	26.33	14.22	24.29	25.13	10.70	28.57	45.65	14.21	8.89	17.07	2.11
	CGCN	27.78	26.95	12.92	23.52	23.96	8.81	35.61	64.81	30.16	28.02	39.04	13.57
	DIVIDE	36.05	36.18	20.22	29.30	34.69	15.13	56.13	73.31	53.82	29.01	49.69	20.92
	CANDY	35.57	37.00	20.71	29.13	33.70	14.87	65.80	82.23	62.52	30.13	49.77	20.06
	ROLL	36.13	36.76	17.99	23.15	24.28	8.39	16.50	40.44	12.16	18.51	39.78	11.63
	Ours	**41.23**	**41.43**	**25.05**	**31.13**	**37.36**	**17.00**	**67.12**	**84.45**	**64.13**	**45.97**	**54.66**	**31.36**
MR 0.2 CR 0.5	DCP	34.31	**37.70**	19.55	17.95	22.13	5.96	36.98	63.14	32.46	7.36	17.71	1.58
	SURE	34.05	35.32	18.37	20.05	23.20	7.40	32.18	58.49	20.47	11.19	25.69	4.19
	GCFAgg	27.85	24.05	12.73	23.24	24.19	9.92	27.57	45.00	14.43	7.77	15.68	1.67
	CGCN	28.36	31.46	16.32	22.24	25.04	9.61	35.83	76.99	41.69	24.07	35.01	10.17
	DIVIDE	33.54	35.40	19.90	27.94	31.81	13.75	57.87	76.59	**58.56**	24.92	46.78	17.61
	CANDY	31.24	34.08	19.00	24.72	28.03	11.27	**62.57**	81.52	55.76	25.00	47.27	17.36
	ROLL	27.03	25.83	9.42	16.40	15.49	3.20	12.97	36.57	9.80	16.12	36.52	9.66
	Ours	**36.48**	37.66	**21.14**	**28.50**	**33.09**	**14.31**	61.19	**82.15**	49.65	**43.54**	**53.66**	**29.07**
MR 0.5 CR 0.2	DCP	33.62	35.05	14.48	24.48	27.57	10.35	38.03	64.81	30.53	9.30	19.71	2.44
	SURE	25.37	26.07	11.48	21.38	24.14	8.08	27.52	53.57	15.64	6.86	15.83	1.58
	GCFAgg	24.26	13.31	6.45	22.00	19.02	7.77	23.83	38.62	10.43	5.24	9.64	0.57
	CGCN	29.65	29.89	15.37	23.57	24.86	9.08	29.22	58.19	26.19	25.08	35.71	11.60
	DIVIDE	32.88	32.87	18.08	29.00	32.49	14.37	43.98	61.51	37.87	23.04	43.28	14.71
	CANDY	34.60	35.31	19.84	27.77	31.46	13.63	58.35	78.55	56.14	27.97	48.24	18.81
	ROLL	35.23	35.79	18.54	23.34	23.99	8.83	14.78	38.46	11.07	17.54	35.48	9.67
	Ours	**39.54**	**39.55**	**23.12**	**31.20**	**36.25**	**16.92**	**66.87**	**84.15**	**67.31**	**38.84**	**50.09**	**24.98**
MR 0.5 CR 0.5	DCP	26.35	31.84	13.42	18.52	23.32	7.40	32.34	58.43	21.55	5.19	10.86	0.54
	SURE	26.91	28.73	12.06	19.57	21.18	6.60	25.90	54.83	18.07	7.00	17.28	1.77
	GCFAgg	22.27	14.13	6.68	20.57	17.30	6.72	21.56	37.88	9.61	4.61	8.88	0.42
	CGCN	27.27	30.11	14.68	19.67	22.51	7.38	33.15	59.86	24.95	20.74	32.53	8.41
	DIVIDE	30.27	31.25	16.31	26.13	29.12	12.30	48.07	68.23	44.69	20.67	42.07	12.52
	CANDY	29.44	32.67	17.09	24.08	27.21	11.01	51.28	75.16	41.70	24.70	46.58	17.19
	ROLL	26.29	24.98	9.41	14.62	13.00	2.19	13.82	36.54	10.30	14.76	32.84	7.71
	Ours	**36.19**	**36.84**	**20.83**	**28.72**	**32.54**	**14.50**	**57.06**	**80.34**	**45.37**	**37.26**	**49.30**	**23.25**

Table 1 reports results under different MR. Our method consistently achieves the best performance, benefiting from its generative objective and robust correspondence discovery, which remain effective even with few aligned pairs. Table 2 further evaluates scenarios with both alignable and unalignable mismatches. While all baselines degrade severely as MR and CR increase, our method maintains strong performance by jointly leveraging GMM-based marginals to down-weight noisy samples and virtual samples to absorb unalignable ones, mitigating the influence of low-quality pairs.

4.3 Posterior distribution Visualization (Q2)

We next investigate whether CorreGen can uncover the latent correspondences across views. On Caltech101 with MR=0.2 and CR=0.0, we sample a mini-batch and estimate their posterior distributions at different training stages, comparing them with the true category-level ground truth.

As shown in Figure 3, the category-level correlations are weak in the early training phase. By mid training, the estimated posterior distributions already resemble the ground truth, and the gap further narrows in the later stages. These results demonstrate that CorreGen progressively uncovers the latent class-level correspondences, thereby effectively alleviating category-level mismatches.

5 Conclusion

In this paper, we propose a novel generative framework for multi-view clustering under the noisy correspondence challenge. Unlike existing discriminative approaches that rely heavily on off-the-shelf pairwise alignments, our method models cross-view dependencies by maximizing the joint likelihood of observed data, thereby uncovering latent correspondences in a principled manner. Extensive experiments across multiple datasets demonstrate that our approach not only achieves superior clustering performance but also exhibits strong robustness to sample-level and category-level mismatches. In the future, we plan to extend this framework to unpaired multi-modal learning and apply it to cross-modal retrieval tasks with large-scale noisy data.

REFERENCES

Laetitia Chapel, Mokhtar Z Alaya, and Gilles Gasso. Partial optimal tranport with applications on positive-unlabeled learning. *Advances in Neural Information Processing Systems*, 33:2903–2913, 2020.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

Ruiming Guo, Mouxing Yang, Yijie Lin, Xi Peng, and Peng Hu. Robust contrastive multi-view clustering against dual noisy correspondence. 37, 2024.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=OOsR8BzCnl5.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Zhenyu Huang, Peng Hu, Joey Tianyi Zhou, Jiancheng Lv, and Xi Peng. Partially view-aligned clustering. In *Proceedings of the 34th Conference on Neural Information Processing Systems , NeurIPS'2020*, 2020.

Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 29406–29419. Curran Associates, Inc., 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Yeqing Li, Feiping Nie, Heng Huang, and Junzhou Huang. Large-scale multi-view spectral clustering via bipartite graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Zhaoyang Li, Qianqian Wang, Zhiqiang Tao, Quanxue Gao, Zhaohua Yang, et al. Deep adversarial multi-view clustering network. In *IJCAI*, volume 2, pp. 4, 2019.

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

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

Yijie Lin, Mouxing Yang, Jun Yu, Peng Hu, Changqing Zhang, and Xi Peng. Graph matching with bi-level noisy correspondence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 23362–23371, 2023.

Yijie Lin, Jie Zhang, Zhenyu Huang, Jia Liu, Zujie Wen, and Xi Peng. Multi-granularity correspondence learning from long-term noisy videos. *arXiv preprint arXiv:2401.16702*, 2024.

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.

James B McQueen. Some methods of classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.*, pp. 281–297, 1967.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Yuan Sun, Yang Qin, Yongxiang Li, Dezhong Peng, Xi Peng, and Peng Hu. Robust multi-view clustering with noisy correspondence. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Yuan Sun, Yongxiang Li, Zhenwen Ren, Guiduo Duan, Dezhong Peng, and Peng Hu. Roll: Robust noisy pseudo-label learning for multi-view clustering with noisy correspondence. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30732–30741, 2025.

Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frédéric Precioso. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6, 2015. doi: 10.1109/ICMEW.2015.7169757.

Yiming Wang, Dongxia Chang, Zhiqiang Fu, Jie Wen, and Yao Zhao. Partially view-aligned representation learning via cross-view graph contrastive network. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):7272–7283, 2024.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision, 2020.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19863–19872, 2023.

Xiaoqiang Yan, Shizhe Hu, Yiqiao Mao, Yangdong Ye, and Hui Yu. Deep multi-view learning methods: A review. *Neurocomputing*, 448:106–129, 2021. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2021.03.090. URL https://www.sciencedirect.com/science/article/pii/S0925231221004768.

Mouxing Yang, Yunfan Li, Zhenyu Huang, Zitao Liu, Peng Hu, and Xi Peng. Partially view-aligned representation learning with noise-robust contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1134–1143, June 2021.

Mouxing Yang, Zhenyu Huang, Peng Hu, Taihao Li, Jiancheng Lv, and Xi Peng. Learning with twin noisy labels for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14308–14317, 2022a.

Mouxing Yang, Yunfan Li, Peng Hu, Jinfeng Bai, Jian Cheng Lv, and Xi Peng. Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022b.

Mouxing Yang, Zhenyu Huang, and Xi Peng. Robust object re-identification with coupled noisy labels. *International Journal of Computer Vision*, 132(7):2511–2529, 2024.

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

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

APPENDIX

A EFFECIENT SOLVER FOR JOINT DISTRIBUTION ESTIMATION (PROOF OF PROPOSITION 1)

In this section, we briefly prove the efficiently solver for Eq. (11). With the entropy regularization item $\mathcal{H}(\boldsymbol{P})$, the objective function can be formulated as:

$$\arg\max_{\boldsymbol{P}\geq 0} \langle \boldsymbol{P}, \boldsymbol{S}\rangle - \lambda\mathcal{H}(\boldsymbol{P}) \quad \text{s.t.} \quad \boldsymbol{P}\mathbf{1} = \boldsymbol{p}^{(v_1)},\ \boldsymbol{P}^\top\mathbf{1} = \boldsymbol{p}^{(v_1)}, \tag{19}$$

where $\mathcal{H}(\boldsymbol{P}) = -\sum_{i,j}\boldsymbol{P}_{ij}\log\boldsymbol{P}_{ij}$ and $\lambda > 0$. Compared with the original formulation, Eq. (19) is both convex and smooth. To proceed, we introduce the Lagrangian together with dual multipliers $\alpha \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^m$ for the row and column constraints, respectively.

$$\mathcal{L}(\boldsymbol{P}, \alpha, \beta) = \langle \boldsymbol{P}, \boldsymbol{S}\rangle - \lambda\sum_{i,j}\boldsymbol{P}_{ij}\log\boldsymbol{P}_{ij} + \alpha^\top\big(\boldsymbol{p}^{(v_1)} - \boldsymbol{P}\mathbf{1}\big) + \beta^\top\big(\boldsymbol{p}^{(v_2)} - \boldsymbol{P}^\top\mathbf{1}\big). \tag{20}$$

Take first-order optimality w.r.t. $\boldsymbol{P}_{ij}$. For any i, j we have

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{P}_{ij}} = \boldsymbol{S}_{ij} - \lambda(1 + \log\boldsymbol{P}_{ij}) - \alpha_i - \beta_j = 0. \tag{21}$$

Rearranging gives

$$\log\boldsymbol{P}_{ij} = \frac{\boldsymbol{S}_{ij} - \alpha_i - \beta_j}{\lambda} - 1. \tag{22}$$

The constants 1 can then be re-parameterized by incorporating them into the exponentials of the dual variables, resulting in a multiplicative scaling form of the solution:

$$\boldsymbol{P}_{ij} = \boldsymbol{u}_i\exp(\boldsymbol{S}_{ij}/\lambda)\boldsymbol{v}_j, \tag{23}$$

where $\boldsymbol{u}_i := \exp(-\alpha_i/\lambda)$ and $\boldsymbol{v}_j := \exp(-\beta_j/\lambda)$ are strictly positive. In matrix form

$$\boldsymbol{P} = \mathrm{Diag}(\boldsymbol{u})\ \exp(\boldsymbol{S}/\lambda)\ \mathrm{Diag}(\boldsymbol{v}). \tag{24}$$

Impose the marginal constraints $\boldsymbol{P}\mathbf{1} = \boldsymbol{p}^{(v_1)}$ and $\boldsymbol{P}^\top\mathbf{1} = \boldsymbol{p}^{(v_2)}$ yields

$$\mathrm{Diag}(\boldsymbol{u})\,\big(\exp(\boldsymbol{S}/\lambda)\,\boldsymbol{v}\big) = \boldsymbol{p}^{(v_1)}, \qquad \mathrm{Diag}(\boldsymbol{v})\,\big(\exp(\boldsymbol{S}^\top/\lambda)\,\boldsymbol{u}\big) = \boldsymbol{p}^{(v_2)}. \tag{25}$$

Solving these equations componentwise leads to the alternating updates (Cuturi, 2013):

$$\boldsymbol{u} \leftarrow \boldsymbol{p}^{(v_1)} \,/\, \big(\exp(\boldsymbol{S}/\lambda)\,\boldsymbol{v}\big), \qquad \boldsymbol{v} \leftarrow \boldsymbol{p}^{(v_2)} \,/\, \big(\exp(\boldsymbol{S}^\top/\lambda)\,\boldsymbol{u}\big). \tag{26}$$

Since $\exp(\boldsymbol{S}/\lambda)_{ij} > 0$ for finite $\boldsymbol{S}$ and $\lambda > 0$, the kernel is strictly positive. By the Sinkhorn–Knopp theorem the alternating row/column scaling converges to positive vectors $(\boldsymbol{u}, \boldsymbol{v})$ that enforce the prescribed marginals. The resulting $\boldsymbol{P}^*$ is therefore the unique maximizer of the entropy-regularized problem.

In the case of our extended marginals, we introduce a virtual node with mass ρ and extend the transport plan to $\tilde{\boldsymbol{P}} \in \mathbb{R}^{(N+1)\times(N+1)}$ such that

$$\tilde{\boldsymbol{P}}\mathbf{1}_{N+1} = [\boldsymbol{p}^{(v_1)}; \rho], \quad \tilde{\boldsymbol{P}}^\top\mathbf{1}_{N+1} = [\boldsymbol{p}^{(v_2)}; \rho]. \tag{27}$$

Following Chapel et al. (2020), we only need to extend the correlation matrix as:

$$\tilde{\boldsymbol{S}} = \begin{bmatrix} \boldsymbol{S} & \mathbf{0}_{N\times 1} \\ \mathbf{0}_{1\times N} & A \end{bmatrix} \tag{28}$$

where A is a constant chosen such that $A > \max(\boldsymbol{S}_{ij})$. Therefore, the above alternating scaling algorithm can then be directly applied to $\tilde{\boldsymbol{P}}$ to efficiently compute the entropy-regularized solution under extended marginal.

B Contrastive Learning as a Special Case of CorrGen (Proof of Proposition 2)

Starting from our generative objective in Eq. (8):

$$\theta^* = \arg\max_{\theta} \sum_{i=1}^{N} \sum_{j=1}^{N} p(\boldsymbol{x}_j^{(v_2)}; \boldsymbol{x}_i^{(v_1)}, \theta^t) \log p(\boldsymbol{x}_i^{(v_1)}, \boldsymbol{x}_j^{(v_2)}; \theta). \tag{29}$$

Under the assumption that the posterior collapses to $p(\boldsymbol{x}_i^{(v_1)}; \boldsymbol{x}_i^{(v_2)}, \theta) = 1$, the summation over j reduces to

$$\theta^* = \arg\max_{\theta} \sum_{i=1}^{N} \log p(\boldsymbol{x}_i^{(v_1)}, \boldsymbol{x}_i^{(v_2)}; \theta). \tag{30}$$

Further decomposing the joint probability gives

$$p(\boldsymbol{x}_i^{(v_1)}, \boldsymbol{x}_i^{(v_2)}; \theta) = p(\boldsymbol{x}_i^{(v_2)}; \boldsymbol{x}_i^{(v_1)}, \theta)\, p(\boldsymbol{x}_i^{(v_1)}; \theta). \tag{31}$$

If the marginal $p(\boldsymbol{x}_i^{(v_1)}; \theta)$ is uniform, *i.e.*, $p(\boldsymbol{x}_i^{(v_1)}; \theta) = \frac{1}{N}$, it contributes only a constant independent of θ, which can be omitted. Thus, the objective simplifies to

$$\theta^* = \arg\max_{\theta} \sum_{i=1}^{N} \log p(\boldsymbol{x}_i^{(v_2)}; \boldsymbol{x}_i^{(v_1)}, \theta), \tag{32}$$

After parameterizing the conditional probability with similarity in the embedding space, yields exactly the InfoNCE objective (He et al., 2020):

$$\theta^* = \arg\max_{\theta} \sum_{i=1}^{N} \log \frac{\exp(s(\boldsymbol{z}_i^{(v_1)}, \boldsymbol{z}_i^{(v_2)})/\tau)}{\sum_{n=1}^{N} \exp(s(\boldsymbol{z}_i^{(v_1)}, \boldsymbol{z}_n^{(v_2)})/\tau)}. \tag{33}$$

C Implementation Details

Implementation of CorreGen. CorreGen is implemented on top of DIVIDE (Lu et al., 2024). Specifically, we replace the original contrastive objective in DIVIDE with our generative objective, while retaining its feature extraction structure as the mapping function f_θ. For the within-view contrastive module (*i.e.*, between features and their momentum counterparts), we fuse the estimated posterior matrix $\boldsymbol{Q}$ with the identity matrix $\boldsymbol{I}$ at a ratio of $\beta = 0.5$. For the cross-view learning module, we directly use the estimated posterior matrix without modification. To ensure stable training, we initialize the EM algorithm with the identity matrix $\boldsymbol{I}$ as the posterior estimate in the first few iterations, which serves as a warm start to avoid poor local optima. After this warmup phase, we switch to the adaptive posterior estimation strategy described in our method, thereby uncovering latent correspondences across views.

Training Setup. We implement CorreGen with PyTorch 2.1.2 and optimize it using Adam optimizer (Kingma & Ba, 2014) with the learning rate of 0.002. The batch size is set to 512 for smaller datasets (*e.g.*, Scene15, LandUse21) and 1024 for larger ones (*e.g.*, Caltech101, UMPC-Food101). All experiments are conducted on Ubuntu 20.04 with NVIDIA 3090 GPUs. We set the maximum warmup phase to 50 epochs and train for a total of 200 epochs. The regularization parameter $\lambda = 0.03$, and the noise rate for the virtual sample in Eq. (12) is set to $\rho = 0.2$ across all experiments.

Datasets. We evaluate our method on four widely used multi-view benchmarks:

- **Scene15** (Fei-Fei & Perona, 2005) contains 4,485 natural images spanning 15 scene categories, covering both indoor and outdoor scenarios. We extract two types of hand-crafted features for each image, namely, PHOG and GIST descriptors.
- **Caltech101** (Li et al., 2015) includes 8,677 images from 101 object categories. To form two distinct views, we adopt deep representations obtained from DECAF and VGG19 networks, consistent with Han et al. (2021).

- **LandUse-21** (Yang & Newsam, 2010) contains 2,100 satellite imagery samples in 21 categories. We follow Lin et al. (2022) to construct two views by extracting PHOG and LBP descriptors.
- **UMPC-Food101** (Wang et al., 2015) consists of paired food images and textual recipes, with 60,000 samples for training and 20,000 samples for testing across 101 categories. We use the test split for clustering evaluation. Visual features are extracted using a ViT(Wu et al., 2020) pretrained on ImageNet, while textual features are obtained with BERT (Devlin et al., 2018). Notably, the recipe descriptions often contain irrelevant or noisy information, making UMPC-Food101 a realistic benchmark for studying noisy correspondence.

Simulation of sample-level mismatch. To evaluate robustness under different conditions, we simulate two types of sample-level mismatches: i) *Alignable mismatch*: a fraction of instances (each with multiple views) are randomly permuted across views. The fraction is controlled by the *Mismatch Ratio (MR)*. ii) *Unalignable mismatch*: a fraction of view samples are corrupted with random Gaussian noise, with the fraction defined as the *Corruption Ratio (CR)*.

D Performance Visualization with varing MR and CR value (Q3)

Previous comparisons in Section 4.2 focused on specific MR and CR values, which do not fully reveal robustness across different mismatch levels. Here, we fix MR at two representative values and vary CR continuously, visualizing clustering performance of CorreGen and four state-of-the-art baselines to examine their robustness.

For evaluation, we re-align samples across views using a nearest-neighbor principle following Guo et al. (2024); Sun et al. (2025). To quantify category-level consistency, we report the *Category-level Alignment Rate* (CAR) (Yang et al., 2021), defined as

$$\text{CAR} = \frac{1}{N}\sum_{i=1}^{N}\delta\left(C(\boldsymbol{x}_i^{(v_1)}),,C(\boldsymbol{x}_{\pi(i)}^{(v_2)})\right), \tag{34}$$

where $C(\cdot)$ is the oracle category label, $\pi(i)$ is the re-aligned counterpart of $\boldsymbol{x}_i^{(v_1)}$, and $\delta(\cdot)$ is the indicator function. As shown in Figure 4, on UMPC-Food101 CorreGen demonstrates substantially lower performance degradation as CR increases, consistently outperforming all baselines. Even under severe mismatches (*e.g.*, MR=0.5), CorreGen maintains a stable CAR score, highlighting its ability to recover reliable category-level correspondences despite high noise.

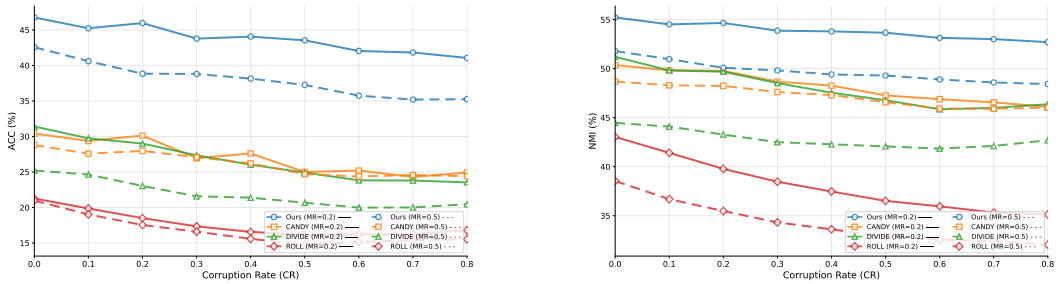
E Parameters Analysis (Q4)

In this section, we analyze the sensitivity of CorreGen on two key parameters in the E-step: the curve-shaping parameter m and the noise rate ρ. Figure 5 reports the results under $\text{MR} = 0.2$ and $\text{CR} = 0.2$. For ρ, we observe that the performance remains stable across a wide range of ρ values. For m, the performance is consistently strong when $m \leq 10$, where the marginal probabilities remain moderately discriminative. As m increases further, the probability distribution becomes overly smoothed, leading to a slight decline in performance. These results confirm that CorreGen is not overly sensitive to hyperparameter choices.

F Ablation Studies (Q5)

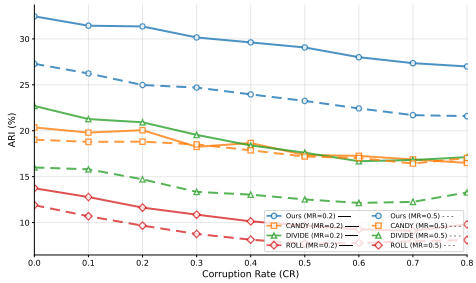
In this section, we conduct ablation studies on Scene15 and UMPC-Food101 to evaluate the effectiveness of each component. We also compare our method CorreGen with the standard InfoNCE objective. Experiments are performed under two settings: $(\text{MR} = 0.0, \text{CR} = 0.0)$ and $(\text{MR} = 0.2, \text{CR} = 0.2)$.

As shown in Table 3, the results lead to three key observations: i) On relatively clean datasets, the effect of the Virtual Sample module is not significant, and using a smaller ρ may yield better results; ii) The GMM-guided marginal estimation consistently enhances clustering accuracy by assigning higher probabilities to informative samples, thereby improving joint distribution estimation. iii)

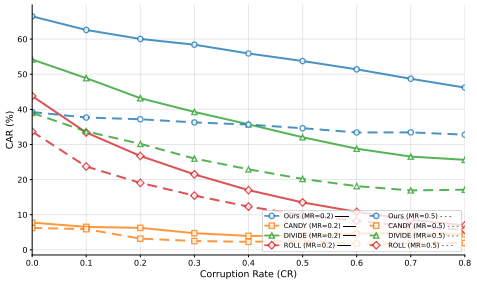


(a) ACC

(b) NMI

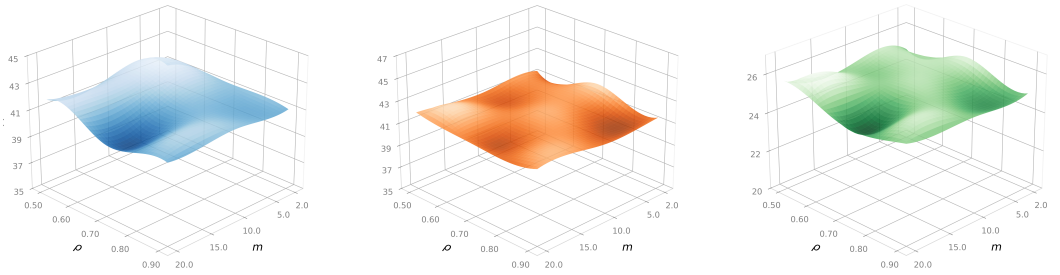


(c) ARI



(d) CAR

Figure 4: The clustering performance under varying CR value. Solid lines indicate results with MR $= 0.2$, while dashed lines correspond to MR $= 0.5$. The CR values varies from 0.0 to 0.8.



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

Figure 5: Parameters Analysis.

Training with vanilla InfoNCE fails to capture latent sample- and category-level correspondences, resulting in significant performance degradation under noisy conditions.

G IMAGE-TEXT PAIR EXAMPLE OF UMPC-FOOD101

UMPC-Food101 (Wang et al., 2015) is constructed by crawling food images with textual recipes collected from the web. As shown in Figure 6, the texts often contain irrelevant descriptions, hyperlinks, or noisy information unrelated to the visual content, making it a realistic benchmark for studying noisy correspondence in multi-view clustering.

H THE USE OF LARGE LANGUAGE MODELS

In this paper, LLMs were used to refine the writing in the Introduction, Related Work, and Experiments sections, as well as to verify the clarity of mathematical derivations.

Table 3: Abaltion study of CorreGen on Scene15 and UMPC-Food101, where *w/o* denotes the component is not adopted. "Virtual" refers to the Virtual Sample module, "Guide" refers to the GMM-guided marginal estimation, and "Vanilla InfoNCE" denotes training with the standard contrastive objective.

Setting	MR=0.0, CR=0.0						MR=0.2, CR=0.2					
	Scene15			UMPC-Food101			Scene15			UMPC-Food101		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
CorreGen	**50.25**	**48.92**	**32.87**	**49.77**	58.36	35.73	**41.78**	**41.67**	**25.50**	**45.97**	**54.66**	**31.36**
w/o Virtual	49.44	48.38	32.15	49.45	**59.22**	**36.65**	41.10	41.12	24.77	44.01	53.92	30.36
w/o Guide	49.06	48.01	31.98	49.44	57.95	35.37	40.98	41.21	24.77	44.59	54.03	30.67
w/o Virtual & Guide	49.00	48.33	31.83	48.92	58.42	35.61	40.52	40.95	24.66	43.68	53.41	29.78
Vanilla InfoNCE	47.83	47.81	31.37	48.47	57.82	34.73	38.36	37.60	21.96	43.84	52.76	29.15

Figure 6: Examples of noisy image-text pair in UMPC-Food101 datasets.