DUAL KERNEL CONTRASTIVE LEARNING FOR PARTIALLY-ALIGNED MULTI-VIEW CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Although multi-view clustering has achieved significant progress, most existing methods assume that all views are fully aligned. In real-world scenarios, however, temporal asynchrony, spatial misalignment, and heterogeneous acquisition conditions often lead to partially aligned views, posing substantial challenges to effective clustering. Existing partially-aligned multi-view clustering methods mainly suffer from two limitations: they either insufficiently exploit abundant unaligned samples or rely on cumbersome and error-prone explicit sample-level realignment, which may lead to suboptimal clustering performance. To address these issues, we propose a simple yet effective method, termed Dual Kernel Contrastive Learning for Partially-Aligned Multi-View Clustering (DKLPVC). Specifically, DKLPVC first exploits cross-view semantic similarity matrices computed from aligned samples to alleviate view distribution shifts and further uses them as prior information to construct a kernel alignment graph. This graph serves as a semantic bridge between sparse aligned pairs and abundant unaligned samples, thereby avoiding explicit sample-level realignment. Under the guidance of the learned graph, DKLPVC performs kernel contrastive learning at both the feature and cluster levels in a kernel-induced Hilbert space. The feature-level objective enhances local cross-view semantic consistency, while the cluster-level objective corrects local semantic bias from a global clustering perspective by encouraging intra-cluster compactness and inter-cluster separation. Extensive experiments on eight benchmark datasets demonstrate that DKLPVC consistently outperforms existing methods in partially-aligned multi-view clustering tasks.

1 INTRODUCTION

In practical applications, real-world data are often collected from multiple modalities or heterogeneous sources, giving rise to multi-view data Zhao et al. (2024a); Wang et al. (2024a). Multi-view clustering (MVC) aims to improve clustering performance by exploiting the consistency and complementarity among different views. Although existing MVC methods have achieved remarkable progress, most of them are developed under the assumption that all views are complete and fully aligned Peng et al. (2019); Guan et al. (2024); Liu et al. (2024). In practice, however, temporal asynchrony, spatial misalignment, and imperfect data acquisition often make this assumption unrealistic, leading to partially aligned multi-view data Huang et al. (2020); Wang et al. (2024c); He et al. (2024); Wang et al. (2024d); Zhao et al. (2024b). This setting gives rise to the partially-aligned multi-view clustering (PVC) problem, whose core challenge lies in how to effectively exploit both shared consistency and complementary information across views when only a small subset of samples is aligned.

To address PVC, a series of methods have been proposed. Early studies mainly formulate PVC as an explicit cross-view alignment problem, often relying on matching algorithms such as the Hungarian algorithm Kuhn (1955). For instance, Huang et al. Huang et al. (2020) learned consistent representations by minimizing the Euclidean distance between partially aligned samples and further introduced a differentiable surrogate of the Hungarian algorithm to recover cross-view correspondences. Despite their effectiveness, these methods heavily rely on sample-level matching, which is often cumbersome, error-prone, and difficult to scale in deep learning frameworks. More recent methods Wen et al. (2023); Wang et al. (2024c;d); Gao et al. (2025) attempt to avoid explicit matching by implicitly learning cross-view relationships through shared representations, graph structures,

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.

or distribution-level modeling. For example, Wen et al. Wen et al. (2023) proposed a parameter-free graph clustering framework that exploits structural information from each view to optimize cross-view correspondences and generate a consensus graph, while Wang et al. Wang et al. (2024d) introduced a graph contrastive learning framework to jointly facilitate alignment and latent representation learning. These approaches alleviate the dependence on explicit correspondence estimation and are more amenable to end-to-end optimization.

However, existing PVC methods still face a more fundamental challenge beyond correspondence recovery itself: how to transfer reliable semantic priors from a sparse aligned subset to abundant unaligned samples without introducing noisy instance-level matching Yang et al. (2022; 2021); He et al. (2024). This challenge involves three closely coupled aspects. First, the semantic relations inferred from aligned pairs are often highly nonlinear and cannot be adequately characterized by simple pointwise similarity in the original latent space. Second, even when reliable semantic priors are available from the aligned subset, most existing methods lack an effective mechanism to propagate them to unaligned samples, causing the abundant unaligned data to be insufficiently exploited. Third, feature-level semantic alignment alone is insufficient, because local pairwise consistency in an unsupervised setting may introduce semantic bias and lead to suboptimal global clustering structures. Therefore, a desirable PVC framework should simultaneously answer three questions: how to encode cross-view semantic relations in a more expressive space, how to propagate semantic priors from aligned to unaligned data without explicit realignment, and how to correct local semantic bias at the global clustering level.

To this end, we propose a simple yet effective framework, termed *Dual Kernel Contrastive Learning for Partially-Aligned Multi-View Clustering* (DKLPVC). As illustrated in Fig. 1, DKLPVC first performs view distribution alignment by leveraging cross-view similarity matrices computed from aligned samples, thereby alleviating view heterogeneity and distribution shift across different views. Based on the aligned semantic priors, we then construct a cross-view kernel alignment graph in a reproducing kernel Hilbert space (RKHS) for both aligned and unaligned samples. Compared with the conventional Euclidean space, the kernel-induced Hilbert space is more capable of capturing high-order nonlinear semantic structures, thereby providing more robust guidance for cross-view association. Under the guidance of the learned kernel alignment graph, DKLPVC further performs feature-level kernel contrastive learning to enhance semantic consistency among cross-view samples. Instead of relying on explicit instance-level matching, the learned graph serves as a semantic alignment bridge, enabling unaligned samples to participate in cross-view consistency learning. Furthermore, to alleviate the semantic bias caused by local feature-level contrastive learning, we introduce a cluster-level kernel contrastive objective that encourages samples assigned to the same cluster to be compact while separating samples assigned to different clusters. In this way, DKLPVC learns more robust and semantically consistent representations, leading to improved clustering performance without requiring explicit correspondence recovery. The main contributions of this work are summarized as follows:

- We propose an end-to-end semantic alignment framework for partially-aligned multi-view clustering, which transfers semantic priors from sparse aligned pairs to abundant unaligned samples without requiring explicit sample-level realignment.
- We introduce a prior-driven kernel alignment graph that lifts aligned semantic priors into a reproducing kernel Hilbert space and serves as a semantic bridge for robust cross-view association over both aligned and unaligned samples.
- We develop a dual kernel contrastive learning strategy, where the feature-level objective performs graph-guided local semantic alignment, while the cluster-level objective further corrects local semantic bias from a global clustering perspective.

2 METHODS

2.1 PRELIMINARIES

In partially view-aligned multi-view learning, only a subset of samples are aligned across different views. Specifically, let $\{\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(V)}\}$ denote a partially view-aligned multi-view dataset consisting of N samples and V views. For the v-th view, the feature matrix is defined as

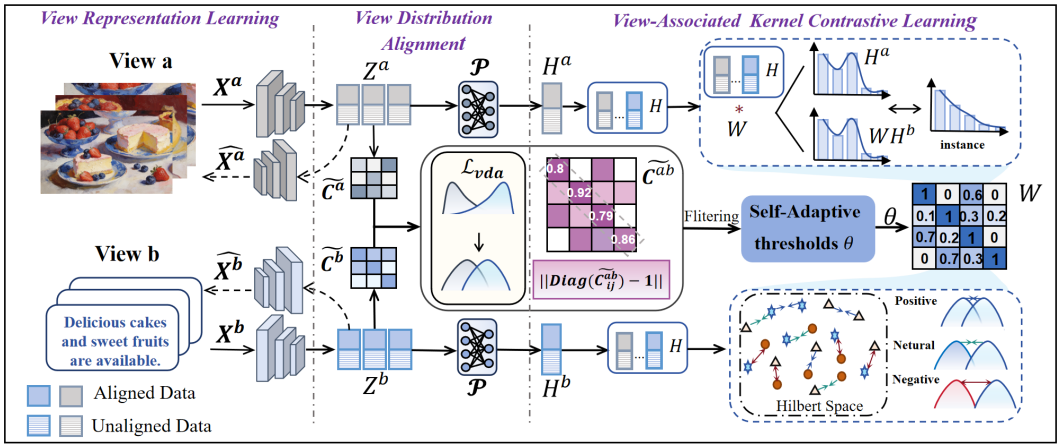


Figure 1: The view distribution alignment module minimizes the discrepancies between view-specific semantic matrices. The kernel alignment graph learns a reliable semantic guidance graph $\boldsymbol{W}$ to perform semantic-level kernel contrastive learning, capturing consistent semantic information from both aligned and unaligned data. Meanwhile, the semantic guidance graph performs view semantic alignment at the cluster level through kernel contrastive learning, alleviating semantic bias.

$\mathbf{X}^v = [\mathbf{x}_1^v, \mathbf{x}_2^v, \ldots, \mathbf{x}_N^v] \in \mathbb{R}^{N \times D_v}$, where N is the number of samples and D_v represents the feature dimensionality of view v. The dataset containing N_a aligned samples and N_u unaligned samples, respectively, satisfying $N = N_a + N_u$.

2.2 View-Specific Autoencoder

To capture these view-specific features, we employ view-specific autoencoders. For each view, a view-dependent encoder $\mathcal{E}^v$ is used to map the original input data into a shared latent space:

$$\mathbf{Z}^v = \mathcal{E}^v(\mathbf{X}^v) \in \mathbb{R}^{N \times d},$$

After obtaining the latent representation for each view, a view-specific decoder $\mathcal{D}^v$ is employed to reconstruct the original input from the latent embedding:

$$\widehat{\mathbf{X}}^v = \mathcal{D}^v(\mathbf{Z}^v).$$

To retain the view-specific intrinsic structure and semantic information, we introduce a reconstruction loss to quantify the difference between the original data $\mathbf{X}^v$ and the reconstructed output $\widehat{\mathbf{X}}^v$:

$$\mathcal{L}_{rec} = \sum_{v=1}^{V} \left\| \mathbf{X}^v - \widehat{\mathbf{X}}^v \right\|_2^2. \tag{1}$$

2.3 View Distribution Alignment

In partially view-aligned multi-view scenarios, although samples from different views correspond to the same instance, their latent representations may vary significantly due to scale differences and distribution shifts. To effectively align the view distributions while preserving semantic structures, we design a cross-view semantic alignment mechanism that incorporates both instance-level and structure-level consistency.

Specifically, given the aligned latent representations $\widetilde{\mathbf{Z}}^v \in \mathbb{R}^{N_a \times d}$, the cross-view semantic similarity between samples $\widetilde{\mathbf{z}}_i^a$ and $\widetilde{\mathbf{z}}_j^b$ is computed using cosine similarity:

$$\widetilde{\mathcal{C}}_{ij}^{ab} = \mathrm{sim}(\widetilde{\mathbf{z}}_i^a, \widetilde{\mathbf{z}}_j^b) = \frac{\langle \widetilde{\mathbf{z}}_i^a, \widetilde{\mathbf{z}}_j^b \rangle}{\|\widetilde{\mathbf{z}}_i^a\| \, \|\widetilde{\mathbf{z}}_j^b\|}, \tag{2}$$

where the diagonal elements correspond to aligned instances across views. To enforce semantic consistency for aligned samples, we define the cross-view feature alignment loss as

$$\mathcal{L}_{cfa} = \frac{1}{N_a} \left\| \mathrm{diag}(\widetilde{\mathcal{C}}^{ab}) - \mathbf{1} \right\|_2^2, \tag{3}$$

where $\mathrm{diag}(\cdot)$ extracts the diagonal elements and $\mathbf{1} \in \mathbb{R}^{N_a}$ is an all-ones vector. To preserve higher-order semantic structures across views, we construct the intra-view normalized semantic matrix $\widetilde{\mathcal{C}}^v \in \mathbb{R}^{N_a \times N_a}$:

$$\widetilde{\mathcal{C}}^v_{ij} = \mathrm{sim}(\widetilde{\mathbf{z}}^v_i, \widetilde{\mathbf{z}}^v_j), \tag{4}$$

and enforce consistency between corresponding semantic matrices across views by minimizing

$$\mathcal{L}_{cma} = \frac{1}{N_a} \left\| \widetilde{\mathcal{C}}^{aa} - \widetilde{\mathcal{C}}^{bb} \right\|_2^2. \tag{5}$$

Finally, the view distribution alignment(VDA) objective jointly optimizes instance-level and structure-level alignment:

$$\begin{aligned} \mathcal{L}_{vda} &= \mathcal{L}_{cfa} + \mathcal{L}_{cma} \\ &= \frac{1}{N_a} \Big(\left\| \mathrm{diag}(\widetilde{\mathcal{C}}^{ab}) - \mathbf{1} \right\|_2^2 + \left\| \widetilde{\mathcal{C}}^{aa} - \widetilde{\mathcal{C}}^{bb} \right\|_2^2 \Big). \end{aligned} \tag{6}$$

2.4 Feature-level Kernel Contrastive Learning

2.4.1 Kernel-Aligned Guidance Graph Construction

Unlike generic graph-based PVC methods that construct graphs directly from latent features or rely on estimated correspondences, our alignment graph is prior-driven. Specifically, it is derived from semantic priors learned on the aligned subset after view distribution alignment, and then propagated to all samples through adaptive graph construction in the kernel space. In this sense, the proposed graph is not merely a structural regularizer, but a semantic bridge that transfers reliable cross-view supervision from sparse aligned pairs to abundant unaligned data. Based on the cross-view semantic correlations obtained from aligned samples in Eq. (2), we further construct a cross-view kernel alignment guidance graph $\boldsymbol{W}^{ab} \in \mathbb{R}^{N \times N}$ using the normalized semantic matrix $\mathcal{C}^{ab}$ defined over all samples. Specifically, each edge weight ω^{ab}_{ij} is determined as

$$\omega^{ab}_{ij} = \begin{cases} 1, & i = j \text{ and } \mathbf{x}^a_i \sim \mathbf{x}^b_i \\ \mathcal{C}^{ab}_{ij}, & \mathcal{C}^{ab}_{ij} > \theta^{ab} \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

where $0 \le \omega^{ab}_{ij} \le 1$, and $\mathbf{x}^a_i \sim \mathbf{x}^b_i$ denotes that $\mathbf{x}^a_i$ and $\mathbf{x}^b_i$ form a known aligned instance pair across views.

Given the similarity scores of aligned pairs $\mathrm{diag}(\widetilde{\mathcal{C}}^{ab})$, we estimate an adaptive threshold from their empirical distribution rather than using a fixed constant. Since normalized similarities lie in $(0, 1)$, we model them with a Beta distribution, i.e., $\mathcal{C}^{ab}_{ij} \sim \mathrm{Beta}(\alpha, \beta)$. Let $\hat{\mu}$ and $\hat{\sigma}^2$ denote the sample mean and variance of the clamped similarities. The Beta parameters are estimated as

$$\hat{\alpha} = \hat{\mu} \left(\frac{\hat{\mu}(1-\hat{\mu})}{\hat{\sigma}^2} - 1 \right), \qquad \hat{\beta} = (1-\hat{\mu}) \left(\frac{\hat{\mu}(1-\hat{\mu})}{\hat{\sigma}^2} - 1 \right), \tag{8}$$

We then define the threshold as the lower quantile of the estimated positive similarity distribution:

$$\theta^{ab} = \left(F^{-1}(0.5; \hat{\alpha}, \hat{\beta}) \right), \tag{9}$$

where F^{-1} is the inverse Beta cumulative distribution function. This design yields a robust and data-adaptive cutoff, which is less sensitive to extreme low-similarity outliers.

2.4.2 KERNEL CONTRASTIVE LOSS

To reduce view-specific interference and redundant information, we introduce a shared multi-layer perceptron (MLP) projector to transform the latent representations of different views into a more abstract semantic space: $\mathbf{H}^v = \mathcal{P}(\mathbf{Z}^v)$ where $\mathbf{H}^v \in \mathbb{R}^{N\times d}$ denotes the projected representation of the v-th view, and the projector $\mathcal{P}$ is shared by all views. Based on this semantic space, the kernel alignment graph $\mathbf{W}^{ab}$ is further utilized to guide contrastive learning over all samples. In particular, known aligned cross-view instance pairs are assigned as strong positive pairs, whereas graph-connected neighboring samples are regarded as soft positive pairs weighted by ω_{ij}^{ab}, which helps reduce the adverse effect of noisy neighbors. The remaining pairs are treated as negative pairs. The kernel contrastive loss is defined as

$$\mathcal{L}_{fkcl} = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{\sum_{n\in\mathcal{N}_i^{ab}\cup\{i\}} \omega_{in}^{ab}\, \exp\left(k\big(\mathbf{h}_i^a, \mathbf{h}_n^b\big)/\tau_f\right)}{\sum_{j=1}^{N} \exp\left(k(\mathbf{h}_i^a, \mathbf{h}_j^b)/\tau_f\right)}, \tag{10}$$

where $\mathcal{N}_i^{ab}$ denotes the neighborhood of the i-th sample in W^{ab}, τ_f is a temperature parameter, $\mathbf{h}_i$ is a column vector of $\mathbf{H}^v$ and $k(\cdot)$ is a Gaussian kernel defined as

$$k(\mathbf{h}_i^a, \mathbf{h}_n^b) = \exp\left(\frac{-\left\|\frac{f_{\mathbf{I}}(\mathbf{h}_i^a)}{\|f_{\mathbf{I}}(\mathbf{h}_i^a)\|_2} - \frac{f_{\mathbf{I}}(\mathbf{h}_n^b)}{\|f_{\mathbf{I}}(\mathbf{h}_n^b)\|_2}\right\|^2}{2\phi_I^2}\right), \tag{11}$$

with $f_{\mathbf{I}}(\cdot)$ denoting a learnable mapping function. ϕ_I represents the variance parameter in the Gaussian kernel. In this way, robust semantic alignment can be achieved for both aligned and unaligned data.

2.5 SEMANTIC MATCHING FEATURE FUSION

Instead of adopting complicated sample-to-sample realignment schemes, we directly perform semantic alignment based on the semantic graph $\boldsymbol{W}^{ab}$. For each sample, the features from view b are aggregated through weighted neighborhood matching:

$$\widehat{\mathbf{h}}_i^b = \sum_{k\in\mathcal{N}_i^{ab}} \omega_{ik}^{ab}\, \mathbf{h}_k^b, \tag{12}$$

and the final fused embedding is:

$$\mathbf{h}_i = \mathrm{cat}\big(\mathbf{h}_i^1, \mathrm{norm}(\widehat{\mathbf{h}}_i^2), \ldots, \mathrm{norm}(\widehat{\mathbf{h}}_i^V)\big), \tag{13}$$

where $\mathrm{cat}(\cdot)$ denotes concatenation and $\mathrm{norm}(\cdot)$ denotes ℓ_2 normalization. The fused representation matrix is denoted by $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N]$. The fused representation $\mathbf{H}$ is finally fed into K-means for clustering.

2.6 CLUSTER-LEVEL KERNEL CONTRASTIVE LOSS

Feature-level graph-guided alignment mainly operates on local semantic relations. Although it enables unaligned samples to participate in cross-view consistency learning, it is still essentially driven by local pairwise semantics and may therefore introduce semantic bias in the absence of labels. To address this issue, we further introduce a cluster-level kernel contrastive objective. At the cluster level, we aim to encourage consistent clustering structures across multiple views. To this end, a shared clustering MLP $\mathcal{F}^C$ is introduced to extract clustering-consistent features $\{\mathbf{Q}^v \in \mathbb{R}^{N\times c}\}_{v=1}^V$ from the structural consistency representations $\{\mathbf{Z}^v\}_{v=1}^V$. Formally, this process is defined as $\mathbf{Q}^v = \mathrm{Softmax}(\mathcal{F}^C(\mathbf{Z}^v))$. Here, C denotes the number of clusters. The matrix $\mathbf{Q}^v$ characterizes the cluster assignment distribution of samples in the v-th view, while $\hat{\mathbf{Q}}$ denotes the clustering distribution inferred from the other views. In particular, each element $q_{i,j}^v$ in $\mathbf{Q}^v$ represents the probability that the i-th sample in the v-th view belongs to the j-th cluster. Inspired by Li et al. (2021), we adopt a self-supervised contrastive learning strategy to optimize the cluster assignment vectors, leading to the following cluster-level Kernel contrastive loss:

$$\begin{aligned}\mathcal{L}_{\text{ckcl}} = &-\frac{1}{C}\sum_{v=1}^{V}\sum_{i=1}^{C}\log\frac{\exp\left(k_c(\mathbf{q}_i^v, \hat{\mathbf{q}}_i)/\tau_c\right)}{\sum_{j=1}^{C}\exp\left(k_c(\mathbf{q}_i^v, \hat{\mathbf{q}}_j)/\tau_c\right) - \exp\left(k_c(\mathbf{q}_i^v, \hat{\mathbf{q}}_i)/\tau_c\right)} \\ &+\sum_{v=1}^{V}\sum_{j=1}^{C}\left(\frac{1}{N}\sum_{i=1}^{N} q_{i,j}^v\right)\log\left(\frac{1}{N}\sum_{i=1}^{N} q_{i,j}^v\right),\end{aligned} \tag{14}$$

$$k_c(\mathbf{q}_i^a, \mathbf{q}_n^b) = \exp\left(\frac{-\left\| \frac{f_{\mathbf{C}}(\mathbf{q}_i^a)}{\|f_{\mathbf{C}}(\mathbf{q}_i^a)\|_2} - \frac{f_{\mathbf{C}}(\mathbf{q}_n^b)}{\|f_{\mathbf{C}}(\mathbf{q}_n^b)\|_2} \right\|^2}{2\phi_c^2}\right), \tag{15}$$

where τ_C denotes the temperature parameter and C is the number of clusters. ϕ_C represents the variance parameter in the Gaussian kernel. The final training objective integrates reconstruction, view distribution alignment, and dual kernel contrastive learning:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \mathcal{L}_{vda} + \alpha\,\mathcal{L}_{fkcl} + \beta\mathcal{L}_{ckcl}, \tag{16}$$

where α and β balance the different components.

Table 1: Clustering performance on a half aligned (50%) multi-view dataset.

Methods		HandWritten			BDGP			WIKI			MNIST-USPS		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
CVPR'2022	MFLVC	34.77 ± 1.60	24.37 ± 0.95	15.83 ± 0.64	73.47 ± 7.71	60.55 ± 6.91	55.68 ± 9.68	24.79 ± 2.73	9.94 ± 1.45	4.66 ± 1.50	60.02 ± 1.29	37.00 ± 1.30	32.99 ± 1.04
MM'2023	DealMVC	37.03 ± 1.11	25.61 ± 0.65	17.25 ± 0.30	60.74 ± 4.37	47.40 ± 2.72	37.65 ± 3.00	32.63 ± 0.51	14.49 ± 0.14	10.39 ± 0.29	60.48 ± 1.76	35.65 ± 1.00	32.53 ± 1.28
CVPR'2023	GCFAgg	39.04 ± 0.30	26.55 ± 0.70	17.05 ± 0.35	46.82 ± 6.23	31.01 ± 5.40	21.87 ± 6.24	29.18 ± 1.67	11.15 ± 0.68	7.09 ± 0.54	56.84 ± 0.44	30.60 ± 0.49	27.36 ± 0.54
AAAI'2024	ICMVC	48.28 ± 3.36	36.34 ± 1.04	25.63 ± 1.67	63.90 ± 2.24	40.44 ± 3.94	37.99 ± 4.31	23.96 ± 0.94	8.34 ± 0.84	5.44 ± 0.68	70.32 ± 0.62	46.36 ± 0.83	45.11 ± 0.94
AAAI'2024	Surer	72.76 ± 5.54	63.66 ± 5.57	49.90 ± 9.02	42.50 ± 5.82	17.26 ± 7.50	12.22 ± 5.01	32.77 ± 0.80	15.11 ± 0.45	10.56 ± 0.56	32.31 ± 3.91	23.10 ± 7.42	12.18 ± 3.36
CVPR'2024	MVCAN	40.37 ± 1.88	29.27 ± 0.72	19.83 ± 0.60	73.25 ± 18.63	55.40 ± 21.79	53.06 ± 22.16	31.95 ± 0.24	15.05 ± 0.08	10.26 ± 0.08	61.50 ± 4.61	51.51 ± 5.41	43.53 ± 6.96
NIPS'2022	PVC	84.16 ± 0.51	78.04 ± 0.38	74.34 ± 0.52	85.53 ± 0.03	64.77 ± 0.06	67.18 ± 0.06	28.19 ± 0.11	18.08 ± 0.17	10.55 ± 0.13	75.85 ± 0.41	71.05 ± 0.31	61.34 ± 0.39
CVPR'2021	MvCLN	63.66 ± 0.55	58.98 ± 0.55	47.77 ± 0.52	80.61 ± 1.03	58.26 ± 2.05	58.49 ± 1.97	35.52 ± 0.48	17.22 ± 0.43	12.82 ± 0.55	85.60 ± 0.49	72.60± 0.71	71.23 ± 0.82
PAMI'2022	SURE	74.22 ± 0.80	70.48 ± 0.92	60.20 ± 1.02	82.46 ± 0.76	66.38 ± 0.97	59.52 ± 1.34	34.86 ± 0.07	16.22 ± 0.29	12.27 ± 0.27	93.35 ± 0.29	85.40 ± 0.57	85.98 ± 0.60
MM'2024	CGDA	83.37 ± 3.58	77.91 ± 3.88	71.88 ± 4.56	90.56 ± 0.15	77.53 ± 0.68	76.83 ± 0.97	40.40 ± 1.42	34.52 ± 1.04	26.83 ± 0.33	80.11 ± 4.51	80.21 ± 1.75	72.07 ± 3.97
MM'2024	VITAL	89.61 ± 0.61	80.16 ± 0.95	78.31 ± 1.05	82.91 ± 6.41	70.34 ± 6.07	66.51 ± 7.40	36.44 ± 0.17	20.70 ± 0.65	15.24 ± 0.44	94.28 ± 0.45	86.16 ± 0.83	87.75 ± 0.92
MM'2024	CGCN	71.18 ± 2.22	72.55 ± 1.30	62.24 ± 2.28	88.06 ± 0.27	70.14 ± 0.51	72.86 ± 0.57	35.71 ± 0.34	19.01 ± 0.23	13.85 ± 0.40	94.09 ± 0.03	88.96 ± 0.03	89.59 ± 0.06
TCSVT'2025	TCLPVC	92.68 ± 1.27	86.38 ± 1.24	85.62 ± 1.42	91.78 ± 0.24	77.94 ± 0.78	80.59 ± 0.56	50.00 ± 3.04	47.72 ± 2.38	35.42 ± 3.71	93.70 ± 0.26	86.12 ± 0.42	86.51 ± 0.53
MM'2025	BIN	91.95±1.19	80.22±1.28	81.11±2.84	91.88±1.87	80.22±1.28	81.11±2.32	53.40±2.25	48.21±2.14	45.32±1.41	88.65±2.65	78.44±1.57	76.93±1.35
PAMI'2026	CAMERA	88.51±7.51	84.35±1.92	76.81±5.42	90.07±1.15	77.54±2.18	77.00±2.52	38.08±0.61	22.11±1.10	16.03±0.84	95.48±0.67	89.05±1.27	90.26±1.40
Ours	**DKLPVC**	**94.53±0.83**	**89.25±1.27**	**88.59±1.76**	**95.52±2.08**	**89.19±3.09**	**89.40±4.52**	**60.80±1.22**	**53.60±1.51**	**45.65±1.48**	**96.16±2.47**	**91.47±1.05**	**91.58±1.81**
Methods		NUS-WIDE			Hdigit			Deep Animal			Reuters		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
CVPR'2022	MFLVC	16.45 ± 0.34	3.86 ± 0.54	2.35 ± 0.18	78.48 ± 1.24	59.98 ± 1.68	58.55 ± 2.03	12.01 ± 1.08	20.87 ± 1.15	6.39 ± 0.42	37.54 ± 1.49	13.73 ± 0.60	12.60 ± 0.60
MM'2023	DealMVC	26.98 ± 1.64	9.21 ± 0.73	6.02 ± 0.63	75.53 ± 5.89	62.50 ± 3.23	59.61 ± 4.95	29.09 ± 1.35	38.05 ± 0.70	16.64 ± 0.61	44.89 ± 2.94	14.75 ± 0.64	15.25 ± 0.78
CVPR'2023	GCFAgg	21.84 ± 3.86	9.57 ± 3.71	5.32 ± 2.34	79.68 ± 2.04	62.12 ± 3.23	60.57 ± 3.53	18.96 ± 1.78	31.27 ± 1.68	10.40 ± 1.47	35.65 ± 1.55	14.28 ± 0.47	11.36 ± 0.68
AAAI'2024	ICMVC	42.45 ± 0.57	20.71 ± 0.49	16.98 ± 0.49	78.99 ± 0.81	58.81 ± 1.28	59.00 ± 1.39	43.51 ± 2.01	53.73 ± 0.84	31.67 ± 1.75	39.00 ± 1.53	16.07 ± 0.61	13.10 ± 0.84
AAAI'2024	Surer	23.07 ± 2.72	13.58 ± 2.88	5.63 ± 2.25	29.12 ± 2.49	17.85 ± 3.43	8.03 ± 0.98	13.41 ± 1.22	19.11 ± 1.23	3.92 ± 0.90	39.29 ± 4.31	10.05 ± 1.89	8.34 ± 4.25
CVPR'2024	MVCAN	26.27 ± 0.78	8.97 ± 0.53	5.48 ± 0.47	39.65 ± 2.94	18.71 ± 1.24	14.36 ± 1.44	40.10 ± 2.55	57.39 ± 2.46	34.16 ± 3.03	33.01 ± 1.42	9.63 ± 0.96	9.72 ± 0.80
NIPS'2022	PVC	44.01 ± 0.51	39.43 ± 0.45	32.78 ± 0.24	81.92 ± 0.03	81.22 ± 0.04	80.26 ± 0.05	3.83 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	40.34 ± 0.03	14.23 ± 0.04	12.41 ± 0.05
CVPR'2021	MvCLN	62.68 ± 0.35	44.73 ± 0.52	40.52 ± 0.58	94.94 ± 0.42	88.15 ± 0.89	89.12 ± 0.89	36.72 ± 0.30	55.41 ± 0.36	31.60 ± 0.33	55.51 ± 0.46	34.35 ± 1.01	28.29 ± 0.93
PAMI'2022	SURE	59.37 ± 0.33	41.39 ± 0.43	37.83 ± 0.36	93.04 ± 0.33	84.14 ± 0.56	85.21 ± 0.65	28.44 ± 0.17	40.16 ± 0.31	19.11 ± 0.36	53.33 ± 1.49	33.41 ± 0.48	27.02 ± 0.79
MM'2024	CGDA	42.96 ± 0.10	34.90 ± 0.14	28.05 ± 0.06	74.49 ± 3.01	81.95 ± 4.11	70.62 ± 4.71	21.68 ± 0.65	35.24 ± 0.24	15.17 ± 0.63	55.60 ± 1.16	35.23 ± 0.69	28.94 ± 0.90
MM'2024	VITAL	62.85 ± 1.00	47.75 ± 1.16	42.08 ± 1.41	96.44 ± 0.17	90.60 ± 0.40	92.25 ± 0.36	44.79 ± 0.48	49.68 ± 0.59	30.11 ± 0.49	50.33 ± 1.63	32.47 ± 1.37	24.25 ± 1.24
MM'2024	CGCN	56.45 ± 1.50	39.47 ± 0.31	34.01 ± 0.90	96.55 ± 0.10	91.34 ± 0.22	92.52 ± 0.22	38.76 ± 1.02	46.53 ± 0.42	24.96 ± 0.92	45.84 ± 0.82	25.91 ± 0.93	21.36 ± 0.66
TCSVT'2025	TCLPVC	44.47 ± 3.04	33.28 ± 1.42	25.53 ± 1.76	95.54 ± 0.25	89.52 ± 0.41	90.37 ± 0.52	55.12 ± 1.27	68.08 ± 0.39	46.50 ± 1.32	46.57 ± 2.17	32.03 ± 1.88	25.49 ± 1.34
MM'2025	BIN	34.02 ± 1.24	23.10 ± 1.33	7.16 ± 1.03	92.96 ± 1.44	85.22 ± 1.51	85.26 ± 2.67	37.60± 1.57	44.24± 2.25	23.20± 1.98	34.82 ± 2.85	6.59 ± 1.79	25.49 ± 1.34
PAMI'2026	CAMERA	63.17 ± 1.33	51.21 ± 0.74	44.44 ± 1.48	96.88 ± 0.47	92.92 ± 0.78	92.56 ± 1.06	38.08 ± 0.61	22.11 ± 1.10	16.13 ± 0.84	46.65 ± 1.66	26.67 ± 2.78	21.69 ± 2.54
Ours	**DKLPVC**	**66.63 ± 1.44**	**51.78 ± 1.39**	**46.83 ± 1.95**	**98.70 ± 0.45**	**96.37 ± 1.37**	**96.75 ± 1.14**	**63.53 ± 0.93**	**68.30 ± 0.67**	**51.80 ± 0.47**	**61.35 ± 2.10**	**37.18 ± 1.47**	**36.08 ± 1.51**

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUPS

Datasets and Evaluation Metrics. We evaluate the clustering performance of the proposed method on eight widely used benchmark multi-view datasets, namely HandWritten Wen et al. (2023), BDGP Cai et al. (2012), WIKI Rasiwasia et al. (2010), MNIST-USPS Peng et al. (2019), NUS-WIDE Chua et al. (2009), Reuters Amini et al. (2009), Hdigit Chen et al. (2022), and Deep Animal Lampert et al. (2013). The alignment rate is defined as $\eta = \frac{N_a}{N}$, where N_a denotes the number of aligned samples and N is the total number of samples. Clustering performance is evaluated using three standard metrics, namely ACC, NMI, and ARI Guan et al. (2025a;c;b); Zhang et al. (2021); Zhu et al. (2026).

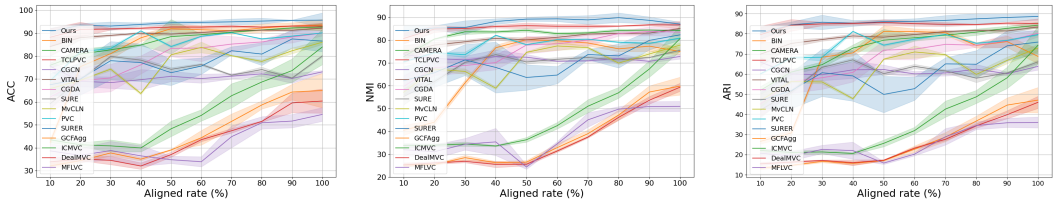
Compared Methods. We compare the proposed method with a range of representative state-of-the-art approaches. Specifically, the general multi-view clustering baselines include MFLVC Xu et al. (2022), DealMVC Yang et al. (2023), GCFAgg Yan et al. (2023), ICMVC Chao et al. (2024), Surer Wang et al. (2024b), and MVCAN Xu et al. (2024). The methods tailored for partially view-aligned clustering include PVC Huang et al. (2020), MvCLN Yang et al. (2021), SURE Yang et al. (2022), CGDA Wang et al. (2024c), CGCN Wang et al. (2024d), TCLPVC Gao et al. (2025), BIN Deng et al. (2025), CAMERA Li et al. (2026), and VITAL He et al. (2024). To ensure a fair and reliable evaluation, each competing method is independently run five times, and the mean performance together with the corresponding standard deviation is reported.

Implementation Details. All experiments are conducted on a workstation running Ubuntu and equipped with an NVIDIA RTX 5090 GPU. The proposed method is implemented in PyTorch and optimized using the Adam optimizer Adam et al. (2014). The initial learning rate is set to 1×10^{-5}, and the model is trained for 500 epochs until convergence. During training, all loss terms are jointly optimized to update the network parameters. The temperature parameter τ_f and τ_c used in contrastive learning is fixed to 1.0. For a fair comparison, all baseline methods were reproduced from their publicly available implementations under the same experimental conditions. For simplicity,

Table 2: Clustering performance on a fully aligned (100%) multi-view dataset.

Methods		HandWritten			BDGP			WIKI			MNIST-USPS		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
CVPR'2022	MFLVC	54.55 ± 2.99	50.99 ± 2.21	36.02 ± 2.70	91.91 ± 7.06	86.44 ± 9.41	83.81 ± 12.91	40.10 ± 3.27	33.83 ± 3.16	24.27 ± 3.08	99.54 ± 0.14	98.69 ± 0.35	99.01 ± 0.29
MM'2023	DealMVC	60.39 ± 5.28	59.41 ± 1.77	45.99 ± 3.01	70.39 ± 14.49	64.16 ± 15.07	52.94 ± 18.09	56.22 ± 2.19	49.36 ± 1.30	40.30 ± 1.41	93.34 ± 11.59	95.76 ± 4.57	92.95 ± 10.44
CVPR'2023	GCFAgg	64.96 ± 7.51	59.88 ± 3.93	47.08 ± 6.29	96.71 ± 0.90	92.69 ± 1.59	92.19 ± 2.06	53.43 ± 2.10	44.86 ± 1.11	35.70 ± 1.22	98.48 ± 0.17	96.37 ± 0.23	96.66 ± 0.35
AAAI'2024	ICMVC	83.97 ± 7.69	80.72 ± 5.65	74.32 ± 9.42	89.17 ± 7.27	83.83 ± 8.27	81.22 ± 11.39	41.20 ± 2.19	32.00 ± 1.01	24.56 ± 1.95	99.32 ± 0.05	98.04 ± 0.12	98.50 ± 0.11
AAAI'2024	Surer	86.54 ± 4.60	82.36 ± 2.33	75.88 ± 5.64	84.06 ± 3.71	74.12 ± 2.22	69.37 ± 3.92	54.95 ± 0.87	51.37 ± 0.92	40.55 ± 0.99	83.16 ± 1.71	79.72 ± 1.76	74.51 ± 1.87
CVPR'2024	MVCAN	75.86 ± 1.44	77.51 ± 1.61	68.81 ± 2.27	96.81 ± 1.09	92.02 ± 2.34	92.36 ± 2.59	55.04 ± 1.04	54.60 ± 0.79	42.79 ± 1.13	82.42 ± 4.29	81.11 ± 1.67	75.26 ± 3.20
NIPS'2022	PVC	90.12 ± 0.17	80.90 ± 0.28	79.58 ± 0.34	95.86 ± 0.02	90.61 ± 0.05	90.17 ± 0.05	44.31 ± 0.14	46.14 ± 0.46	30.16 ± 0.19	89.11 ± 0.07	80.10 ± 0.08	77.55 ± 0.13
CVPR'2021	MvCLN	82.31 ± 1.26	73.96 ± 0.91	66.81 ± 1.64	81.83 ± 2.46	70.64 ± 1.27	65.26 ± 2.34	53.20 ± 0.31	44.60 ± 1.44	32.52 ± 0.74	99.24 ± 0.02	97.80 ± 0.06	98.32 ± 0.05
PAMI'2022	SURE	72.62 ± 1.09	68.94 ± 1.07	57.51 ± 1.48	78.24 ± 5.38	63.47 ± 3.59	58.41 ± 7.52	56.27 ± 0.14	46.96 ± 1.15	35.91 ± 0.29	99.15 ± 0.02	97.61 ± 0.05	98.13 ± 0.04
MM'2024	CGDA	90.06 ± 1.23	83.52 ± 1.65	80.29 ± 1.85	91.44 ± 2.70	82.25 ± 2.46	80.82 ± 3.72	56.80 ± 0.12	48.91 ± 0.13	40.29 ± 0.04	97.48 ± 0.88	94.19 ± 0.67	94.50 ± 0.63
MM'2024	VITAL	92.23 ± 0.19	85.12 ± 0.35	83.39 ± 0.32	84.45 ± 6.76	75.15 ± 7.99	69.85 ± 11.73	55.03 ± 0.51	53.58 ± 0.54	41.81 ± 0.81	99.80 ± 0.06	99.40 ± 0.14	99.55 ± 0.12
MM'2024	CGCN	73.57 ± 1.15	73.19 ± 1.95	64.26 ± 2.69	96.77 ± 1.57	91.20 ± 2.26	92.40 ± 3.38	54.57 ± 0.57	51.81 ± 0.57	41.58 ± 0.74	97.48 ± 0.88	94.19 ± 0.67	94.50 ± 0.63
TCSVT'2025	TCLPVC	93.09 ± 1.37	86.71 ± 1.58	85.09 ± 2.37	97.13 ± 0.94	93.92 ± 1.31	94.17 ± 1.57	48.64 ± 2.32	45.06 ± 1.98	33.15 ± 2.74	98.50±0.13	95.98 ± 0.32	96.69 ± 0.27
MM'2025	BIN	82.40 ± 1.54	75.05 ± 2.31	68.77 ± 3.72	94.74 ± 1.49	89.27 ± 2.51	88.77 ± 3.42	55.16 ± 1.07	46.35 ± 2.41	39.1 ± 1.63	94.74 ± 1.72	89.27 ± 1.52	88.77 ± 2.17
PAMI'2026	CAMERA	92.72 ± 1.54	84.28 ± 2.42	84.25 ± 1.57	95.75 ± 2.25	89.41 ± 2.17	91.75 ± 2.17	54.84 ± 0.58	53.29 ± 0.63	41.42 ± 0.61	95.08 ± 2.09	89.99 ± 1.36	89.52 ± 2.38
Ours	**DKLPVC**	**95.68 ± 3.16**	**87.07 ± 3.81**	**88.37 ± 5.86**	**98.04±0.79**	**95.87±1.82**	**96.21±1.15**	**62.14 ± 0.98**	**56.97 ± 0.49**	**49.97 ± 0.62**	**99.85 ± 0.03**	**99.49 ± 0.08**	**99.62 ± 0.06**
Methods		NUS-WIDE			Hdigit			Deep Animal			Reuters		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
CVPR'2022	MFLVC	27.12 ± 0.66	18.28 ± 0.85	12.34 ± 0.67	**99.79±0.01**	97.61±0.11	**99.52±0.02**	33.50 ± 1.54	44.80 ± 0.99	21.40 ± 0.88	44.74 ± 1.76	28.99 ± 0.62	23.74 ± 1.60
MM'2023	DealMVC	47.36 ± 9.68	33.71 ± 7.34	27.60 ± 7.48	91.16±1.48	80.74±2.14	81.73±2.69	37.16 ± 0.77	50.47 ± 0.71	27.77 ± 0.62	50.61 ± 6.24	32.64 ± 4.99	25.06 ± 4.95
CVPR'2023	GCFAgg	36.22 ± 1.01	27.03 ± 2.06	18.64 ± 1.68	96.43±3.71	94.86±2.40	93.99±4.52	19.92 ± 0.65	32.49 ± 1.54	13.66 ± 1.32	48.93 ± 3.88	32.42 ± 0.74	24.72 ± 1.82
AAAI'2024	ICMVC	65.76 ± 1.38	51.63 ± 0.85	47.03 ± 1.55	99.35±0.11	97.96±0.29	98.55±0.23	47.16 ± 0.65	62.85 ± 0.23	36.90 ± 0.68	48.07 ± 2.30	30.17 ± 1.56	23.07 ± 1.80
AAAI'2024	Surer	32.40 ± 6.31	25.79 ± 7.96	11.81 ± 4.99	86.87±4.59	81.73±4.66	76.57±6.58	21.06 ± 1.05	32.47 ± 1.13	10.37 ± 1.18	46.25 ± 2.86	14.65 ± 2.02	13.52 ± 4.55
CVPR'2024	MVCAN	41.91 ± 4.52	34.48 ± 1.97	24.91 ± 2.55	95.80±0.40	90.27±0.58	90.98±0.80	58.93 ± 1.03	**70.71 ± 0.89**	51.70 ± 1.04	46.33 ± 2.98	24.54 ± 4.11	23.21 ± 4.69
NIPS'2022	PVC	52.12 ± 0.96	42.16 ± 0.73	37.94 ± 0.68	87.45±0.01	92.49±0.03	93.53±0.03	3.83 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	44.06 ± 0.13	20.26 ± 0.12	13.92 ± 0.18
CVPR'2021	MvCLN	62.58 ± 0.72	46.64 ± 1.15	41.59 ± 0.98	98.61±0.11	96.09±0.24	96.95±0.24	40.11 ± 0.43	53.59 ± 0.68	32.28 ± 0.79	54.02 ± 2.84	33.87 ± 3.16	26.78 ± 2.89
PAMI'2022	SURE	63.49 ± 0.23	49.58 ± 0.16	44.87 ± 0.27	97.96±0.02	94.31±0.07	95.51±0.05	41.44 ± 0.08	54.92 ± 0.21	32.13 ± 0.23	52.30 ± 0.92	31.65 ± 2.03	26.41 ± 1.30
MM'2024	CGDA	47.92 ± 2.74	43.23 ± 3.10	26.32 ± 2.68	96.91±0.54	92.80±1.07	93.29±1.14	25.44 ± 1.06	41.90 ± 0.90	18.34 ± 0.78	57.75 ± 0.15	37.49 ± 0.03	32.58 ± 0.04
MM'2024	VITAL	66.07 ± 1.28	54.02 ± 0.59	48.22 ± 1.31	99.75±0.03	**99.19±0.07**	99.44±0.05	55.56 ± 1.41	64.64 ± 0.47	44.48 ± 1.40	53.10 ± 2.30	36.32 ± 1.87	25.60 ± 2.34
MM'2024	CGCN	60.79 ± 0.42	47.60 ± 0.40	40.50 ± 0.41	98.95 ± 0.03	96.94 ± 0.09	97.69 ± 0.07	43.99 ± 2.36	55.02 ± 2.42	31.79 ± 2.27	45.53 ± 0.84	28.61 ± 1.20	23.41 ± 0.89
TCSVT'2025	TCLPVC	46.40 ± 1.72	33.61 ± 1.11	26.36 ± 1.20	99.26 ± 0.08	97.85 ± 0.20	98.37 ± 0.18	50.64 ± 1.51	67.18 ± 0.96	43.02 ± 2.20	46.82 ± 2.55	33.70 ± 2.84	26.27 ± 2.15
MM'2025	BIN	58.54 ± 1.57	48.13 ± 2.31	48.55 ± 1.18	97.61 ± 1.21	94.21 ± 1.57	94.8 ± 1.42	54.5 ± 2.41	66.06 ± 1.02	42.79 ± 2.51	46.26 ± 2.14	18.5 ± 3.27	14.31 ± 1.58
PAMI'2026	CAMERA	63.07 ± 1.33	51.21 ± 0.74	44.44 ± 1.48	98.67 ± 0.45	97.23 ± 0.85	97.15 ± 0.96	48.59 ± 2.52	59.42 ± 1.23	36.48 ± 2.36	48.77 ± 1.72	29.37 ± 2.83	23.34 ± 2.93
Ours	**DKLPVC**	**68.88 ± 0.87**	**55.80 ± 1.09**	**50.57 ± 1.19**	99.59 ± 0.06	98.92 ± 0.16	99.08 ± 0.03	**64.34 ± 0.81**	69.91 ± 1.07	**52.99 ± 0.27**	**61.44 ± 1.53**	**38.32 ± 1.05**	**38.43 ± 1.47**

DKLPVC is presented in the two-view case, while its extension to the general V-view setting is straightforward. Detailed network architectures and dataset-specific hyperparameter settings are provided in the Appendix. The computational complexity analysis is also included in the Appendix.



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

Figure 2: Clustering performance on HandWritten datasets across alignment rates.

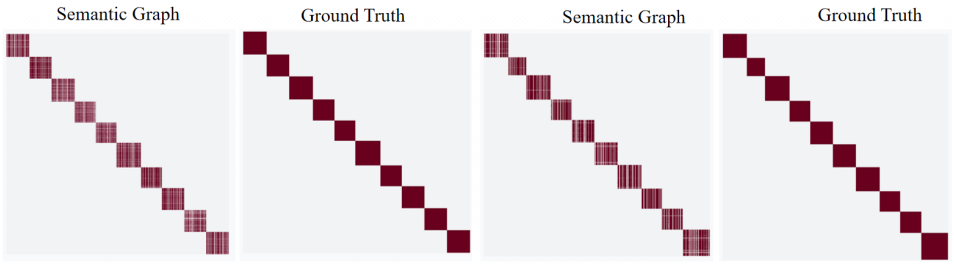
Table 3: Ablation studies of our method.

Aligned	Variants	HandWritten			BDGP			MNIST-USPS			Hdigit		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Partially	$\mathcal{L}_{rec}$	71.42 ± 2.57	61.51 ± 3.27	50.48 ± 4.56	81.96 ± 4.06	64.71 ± 2.41	61.25 ± 4.96	52.12 ± 3.25	44.02 ± 2.86	33.12 ± 2.98	51.21 ± 2.97	39.82 ± 4.26	26.84 ± 3.44
	$\mathcal{L}_{rec} + \mathcal{L}_{vda}$	73.58 ± 0.81	67.21 ± 1.77	55.23 ± 1.66	88.21±3.52	77.25 ± 4.49	74.82 ± 7.27	90.24 ± 1.29	81.22 ± 1.25	80.36 ± 2.35	95.37 ± 0.41	90.37 ± 1.23	90.66 ± 1.5
	$\mathcal{L}_{rec} + \mathcal{L}_{vda} + \mathcal{L}_{fkcl}$	93.62 ± 0.48	86.84 ± 0.74	86.29 ± 0.99	92.34 ± 2.06	85.12 ± 3.12	82.78 ± 4.30	94.62 ± 2.18	89.88 ± 0.12	88.95 ± 2.85	97.90 ± 0.67	94.65 ± 1.18	95.43 ± 0.39
	$\mathcal{L}_{rec} + \mathcal{L}_{vda} + \mathcal{L}_{fkcl} + \mathcal{L}_{ckcl}$	**94.53±0.83**	**89.25±1.27**	**88.59±1.76**	**95.52±2.08**	**89.19±3.09**	**89.40±4.52**	**96.16±2.47**	**91.47±1.05**	**91.58±1.81**	**98.70 ± 0.45**	**96.37 ± 1.37**	**96.75 ± 1.14**
Fully	$\mathcal{L}_{rec}$	71.14 ± 3.57	60.96 ± 3.21	49.21 ± 4.87	81.74 ± 3.51	65.24 ± 3.52	61.25 ± 5.91	52.15 ± 3.52	44.31 ± 2.54	32.95 ± 2.17	51.43 ± 3.05	39.57 ± 4.42	26.72 ± 2.57
	$\mathcal{L}_{rec} + \mathcal{L}_{vda}$	72.47 ± 0.75	66.48 ± 2.15	55.05 ± 1.87	96.62±0.82	91.54 ± 1.72	91.87 ± 1.89	93.75 ± 0.65	88.22 ± 1.25	86.78 ± 1.59	97.52 ± 0.48	94.11 ± 0.85	94.62 ± 0.91
	$\mathcal{L}_{rec} + \mathcal{L}_{vda} + \mathcal{L}_{fkcl}$	93.42 ± 0.80	86.76 ± 1.22	85.93 ± 1.57	97.82 ± 0.14	94.12 ± 0.28	95.11 ± 0.42	99.62 ± 0.17	98.25 ± 0.61	99.15 ± 0.12	99.56 ± 0.45	98.90 ± 0.85	99.05 ± 0.96
	$\mathcal{L}_{rec} + \mathcal{L}_{vda} + \mathcal{L}_{fkcl} + \mathcal{L}_{ckcl}$	**95.68 ± 3.16**	**87.07 ± 3.81**	**88.37 ± 5.86**	**98.04±0.79**	**95.87±1.82**	**96.21±1.15**	**99.85 ± 0.03**	**99.49 ± 0.08**	**99.62 ± 0.06**	**99.59±0.06**	**98.92±0.16**	**99.08±0.13**

3.2 Comparative Analysis

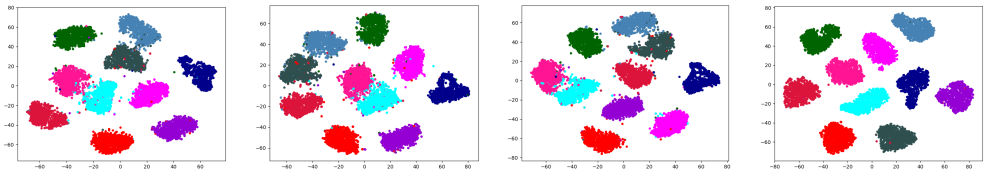
To evaluate the performance of DKLPVC under different view alignment conditions, we conducted systematic experiments under both partial alignment (50%) and full alignment (100%) settings, with results summarized in Tables 1 and 2. In the partial alignment scenario, DKLPVC consistently outperforms existing partial-view clustering methods as well as conventional multi-view clustering baselines across all datasets and evaluation metrics. The performance advantage is particularly pronounced on multi-modal datasets with substantial cross-view distribution discrepancies, such as BDGP and WIKI. These results suggest that by jointly leveraging view distribution alignment and semantic consistency guidance, DKLPVC can effectively integrate information from both aligned and unaligned samples, thereby fully exploiting the underlying multi-view structure. Under the full alignment setting, DKLPVC remains highly competitive. It significantly outperforms general multi-view clustering methods on datasets such as HandWritten, BDGP, and Reuters, and achieves performance comparable to state-of-the-art methods on MNIST-USPS (ACC = 99.85%) and Hdigit (ACC = 99.59%). Further analysis (Fig. 2) shows that DKLPVC maintains stable performance

across varying alignment ratios; even as the alignment ratio decreases, performance degradation is minimal, and the low standard deviations indicate strong robustness and reliability. In summary, by simultaneously aligning view distributions and learning semantic consistency, DKLPVC provides a robust solution for multi-view clustering under both partial and full alignment scenarios.



(a) Hdigit (b) MNIST-USPS

Figure 3: Visualization of the reordered semantic guidance graphs $\mathbf{W}$ together with the ground-truth category matrices on the complete Hdigit and MNIST-USPS datasets in view-aligned (50%) scenario. The learned semantic graphs show strong agreement with the true class structure, where high correlation values are concentrated in regions aligned with the ground-truth categories.



(a) BIN. (b) TCLPVC. (c) CAMERA. (d) Ours.

Figure 4: The t-SNE visualization on the Hdigit datasets in partially view-aligned (50%) scenario.

3.3 T-SNE VISUALIZATION

Figure 4 presents the t-SNE Maaten & Hinton (2008) visualization of the Hdigit dataset under the partially view-aligned setting with 50% alignment. Overall, the conventional PVC multi-view clustering baseline BIN, TCLPVC, and CAMERA achieve improved intra-cluster compactness and inter-cluster separation. In comparison, DKLPVC demonstrates the most well-structured clustering. The learned representations form tighter intra-cluster distributions, clearer inter-cluster boundaries, and contain fewer outliers. This superiority can be attributed to the synergistic effect of view distribution alignment and semantic matching contrastive learning, which jointly promote intra-cluster semantic consistency and inter-cluster discriminability. These visualization results further indicate that DKLPVC is capable of learning robust and discriminative representations from both partially and fully aligned multi-view data, confirming its effectiveness and adaptability in complex, real-world scenarios.

3.4 ABLATION STUDIES

To investigate the contribution of each component, we conducted ablation studies on the view distribution alignment loss ($\mathcal{L}_{vda}$), feature kernel contrastive learning loss ($\mathcal{L}_{fkcl}$), and cluster-level kernel contrastive loss ($\mathcal{L}_{ckcl}$), with results summarized in Table 3. Training with only intra-view reconstruction yields limited performance, as reconstruction alone cannot capture cross-view semantic consistency or distributional relationships. Adding $\mathcal{L}_{vda}$ aligns multi-view representations in the kernel space, mitigating distribution shifts caused by view heterogeneity. Incorporating $\mathcal{L}_{fkcl}$ further reinforces category-level semantic consistency, resulting in consistent improvements across all metrics. Introducing $\mathcal{L}_{ckcl}$ provides additional gains by alleviating semantic bias at the cluster level, enhancing the model's ability to capture inter-class relationships. These findings demonstrate that distribution alignment lays a necessary foundation, while semantic and cluster-level contrastive learning progressively improve representation quality. Under partial alignment, removing $\mathcal{L}_{fkcl}$

or $\mathcal{L}_{ckcl}$ leads to noticeable performance drops, highlighting the importance of jointly leveraging aligned and unaligned samples for semantic learning. Visualization of the learned kernel alignment graph (Fig. 3) further shows strong correspondence with the ground-truth class affinity matrix, and removing this structural guidance degrades performance under both partial (50%) and full (100%) alignment settings. This confirms that the kernel alignment graph serves as a crucial bridge between distribution alignment and kernel-based contrastive learning.

3.5 Comparison Between Kernel Alignment and Correspondence-Based Learning

To evaluate the effectiveness of cross-view feature integration under heterogeneous scenarios, we compare the proposed DKLPVC, which employs a kernel alignment strategy, with a conventional correspondence-based variant. The correspondence-based method establishes explicit one-to-one correspondences across views by searching for nearest neighbors based on Euclidean distance. However, in datasets with strong cross-view distribution discrepancies or significant noise, such explicit matching is prone to errors, which can adversely affect clustering performance. As shown in Fig. 5, DKLPVC consistently outperforms the correspondence-based variant across all metrics on multi-modal datasets such as BDGP, WIKI, and Deep Animal. This indicates that the kernel alignment strategy effectively integrates cross-view information without relying on precise one-to-one correspondences, thereby mitigating the negative impact of incorrect alignments. By enforcing consistency at the representation level rather than enforcing explicit matches, DKLPVC achieves more robust and reliable feature fusion. Overall, these results demonstrate that kernel-based alignment provides an efficient and resilient mechanism for cross-view integration, enhancing clustering stability and robustness in scenarios with high view heterogeneity and distribution shifts.

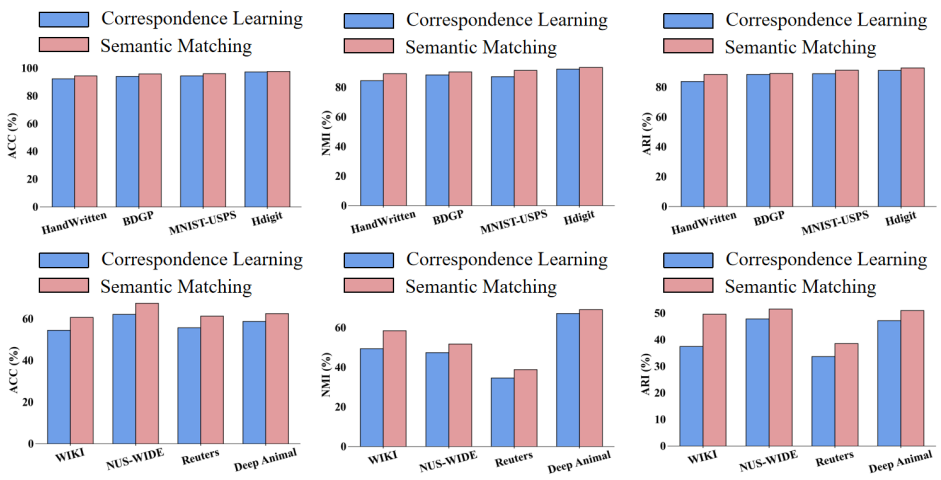


Figure 5: Performance comparison of the semantic matching against correspondence learning on eight datasets.

4 Conclusion

This paper proposes a novel representation Learning framework to address the challenges in partially aligned multi-view learning, specifically the difficulty of establishing correspondences among heterogeneous multi-source data and the issue of view shifts. Specifically, we design a Dual Kernel Contrastive Learning module that performs semantic-level view realignment at the feature level. To further mitigate semantic bias introduced by contrastive learning, we incorporate cluster-level contrastive learning. We conduct extensive comparison experiments and ablation studies to validate the superiority of the proposed model and the effectiveness of each component. In the future, we plan to further address the noise in the graphs and optimize the performance of partially aligned multi-view clustering.

5 AI USE STATEMENT

In this work, we used generative AI tools solely for language polishing of the manuscript and for no other purpose. All AI-polished content was reviewed and revised by the authors, and we take responsibility for the final content of this work.

REFERENCES

Kingma DP Ba J Adam et al. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1412(6), 2014.

Massih R Amini, Nicolas Usunier, and Cyril Goutte. Learning from multiple partially observed views-an application to multilingual text categorization. *Advances in neural information processing systems*, 22, 2009.

Xiao Cai, Hua Wang, Heng Huang, and Chris Ding. Joint stage recognition and anatomical annotation of drosophila gene expression patterns. *Bioinformatics*, 28(12):i16–i24, 2012.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

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

Man-Sheng Chen, Jia-Qi Lin, Xiang-Long Li, Bao-Yu Liu, Chang-Dong Wang, Dong Huang, and Jian-Huang Lai. Representation learning in multi-view clustering: A literature review. *Data Science and Engineering*, 7(3):225–241, 2022.

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

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pp. 1–9, 2009.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020.

Shanghui Deng, Xiao Zheng, Chang Tang, Kun Sun, Yuanyuan Liu, and Xinwang Liu. Find true collaborators: Banzhaf index-based cross view alignment for partially view-aligned clustering. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2447–2456, 2025.

Lele Fu, Zhaoliang Chen, Yongyong Chen, and Shiping Wang. Unified low-rank tensor learning and spectral embedding for multi-view subspace clustering. *IEEE Transactions on Multimedia*, 25:4972–4985, 2022.

Hang Gao, Cheng Liu, Zuosong Cai, Hongming Sun, Gaoyang Li, Ying Li, and Wei Du. A novel approach for effective partially view-aligned clustering with triple-consistency. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Renxiang Guan, Zihao Li, Wenxuan Tu, Jun Wang, Yue Liu, Xianju Li, Chang Tang, and Ruyi Feng. Contrastive multiview subspace clustering of hyperspectral images based on graph convolutional networks. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

Renxiang Guan, Tianrui Liu, Wenxuan Tu, Chang Tang, Wenhan Luo, and Xinwang Liu. Sampling enhanced contrastive multi-view remote sensing data clustering with long-short range information mining. *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–15, 2025a.

Renxiang Guan, Wenxuan Tu, Dayu Hu, Weixuan Liang, Ke Liang, Yaowen Hu, Yue Liu, and Xinwang Liu. Prototype-driven multi-view attribute-missing graph clustering. *IEEE Transactions on Multimedia*, 2025b.

Renxiang Guan, Wenxuan Tu, Siwei Wang, Jiyuan Liu, Dayu Hu, Chang Tang, Yu Feng, Junhong Li, Baili Xiao, and Xinwang Liu. Structure-adaptive multi-view graph clustering for remote sensing data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16933–16941, 2025c.

Changhao He, Hongyuan Zhu, Peng Hu, and Xi Peng. Robust variational contrastive learning for partially view-unaligned clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 4167–4176, 2024.

Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Sharable and individual multi-view metric learning. *IEEE transactions on pattern analysis and machine intelligence*, 40(9):2281–2288, 2017.

Shizhe Hu, Binyan Tian, Weibo Liu, and Yangdong Ye. Self-supervised trusted contrastive multi-view clustering with uncertainty refined. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17305–17313, 2025.

Zhenyu Huang, Peng Hu, Joey Tianyi Zhou, Jiancheng Lv, and Xi Peng. Partially view-aligned clustering. *Advances in Neural Information Processing Systems*, 33:2892–2902, 2020.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):453–465, 2013.

Haobin Li, Yijie Lin, Peng Hu, Mouxing Yang, and Xi Peng. Community-aware multi-view representation learning with incomplete information. *IEEE Trans. Pattern Anal. Mach. Intell.*, 48 (3):3517–3529, 2026. doi: 10.1109/TPAMI.2025.3639582. URL https://doi.org/10.1109/TPAMI.2025.3639582.

Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 8547–8555, 2021.

Suyuan Liu, Qing Liao, Siwei Wang, Xinwang Liu, and En Zhu. Robust and consistent anchor graph learning for multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 36 (8):4207–4219, 2024.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

HU Peng, ZHEN Liangli, PENG Xi, et al. Deep supervised multi-view learning with graph priors [j]. *IEEE Transactions on Image Processing*, 33:123–133, 2024.

Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. Comic: Multi-view clustering without parameter selection. In *International conference on machine learning*, pp. 5092–5101. PMLR, 2019.

Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 251–260, 2010.

Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, pp. 1073–1080, 2009.

Fangdi Wang, Jiaqi Jin, Jingtao Hu, Suyuan Liu, Xihong Yang, Siwei Wang, Xinwang Liu, and En Zhu. Evaluate then cooperate: Shapley-based view cooperation enhancement for multi-view clustering. *Advances in Neural Information Processing Systems*, 37:135355–135379, 2024a.

Jing Wang, Songhe Feng, Gengyu Lyu, and Jiazheng Yuan. Surer: Structure-adaptive unified graph neural network for multi-view clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 15520–15527, 2024b.

Xibiao Wang, Hang Gao, Xindian Wei, Liang Peng, Rui Li, Cheng Liu, Si Wu, and Hau-San Wong. Contrastive graph distribution alignment for partially view-aligned clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5240–5249, 2024c.

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, 2024d.

Yi Wen, Siwei Wang, Qing Liao, Weixuan Liang, Ke Liang, Xinhang Wan, and Xinwang Liu. Unpaired multi-view graph clustering with cross-view structure matching. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Song Wu, Yan Zheng, Yazhou Ren, Jing He, Xiaorong Pu, Shudong Huang, Zhifeng Hao, and Lifang He. Self-weighted contrastive fusion for deep multi-view clustering. *IEEE Transactions on Multimedia*, 26:9150–9162, 2024.

Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16051–16060, 2022.

Jie Xu, Yazhou Ren, Xiaolong Wang, Lei Feng, Zheng Zhang, Gang Niu, and Xiaofeng Zhu. Investigating and mitigating the side effects of noisy views for self-supervised clustering algorithms in practical multi-view scenarios. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22957–22966, 2024.

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.

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*, pp. 1134–1143, 2021.

Mouxing Yang, Yunfan Li, Peng Hu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1055–1069, 2022.

Xihong Yang, Jin Jiaqi, Siwei Wang, Ke Liang, Yue Liu, Yi Wen, Suyuan Liu, Sihang Zhou, Xinwang Liu, and En Zhu. Dealmvc: Dual contrastive calibration for multi-view clustering. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 337–346, 2023.

Tiejian Zhang, Xinwang Liu, Lei Gong, Siwei Wang, Xin Niu, and Li Shen. Late fusion multiple kernel clustering with local kernel alignment maximization. *IEEE Transactions on Multimedia*, 25:993–1007, 2021.

Liang Zhao, Xiao Wang, Zhenjiao Liu, Ziyue Wang, and Zhikui Chen. Learnable graph guided deep multi-view representation learning via information bottleneck. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024a.

Liang Zhao, Qiongjie Xie, Zhengtao Li, Songtao Wu, and Yi Yang. Dynamic graph guided progressive partial view-aligned clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 2024b.

Zhengzhong Zhu, Pei Zhou, Weihong Du, Shiquan Min, and Jiangping Zhu. Unsupervised semantic discovery via global and local semantic alignment in multimodal clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35248–35256, 2026.

A Related work

A.1 Multi-view Learning

Most existing multi-view learning methods assume that samples from different views are completely observed and strictly aligned at the instance level Hu et al. (2017); Peng et al. (2024; 2019); Wu et al. (2024). However, temporal asynchrony, spatial misalignment, and heterogeneous acquisition conditions often break this assumption in real-world scenarios, resulting in partially aligned multi-view data and giving rise to the partially view-unaligned problem (PVP). Existing PVP methods mainly fall into two categories. The first category focuses on recovering cross-view instance correspondences. Representative methods, such as PVC Huang et al. (2020), TCLPVC Gao et al. (2025), CGDA Wang et al. (2024c), and BIN Deng et al. (2025), learn view-invariant representations and estimate sample-level matching relationships through differentiable graph matching, optimal transport, or contribution-aware alignment. Although effective, these methods still rely heavily on correspondence estimation, making them sensitive to the quantity and reliability of aligned samples. The second category avoids explicit instance matching and instead performs category-level or semantic-level alignment. Methods such as MvCLN Yang et al. (2021), SURE Yang et al. (2022), VITAL He et al. (2024), and CAMERA Li et al. (2026) improve robustness by exploiting category alignment, reconstruction, common-specific disentanglement, or view diversity preservation. Nevertheless, they usually align views in the feature or category space, but lack an explicit mechanism to transfer reliable semantic priors from sparse aligned pairs to abundant unaligned samples.

Different from existing methods, our approach neither explicitly recovers instance-level correspondences nor relies solely on featur-level alignment. Instead, we construct a prior-driven kernel alignment graph from the aligned subset to serve as a semantic bridge for both aligned and unaligned samples. Moreover, we further combine feature-level and cluster-level kernel contrastive learning, enabling local semantic consistency while correcting semantic bias from the global clustering perspective.

A.2 Contrastive Learning

Contrastive learning has become a fundamental paradigm for representation learning because of its strong empirical performance, and it has been extensively incorporated into multi-view and multimodal learning frameworks Hu et al. (2025); Fu et al. (2022); Liu et al. (2024). In the standard SimCLR framework, two augmented versions of the same sample are treated as a positive pair, whereas all other samples within the mini-batch are regarded as negative pairs Chen et al. (2020). Although this formulation is effective in conventional settings, it is not directly suitable for partially view-unaligned clustering. On the one hand, under the Partially View-unaligned Problem (PVP), the correspondence between samples across different views is often incomplete or unavailable. As a result, cross-view samples that should represent the same instance cannot be reliably identified, making the direct construction of positive pairs highly uncertain. On the other hand, in unsupervised scenarios, multiple instances within a mini-batch may belong to the same latent semantic category. Treating all of them as negatives inevitably introduces False Negative Pair (FNPs), which can mislead representation learning and weaken the effectiveness of the contrastive objective. To alleviate this issue, existing studies have developed various strategies for filtering or correcting false negatives. SURE identifies and removes FNPs by using an adaptive threshold estimated from the distances between positive and negative pairs Yang et al. (2022). DCL Chuang et al. (2020) and SwAV Caron et al. (2020) improve robustness by reweighting training pairs and emphasizing reliable positive relations. VITAL further designs a robust contrastive loss that incorporates FNP rectification into the learning process, thereby reducing the disturbance caused by incorrectly assigned negatives. Nevertheless, despite these improvements, most existing methods still formulate contrastive learning based on a hard binary distinction between positive and negative pairs. Such a rigid partition is difficult to maintain in PVC settings, where alignment information is limited and cross-view relationships are often noisy or ambiguous. Consequently, the representation quality and clustering performance of these methods remain constrained.

Algorithm 1 Optimization of DKLPVC

Input: Partially view-aligned data $\{\mathbf{X}^v\}_{v=1}^V$; iteration number I; batch size B; embedding dimension d; hyper-parameters α , β and ϕ.
Output: Clustering assignments $\mathbf{C}$.
for $epoch = 1$ **to** I **do**
  Encode each view $\mathbf{X}^v$ via the view-specific encoder $\mathcal{E}^v$ to obtain latent representations $\mathbf{Z}^v$.
  Reconstruct $\hat{\mathbf{X}}^v$ from $\mathbf{Z}^v$ using decoder $\mathcal{D}^v$ and compute the reconstruction loss $\mathcal{L}_{rec}$ according to Eq. (1).
  Compute cross-view and intra-view normalized similarity matrices $\widetilde{\mathbf{C}}^{ab}$ and $\widetilde{\mathbf{C}}^{v}$ for aligned samples via Eq. (2) and Eq. (4), and evaluate the view distribution alignment loss $\mathcal{L}_{vda}$ using Eq. (6).
  Construct the semantic guidance graph w^{ab} over all samples from $\widetilde{\mathbf{C}}^{ab}$ with a threshold θ following Eq. (9).
  Project $\mathbf{Z}^v$ into high-level representations $\mathbf{H}^v$ and compute the feature level kernel contrastive loss $\mathcal{L}_{fkcl}$ via Eq. (10).
  Compute clustering level kernel contrastive loss $\mathcal{L}_{ckcl}$ via Eq. (14)
  Update model parameters by jointly minimizing $\mathcal{L}_{rec}$, $\mathcal{L}_{vda}$, $\mathcal{L}_{ckcl}$ and $\mathcal{L}_{fkcl}$.
end for
The fused representation $\mathbf{H}$ is applied to cluster by K-means. to obtain final cluster assignments $\mathbf{C}$.

B Experiment Details

B.1 Datasets

Table 4: Summary of dataset statistics.

Dataset	#Sample	#Class	#View	Type
HandWritten	2000	10	6	Image
BDGP	2500	5	2	Multimodal
WIKI	2866	10	2	Multimodal
MNIST-USPS	5000	10	2	Image
NUS-WIDE	9000	10	2	Multimodal
Reuters	9379	6	5	Text
Hdigit	10000	10	2	Image
Deep Animal	10158	50	7	Multimodal

DKLPVC is evaluated on eight multi-view datasets: **HandWritten**, **BDGP**, **WIKI**, **MNIST-USPS**, **NUS-WIDE**, **Reuters**, **Hdigit**, and **Deep Animal**. The details of these datasets are reported in Table 4. The proposed DKLPVC method adopts an auto-encoder architecture, the details of which are presented in Table 5. Meanwhile, for the convenience of reproducibility, we provide the hyperparameter details in Table 6. Algorithm 1 illustrates the procedure of the proposed method.

Table 5: The architecture of the encoders and decoders. dim(v) denotes the dimension of the the v-th view.

Dataset	Encoder	Decoder
HandWritten	Linear(dim(v),100),batchNorm,ReLU	Linear(100,dim(v)),batchNorm,ReLU
BDGP	Linear(dim(v),10)),batchNorm,ReLU	Linear(10,dim(v)),batchNorm,ReLU
MNIST-USPS	Linear(dim(v),10)),batchNorm,ReLU	Linear(10,dim(v)),batchNorm,ReLU
Hdigit	Linear(dim(v),10)),batchNorm,ReLU	Linear(10,dim(v)),batchNorm,ReLU
NUS-WIDE	Linear(dim(v),30)),batchNorm,ReLU	Linear(30,dim(v)),batchNorm,ReLU
WIKI	Linear(dim(v),20)),batchNorm,ReLU	Linear(20,dim(v)),batchNorm,ReLU
Reuters	Linear(dim(v),10)),batchNorm,ReLU	Linear(10,dim(v)),batchNorm,ReLU
DeepAnimal	Linear(dim(v),100)),batchNorm,ReLU	Linear(100,dim(v)),batchNorm,ReLU

Table 6: Hyperparameter details.

	HandWritten	BDGP	MNIST-USPS	Hdigit	NUS-WIDE	WIKI	Reuters	DeepAnimal
ϕ_C and ϕ_I	1	1	1	1	1	1	1	1
B	200	200	200	200	200	200	200	200
d	50	32	32	32	32	32	50	50
α	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
β	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
τ_c and τ_f	1	1	1	1	1	1	1	1

B.2 Theoretical Analysis

Theorem 1 (VDA alleviates view shift)**.** *Let the normalized latent representations of the aligned samples in two views be*

$$u_i = \frac{\tilde{z}_i^a}{\|\tilde{z}_i^a\|_2}, \qquad v_i = \frac{\tilde{z}_i^b}{\|\tilde{z}_i^b\|_2}, \qquad i = 1, \ldots, N_a. \tag{17}$$

Define

$$C_{ij}^{ab} = u_i^\top v_j, \qquad C_{ij}^{aa} = u_i^\top u_j, \qquad C_{ij}^{bb} = v_i^\top v_j. \tag{18}$$

Then minimizing L_{vda} alleviates view shift on the aligned subset in the sense that

$$\Delta_{\text{ins}} := \frac{1}{N_a} \sum_{i=1}^{N_a} \|u_i - v_i\|_2^2 \le 2\sqrt{L_{vda}}, \tag{19}$$

$$\Delta_{\text{str}} := \frac{1}{\sqrt{N_a}} \|D^a - D^b\|_F \le 2\sqrt{L_{vda}}, \tag{20}$$

where

$$D_{ij}^a = \|u_i - u_j\|_2^2, \qquad D_{ij}^b = \|v_i - v_j\|_2^2. \tag{21}$$

Therefore, the proposed VDA loss simultaneously reduces the instance-wise shift and the structural shift across views.

Proof. By the normalization of u_i and v_i, one has

$$\|u_i - v_i\|_2^2 = \|u_i\|_2^2 + \|v_i\|_2^2 - 2u_i^\top v_i = 2 - 2u_i^\top v_i = 2(1 - C_{ii}^{ab}). \tag{22}$$

$$\Delta_{\text{ins}} = \frac{1}{N_a} \sum_{i=1}^{N_a} \|u_i - v_i\|_2^2 = \frac{2}{N_a} \sum_{i=1}^{N_a} (1 - C_{ii}^{ab}). \tag{23}$$

Applying the Cauchy–Schwarz inequality yields

$$\frac{1}{N_a} \sum_{i=1}^{N_a} (1 - C_{ii}^{ab}) \le \sqrt{\frac{1}{N_a} \sum_{i=1}^{N_a} (1 - C_{ii}^{ab})^2}. \tag{24}$$

Therefore,

$$\Delta_{\text{ins}} \le 2\sqrt{\frac{1}{N_a} \sum_{i=1}^{N_a} (1 - C_{ii}^{ab})^2} = 2\sqrt{L_{cfa}} \le 2\sqrt{L_{vda}}. \tag{25}$$

Next, for the pairwise distance matrices, again using the normalization property, we have

$$D_{ij}^a = \|u_i - u_j\|_2^2 = 2 - 2u_i^\top u_j = 2 - 2C_{ij}^{aa}, \tag{26}$$

$$D_{ij}^b = \|v_i - v_j\|_2^2 = 2 - 2v_i^\top v_j = 2 - 2C_{ij}^{bb}. \tag{27}$$

Subtracting the above two equalities gives

$$D^a - D^b = -2(C^{aa} - C^{bb}). \tag{28}$$

Taking the Frobenius norm on both sides, we obtain

$$\|D^a - D^b\|_F = 2\|C^{aa} - C^{bb}\|_F = 2\sqrt{N_a L_{cma}}. \tag{29}$$

Dividing both sides by $\sqrt{N_a}$ yields

$$\Delta_{\text{str}} = \frac{1}{\sqrt{N_a}}\|D^a - D^b\|_F = 2\sqrt{L_{cma}} \le 2\sqrt{L_{vda}}. \tag{30}$$

Combining the above results, minimizing L_{vda} simultaneously suppresses the discrepancy between aligned cross-view sample pairs and the discrepancy between intra-view semantic structures. Hence, the VDA loss alleviates view shift. □

Proposition 1. *Minimizing the feature-level kernel contrastive loss maximizes a lower surrogate of the InfoNCE objective. Let*

$$\rho_{ij} = \frac{k(\boldsymbol{h}_i^{(1)}, \boldsymbol{h}_j^{(2)})}{\tau}, \tag{31}$$

and let $\mathcal{N}_i^+$ denote the positive neighbor set of anchor i and $i \notin \mathcal{N}_i^+$ for all anchors i, with $d_i = |\mathcal{N}_i^+|$. Define the InfoNCE objective as

$$I_{\text{NCE}} = \frac{1}{N}\sum_{i=1}^{N} \log \frac{N\exp(\rho_{ii})}{\sum_{j=1}^{N}\exp(\rho_{ij})}. \tag{32}$$

For every anchor i and every $n \in \mathcal{N}_i^+$, $0 \le w_{in} \le 1$, $\rho_{in} \le \rho_{ii}$. Then the feature-level kernel contrastive loss satisfies

$$-\mathcal{L}_{\text{fkcl}} \le I_{\text{NCE}} - \frac{1}{N}\sum_{i=1}^{N} \log \frac{N}{d_i + 1} \le I_{\text{NCE}} \le I(\boldsymbol{H}^{(1)}; \boldsymbol{H}^{(2)}). \tag{33}$$

Therefore, minimizing $\mathcal{L}_{\text{fkcl}}$ maximizes a valid lower surrogate of the InfoNCE objective.

Proof. By definition,

$$\mathcal{L}_{\text{fkcl}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\rho_{ii}) + \sum_{n\in\mathcal{N}_i^+} w_{in}\exp(\rho_{in})}{\sum_{j=1}^{N}\exp(\rho_{ij})}. \tag{34}$$

$$-\mathcal{L}_{\text{fkcl}} = \frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\rho_{ii}) + \sum_{n\in\mathcal{N}_i^+} w_{in}\exp(\rho_{in})}{\sum_{j=1}^{N}\exp(\rho_{ij})}. \tag{35}$$

Since $0 \le w_{in} \le 1$, we obtain

$$-\mathcal{L}_{\text{fkcl}} \le \frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\rho_{ii}) + \sum_{n\in\mathcal{N}_i^+} \exp(\rho_{in})}{\sum_{j=1}^{N}\exp(\rho_{ij})}. \tag{36}$$

Under the assumption $\rho_{in} \le \rho_{ii}$ for all $n \in \mathcal{N}_i^+$,

$$\exp(\rho_{ii}) + \sum_{n\in\mathcal{N}_i^+} \exp(\rho_{in}) \le (d_i + 1)\exp(\rho_{ii}). \tag{37}$$

$$-\mathcal{L}_{\text{fkcl}} \le \frac{1}{N}\sum_{i=1}^{N} \log \frac{(d_i + 1)\exp(\rho_{ii})}{\sum_{j=1}^{N}\exp(\rho_{ij})}. \tag{38}$$

Rearranging the above inequality gives

$$-\mathcal{L}_{\text{fkcl}} \le \frac{1}{N}\sum_{i=1}^{N} \left(\log \frac{N\exp(\rho_{ii})}{\sum_{j=1}^{N}\exp(\rho_{ij})} - \log \frac{N}{d_i + 1} \right). \tag{39}$$

$$-\mathcal{L}_{\text{fkcl}} \le I_{\text{NCE}} - \frac{1}{N}\sum_{i=1}^{N} \log \frac{N}{d_i + 1}. \tag{40}$$

Since $d_i + 1 \leq N$, we have

$$\log \frac{N}{d_i + 1} \geq 0, \tag{41}$$

which further implies

$$-\mathcal{L}_{\text{fkcl}} \leq I_{\text{NCE}}. \tag{42}$$

Finally, by the standard InfoNCE bound,

$$I_{\text{NCE}} \leq I(\boldsymbol{H}^{(1)}; \boldsymbol{H}^{(2)}). \tag{43}$$

The proof is complete. □

B.3 Computational Complexity Analysis

We analyze the computational complexity of the four loss terms under the mini-batch training setting with batch size B. For the reconstruction loss $\mathcal{L}_{rec}$, each sample is reconstructed independently, and thus its complexity is linear with respect to the batch size, i.e., $\mathcal{O}(B)$. For the view distribution alignment loss $\mathcal{L}_{vda}$, it is computed on the aligned samples within the current mini-batch. If B_a denotes the number of aligned samples in the mini-batch ($B_a \leq B$), its complexity is $\mathcal{O}(B_a^2)$, which is upper bounded by $\mathcal{O}(B^2)$. For the feature-level contrastive loss $\mathcal{L}_{fkcl}$, it requires pairwise comparisons among samples within the mini-batch, leading to a complexity of $\mathcal{O}(B^2)$. Similarly, the cluster-level contrastive loss $\mathcal{L}_{ckcl}$ is also computed via pairwise comparisons among mini-batch samples together with their cluster assignments, and therefore has complexity $\mathcal{O}(B^2)$. Accordingly, the overall per-mini-batch training complexity is $\mathcal{O}(B) + \mathcal{O}(B_a^2) + \mathcal{O}(B^2) + \mathcal{O}(B^2) = \mathcal{O}(B^2)$. Therefore, the proposed method has the same computational order as standard mini-batch contrastive learning objectives, in Table 7.

Table 7: Time complexity comparison of different methods.

Method	Time complexity
PVC	$\mathcal{O}(B^3)$
MvCLN	$\mathcal{O}(B^2)$
SURE	$\mathcal{O}(B^2)$
VITAL	$\mathcal{O}(B^2)$
CGCN	$\mathcal{O}(B^2)$
TCLPVC	$\mathcal{O}(B^2)$
CGDA	$\mathcal{O}(B^3)$
BIN	$\mathcal{O}(B^2)$
CAMERA	$\mathcal{O}(B^2)$
Ours(DKLPVC)	$\mathcal{O}(B^2)$

C More experiments

C.1 Additional Ablation Studies

To gain a deeper understanding of the internal mechanism of the view distribution alignment loss $\mathcal{L}_{vda}$, we further investigate the roles of its two key components, namely the *cross-view feature alignment loss* $\mathcal{L}_{cfa}$ and the *covariance matching alignment loss* $\mathcal{L}_{cma}$. To this end, we conduct additional ablation experiments by removing each of these two loss terms individually and evaluate the clustering performance on four representative datasets. The quantitative results are summarized in Table 8. The results show that removing $\mathcal{L}_{cfa}$ leads to a significant performance degradation, indicating that this term plays a critical role in enforcing cross-view feature alignment and alleviating distribution discrepancies across views. In comparison, removing $\mathcal{L}_{cma}$ also causes noticeable performance drops, but to a relatively smaller extent, suggesting that its main contribution lies in stabilizing the latent representation space by preserving higher-order structural consistency. To assess the contribution of the semantic guidance graph, we compare DKLPVC with a guidance-free variant, denoted as w/o graph refinement, as reported in Table 8. This variant removes the semantic

relationship modeling component and relies only on positive and negative sample pairs, with training performed exclusively on the aligned data. The results show that eliminating the semantic graph consistently reduces performance under both the 50% partially aligned and 100% fully aligned settings, demonstrating its essential contribution to the effectiveness of the proposed method. To further

Table 8: Ablation studies of our method. The best results are indicated in bold.

Aligned	Variants	HandWritten			BDGP			MNIST-USPS			Hdigit		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Partially(50%)	w/o $\mathcal{L}_{cfa}$	76.01±2.77	67.93±2.54	59.17±3.46	61.91±2.41	36.14±2.21	30.18±3.42	40.55±1.42	30.27±2.16	20.74±1.53	32.17±2.15	21.14±2.54	12.89±1.55
	w/o $\mathcal{L}_{cma}$	89.87±2.41	86.82±2.84	83.17±3.25	87.56±5.44	79.88±5.15	75.05±9.21	73.81±5.21	72.12±3.17	62.23±4.55	88.31±3.27	73.81±3.51	77.94±4.55
	w/o graph refinement	91.21±0.45	87.37±2.59	86.42±2.16	92.37±1.85	84.52±2.57	86.15±3.25	94.45±1.59	89.25±2.51	89.35±2.71	95.37 ± 2.12	94.25 ± 1.27	94.31 ± 4.27
	Ours	**94.53±0.83**	**89.25±1.27**	**88.59±1.76**	**95.52±2.08**	**89.19±3.09**	**89.40±4.52**	**96.16±2.47**	**91.47±1.05**	**91.58±1.81**	**98.70 ± 0.45**	**96.37 ± 1.37**	**96.75 ± 1.14**
Fully(100%)	w/o $\mathcal{L}_{cfa}$	76.21±1.42	68.11±1.42	60.37±2.42	62.23±3.51	43.15±2.85	36.74±2.25	41.55±2.21	31.21±2.47	21.16±4.24	34.17±1.25	24.26±1.24	15.26±1.55
	w/o $\mathcal{L}_{cma}$	89.94±1.55	86.57±1.48	81.31±4.33	90.15±4.37	85.31±3.52	83.22±5.72	73.27±3.54	69.85±4.33	59.15±2.55	89.68±5.22	83.95±1.77	79.89±2.55
	w/o graph refinement	93.25 ± 1.64	85.75±2.19	85.92±2.39	96.42±2.57	92.31±2.82	92.37±2.85	96.55 ± 1.35	95.25±1.82	96.23 ± 1.65	97.28±1.39	96.81±2.62	96.59±2.51
	Ours	**95.68 ± 3.16**	**87.07 ± 3.81**	**88.37 ± 5.86**	**98.04±0.79**	**95.87±1.82**	**96.21±1.15**	**99.85 ± 0.03**	**99.49 ± 0.08**	**99.62 ± 0.06**	**99.59±0.06**	**98.92±0.16**	**99.08±0.13**

examine its effectiveness, we visualize the learned kernel alignment graph as a heatmap, as shown in Fig. 6. The visualization exhibits a strong correspondence with the ground-truth class affinity matrix, where higher correlation values are mainly concentrated in the regions corresponding to the true classes. These quantitative and qualitative results indicate that the semantic guidance graph can effectively capture meaningful cross-view semantic relationships and provide reliable supervision for contrastive learning.

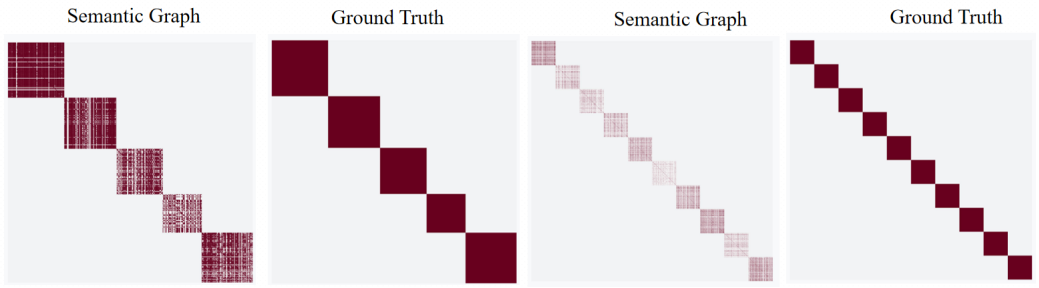


Figure 6: Visualization of the reordered semantic guidance graphs $\mathbf{W}$ together with the ground-truth category matrices on the complete BDGP and HandWritten datasets in view-aligned (50%). The learned semantic graphs demonstrate strong consistency with the true class structure, where high correlation values are concentrated in regions corresponding to the ground-truth categories.

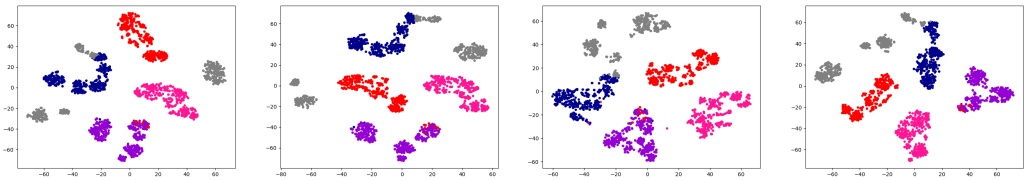
C.2 Performance under Extremely Limited Aligned Data

To further evaluate the robustness of the proposed method under extreme view alignment conditions, we compare DKLPVC with several state-of-the-art partially view-aligned (PVC) methods at alignment ratios of 1% and 0.1%. Under these settings, no more than ten aligned instances are available in each dataset, which empirically approximate the fully unaligned regime. The corresponding quantitative results are reported in Table 9.

Overall, compared with the 50% and 100% alignment settings, all PVC methods suffer from significant performance degradation at extremely low alignment ratios. In particular, methods such as PVC, BIN, and CAMERA, which heavily rely on sufficient aligned sample pairs, fail to produce valid clustering results under these conditions. This indicates that when alignment information is severely limited, correspondence-based approaches struggle to reliably capture category-level semantic structures, leading to weakened cross-view semantic discrimination. In contrast, DKLPVC consistently achieves the best performance across all datasets under near-zero alignment conditions, demonstrating strong robustness. This advantage can be attributed to its semantic matching contrastive learning mechanism, which effectively exploits unaligned samples to mitigate the adverse effects of extreme misalignment and enables the extraction of richer and more representative semantic information. These results further highlight the superiority of DKLPVC in scenarios with extremely limited or even absent view correspondences.

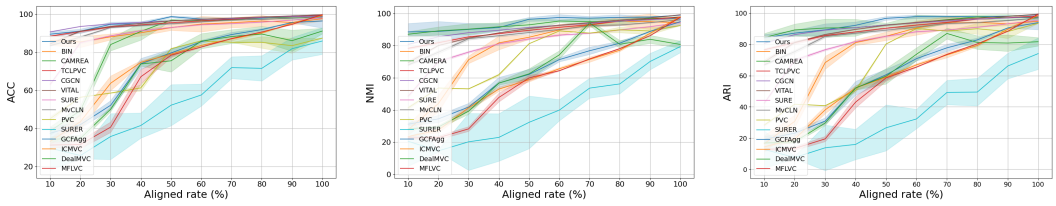
Table 9: Clustering performance (%) under extreme alignment scenarios (1% and 0.1%). The best results are highlighted in bold.

Alignment Rates	Methods	HandWritten ACC	HandWritten NMI	HandWritten ARI	BDGP ACC	BDGP NMI	BDGP ARI	MNIST-USPS ACC	MNIST-USPS NMI	MNIST-USPS ARI	Hdigit ACC	Hdigit NMI	Hdigit ARI
1%	PVC	31.20 ± 1.19	25.94±3.99	13.15±2.69	58.35±2.69	44.80±3.63	**40.78±3.63**	37.95±0.85	35.76±0.19	18.94±0.32	32.36±0.48	29.07±0.18	14.62±0.11
	VITAL	62.52±3.92	52.97±2.62	42.73±3.38	52.21±4.48	30.15±3.64	22.48±5.07	43.26±3.69	34.80±3.23	24.67±3.24	49.35±3.20	39.46±2.63	29.61±3.32
	BIN	54.50±2.39	48.25±2.77	35.67±3.55	42.52±1.28	22.73±2.17	12.29±1.67	39.88±1.34	34.59±3.01	20.82±1.84	45.10±4.08	40.56±3.46	26.29±2.11
	CAMERA	68.45±6.26	62.84±4.20	51.75±6.20	50.36±3.02	27.04±3.27	22.76±4.10	49.64±2.60	39.51±1.41	29.83±1.52	46.01±1.07	32.74±1.76	31.90±1.94
	Ours	**69.37±2.53**	**63.47±2.28**	**53.41±3.57**	**69.39±1.81**	**50.34±2.55**	33.41±1.83	**50.51±2.17**	**41.62±2.24**	**29.91±2.74**	**55.11±2.25**	**41.16±1.57**	**32.17±2.63**
0.10%	PVC	27.20±1.02	22.23±2.83	9.52±1.68	48.51±4.76	37.95±5.13	36.11±4.64	31.52±2.41	31.87±1.90	15.18±2.55	29.13±0.07	24.87±0.05	10.81±0.10
	VITAL	25.01±3.72	18.30±4.69	9.38±3.05	32.36±4.86	16.71±7.27	6.79±3.33	27.60±3.73	19.29±3.27	9.36±3.15	33.45±1.69	23.35±1.46	14.37±1.16
	BIN	44.80±3.76	26.50±1.85	14.61±1.74	33.76±2.38	13.68±2.94	5.78±2.45	29.56±4.51	24.68±3.87	10.50±2.41	35.10±4.08	30.56±3.46	16.29±2.11
	CAMERA	18.48±1.85	5.37±1.18	1.95±0.74	24.84±0.78	1.11±0.28	0.70±0.21	40.06±2.74	34.10±2.15	21.85±2.21	40.45±3.53	31.85±3.23	21.28±3.21
	Ours	**65.33±1.51**	**62.47±3.01**	**50.16±2.35**	**65.98±2.41**	**46.52±2.41**	**38.42±5.27**	**42.31±0.52**	**34.55±1.89**	**25.55±2.98**	**43.58±3.91**	**30.58±2.59**	**22.20±3.21**



(a) CAMERA. (b) TCLPVC. (c) BIN. (d) Ours.

Figure 7: The t-SNE visualization on the BDGP dataset in partially view-aligned (50%) scenario.



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

Figure 8: Clustering performance on the Hdigit dataset under different alignment rates (mean ± std.). Solid curves represent the mean results, and shaded areas denote the standard deviation.

C.3 Robustness and t-SNE Visualization of DKLPVC

DKLPVC demonstrates strong robustness in partially aligned multi-view clustering by effectively capturing category-level semantics from both aligned and unaligned samples. As illustrated in Fig. 8, DKLPVC consistently maintains stable clustering performance across different alignment ratios and outperforms both PVC and MVC baselines on all evaluated datasets. Notably, even when the alignment ratio becomes very low, the performance degradation of DKLPVC remains relatively small, especially on the Hdigit dataset. In addition, the smaller standard deviations further indicate that DKLPVC achieves more stable optimization under challenging partially aligned settings. To provide a more intuitive comparison of the learned representations, Fig. 7 presents the t-SNE visualizations on the BDGP dataset under both partial and full alignment settings. Under the 50% alignment scenario, methods specifically designed for partially aligned clustering generally produce better cluster structures than the generic CAMERA baseline. Among all methods, DKLPVC shows the most well-separated and semantically coherent clusters, which is consistent with the quantitative results. These visualizations further verify the robustness of DKLPVC and its ability to learn high-quality clustering representations.

C.4 Comparison between Cosine Similarity and Gaussian Kernel

We further conducted an ablation experiment comparing the standard cosine similarity (sim()) used in view distribution alignment with our Gaussian RBF kernel in the kernel contrastive loss. The results (Tab 10) shows that the RBF kernel consistently improves clustering performance across all benchmark datasets. This improvement can be attributed to the kernel's ability to implicitly map features into a high-dimensional Hilbert space, capturing nonlinear and higher-order semantic relationships among samples. In addition, combined with the kernel-alignment graph and the

temperature-scaled contrastive objective, the Gaussian kernel enhances the separation between positive and negative pairs and allows both aligned and unaligned samples to contribute effectively to representation learning. In addition, Table 11 compares the proposed adaptive thresholding strategy with a fixed-threshold alternative. The results show that the adaptive threshold consistently achieves better performance. This advantage mainly comes from the fact that the Beta-based estimation adjusts the threshold according to the similarity distribution of the current data, thereby yielding a more suitable decision boundary than a manually fixed threshold.

Table 10: Comparative experiments.

Datasets	HandWritten			BDGP			MNIST-USPS			Hdigit		
Metrics	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Cosine similarity	92.51±1.31	87.42±1.23	84.52±2.56	93.71±3.72	87.51±1.42	88.71±2.57	95.82±1.59	89.38±1.59	90.55±2.89	97.57±2.89	95.44±4.55	95.52 ±2.91
Ours	**94.53±0.83**	**89.25±1.27**	**88.59±1.76**	**95.52±2.08**	**89.19±3.09**	**89.40±4.52**	**96.16±2.47**	**91.47±1.05**	**91.58±1.81**	**98.70 ± 0.45**	**96.37 ± 1.37**	**96.75 ± 1.14**

Table 11: Comparative experiments.

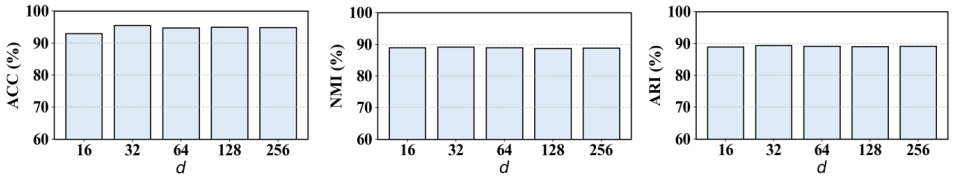
Datasets	HandWritten			BDGP		
Metrics	ACC	NMI	ARI	ACC	NMI	ARI
$\theta = 0.9$	89.14±2.51	82.31±2.51	79.44±5.12	90.11±2.25	84.25±2.74	85.74±4.25
Ours	**94.53±0.83**	**89.25±1.27**	**88.59±1.76**	**95.52±2.08**	**89.19±3.09**	**89.40±4.52**

C.5 Hyper-Parameter Analysis

To further evaluate the robustness of DKLPVC, we analyze the sensitivity of several important hyper-parameters, including the latent dimension d and the batch size B, As shown in Fig. 9, the latent dimension d has a moderate influence on clustering performance. When d is too small, the representation space may be insufficient to capture discriminative semantic information. In contrast, an excessively large d may introduce redundant information and weaken clustering compactness. On both the BDGP and Hdigit datasets, DKLPVC achieves the best performance at a moderate latent dimension (e.g., $d = 32$), while remaining relatively stable in the neighborhood of this setting.

Fig. 10 illustrates the effect of the batch size B. Overall, DKLPVC is not very sensitive to this parameter. On the BDGP dataset, ACC, NMI, and ARI remain highly stable as B varies from 200 to 5000. On the HandWritten dataset, the performance slightly decreases when $B = 200$, but stays consistently strong for $B \in [500, 5000]$. This indicates that the proposed method can maintain reliable optimization behavior under a broad range of batch sizes.

We analyze the sensitivity of the hyperparameters α and β in the total loss (Eq. (16)) on the BDGP dataset. As shown in Fig. 11, when α is too small, the sample representation learning constraint becomes weak, making it difficult to sufficiently regularize the latent space. In contrast, when α is moderately increased (e.g., $\alpha \geq 0.3$) and is comparable to or slightly larger than β, the model achieves better and more stable performance.

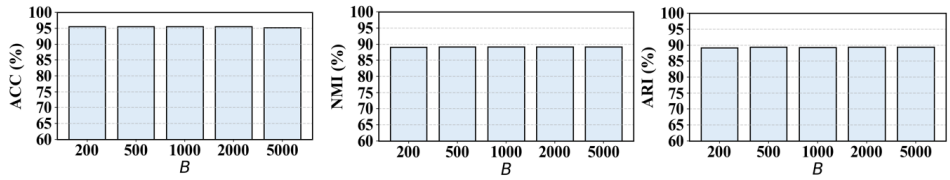


(a) BDGP

Figure 9: Sensitivity analysis of the latent dimension d on the BDGP datasets.

C.6 Loss convergence analysis

In addition, Fig. 12 illustrates the variations of different loss components during the training process of DKLPVC on the BDGP dataset. It can be observed that the total loss decreases consistently and rapidly in the initial stage, and then gradually stabilizes with only minor fluctuations. The individual loss terms also exhibit similar convergence trends. These results demonstrate the favorable convergence behavior and training stability of DKLPVC, and indicate that the model can effectively balance the multiple objectives during joint optimization.



(a) BDGP

Figure 10: Sensitivity analysis of the batchsize B on the BDGP datasets.

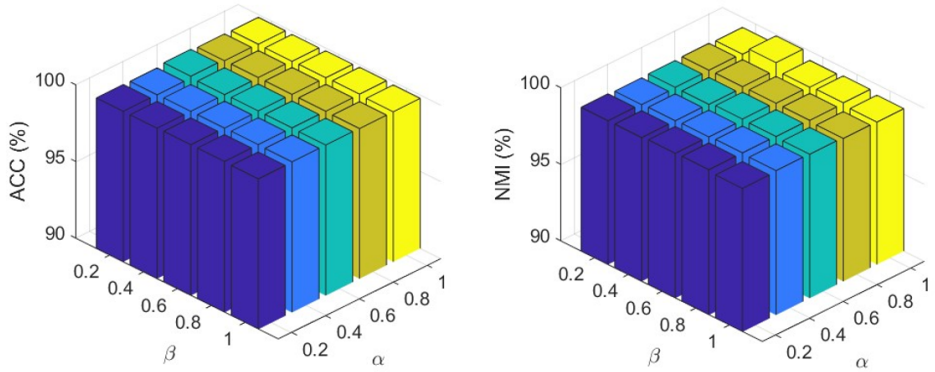


Figure 11: Hyper-parameter Sensitivity Analysis

D Evaluation Metrics

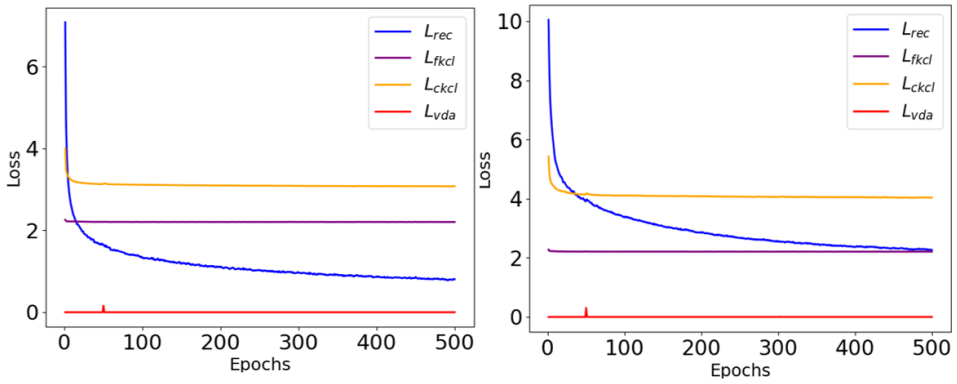
We evaluate clustering performance using three widely adopted metrics: unsupervised clustering accuracy (ACC), normalized mutual information (NMI), and adjusted Rand index (ARI).

Accuracy (ACC): ACC quantifies the proportion of correctly assigned samples after optimally aligning the predicted cluster labels with the ground truth labels. For the i-th data point, let l_i denote its ground truth label and u_i its predicted cluster assignment. ACC is calculated as

$$\mathrm{ACC} = \max_{m} \frac{\sum_{i=1}^{n} \mathbf{1}(l_i = m(u_i))}{n}, \tag{44}$$

where m denotes a one-to-one label mapping between the assignments U and V. The optimal mapping can be obtained using the Hungarian algorithm Kuhn (1955).

Normalized Mutual Information (NMI): NMI Strehl & Ghosh (2002) assesses the agreement between the ground truth and predicted partitions by quantifying their shared information relative to



(a) BDGP (b) HandWritten

Figure 12: Loss convergence analysis

their entropies. For n data points, it is expressed as

$$\mathrm{NMI} = \frac{\sum_{i=1}^{C_U}\sum_{j=1}^{C_V}\frac{|U_i\cap V_j|}{n}\log\frac{n|U_i\cap V_j|}{|U_i||V_j|}}{\mathrm{mean}\left(-\sum_{i=1}^{C_U}|U_i|\log\frac{|U_i|}{n}, -\sum_{j=1}^{C_V}|V_j|\log\frac{|V_j|}{n}\right)}. \tag{45}$$

Here, C_U and C_V are the numbers of clusters in U and V, respectively, while U_i and V_j represent their corresponding clusters.

Adjusted Rand Index (ARI): ARI Vinh et al. (2009) measures pairwise agreement between two partitions while correcting for agreement expected by chance. Its value is invariant to permutations of cluster labels. For a dataset X containing n samples, consider two partitions $U = \{u_1, u_2, \cdots, u_r\}$ and $V = \{v_1, v_2, \cdots, v_s\}$. Their overlap is summarized by $n_{ij} = |U_i \cap V_j|$, where $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, s$. The corresponding cluster sizes are $a_i = \sum_j n_{ij}$ and $b_j = \sum_i n_{ij}$. ARI is then computed as

$$\mathrm{ARI} = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \frac{\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i\binom{a_i}{2} + \sum_j\binom{b_j}{2}\right] - \frac{\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}}{\binom{n}{2}}}. \tag{46}$$