NONLINEAR KERNELS IN LINEAR TIME: ORDER-INVARIANT LARGE-SCALE MULTIVIEW CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multiview clustering (MVC) has garnered considerable attention owing to its capability to uncover intrinsic data structures by integrating complementary information across multiple views. However, existing large-scale MVC methods still suffer from two notable limitations: 1) tensor-based modeling may implicitly exploit the sequential structure of sorted samples, introducing hidden prior information that conflicts with the unsupervised learning setting, 2) they lack the ability to characterize complex nonlinear structures of multiview data under linear computational complexity. To address these issues, we propose a novel large-scale MVC model for effectively handling nonlinear data, which not only avoids employing the latent priors of sorted samples to enhance order invariance in high-order tensor modeling, but also captures complex nonlinear sample relationships while preserving linear scalability in the number of samples. Then, an efficient ADMM-based alternating optimization algorithm with linear complexity is developed to solve the proposed model. More importantly, the convergence guarantee of the algorithm is theoretically established. Extensive experiments conducted under sorted and shuffled settings on diverse benchmark datasets, including large-scale dataset with more than 100,000 samples, demonstrate that our proposed method achieves superior clustering performance against state-of-the-art methods.

1 INTRODUCTION

Multiview clustering (MVC), which endeavors to integrate complementary and consistent information from multiple views to uncover the underlying cluster structure of data, has emerged as an effective unsupervised learning paradigm Wang et al. (2025; 2024b). Owing to its ability to mine potential patterns hidden in heterogeneous data, MVC has attracted significant attention in machine learning Jia et al. (2023) and has been widespreadly applied to various field, including identifying disease Gao et al. (2024); Pang et al. (2010), human action recognition Sheng et al. (2020); Liu et al. (2017), and outlier detection Toller et al. (2023); Zhao et al. (2018).

Existing MVC methods can be broadly categorized into matrix-oriented and tensor-oriented approaches, among which tensor-oriented MVC has demonstrated promising performance in capturing high-order correlations across views Liu et al. (2026); Fu et al. (2026b). Within this line, the tensor singular value decomposition (t-SVD) based tensor nuclear norm (TNN) has been widely employed to characterize the underlying low-rank structure. However, its uniform shrinkage of singular values may over-penalize informative components and yield suboptimal low-rank representations Shan et al. (2024); Wang et al. (2022), motivating the development of numerous nonconvex based MVC methods Xia et al. (2022); Guo et al. (2023); Fu et al. (2026a). Despite demonstrating impressive performance, the tensor-related operations involved in these t-SVD-based methods incur quadratic or even cubic computational costs with the number of samples, limiting their applicability to large-scale data Gu & Feng (2024); Ji & Feng (2023). To this end, factorization and anchor learning strategies have received increasing attention in large-scale MVC, which substantially reduce computational burden but typically sacrifice the complete preservation of pairwise relationships among all samples.

Another concern is that most existing approaches implicitly exploit the sequential information of the sorted samples due to the utilization of the tensor rotation trick, thus violating the unsupervised learning setting. To mitigate this issue, Ji & Feng (2025a) proposed a fast tensorial MVC method,

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.

which performs tensor learning in an anchor-induced space to avoid reliance on prior information hidden in sorted data, improving efficiency on shuffled samples. However, its reliance on a limited set of anchors makes the learned relationships sensitive to anchor quality. Recently, Liu et al. (2025) developed a novel large-scale MVC method by removing the tensor rotation trick and exploiting implicit linear kernels to preserve relationships among all samples with linear complexity, thereby improving robustness to sample permutations without resorting to anchor approximation. Nevertheless, real-world data often exhibit complex nonlinear structures, where samples may be distributed on curved low-dimensional manifolds that cannot be adequately characterized by linear similarities. Directly introducing nonlinear kernels can capture such nonlinear relationships, but typically requires constructing full kernel matrices, undermining the linear scalability of existing large-scale MVC methods. This leads to a more fundamental challenge: how can linear computational complexity and invariance to sample ordering be preserved while incorporating nonlinear similarities among all samples into MVC.

To address the aforementioned challenges, this work proposes a novel MVC model, termed NK-OLMC. Compared with existing models, our model not only preserves informative relationships among all samples and exploits high-order correlations across views, but also enhances both linear and nonlinear similarity modeling while retaining efficiency. Specifically, the proposed model has three prominent advantages. First, it exploits nonlinear kernel modeling to enhance the capability to accommodate both linear and nonlinear data structures, enabling effective characterization of underlying sample relationships. Second, the computational cost of this model scales linearly with the number of samples while retaining relationships among all samples, making it suitable for large-scale MVC tasks. Third, it maintains stable clustering performance across different sample permutations, demonstrating robustness to sample ordering. Overall, the main contributions can be illustrated as follows:

- A novel large-scale MVC model is proposed to capture both linear and nonlinear relationships among all samples while retaining linear complexity with respect to the sample size. Specifically, sample similarities are approximated in an explicit low-dimensional feature space, avoiding the quadratic cost of full kernel construction, while the view-specific latent representations are jointly modeled to exploit high-order correlations without relying on sample ordering.
- An efficient alternating optimization algorithm with linear computational complexity is designed for the resulting nonconvex problem. Furthermore, theoretical convergence analysis guarantee that any accumulation point generated by the proposed algorithm satisfies the first-order Karush–Kuhn–Tucker (KKT) conditions under the orthogonality constraints.
- Extensive experiments on diverse benchmark datasets under both ordered and shuffled settings consistently demonstrate that the proposed method achieves competitive clustering performance against state-of-the-art methods and exhibits order-invariant behavior under sample permutations.

2 Related work

Existing scalable MVC methods mainly alleviate computational burden through factorization strategy Gao et al. (2019); Yang et al. (2021); Li et al. (2023) or anchor learning strategy Gu & Feng (2024); Kang et al. (2020); Ji & Feng (2023). Among them, factorization-based methods decompose the original data or representations into lower-dimensional factors. For example, Gao et al. (2019) incorporated multi-manifold regularization into multiview matrix factorization to preserve local geometry while learning a shared consensus representation, whereas Li et al. (2023) extended factorization to the tensor domain to capture higher-order dependencies across views. Anchor-based methods instead represent samples through a small set of representative anchor points or dictionaries. For instance, Gu & Feng (2024) present a scalable tensor MVC framework to learn compact anchor-based subspace representations with anchor hypergraph regularization, achieving improved scalability while preserving local geometric structures. Although these approaches substantially reduce computational costs, their representations inevitably compress the original sample relationships, whose fidelity may consequently depend on the capacity of the latent factors or the representativeness of the selected anchors.

Another important issue in tensor-based MVC is its sensitivity to sample arrangement. To this end, several methods have emerged to substantially alleviate sample-order dependence while maintaining scalability Liu et al. (2025); Ji & Feng (2025a). However, their similarity modeling remains essentially linear, leaving complex nonlinear geometry insufficiently characterized. Kernel-based MVC provides a natural remedy by representing nonlinear sample relationships in reproducing kernel Hilbert spaces Liu et al. (2023); Chen et al. (2020); Wang et al. (2021), and multiple-kernel methods further integrate view-specific nonlinear kernels to capture complementary structures across views Liu (2024); Wang et al. (2024a). However, explicit kernels generally requires constructing $N \times N$ kernel matrices, incurring at least quadratic storage and computational costs. Therefore, existing studies have made substantial progress in addressing different aspects of these challenges, yet unified model that simultaneously achieves scalability, robustness to sample permutations, and effective nonlinear structure modeling remains insufficiently explored.

3 Proposed method

3.1 Problem formulation

Recent advances have shown that linear kernels can provide an effective and computationally efficient means for large-scale MVC, while their ability to characterize complex nonlinear structures remains limited. This fact motivates us to explore whether nonlinear relationships among samples can be effectively modeled while preserving linear computational complexity. Specifically, given a multiview dataset $\{\mathbf{X}_v\}_{v=1}^V$ with potentially complex nonlinear structures, consisting of N samples observed from V views, let $\mathbf{X}_v = [\mathbf{x}_{v,1}, \ldots, \mathbf{x}_{v,N}] \in \mathbb{R}^{d_v \times N}$ denote the feature matrix of the v-th view and d_v is its feature dimension. To characterize complex nonlinear structures, we employ the Gaussian radial basis function (RBF) kernel, which is universal and models sample relationships through a radial function of pairwise Euclidean distances (Steinwart, 2001), providing a rich characterization of nonlinear structures. It is formulated as

$$k_v(\mathbf{x}_{v,i}, \mathbf{x}_{v,j}) = \exp\left(-\frac{\|\mathbf{x}_{v,i} - \mathbf{x}_{v,j}\|_2^2}{2\sigma_v^2}\right), \tag{1}$$

where $\sigma_v > 0$ controls the bandwidth of the kernel in the v-th view. We set $\sigma_v = \alpha\sqrt{m_v/2}$, where $\alpha > 0$ is a scaling parameter and m_v denotes the median pairwise squared Euclidean distance between distinct samples in the v-th view. Accordingly, the nonlinear kernel matrix $\mathbf{K}_v \in \mathbb{R}^{N\times N}$ is defined by $\mathbf{K}_v(i,j) = k_v(\mathbf{x}_{v,i}, \mathbf{x}_{v,j})$. Although nonlinear kernels enable the characterization of complex nonlinear relationships among samples, explicitly constructing $\mathbf{K}_v$ requires $\mathcal{O}(N^2)$ computational cost, which becomes prohibitive for large-scale data. This naturally raises the question of whether this quadratic kernel bottleneck can be overcome in linear time. The answer is affirmative.

Specifically, we employ random Fourier features Rahimi & Recht (2007) to construct explicit finite-dimensional representation of the Gaussian kernel. For the v-th view, we independently sample

$$\boldsymbol{w}_{v,l} \sim \mathcal{N}\left(\mathbf{0}, \sigma_v^{-2}\mathbf{I}_{d_v}\right), \qquad b_{v,l} \sim \mathcal{U}(0, 2\pi), \tag{2}$$

for $l = 1, \ldots, D_r$, where D_r denotes the dimension of random Fourier features, $\boldsymbol{w}_{v,l} \in \mathbb{R}^{d_v}$ denotes the l-th random frequency vector and $b_{v,l}$ is the corresponding random phase. Here, $\mathcal{N}(\cdot,\cdot)$ and $\mathcal{U}(\cdot,\cdot)$ represent the Gaussian and uniform distributions, respectively. For each sample $\mathbf{x}_{v,i}$, the corresponding random feature mapping $\mathbf{z}_v(\mathbf{x}_{v,i})$ is constructed as

$$\mathbf{z}_v(\mathbf{x}_{v,i}) = \sqrt{\frac{2}{D_r}}[\cos(\boldsymbol{w}_{v,1}^\top \mathbf{x}_{v,i} + b_{v,1}), \ldots, \cos(\boldsymbol{w}_{v,D_r}^\top \mathbf{x}_{v,i} + b_{v,D_r})]^\top \in \mathbb{R}^{D_r}. \tag{3}$$

Let $\boldsymbol{\Omega}_v = [\boldsymbol{w}_{v,1}, \ldots, \boldsymbol{w}_{v,D_r}] \in \mathbb{R}^{d_v \times D_r}$ and $\mathbf{b}_v = [b_{v,1}, \ldots, b_{v,D_r}]^\top$. Collecting the random features of all samples, we define

$$\mathbf{Z}_v = \sqrt{\frac{2}{D_r}} \cos\left(\boldsymbol{\Omega}_v^\top \mathbf{X}_v + \mathbf{b}_v \mathbf{1}_N^\top\right), \tag{4}$$

where $\mathbf{Z}_v$ denotes the explicit nonlinear feature representation of the v-th view. The inner product between two random feature vectors yields a Monte Carlo approximation to the Gaussian kernel

Avron et al. (2016), that is $\mathbf{z}_v(\mathbf{x}_{v,i})^\top \mathbf{z}_v(\mathbf{x}_{v,j}) \approx k_v(\mathbf{x}_{v,i}, \mathbf{x}_{v,j})$. Consequently, the nonlinear kernel matrix can be approximated by

$$\mathbf{K}_v \approx \mathbf{Z}_v^\top \mathbf{Z}_v. \tag{5}$$

Thus, the nonlinear relationships among all samples can be represented through $\mathbf{Z}_v \in \mathbb{R}^{D_r \times N}$ without explicitly storing $\mathbf{K}_v$, reducing the storage from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$ when $D_r \ll N$.

To encode these nonlinear sample relationships, we further introduce a latent representation $\mathbf{H}_v \in \mathbb{R}^{c \times N}$ for the v-th view, where c denotes the number of clusters. Following the kernel alignment formulation in Liu et al. (2025), we characterize the pairwise similarities induced by the latent representation through $\mathbf{H}_v^\top \mathbf{H}_v$. By imposing the orthogonality constraint $\mathbf{H}_v \mathbf{H}_v^\top = \mathbf{I}_c$, we align the latent similarity with the kernel matrix $\mathbf{K}_v$ as

$$\max_{\mathbf{H}_v} \operatorname{Tr}\left(\mathbf{K}_v \mathbf{H}_v^\top \mathbf{H}_v\right), \qquad s.t.\ \mathbf{H}_v \mathbf{H}_v^\top = \mathbf{I}_c. \tag{6}$$

Substituting the random Fourier features approximation $\mathbf{K}_v \approx \mathbf{Z}_v^\top \mathbf{Z}_v$ into (6) yields

$$\max_{\mathbf{H}_v} \operatorname{Tr}\left(\mathbf{Z}_v^\top \mathbf{Z}_v \mathbf{H}_v^\top \mathbf{H}_v\right), \qquad s.t.\ \mathbf{H}_v \mathbf{H}_v^\top = \mathbf{I}_c. \tag{7}$$

Crucially, the matrix $\mathbf{Z}_v^\top \mathbf{Z}_v \in \mathbb{R}^{N \times N}$ need not be explicitly formed since $\operatorname{Tr}\left(\mathbf{Z}_v^\top \mathbf{Z}_v \mathbf{H}_v^\top \mathbf{H}_v\right) = \left\|\mathbf{Z}_v \mathbf{H}_v^\top\right\|_F^2$, which can be evaluated in $\mathcal{O}(N D_r c)$ time and therefore scales linearly with N.

To further exploit high-order correlations across views, we directly stack the view-specific latent representations into a third-order tensor $\mathcal{H} \in \mathbb{R}^{c \times N \times V}$ without applying tensor rotation. Moreover, we introduce the nonconvex weighted tensor nuclear norm (WTNN) to adaptively differentiate the contributions of distinct singulars, effectively characterizing the intrinsic low-rank structure. Combining within-view nonlinear similarity preservation with cross-view high-order correlation, the proposed model is formulated as

$$\begin{aligned} \min_{\mathbf{H}_v} \ & -\frac{1}{2} \sum_{v=1}^{V} \operatorname{Tr}\left(\mathbf{Z}_v^\top \mathbf{Z}_v \mathbf{H}_v^\top \mathbf{H}_v\right) + \lambda \|\mathcal{H}\|_{\omega,\circledast} \\ s.t.\ & \mathbf{H}_v \mathbf{H}_v^\top = \mathbf{I},\ \ v = 1, 2, \ldots, V, \\ & \mathcal{H} = \Phi(\mathbf{H}_1, \mathbf{H}_2, ...\mathbf{H}_V), \end{aligned} \tag{8}$$

where λ is a trade-off parameter and $\Phi(\cdot)$ denotes the tensor construction strategy.

Remark 3.1 (*The superiority of nonlinear similarity learning*). Linear similarities characterize sample relationships through inner products in the original feature space and may be insufficient for data with complex nonlinear geometry. By introducing the Gaussian kernel, the proposed model measures pairwise relationships in a nonlinear feature space, allowing the latent similarity $\mathbf{H}_v^\top \mathbf{H}_v$ to be aligned with nonlinear kernel structure. This extends similarity modeling beyond linear relationships and enables complex nonlinear dependencies among samples to be incorporated into multiview clustering. Consequently, the learned latent representations can better reflect the underlying clustering structure, particularly when the data distribution exhibits pronounced nonlinearity.

Remark 3.2 (*The benefit of scalable nonlinear kernel approximation*). Despite its nonlinear modeling capability, direct kernelization leads to quadratic storage and computational costs with respect to the number of samples. Random Fourier features provide an explicit approximation to the Gaussian kernel, satisfying $\mathbf{Z}_v^\top \mathbf{Z}_v \approx \mathbf{K}_v$. Moreover, the random feature mapping is applied to every sample, nonlinear similarities over the complete sample set can be approximated without anchor selection. It therefore retains comprehensive sample relationships while enhancing the scalability of nonlinear similarity learning for large-scale datasets.

Remark 3.3 (*The benefit of order-invariant modeling*). We directly stack the view-specific latent representations into $\mathcal{H} \in \mathbb{R}^{c \times N \times V}$ along the view dimension and avoid tensor rotation in subsequent tensor low-rank modeling. Since t-SVD applies FFT along this mode, the resulting Fourier domain modeling is naturally devoted to inter-view dependencies rather than sample sequences, enabling the tensor low-rank regularization to capture high-order correlations across views without operating along the sample dimension or exploiting sample-order information. Therefore, the proposed model yields order-invariant behavior while preserving high-order correlations across views.

3.2 OPTIMIZATION

We employ the ADMM to optimize problem (8). Accordingly, by introducing the auxiliary variable $\mathcal{G}$ for the separability of $\mathcal{H}$, the corresponding unconstrained form of optimization problem (8) can be formulated as

$$\begin{aligned} \mathcal{L}\left(\mathbf{H}_v, \mathcal{G}, \mathcal{W}\right) = & -\frac{1}{2}\sum_{v=1}^{V} \mathrm{Tr}\left(\mathbf{Z}_v^\top \mathbf{Z}_v \mathbf{H}_v^\top \mathbf{H}_v\right) + \lambda\|\mathcal{G}\|_{\omega,\circledast} + \langle \mathcal{W}, \mathcal{G} - \mathcal{H}\rangle + \frac{\rho}{2}\|\mathcal{G} - \mathcal{H}\|_F^2 \\ & s.t.\ \ \mathbf{H}_v \mathbf{H}_v^\top = \mathbf{I}, \end{aligned} \tag{9}$$

where $\mathcal{W}$ is Lagrangian multiplier, ρ denotes a positive penalty parameter to control convergence. The complete optimization procedure is shown in the appendix.

3.3 ALGORITHM ANALYSIS

Convergence Analysis. The convergence of algorithms under ADMM framework for nonconvex optimization problems is generally difficult to establish Tan et al. (2025); Wang et al. (2025). In our model, the nonconvex tensor regularization and orthogonality constraints further complicate the theoretical analysis. Nevertheless, we provide a rigorous convergence analysis for Algorithm 1 in Theorem 1 [1], showing that any accumulation point generated by our algorithm satisfies the first-order KKT conditions.

Theorem 1. *Let $\{\mathcal{T}^k = (\boldsymbol{H}_v^k, \mathcal{G}^k; \mathcal{W}^k)\}_{k=1}^{\infty}$ be the sequence generated by the Algorithm 1. If $\lim_{k\to\infty}\{\rho_k(\mathcal{N}_v^{k+1} - \mathcal{N}_v^k)\} = \boldsymbol{0}$ with $\mathcal{N}_v^k = (\boldsymbol{H}_v^k, \mathcal{W}^k)$ holds, the sequence $\{\mathcal{T}^k\}$ satisfies the following properties:*

1) The sequence $\{\mathcal{T}^k\}_{k=1}^{\infty}$ is bounded.

2) Any accumulation point $\mathcal{T}^*$ of $\{\mathcal{T}^k\}_{k=1}^{\infty}$ satisfies first-order KKT conditions.

Computational complexity analysis. The computational requirements of the proposed algorithm mainly lie in the update of variables $\mathbf{H}_v$, $\mathcal{G}$ and $\mathcal{W}$. For updating $\mathbf{H}_v \in \mathbb{R}^{c\times N}$, it takes $\mathcal{O}(tNcD_rV + tc^2NV)$ for all views due to the operation of $(\mathbf{H}_v\mathbf{Z}_v^\top)\mathbf{Z}_v$ and the compact SVD of $\mathbf{M}_v \in \mathbb{R}^{c\times N}$ after t iterations. For tensor $\mathcal{G} \in \mathbb{R}^{c\times N\times V}$, the updating relies on SVD, FFT and inverse FFT operations, which respectively takes $\mathcal{O}(c^2NV)$, $\mathcal{O}(cNVlogV)$ and $\mathcal{O}(cNVlogV)$, accumulating a complexity of $\mathcal{O}(c^2NV + cNVlogV)$. Updating $\mathcal{W} \in \mathbb{R}^{c\times N\times V}$ only costs $\mathcal{O}(cNV)$. Considering the fact $c, V \ll N$, the overall computational complexity of the proposed Algorithm 1 is $\mathcal{O}(TN)$ after T iterations, which is the linear to the number of samples. Furthermore, the space complexity of our model can be expressed as $\mathcal{O}(VN(D_r + c))$, which also maintains the linear growth for N.

4 EXPERIMENT

4.1 EXPERIMENTAL SETTING

In this section, we validate the proposed model on eight widely adopted datasets in comparison with fourteen representative clustering models. All experiments are implemented in MATLAB R2018b and executed on a computing platform with an Intel(R) Xeon(R) Gold 6230R CPU and 256GB RAM. Due to space limitation, a subset of the experiments is shown here, details of the remaining experiments can be found in the appendix.

Table 1: Statistical information of eight datasets.

Dataset	Dimension	Sample	Cluster	View
ProteinFold	694/· · ·/694/694	694	27	12
100leaves	64/64/64	1660	100	3
COIL20	4096/3304/6750	1440	20	3
Caltech101	48/40/254/1984/512/928	9144	102	6
ALOI	77/13/64/125	10800	100	4
Animal	2689/2000/2001/2000	11673	20	4
CIFAR100	512/2048/1024	50000	100	3
YoutubeFace	64/512/64/647/838	101499	31	5

Datasets: For experiments, eight benchmark datasets under various scenarios are selected, including ProteinFold, COIL20, 100leaves, Caltech101, ALOI, Animal, CIFAR100 and YoutubeFace. More detailed statistics are shown in Table

[1] All the proofs can be found in the appendix.

Table 2: The ACC, NMI and ARI values of all methods on eight shuffled datasets.

Metric	Methods	ProteinFold	COIL20	100leaves	Caltech101	ALOI	Animal	CIFAR100	YoutubeFace
ACC(%)	t-SVD-MSC	31.41±1.48	32.01±4.84	65.99±2.55	26.76±1.27	40.67±1.92	OT	OM	OM
	LMVSC	23.42±1.34	62.71±5.96	56.93±1.15	12.12±0.31	59.05±2.15	12.24±0.35	67.46±1.05	11.76±0.26
	TLSpNM	30.59±1.88	73.90±1.73	68.75±1.29	26.24±1.49	33.20±0.57	17.83±1.38	OM	OM
	Orth-NTF	23.06±0.00	45.90±0.00	43.69±0.00	12.47±0.00	30.52±0.00	10.17±0.00	33.34±0.00	OT
	ASR-ETR	31.25±1.13	74.24±2.96	66.36±1.26	27.77±1.02	58.53±1.76	16.34±0.18	84.05±1.07	15.48±0.28
	NLRSC	31.82±2.44	73.54±0.21	72.78±1.54	25.66±1.19	OT	OT	OM	OM
	S^{2}MVTC	18.93±1.03	56.33±8.08	30.14±1.52	14.97±1.23	34.78±2.32	15.31±0.77	82.73±1.06	14.28±0.98
	STONE	32.54±1.34	75.92±3.40	71.58±1.85	25.87±0.79	60.82±1.94	14.25±0.07	90.23±2.38	13.32±0.30
	ALPC	31.01±1.71	60.24±1.40	60.85±2.22	16.97±0.49	15.02±0.20	15.18±0.49	89.22±2.40	14.48±0.36
	ESTMC	30.40±1.47	73.94±1.63	69.86±1.96	25.86±0.27	62.88±1.14	14.82±0.09	91.61±1.64	26.64±0.00
	SE-FTMC	29.48±1.75	70.60±3.67	67.99±1.79	24.92±0.84	67.55±1.59	15.65±0.78	89.34±1.38	14.87±0.75
	LMTC	30.92±1.70	72.94±2.10	71.38±1.63	27.54±1.40	68.97±1.19	17.49±0.67	94.27±1.45	25.66±1.32
	UVELRS	30.32±2.22	75.42±2.13	77.44±1.65	27.26±0.57	43.58±1.35	17.97±0.60	OM	OM
	CAL	32.82±1.24	59.37±2.77	75.68±1.99	22.41±0.69	50.88±1.29	16.67±0.47	OT	OT
	Ours	**32.97±1.62**	**76.05±0.36**	**80.26±1.86**	**28.48±0.90**	**73.39±1.17**	**18.68±0.75**	**96.12±1.04**	**28.00±0.99**
NMI(%)	t-SVD-MSC	38.54±1.05	39.15±5.57	84.47±0.95	38.09±2.19	60.31±1.24	OT	OM	OM
	LMVSC	29.47±1.03	74.68±2.97	77.61±0.51	25.47±0.32	76.03±0.83	9.00±0.13	89.07±1.24	5.20±0.08
	TLSpNM	36.23±1.31	81.43±1.59	84.27±0.58	46.87±0.24	49.34±0.34	13.80±0.62	OM	OM
	Orth-NTF	27.60±0.00	56.42±0.00	67.56±0.00	23.86±0.00	44.78±0.00	3.52±0.00	39.11±0.00	OT
	ASR-ETR	40.42±1.09	81.71±1.34	83.94±0.83	48.55±0.43	73.67±1.10	11.91±0.15	96.26±0.33	13.41±0.09
	NLRSC	38.83±1.69	81.29±0.39	87.50±0.66	49.05±0.38	OT	OT	OM	OM
	S^{2}MVTC	20.05±1.98	71.92±5.24	65.91±1.07	27.22±1.35	57.86±1.92	12.75±1.29	96.17±0.39	13.14±0.76
	STONE	41.11±1.84	**84.88±7.10**	86.74±1.04	46.41±0.67	76.84±0.79	10.01±0.08	97.68±0.68	10.93±0.27
	ALPC	**42.07±1.26**	72.72±0.61	81.06±0.83	33.74±0.38	38.13±0.10	8.73±0.14	97.35±0.72	12.82±0.09
	ESTMC	39.73±1.05	82.33±0.72	85.21±1.15	46.82±0.43	76.95±0.53	11.51±0.12	98.14±0.30	0.03±0.00
	SE-FTMC	39.44±0.87	81.28±1.70	84.36±0.56	46.35±0.46	78.00±0.74	12.08±0.46	97.70±0.29	11.05±0.70
	LMTC	39.47±1.09	82.15±0.98	85.73±0.87	49.00±0.40	79.66±0.51	13.40±0.31	98.70±0.28	22.50±1.13
	UVELRS	38.08±1.57	80.25±1.09	88.88±0.42	48.64±0.30	62.80±1.11	12.96±0.18	OM	OM
	CAL	41.52±2.13	65.38±2.04	89.49±0.54	44.11±0.33	64.95±0.89	13.55±0.41	OT	OT
	Ours	41.21±0.78	84.03±0.28	**91.97±0.58**	**49.50±0.22**	**82.86±0.47**	**14.86±0.57**	**99.10±0.29**	**24.60±0.71**
ARI(%)	t-SVD-MSC	10.99±0.99	9.34±3.89	55.48±2.45	11.00±2.36	14.37±1.44	OT	OM	OM
	LMVSC	8.68±1.02	55.44±5.47	43.83±1.26	8.13±0.66	41.51±2.46	2.39±0.09	38.26±7.95	0.96±0.04
	TLSpNM	8.97±1.32	66.19±2.43	57.81±0.95	16.66±1.29	18.32±0.38	5.20±0.28	OM	OM
	Orth-NTF	5.92±0.00	31.16±0.00	26.93±0.00	3.82±0.00	15.91±0.00	1.03±0.00	17.61±0.00	OT
	ASR-ETR	15.36±1.41	67.95±3.08	56.25±1.59	**21.49±1.63**	37.65±4.49	4.57±0.15	84.63±2.07	2.39±0.10
	NLRSC	12.48±2.17	66.37±0.45	63.92±2.03	18.84±2.15	OT	OT	OM	OM
	S^{2}MVTC	4.99±0.97	49.13±9.33	21.02±1.16	11.04±1.66	23.00±2.14	4.15±0.68	83.36±1.34	1.98±0.27
	STONE	14.51±1.41	71.41±3.94	63.18±2.36	20.65±1.46	50.00±1.54	3.29±0.03	90.02±3.41	**18.55±0.06**
	ALPC	15.55±1.20	54.07±1.14	50.65±2.03	13.13±0.35	7.86±0.03	2.97±0.18	88.27±4.30	2.20±0.09
	ESTMC	15.14±1.20	68.26±1.53	59.79±2.86	18.53±0.43	49.75±1.41	3.88±0.03	92.81±1.22	0.01±0.00
	SE-FTMC	14.25±1.51	66.24±3.24	58.06±1.40	18.52±1.20	37.56±2.19	4.39±0.36	90.95±1.02	1.96±0.20
	LMTC	14.60±1.69	67.88±2.13	61.66±1.92	20.67±1.37	53.32±1.39	5.48±0.28	94.99±0.17	5.74±0.32
	UVELRS	10.30±2.20	67.75±1.76	69.46±1.14	18.52±0.69	24.73±1.95	5.15±0.13	OM	OM
	CAL	14.34±1.68	30.28±3.71	68.21±0.25	16.23±1.00	12.36±1.88	3.40±0.52	OT	OT
	Ours	**16.30±1.43**	**71.44±0.50**	**75.73±1.87**	21.04±1.26	**53.73±2.21**	**6.11±0.32**	**96.39±1.26**	6.52±0.28

1. Since the samples in each dataset are originally ordered by category, we generate the shuffled setting by randomly arranging the sample order across all datasets.

Competitors: To demonstrate the the superiority of the proposed model, fourteen state-of-the-art multi-view clustering models are selected for comparison, including t-SVD-MSC Xie et al. (2018), LMVSC Kang et al. (2020), TLSpNM Guo et al. (2023), Orth-NTF Li et al. (2023), ASR-ETR Ji & Feng (2023), NLRSC Wang et al. (2024b), S^{2}MVTC Long et al. (2024), STONE Gu & Feng (2024), ALPC Chen et al. (2025), ESTMC Ji & Feng (2025a), SE-FTMC Ji & Feng (2025b), LMTC Liu et al. (2025), UVELRS Wang et al. (2026) and CAL Chen et al. (2026).

Evaluation metrics: Five widely-used metrics, i.e, accuracy (ACC), normalized mutual information (NMI), adjusted rank index (ARI), F-score and Precision Manning et al. (2008), are employed to comprehensively evaluate clustering performance. The higher scores indicate the better clustering quality for all metrics. To eliminate randomness, each method is executed multiple times and the average is presented. Additionally, in our tables, OM (out-of-memory) and OT (out-of-time) mean exceeding the maximum time limit and insufficient computing memory, respectively.

Parameter setting: For our proposed method, three parameters that need to be tuned, namely D_r, λ and C. Specifically, we set $D_r \in \{100, 200, \ldots, 4000\}$, $\lambda \in \{10^{-8}, 10^{-7}, \cdots, 10^{-1}\}$ and $C \in \{10^{-4}, 10^{-2}, \cdots, 10^{2}\}$. To ensure the fairness, we use the official source codes and the parameter settings recommended in the articles for all baselines to conduct experiments.

4.2 Performance evaluation

To validate the effectiveness of our method, we evaluate the clustering performance on eight datasets and compare the results with different methods across five metrics. Due to space limitations, only the results in terms of ACC, NMI, and ARI are reported in Table 2, while the results of F-score and

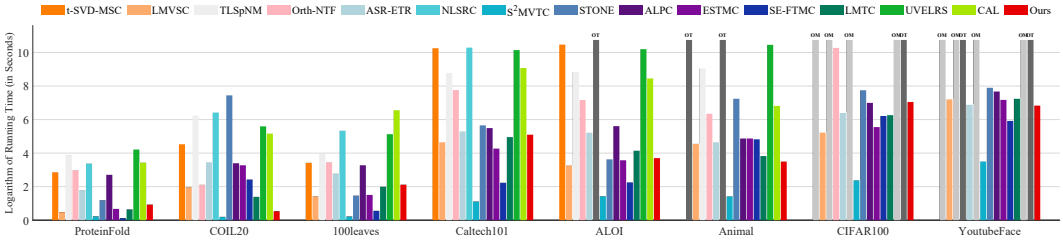


Figure 1: Runtime comparison of all methods on benchmark datasets. To provide a clearer presentation, execution times (in seconds) are shown on a logarithmic scale.

Table 3: The results (ACC%) of different methods on four datasets (ordered and shuffled cases).

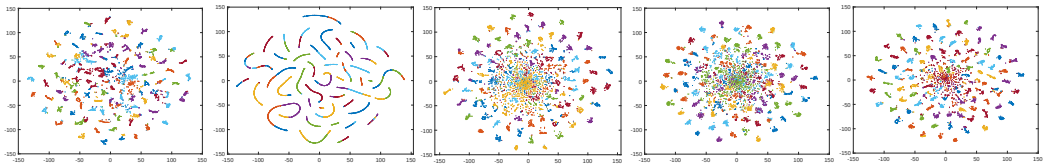
Methods	ProteinFold			COIL20			100leaves			Caltech101		
	Ordered	Shuffled	Gap	Ordered	Shuffled	Gap	Ordered	Shuffled	Gap	Ordered	Shuffled	Gap
t-SVD-MSC	48.12	31.41	16.71	77.11	32.01	45.10	94.14	65.99	28.15	47.39	26.76	20.63
Orth-NTF	36.17	23.06	13.11	85.32	45.90	39.42	70.44	43.69	26.75	23.24	12.47	10.77
STONE	51.18	32.54	18.64	92.29	75.92	16.37	88.56	71.58	16.98	65.11	25.87	39.24
S^2MVTC	35.46	18.93	16.53	75.35	56.33	19.02	44.00	30.14	13.86	44.52	14.97	29.55
SE-FTMC	30.62	29.48	1.14	74.09	70.60	3.49	70.19	67.99	2.20	27.19	24.92	2.27
LMTC	30.99	30.92	0.07	72.56	72.94	0.38	70.66	71.38	0.72	28.24	27.54	0.70
Ours	32.91	32.97	0.06	76.51	76.05	0.46	79.94	80.26	0.32	28.91	28.48	0.43

Precision are provided in the Appendix, where the best-performing and sub-optimal results are emphasized in **bold** and underline, respectively. From the results, we draw the following observations.

- Traditional tensor-based multiview clustering methods, such as t-SVD-MSC, TLSpNM, NLRSC and UVELRS, are generally inferior to several recent large-scale multiview clustering methods. For example, on the ALOI dataset, SE-FTMC improves the ACC over t-SVD-MSC by 26.88%, while our method achieves a more pronounced improvement of 32.72%. The reason is that these methods typically perform FFT along the sample dimension, making the resulting tensor modeling potentially dependent on the input sample arrangement. In contrast, recent methods that avoid such operations often exhibit more stable clustering performance. This observation further motivates removing the tensor rotation trick when exploiting high-order correlations of data.
- Compared with large-scale methods based on factorization or anchor learning, such as Orth-NTF and ASR-ETR, both LMTC and our method achieve better clustering results in most cases. For instance, on the COIL20 dataset, our method outperforms the Orth-NTF across performance metrics ACC and NMI with improvement of 30.15% and 27.61%, respectively. This is mainly because these methods inevitably omit informative relationships among samples through factorization or a limited number of anchors. This demonstrates the importance of retaining all-sample relationships for large-scale multiview clustering.
- More importantly, our method outperforms related competitors LMTC and SE-FTMC, particularly on the COIL20 and 100leaves datasets. This is because LMTC and SE-FTMC rely on linear similarity, while our model incorporates nonlinear similarity modeling to better characterize complex nonlinear structures of data while exploiting the intrinsic low-rank correlations across views, thereby yielding more informative latent representations.
- Overall, the proposed method achieves superior or highly competitive performance across almost all datasets and evaluation metrics. In particular, it obtains the highest ACC values on all eight datasets. These results consistently demonstrate that the proposed model can simultaneously achieve comprehensive sample relationship preservation, nonlinear structure characterization and high-order multiview correlation exploitation, leading to robust clustering performance.

4.3 Efficiency and Stability Analysis

Efficiency assessment: The running times of our method and competing methods on different datasets are intuitively presented in Figure 1 to demonstrate the computational efficiency and feasibility to large-scale data. It can be observed that our method maintains competitive efficiency



(a) LMVSC (b) ALPC (c) SE-FTMC (d) LMTC (e) Ours

Figure 2: Visualization results of five methods on ALOI datasets.

Table 4: Ablation study of the proposed model.

Methods	ProteinFold			COIL20			100leaves			ALOI		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
LMTC	30.92	39.47	14.60	72.94	82.15	67.88	71.38	85.73	61.66	47.39	79.66	53.32
NK-OLMC-v1	32.75	**41.27**	15.77	76.02	83.32	70.98	78.91	90.79	73.03	73.22	82.44	50.05
NK-OLMC	**32.97**	41.21	**16.30**	**76.05**	**84.03**	**71.44**	**80.26**	**91.97**	**75.73**	**73.39**	**82.86**	**53.73**

across datasets of different scales. Although some anchor-based and factorization-based methods, e.g., LMVSC and S^2MVTC, are faster on several datasets, our proposed method achieves superior performance, significantly outperforming these methods across all challenging datasets. In particular, as the data scale increases to CIFAR100 and YoutubeFace, several conventional tensor-based competitors incur substantially higher costs or encounter memory on large-scale datasets, whereas our method remains tractable. This demonstrates that the proposed method mitigates the high computational cost and shows scalability for large-scale multi-view clustering.

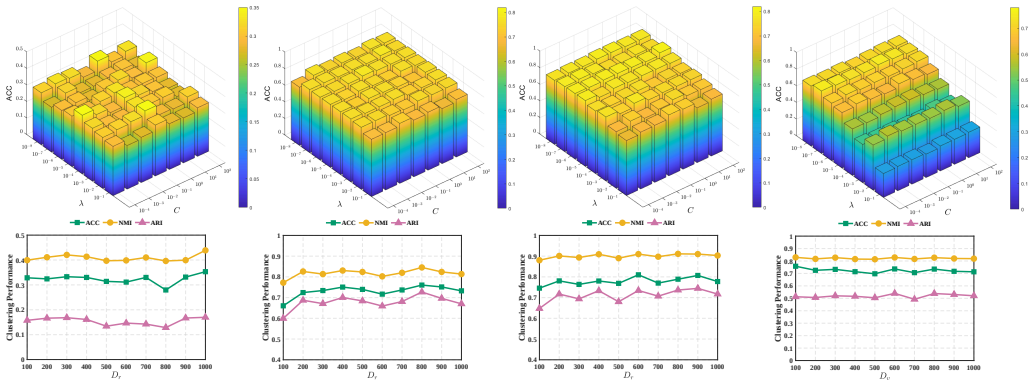
Stability analysis: Here, we investigate the stability of the proposed methods by comparing five methods in two settings, namely ordered data and randomly shuffled data prior to clustering. The clustering results are collected in Table 3. We can see that multi-view tensor clustering methods, e.g., t-SVD-MSC, Orth-NTF and STONE, exhibit substantial performance degradation after sample shuffling, especially on COIL20 dataset. This phenomenon indicates that tensor structures learned from these methods are highly sensitive to sample ordering and may benefit from the structural prior information induced by ordered data. In contrast, our method is nearly invariant to sample permutations, which yields only 0.06%, 0.46%, and 0.32% ACC gaps on ProteinFold, COIL20, and 100leaves, respectively. These results strongly demonstrate that our method is less dependent on sample ordering, thereby yielding stable clustering performance under arbitrary data arrangements.

4.4 Results visualization

To intuitively illustrate the quality of the learned representations, the spectral embeddings results generated by three comparison methods and our method on COIL20, 100leaves and ALOI datasets are visualized by t-SNE van der Maaten & Hinton (2008). As exhibited in Figure 2, the samples in Figure 2(a) - Figure 2(d) exhibit noticeable inter-class overlap, making them difficult to classify correctly. While our method yields a more compact structure with clearer separation and less ambiguity between different clusters as shown in Figure 2(d). These observations suggest that the learned representation better recovers coherent patterns with the discriminative cluster structure, which is consistent with the quantitative clustering results in Table 2.

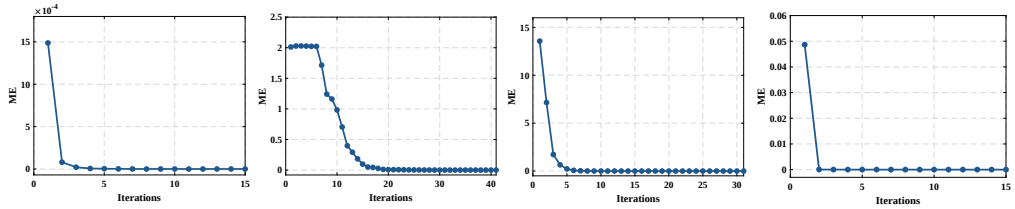
4.5 Model analysis

Ablation Study: The ablation experiments are conducted on four datasets to evaluate the effectiveness of our model. Specifically, LMTC is adopted as the baseline, while NK-OLMC-v1 extends LMTC by incorporating nonlinear kernel modeling. As shown in Table 4, NK-OLMC-v1 substantially improves upon LMTC across most evaluation metrics, while our proposed method further achieves the best results in most cases. This clearly demonstrates that nonlinear kernel effectively captures complex nonlinear sample relationship, while WTNN exploits the underlying low-rank structure, jointly yielding more discriminative representations for clustering.



(a) ProteinFold (b) COIL20 (c) 100leaves (d) ALOI

Figure 3: Sensitivity analysis of λ, C and D_r in terms of clustering performance over four datasets.



(a) ProteinFold (b) COIL20 (c) 100leaves (d) ALOI

Figure 4: The convergence curves on four datasets.

Parameter study: To investigate the impact of parameters λ, C and D_r on our proposed model, the extensive grid search is conducted to analyze their effectiveness and sensitivity. As depicted in Figure 3, the clustering performance remains relatively stable over a broad range of λ and C on ProteinFold, COIL20, and 100leaves datasets, indicating that the proposed method is insensitive to parameter variations. It is worth noting that the variation of D_r results in moderate fluctuations across four datasets. These results demonstrate that our method is stable to parameter perturbations and can achieve satisfactory clustering performance over relatively broad parameter ranges.

Convergence analysis: To further validate the stability of the proposed model, we investigate the iterative behaviors of maximum error (ME) value on four datasets, where ME $= |\max\{obj^k - obj^{k-1}|/|obj^k|, \|\mathcal{G} - \mathcal{H}\|_\infty\}$. As shown in Figure 4, the ME values rapidly drop to 0 within 5 iterations and remain stable thereafter on ProteinFold, 100leaves and ALOI datasets, while slightly more iterations are required on COIL20 and Caltech101 datasets. This provides clear numerical evidence for the efficient and reliable convergence behavior of the proposed algorithm.

5 Conclusion

In this paper, we presented a novel model, termed NK-OLMC, for order-invariant large-scale multi-view clustering. On the one hand, it avoids the reliance of high-order tensor modeling on sequential priors embedded in sorted samples, thereby improving robustness to sample permutations. On the other hand, the model effectively handles large-scale data with complex nonlinear structures in linear time. Furthermore, an efficient alternating optimization algorithm with linear computational complexity and theoretical convergence guarantee is developed to solve the proposed model. Extensive experiments on various large-scale datasets consistently demonstrate the effectiveness and stability of the proposed model.

REFERENCES

Haim Avron, Vikas Sindhwani, Jiyan Yang, and Michael W. Mahoney. Quasi-monte carlo feature maps for shift-invariant kernels. *Journal of Machine Learning Research*, 17(120):1–38, 2016.

Robert G. Bartle and Donald R. Sherbert. *Introduction to Real Analysis*. Wiley, New York, 3rd edition, 2000.

Yawei Chen, Huibing Wang, Jinjia Peng, and Yang Wang. Anchor learning with potential cluster constraints for multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 15939–15947, 2025.

Yongyong Chen, Xiaolin Xiao, and Yicong Zhou. Jointly learning kernel representation tensor and affinity matrix for multi-view clustering. *IEEE Transactions on Multimedia*, 22(8):1985–1997, 2020.

Zhe Chen, Fanhui Meng, Tianyang Xu, and Xiao-Jun Wu. Cluster-aware anchor learning for multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17714–17723, 2026.

Yao Fu, Dong Hu, and Zhi Wang. Nonconvex low-rank tensor representation with deep priors for multiview subspace clustering. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306, 2026a.

Yao Fu, Zhi Wang, Dong Hu, Tao Jia, Chao Gao, and Zhen Wang. A unified framework with capped tensor norm minimization for multiview subspace learning. *IEEE Transactions on Knowledge and Data Engineering*, 38(9):5607–5621, 2026b.

Hang Gao, Wenjun Shen, Rui Li, Cheng Liu, and Si Wu. Collaborative structure-preserved missing data imputation for single-cell rna-seq clustering. *IEEE Transactions on Computational Biology and Bioinformatics*, 21(5):1480–1491, 2024.

Quanxue Gao, Wei Xia, Zhizhen Wan, De-Yan Xie, and Pu Zhang. Tensor-svd based graph learning for multi-view subspace clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 3930–3937, 2020.

Shengxiang Gao, Zhengtao Yu, Taisong Jin, and Ming Yin. Multi-view low-rank matrix factorization using multiple manifold regularization. *Neurocomputing*, 335:143–152, 2019.

Zhibin Gu and Songhe Feng. From dictionary to tensor: a scalable multi-view subspace clustering framework with triple information enhancement. In *Advances in Neural Information Processing Systems*, volume 37, pp. 103545–103573, 2024.

Jipeng Guo, Yanfeng Sun, Junbin Gao, Yongli Hu, and Baocai Yin. Logarithmic schatten-p norm minimization for tensorial multi-view subspace clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3396–3410, 2023.

Jintian Ji and Songhe Feng. Anchor structure regularization induced multi-view subspace clustering via enhanced tensor rank minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19286–19295, 2023.

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

Jintian Ji and Songhe Feng. Anchors bring stability and efficiency: Fast tensorial multi-view clustering on shuffled datasets. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1676–1685, 2025b.

Xiaodong Jia, Xiao-Yuan Jing, Qixing Sun, Songcan Chen, Bo Du, and David Zhang. Human collective intelligence inspired multi-view representation learning — enabling view communication by simulating human communication mechanism. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7412–7429, 2023.

Zhao Kang, Wangtao Zhou, Zhitong Zhao, Junming Shao, Meng Han, and Zenglin Xu. Large-scale multi-view subspace clustering in linear time. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 4412–4419, 2020.

Misha E Kilmer, Karen Braman, Ning Hao, and Randy C Hoover. Third-order tensors as operators on matrices: A theoretical and computational framework with applications in imaging. *SIAM Journal on Matrix Analysis and Applications*, 34(1):148–172, 2013.

Adrian S Lewis and Hristo S Sendov. Nonsmooth analysis of singular values. part i: Theory. *Set-Valued Analysis*, 13:213–241, 2005.

Jing Li, Quanxue Gao, QIANQIAN WANG, Ming Yang, and Wei Xia. Orthogonal non-negative tensor factorization based multi-view clustering. In *Advances in Neural Information Processing Systems*, volume 36, pp. 18186–18202, 2023.

An-An Liu, Yu-Ting Su, Wei-Zhi Nie, and Mohan Kankanhalli. Hierarchical clustering multi-task learning for joint human action grouping and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(1):102–114, 2017.

Jiyuan Liu, Xinwang Liu, Yuexiang Yang, Qing Liao, and Yuanqing Xia. Contrastive multi-view kernel learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9552–9566, 2023.

Jiyuan Liu, Xinwang Liu, Chuankun Li, Xinhang Wan, Hao Tan, Yi Zhang, Weixuan Liang, Qian Qu, Yu Feng, Renxiang Guan, and Ke Liang. Large-scale multi-view tensor clustering with implicit linear kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20727–20736, 2025.

Xinwang Liu. Incomplete multiple kernel alignment maximization for clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3):1412–1424, 2024.

Zhuo Liu, Dong Hu, and Zhi Wang. Nonconvex low-rank tensor representation with self-enhanced structure penalty for multi-view subspace clustering. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Zhen Long, Qiyuan Wang, Yazhou Ren, Yipeng Liu, and Ce Zhu. S2MVTC: A simple yet efficient scalable multi-view tensor clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24088–24097, 2024.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

Feiping Nie, Rui Zhang, and Xuelong Li. A generalized power iteration method for solving quadratic problem on the stiefel manifold. *Science China Information Sciences*, 60(11), 2017.

Chao-Yang Pang, Wei Hu, Ben-Qiong Hu, Ying Shi, Charles R. Vanderburg, Jack T. Rogers, and Xudong Huang. A special local clustering algorithm for identifying the genes associated with alzheimer's disease. *IEEE Transactions on Nanobioscience*, 9(1):44–50, 2010.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, pp. 1177–1184, 2007.

Yiwen Shan, Dong Hu, and Zhi Wang. A novel truncated norm regularization method for multi-channel color image denoising. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(9):8427–8441, 2024.

Biyun Sheng, Jun Li, Fu Xiao, Qun Li, Wankou Yang, and Junwei Han. Discriminative multi-view subspace feature learning for action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(12):4591–4600, 2020.

Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

Yusen Tan, Dong Hu, Yao Fu, Chunming Yang, and Zhi Wang. Generalized nonconvex low-rank approximation and sparse regularizer for hyperspectral image denoising. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–17, 2025.

Maximilian B. Toller, Bernhard C. Geiger, and Roman Kern. Cluster purging: Efficient outlier detection based on rate-distortion theory. *IEEE Transactions on Knowledge and Data Engineering*, 35(2):1270–1282, 2023.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Chang-Dong Wang, Man-Sheng Chen, Ling Huang, Jian-Huang Lai, and Philip S. Yu. Smoothness regularized multiview subspace clustering with kernel learning. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):5047–5060, 2021.

Jun Wang, Zhenglai Li, Chang Tang, Suyuan Liu, Xinhang Wan, and Xinwang Liu. Multiple kernel clustering with adaptive multi-scale partition selection. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6641–6652, 2024a.

Yapeng Wang, Quanxue Gao, Fangfang Li, Yu Yun, and Ming Yang. Unified view extraction with low-rankness and smoothness fusion for multi-view subspace clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26634–26642, 2026.

Zhi Wang, Yu Liu, Xin Luo, Jianjun Wang, Chao Gao, Dezhong Peng, and Wu Chen. Large-scale affine matrix rank minimization with a novel nonconvex regularizer. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9):4661–4675, 2022.

Zhi Wang, Zhuo Liu, Dong Hu, and Tao Jia. Nonconvex multiview subspace clustering framework with efficient method designs and theoretical analysis. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 5162–5170, 2024b.

Zhi Wang, Dong Hu, Zhuo Liu, Chao Gao, and Zhen Wang. Iteratively capped reweighting norm minimization with global convergence guarantee for low-rank matrix learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1923–1940, 2025.

Wei Xia, Xiangdong Zhang, Quanxue Gao, Xiaochuang Shu, Jungong Han, and Xinbo Gao. Multiview subspace clustering by an enhanced tensor nuclear norm. *IEEE Transactions on Cybernetics*, 52(9):8962–8975, 2022.

Yuan Xie, Dacheng Tao, Wensheng Zhang, Yan Liu, Lei Zhang, and Yanyun Qu. On unifying multiview self-representations for clustering by tensor multi-rank minimization. *International Journal of Computer Vision*, 126:1157–1179, 2018.

Ben Yang, Xuetao Zhang, Feiping Nie, Fei Wang, Weizhong Yu, and Rong Wang. Fast multi-view clustering via nonnegative and orthogonal factorization. *IEEE Transactions on Image Processing*, 30:2575–2586, 2021.

Handong Zhao, Hongfu Liu, Zhengming Ding, and Yun Fu. Consensus regularized multi-view outlier detection. *IEEE Transactions on Image Processing*, 27(1):236–248, 2018.

A Appendix

The notations, comprehensive elaboration for the optimization steps, proof of convergence and remaining experimental results are given in the appendix.

A.1 LLM usage

We use LLMs exclusively for language refinement to improve fluency and precision, without applying them to any other aspect of the work.

A.2 NOTATIONS AND PRELIMINARIES

We summarize the notations and definitions frequently used in this paper. For a third-order tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, its v-th frontal slice is denoted by $\mathcal{X}_v = \mathcal{X}(:,:,v)$, where $v = 1, \ldots, n_3$, and $\bar{\mathcal{X}}$ denotes fast Fourier transform (FFT) along the tube fiber, i.e., $\bar{\mathcal{X}} = \text{fft}(\mathcal{X}, [], 3)$ and $\mathcal{X} = \text{ifft}(\bar{\mathcal{X}}, [], 3)$. The identity tensor $\mathcal{I} \in \mathbb{R}^{n_1 \times n_1 \times n_3}$ has the identity matrix as its first frontal slice and zero matrices as all remaining frontal slices. The transpose of tensor $\mathcal{X}$ is $\mathcal{X}^\top \in \mathbb{R}^{n_2 \times n_1 \times n_3}$, which can be obtained by transposing each frontal slice and reversing the order of the transposed slices from the 2 to n_3. A tensor is f-diagonal when each frontal slice is a diagonal matrix. Now, Given tensors $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ and $\mathcal{Y} \in \mathbb{R}^{n_2 \times n_4 \times n_3}$, we introduce the following tensor operations. The block vectorization matrix is constructed by $\mathbf{bvec}(\mathcal{X}) = [\mathcal{X}_1; \ldots; \mathcal{X}_{n_3}]$, and the inverse operation is $\mathbf{bvfold}(\mathbf{bvec}(\mathcal{X})) = \mathcal{X}$. The block circular matrix is defined as

$$\mathbf{bcirc}(\mathcal{X}) = \begin{bmatrix} \mathcal{X}_1 & \mathcal{X}_{n_3} & \ldots & \mathcal{X}_2 \\ \mathcal{X}_2 & \mathcal{X}_1 & \ldots & \mathcal{X}_3 \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{X}_{n_3} & \mathcal{X}_{n_3-1} & \ldots & \mathcal{X}_1 \end{bmatrix}. \tag{10}$$

The t-product between $\mathcal{X}$ and $\mathcal{Y}$ can be defined as $\mathcal{Z} = \mathcal{X} * \mathcal{Y} = \mathbf{bvfold}\{\mathbf{bcirc}(\mathcal{X})\mathbf{bvec}(\mathcal{Y})\}$. Tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_1 \times n_3}$ is orthogonal when it satisfies $\mathcal{X}^T * \mathcal{X} = \mathcal{X} * \mathcal{X}^T = \mathcal{I}$. Based on the above operations, we have the following definitions.

Definition 1 (t-SVD Kilmer et al. (2013))**.** *Given a tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, the tensor Singular Value Decomposition (t-SVD) of $\mathcal{X}$ can be given by*

$$\mathcal{X} = \mathcal{U} * \mathcal{S} * \mathcal{V}^T, \tag{11}$$

where $\mathcal{U} \in \mathbb{R}^{n_1 \times n_1 \times n_3}$ and $\mathcal{V} \in \mathbb{R}^{n_2 \times n_2 \times n_3}$ are orthogonal, $\mathcal{S} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is an f-diagonal tensor.

Definition 2 (WTNN Gao et al. (2020))**.** *Given a tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, the weighted tensor nuclear norm of $\mathcal{X}$ is defined by*

$$\|\mathcal{X}\|_{\omega,\circledast} = \sum_{i=1}^{min(n_1,n_2)} \sum_{v=1}^{n_3} \omega_i * \bar{\mathcal{S}}_v(i,i). \tag{12}$$

where f-diagonal tensor $\bar{\mathcal{S}}$ is obtained by t-SVD of tensor $\mathcal{X}$ in Fourier domain. The notations used in this work are outlined in Table 5.

A.3 DETAILED OPTIMIZATION PROCEDURE

The corresponding unconstrained form of optimization problem (8) can be obtain as

$$\begin{aligned} \mathcal{L}(\mathbf{H}_v, \mathcal{G}, \mathcal{W}) = -\frac{1}{2} \sum_{v=1}^{V} \text{Tr}\left(\mathbf{Z}_v^\top \mathbf{Z}_v \mathbf{H}_v^\top \mathbf{H}_v\right) + \lambda \|\mathcal{G}\|_{\omega,\circledast} + \langle \mathcal{W}, \mathcal{G} - \mathcal{H} \rangle + \frac{\rho}{2} \|\mathcal{G} - \mathcal{H}\|_F^2 \\ s.t.\ \ \mathbf{H}_v \mathbf{H}_v^\top = \mathbf{I}, \end{aligned} \tag{13}$$

where $\mathcal{W}$ is Lagrangian multiplier, ρ denotes a positive penalty parameter to control convergence.

$\mathbf{H}^{(v)}$-subproblem. Fixing the other variables, according to $\Phi_v^{-1}(\mathcal{H}) = \mathbf{H}_v$, $\Phi_v^{-1}(\mathcal{G}) = \mathbf{G}_v$ and $\Phi_v^{-1}(\mathcal{W}) = \mathbf{W}_v$, the optimization subproblem for variable $\mathbf{H}_v$ can be written as

$$\begin{aligned} \mathbf{H}_v = \arg\min_{\mathbf{H}_v} -\frac{1}{2} \text{Tr}\left(\mathbf{Z}_v^\top \mathbf{Z}_v \mathbf{H}_v^\top \mathbf{H}_v\right) + \langle \mathbf{W}_v, \mathbf{G}_v - \mathbf{H}_v \rangle + \frac{\rho}{2} \|\mathbf{G}_v - \mathbf{H}_v\|_F^2, \\ s.t.\ \mathbf{H}_v \mathbf{H}_v^\top = \mathbf{I}, \end{aligned} \tag{14}$$

By expanding the quadratic penalty term, exploiting the orthogonality constraint $\mathbf{H}_v \mathbf{H}_v^\top = \mathbf{I}$ and omitting the terms independent of $\mathbf{H}_v$, problem (14) can be equivalently transformed into

$$\begin{aligned} \mathbf{H}_v = \arg\max_{\mathbf{H}_v} \ \text{Tr}\left(\mathbf{H}_v \mathbf{Z}_v^\top \mathbf{Z}_v \mathbf{H}_v^\top\right) + 2\,\text{Tr}\left(\mathbf{H}_v^\top (\rho \mathbf{G}_v + \mathbf{W}_v)\right), \\ s.t.\ \mathbf{H}_v \mathbf{H}_v^\top = \mathbf{I}, \end{aligned} \tag{15}$$

Table 5: The Summarization of Notations

Notation	Description
x, $\mathbf{x}$, $\mathbf{X}$, $\mathcal{X}$	Scalar, vector, matrix, tensor
v, V	The v-th view and number of views
d_v, N	The dimension of v-th view and number of sample
c, D_r	The number of cluster and dimension of random Fourier features
$\mathbf{Z}_v$	The random Fourier features of the v-th view
$\mathbf{H}_v$	The latent representation of the v-th view
$\mathbf{K}_v$	The nonlinear kernel matrix of the v-th view
$\mathbf{X}(i,j)$	The (i,j)-th element of matrix $\mathbf{X}$
$\mathbf{X}^T$	The transpose of matrix $\mathbf{X}$
$\sigma_i(X)$	The v-th leading singular value of matrix $\mathbf{X}$
Tr($\mathbf{X}$)	The trace of matrix $\mathbf{X}$
$sign(x)$	The sign function about saclar x
$\mathcal{X}_k$	The k-th frontal slice of the tensor $\mathcal{X}$
$\mathbf{X}_v$	The feature matrix $\mathbf{X}$ of the v-the view
$\mathbf{I}_n$	The $n \times n$ identity matrix
$\|\mathbf{X}\|_* = \sum_i \sigma_i(\mathbf{X})$	The nuclear norm of matrix $\mathbf{X}$
$\|\mathbf{X}\|_F = (\text{Tr}(\mathbf{X}^\top \mathbf{X}))^{1/2}$	The Frobenius norm of $\mathbf{X}$
$\|\mathbf{X}\|_\infty = max_{ij}\|X(i,j)\|$	The infinity norm of matrix $\mathbf{X}$

The problem (15) is a quadratic optimization problem on the Stiefel manifold Nie et al. (2017). We employ the generalized power iteration (GPI) strategy to solve this subproblem. With the other variables fixed, given the current iterate $\mathbf{H}_v^t$, we construct

$$\mathbf{M}_v^t = \mathbf{H}_v^t \mathbf{Z}_v^\top \mathbf{Z}_v + \rho_k \mathbf{G}_v + \mathbf{W}_v \tag{16}$$

and update

$$\mathbf{H}_v^{t+1} = \underset{\mathbf{H}_v \mathbf{H}_v^\top = \mathbf{I}_c}{\arg\max} \text{Tr}\left(\mathbf{H}_v^\top \mathbf{M}_v^t\right). \tag{17}$$

Let the compact SVD of $\mathbf{M}_v^t$ be

$$\mathbf{M}_v^t = \mathbf{U}_v^t \mathbf{\Sigma}_v^t (\mathbf{V}_v^t)^\top. \tag{18}$$

Then, problem (17) admits the closed-form solution

$$\mathbf{H}_v^{t+1} = \mathbf{U}_v^t (\mathbf{V}_v^t)^\top, \tag{19}$$

where $\mathbf{U}_v^t$ and $\mathbf{V}_v^t$ denote the left and right singular vector matrices of $\mathbf{M}_v^t$, respectively. The above iterations are repeated until convergence, and the converged solution is taken as $\mathbf{H}_v^*$.

$\mathcal{G}$-subproblem. Fixing the other variables, the optimization subproblem for variable $\mathcal{G}$ can be reduced to

$$\begin{aligned} \mathcal{G} &= \arg\min_{\mathcal{G}} \lambda \|\mathcal{G}\|_{\omega,\circledast} + \langle \mathcal{W}, \mathcal{G} - \mathcal{H} \rangle + \frac{\rho}{2} \|\mathcal{G} - \mathcal{H}\|_F^2 \\ &= \arg\min_{\mathcal{G}} \frac{\lambda}{\rho} \|\mathcal{G}\|_{\omega,\circledast} + \frac{1}{2} \left\| \mathcal{G} - \left(\mathcal{H} - \frac{\mathcal{W}}{\rho} \right) \right\|_F^2. \end{aligned} \tag{20}$$

The closed-form solution of (20) can be obtained by the following lemma.

Lemma 1. *Suppose $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ with t-SVD $\mathcal{X} = \mathcal{U} * \mathcal{S} * \mathcal{V}^T$ and $\tau > 0$. Then, the optimal solution to*

$$\arg\min_{\mathcal{G}} \frac{1}{2} \|\mathcal{G} - \mathcal{X}\|_F^2 + \tau \|\mathcal{G}\|_{\omega,\circledast} \tag{21}$$

is given by

$$\mathcal{G}^* = \Gamma_{\tau * \omega}(\mathcal{X}) = \mathcal{U} * \text{ifft}(\mathcal{P}_{\tau * \omega}(\bar{\mathcal{X}}), [], 3) * \mathcal{V}^T, \tag{22}$$

where $\text{ifft}(\mathcal{P}_{\tau*\omega}(\bar{\mathcal{X}}), [], 3) \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is an f-diagonal tensor, whose the i-th diagonal elements of the v-th frontal slice can be found by $\mathcal{P}_{\tau*\omega}(\bar{X}_v)(i,i) = sign(\bar{S}_v(i,i)) \max(\bar{S}_v(i,i) - \tau * \omega_i, 0)$. Here, considering the influence of the size of matrix $G_v \in \mathbb{R}^{n_1 \times n_2}$, we utilize the weighting strategy $\omega_i = \frac{C\sqrt{n_1 n_2}}{\bar{S}_v(i,i) + \varepsilon}$, where C is a positive parameter.

Algorithm 1 The algorithm for solving problem (8).

Input: Random Fourier feature matrices $\{\mathbf{Z}_v\}_{v=1}^V$ generated from the multi-view data $\{\mathbf{X}_v\}_{v=1}^V$, parameter λ, C and cluster number c.

Initialized: Initialize $\{\mathbf{H}_v\}_{v=1}^V$ such taht $\mathbf{H}_v\mathbf{H}_v^\top = \mathbf{I}$, $\mathcal{G} = \mathcal{W} = \mathbf{0}$, ρ=10^{-5}, ρ_{max}=10^{10}, ϵ_1=10^{-5}, ϵ_2=10^{-7}.

1: **while** not converge **do**
2: Update $\{\mathbf{H}_v^{k+1}\}_{v=1}^V$ by the GPI procedure in Eqs. (16)–(19);
3: Update $\mathcal{G}$ by (23);
4: Update Lagrange multiplier $\mathcal{W}$, penalty parameter ρ by (24) - (25);
5: Compute the objective value obj^k with model (8);
6: Check the convergence conditions:
$|obj^k - obj^{k-1}|/|obj^k| < \epsilon_1$ and $\|\mathcal{G} - \mathcal{H}\|_\infty < \epsilon_2$;
7: **end while**

Output: Perform k-means on the concatenation of $\{\mathbf{H}_v\}_{v=1}^V$ to obtain final clustering results.

According to Lemma 1, the optimal solution of problem (20) can be obtained by

$$\mathcal{G}^* = \Gamma_{\frac{\lambda}{\rho}*\omega}(\mathcal{H} - \frac{\mathcal{W}}{\rho}). \tag{23}$$

$\mathcal{W}$-subproblem and ρ-subproblem. Finally, multiplier $\mathcal{W}$ and the penalty parameter ρ can be updated by

$$\begin{cases} \mathcal{W} &= \mathcal{W} + \rho(\mathcal{G} - \mathcal{H}), & (24) \\ \rho &= \min(\eta\rho, \rho_{max}), & (25) \end{cases}$$

where $\eta > 1$ is used to accelerate convergence and ρ_{max} denotes the upper bound of ρ. Given the space limitations, the pseudocode of the comprehensive optimization algorithm is outlined in the supplementary material.

A.4 Proofs of Theorem 1

Before proving Theorem 1, a basic lemmas is firstly presented.

Lemma 2. *Lewis & Sendov (2005) Let $F(\cdot) : \mathbb{R}^{m\times n} \to \mathbb{R}$ be represented as $F(X) = f \circ \sigma(X) = f(\sigma_1(X), \sigma_2(X), .., \sigma_r(X))$, where the function $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$ is differentiable, $X = Udiag(\sigma_1, \sigma_2, ..., \sigma_r)V^T$ is SVD of $X \in \mathbb{R}^{m\times n}$, and $r = min(m, n)$. Then the gradient of $F(X)$ at matrix X is*

$$\frac{\partial F(X)}{\partial X} = \partial(f \circ \sigma)(X) = Udiag(\Delta)V^T, \tag{26}$$

where $\Delta = \partial f(\sigma(X)) = (\frac{\partial f(\sigma_1(X))}{\partial X}, ..., \frac{\partial f(\sigma_r(X))}{\partial X})$ and $\frac{\partial f(\sigma_i(X))}{\partial X} = \partial|\sigma_i(X)|f'[\sigma_i(X)]$, $i = 1, ..., r$.

Then, we can prove the Theorem 1.

Proof. 1) Firstly, we prove the boundedness of sequence $\{\mathcal{T}^k\}_{k=1}^\infty$. According to the orthogonality constraint $\mathbf{H}_v^k(\mathbf{H}_v^k)^\top = \mathbf{I}$ in updating rule of $\mathbf{H}_v$, we have

$$\|\mathbf{H}_v^k\|_F^2 = \mathrm{Tr}\left(\mathbf{H}_v^k(\mathbf{H}_v^k)^\top\right) = \mathrm{Tr}(\mathbf{I}_c) = c. \tag{27}$$

Hence, $\{\mathbf{H}_v^k\}$ is bounded for every v, and consequently $\|\mathcal{H}^k\|_F = \sqrt{Vc}$.

Next, we verify the boundedness of $\{\mathcal{W}^k\}$. Based on the updating rule of $\mathcal{G}^{k+1}$ in Algorithm 1 and the first-order optimal condition, we can obtain

$$\mathbf{0} \in \lambda\partial\|\mathcal{G}^{k+1}\|_{\omega,\circledast} + \mathcal{W}^k + \rho_k(\mathcal{G}^{k+1} - \mathcal{H}^{k+1}). \tag{28}$$

According to the multiplier update $\mathcal{W}^{k+1} = \mathcal{W}^k + \rho_k(\mathcal{G}^{k+1} - \mathcal{H}^{k+1})$, we can deduce

$$\mathcal{W}^{k+1} \in -\lambda\partial\|\mathcal{G}^{k+1}\|_{\omega,\circledast}. \tag{29}$$

Assume that the t-SVD of tensor $\mathcal{G} \in \mathbb{R}^{c\times N\times V}$ is $\mathcal{U} * \mathcal{S} * \mathcal{V}^T$ and $\bar{\mathcal{G}} = \text{fft}(\mathcal{G}, [], 3)$. By utilizing Lemma 1, we have

$$\begin{aligned}\|\partial\|\mathcal{G}^{k+1}\|_{\omega,\circledast}\|_F^2 =& \left\|\frac{1}{V}\mathcal{U} * \text{ifft}(\partial f(\bar{\mathcal{S}}), [], 3) * \mathcal{V}^T\right\|_F^2 \\ =& \frac{1}{V^2}\|\text{ifft}(\partial f(\bar{\mathcal{S}}), [], 3)\|_F^2 \\ =& \frac{1}{V^2}\frac{1}{V}\|\partial f(\bar{\mathcal{S}})\|_F^2 \\ =& \frac{1}{V^3}\sum_{j=1}^{V}\sum_{i=1}^{c}[\partial f(\bar{\mathcal{S}}_j(i,i))]^2 \\ \leq& \frac{1}{V^3}\sum_{j=1}^{V}\sum_{i=1}^{c}(\omega_j)^2 \\ \leq& \frac{1}{V^3}\sum_{j=1}^{V}\sum_{i=1}^{c}\left(\frac{C\sqrt{cN}}{\epsilon}\right)^2 = \frac{C^2cN}{V^2\epsilon^2} < +\infty,\end{aligned} \tag{30}$$

where $c \ll N$ and the inequalities are derived from the fact that $|\partial f(\bar{\mathcal{S}}_j(i,i))| \leq \omega_j = \frac{C\sqrt{cN}}{\bar{\mathcal{S}}(i,i)+\epsilon} \leq \frac{C\sqrt{cN}}{\epsilon}$. Thus, it can be concluded that $\partial\|\mathcal{G}^{k+1}\|_{\omega,\circledast}$ and $\{\mathcal{W}^k\}$ are bounded.

According to the fact that $\rho_k = \eta\rho_{k-1}$ and the boundedness of $\{\mathcal{W}^k\}$, we have $\rho_k > \rho_0 > 0$ and

$$\begin{aligned}\|\mathcal{G}^{k+1}\|_F \leq& \left\|\mathcal{H}^{k+1}\right\|_F + \frac{\|\mathcal{W}^{k+1} - \mathcal{W}^k\|_F}{\rho_k} \\ \leq& \sqrt{Vc} + \frac{R}{\rho_0} < +\infty,\end{aligned} \tag{31}$$

where R is the upper boundary of $\|\mathcal{W}^{k+1} - \mathcal{W}^k\|_F$ for all k. Therefore, $\{\mathcal{G}^k\}$ is also bounded.

In summary, we can infer the boundedness of the sequence $\{\mathcal{T}^k\}_{k=1}^{\infty}$.

2)Based on the Bolzano-Weierstrass theorem Bartle & Sherbert (2000), the bounded sequence $\{\mathcal{T}^k\}_{k=1}^{\infty}$ exists at least one convergent subsequence. Without loss of generality, we assume that the subsequence $\{\mathcal{T}^k\}_{k=1}^{\infty}$ is represented by itself and converges to an accumulation point, which can be denoted as $\mathcal{T}^*$. Thus, we have

$$\lim_{k\to\infty}(\mathbf{H}_v^k, \mathcal{G}^k; \mathcal{W}^k) = (\mathbf{H}_v^*, \mathcal{G}^*; \mathcal{W}^*). \tag{32}$$

According to the updating rules (24) and the boundedness of $\{\mathcal{W}^k\}$, we get

$$\lim_{k\to\infty}(\mathcal{G}^{k+1} - \mathcal{H}^{k+1}) = \lim_{k\to\infty}\frac{\mathcal{W}^{k+1} - \mathcal{W}^k}{\rho_k} = 0 \tag{33}$$

due to the fact that $\lim_{k\to\infty}\rho_k(\mathcal{W}^{k+1} - \mathcal{W}^k) = \mathbf{0}$. Then, the feasibility constraint holds

$$\mathcal{G}^* - \mathcal{H}^* = 0. \tag{34}$$

Meanwhile, since the orthogonality constraint $\mathbf{H}_v^k(\mathbf{H}_v^k)^\top = \mathbf{I}$ holds for each k, we have

$$\mathbf{H}_v^*(\mathbf{H}_v^*)^\top = \mathbf{I}, \ \ v = 1, 2, \ldots, V. \tag{35}$$

Therefore, the accumulation point satisfies all the primal feasibility constraints.

Next, $\mathcal{G}^{k+1}$ satisfy the first-order optimality condition. When $k \to \infty$, we can drive that

$$\mathbf{0} \in \lambda\partial\|\mathcal{G}^*\|_{\omega,\circledast} + \mathcal{W}^*. \tag{36}$$

Moreover, we establish the stationarity condition with respect to $\mathbf{H}_v$. For simplicity, denote $\mathbf{A}_v = \mathbf{Z}_v^\top \mathbf{Z}_v$. Since $\mathbf{H}_v^{k,t+1}$ is the optimal solution of the inner GPI surrogate problem, we introduce a symmetric multiplier $\bar{\Lambda}_v^{k,t+1}$ and consider

$$\mathcal{Q}_v^{k,t}(\mathbf{H}_v, \bar{\Lambda}_v) = -\operatorname{Tr}\left(\mathbf{H}_v^\top \mathbf{M}_v^{k,t}\right) + \frac{1}{2}\operatorname{Tr}\left[\bar{\Lambda}_v\left(\mathbf{H}_v\mathbf{H}_v^\top - \mathbf{I}_c\right)\right]. \tag{37}$$

The first-order optimality condition with respect to $\mathbf{H}_v$ gives

$$-\mathbf{M}_v^{k,t} + \bar{\Lambda}_v^{k,t+1}\mathbf{H}_v^{k,t+1} = \mathbf{0}, \tag{38}$$

where $\bar{\Lambda}_v^{k,t+1} = (\bar{\Lambda}_v^{k,t+1})^\top$. Substituting $\mathbf{M}_v^{k,t} = \mathbf{H}_v^{k,t}\mathbf{A}_v + \rho_k\mathbf{G}_v^k + \mathbf{W}_v^k$ yields

$$\mathbf{H}_v^{k,t}\mathbf{A}_v + \rho_k\mathbf{G}_v^k + \mathbf{W}_v^k = \bar{\Lambda}_v^{k,t+1}\mathbf{H}_v^{k,t+1}. \tag{39}$$

Since the inner GPI procedure is performed until convergence, $\mathbf{H}_v^{k,t}$ and $\mathbf{H}_v^{k,t+1}$ converge to the same fixed point $\mathbf{H}_v^{k+1}$. Passing to the limit $t \to \infty$, we obtain

$$\mathbf{H}_v^{k+1}\mathbf{A}_v + \rho_k\mathbf{G}_v^k + \mathbf{W}_v^k = \bar{\Lambda}_v^{k+1}\mathbf{H}_v^{k+1}. \tag{40}$$

Define

$$\Lambda_v^{k+1} = \bar{\Lambda}_v^{k+1} - \rho_k\mathbf{I}_c. \tag{41}$$

Then Λ_v^{k+1} is symmetric, and (40) yields

$$-\mathbf{H}_v^{k+1}\mathbf{A}_v - \mathbf{W}_v^k + \rho_k(\mathbf{H}_v^{k+1} - \mathbf{G}_v^k) + \Lambda_v^{k+1}\mathbf{H}_v^{k+1} = \mathbf{0}. \tag{42}$$

Multiplying both sides of (40) from the right by $(\mathbf{H}_v^{k+1})^\top$ and using $\mathbf{H}_v^{k+1}(\mathbf{H}_v^{k+1})^\top = \mathbf{I}_c$, we have

$$\bar{\Lambda}_v^{k+1} = \left(\mathbf{H}_v^{k+1}\mathbf{A}_v + \rho_k\mathbf{G}_v^k + \mathbf{W}_v^k\right)(\mathbf{H}_v^{k+1})^\top. \tag{43}$$

Therefore,

$$\begin{aligned}\|\bar{\Lambda}_v^{k+1}\|_F &\le \|\mathbf{A}_v\|_F + \rho_k\|\mathbf{G}_v^k\|_F + \|\mathbf{W}_v^k\|_F \\ &\le \|\mathbf{A}_v\|_F + \rho_{\max}\|\mathbf{G}_v^k\|_F + \|\mathbf{W}_v^k\|_F.\end{aligned} \tag{44}$$

Since $\{\mathbf{G}_v^k\}$ and $\{\mathbf{W}_v^k\}$ are bounded, $\{\bar{\Lambda}_v^k\}$ is bounded. Furthermore,

$$\|\Lambda_v^{k+1}\|_F \le \|\bar{\Lambda}_v^{k+1}\|_F + \rho_{\max}\sqrt{c}, \tag{45}$$

and hence $\{\Lambda_v^k\}$ is also bounded.

By the assumption $\lim_{k\to\infty}\rho_k(\mathbf{H}_v^{k+1} - \mathbf{H}_v^k) = 0$ and $\rho_k \ge \rho_0 > 0$, we have

$$\mathbf{H}_v^{k+1} - \mathbf{H}_v^k \to \mathbf{0},\ k \to \infty. \tag{46}$$

Together with $\mathbf{G}_v^k - \mathbf{H}_v^k \to \mathbf{0}$ by (33), this gives

$$\mathbf{H}_v^{k+1} - \mathbf{G}_v^k = (\mathbf{H}_v^{k+1} - \mathbf{H}_v^k) + (\mathbf{H}_v^k - \mathbf{G}_v^k) \to \mathbf{0},\ k \to \infty. \tag{47}$$

Since $\rho_k \le \rho_{\max}$,

$$\rho_k(\mathbf{H}_v^{k+1} - \mathbf{G}_v^k) \to \mathbf{0},\ k \to \infty. \tag{48}$$

Let $\{\mathcal{T}^{k_j}\}$ be a subsequence converging to an arbitrary accumulation point $\mathcal{T}^*$. Since $\{\Lambda_v^{k_j+1}\}$ is bounded, there exists a further subsequence, still denoted by $\{k_j\}$, such that

$$\Lambda_v^{k_j+1} \to \Lambda_v^*,\ k \to \infty, \qquad \Lambda_v^* = (\Lambda_v^*)^\top.$$

Moreover, $\mathbf{H}_v^{k_j+1} - \mathbf{H}_v^{k_j} \to \mathbf{0}$ implies $\mathbf{H}_v^{k_j+1} \to \mathbf{H}_v^*$ when $k \to \infty$. Taking the limit in (42) and using (48) yields

$$-\mathbf{H}_v^*\mathbf{Z}_v^\top\mathbf{Z}_v - \mathbf{W}_v^* + \Lambda_v^*\mathbf{H}_v^* = \mathbf{0}. \tag{49}$$

Therefore, Eqs. (34), (35), (36), and (49) jointly yield

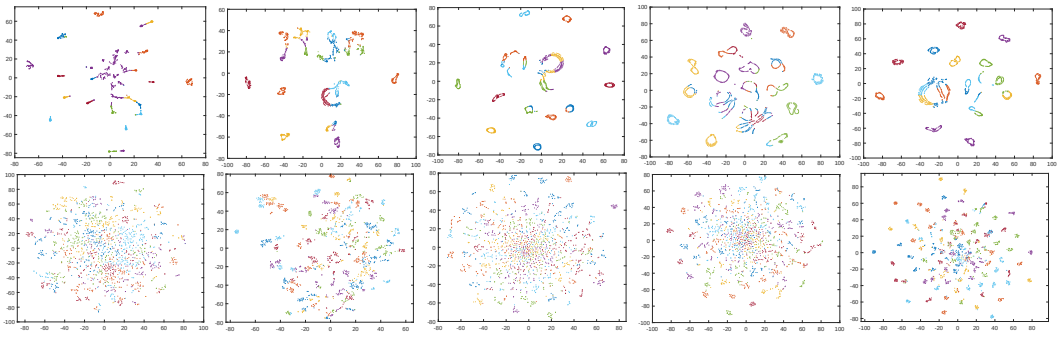
$$\begin{cases}\mathcal{G}^* = \mathcal{H}^*, \\ \mathbf{H}_v^*(\mathbf{H}_v^*)^\top = \mathbf{I}_c, \quad v = 1, \dots, V, \\ \mathbf{0} \in \lambda\partial\|\mathcal{G}^*\|_{\omega,\circledast} + \mathcal{W}^*, \\ -\mathbf{H}_v^*\mathbf{Z}_v^\top\mathbf{Z}_v - \mathbf{W}_v^* + \Lambda_v^*\mathbf{H}_v^* = \mathbf{0}, \quad v = 1, \dots, V,\end{cases} \tag{50}$$

with $\Lambda_v^* = (\Lambda_v^*)^\top$. These conditions are exactly the first-order KKT conditions of problem (8). Consequently, any accumulation point $\mathcal{T}^*$ of the sequence generated by Algorithm 1 is a stationary point of problem (8).

□

Table 6: The F-score and Precision values of all methods on eight shuffled datasets (mean ± standard deviation).

Metric	Methods	ProteinFold	COIL20	100leaves	Caltech101	ALOI	Animal	CIFAR100	YoutubeFace
F-score(%)	t-SVD-MSC	17.02±0.81	17.18±3.29	55.99±2.41	14.77±2.09	15.81±1.39	OT	OM	OM
	LMVSC	5.53±1.45	57.89±5.09	44.42±1.25	4.93±0.47	48.06±1.90	8.66±0.18	52.21±7.70	0.78±0.05
	TLSpNM	15.19±1.03	67.95±2.32	58.24±0.94	19.03±1.26	19.33±0.37	10.92±0.25	OM	OM
	Orth-NTF	9.93±0.00	34.97±0.00	27.64±0.00	2.34±0.00	16.77±0.00	6.45±0.00	17.13±0.00	OT
	ASR-ETR	19.15±1.32	69.58±2.91	56.71±1.57	**23.09±1.58**	38.49±4.38	9.85±0.14	84.80±2.04	7.22±0.12
	NLRSC	17.78±1.75	68.11±0.41	64.30±2.00	20.83±2.07	OT	OT	OM	OM
	S^2MVTC	13.26±0.79	52.14±8.49	22.20±1.13	13.62±1.62	24.05±2.10	11.02±0.60	83.55±1.61	8.32±0.59
	STONE	18.59±1.33	**72.87±3.72**	63.55±2.33	22.24±1.44	50.58±1.52	9.03±0.03	90.13±3.37	7.23±0.06
	ALPC	19.40±1.19	56.50±1.08	51.17±2.01	15.14±0.37	9.03±0.02	10.22±0.16	88.40±4.25	7.51±0.10
	ESTMC	18.81±1.15	69.87±1.44	60.19±2.83	19.81±0.44	50.29±1.39	9.10±0.03	92.88±1.21	**16.42±0.00**
	SE-FTMC	17.94±1.46	67.96±3.05	58.48±1.39	19.79±1.18	38.42±2.14	9.58±3.43	91.04±1.01	7.39±0.27
	LMTC	18.42±1.62	69.49±2.02	62.04±1.90	21.94±1.36	53.84±1.37	10.47±0.28	95.04±1.15	11.01±0.29
	UVELRS	16.10±1.94	69.39±1.66	69.76±1.13	19.84±0.69	25.81±1.89	10.40±0.12	OM	OM
	CAL	19.35±1.36	35.24±3.25	68.54±2.44	17.93±0.95	13.89±1.82	**11.42±0.44**	OT	OT
	Ours	**19.97±1.37**	**72.87±0.47**	**75.97±1.85**	22.30±1.26	**54.28±2.18**	11.34±0.31	**96.43±1.25**	11.42±0.31
Precision(%)	t-SVD-MSC	12.74±0.89	9.88±2.23	46.63±3.16	9.88±1.91	9.33±0.99	OT	OM	OM
	LMVSC	12.37±1.18	64.99±3.80	51.05±1.34	7.36±0.46	**48.66±1.86**	9.59±0.49	52.88±7.56	10.11±0.24
	TLSpNM	11.31±1.22	64.57±2.02	53.42±1.05	18.80±1.19	15.51±0.45	9.85±0.24	OM	OM
	Orth-NTF	11.82±0.00	31.23±0.00	26.74±0.00	9.50±0.00	16.10±0.00	6.90±0.00	18.13±0.00	OT
	ASR-ETR	21.70±1.70	67.61±3.57	50.33±1.54	31.83±2.67	28.62±4.83	10.42±0.17	76.75±3.11	13.06±0.14
	NLRSC	15.02±2.42	65.24±0.97	58.10±2.36	23.83±2.07	OT	OT	OM	OM
	S^2MVTC	7.57±0.56	45.04±11.04	14.04±0.92	13.25±1.62	17.39±1.80	9.07±0.78	75.33±2.33	11.53±0.41
	STONE	19.77±1.54	70.19±4.54	58.62±2.30	31.21±1.97	46.15±1.75	8.91±0.03	84.51±5.42	11.81±0.12
	ALPC	21.54±1.11	52.74±1.02	45.56±2.29	18.53±0.74	7.01±0.10	8.01±0.14	81.81±6.64	12.37±0.14
	ESTMC	22.02±1.33	67.86±2.15	56.49±3.36	35.37±0.59	45.77±1.85	9.82±0.03	88.97±1.69	8.95±0.00
	SE-FTMC	21.12±1.66	65.96±3.93	54.98±1.45	35.68±2.20	27.92±2.23	10.33±0.37	86.33±1.38	11.94±0.27
	LMTC	20.87±1.81	68.66±2.30	58.67±1.83	38.79±2.16	47.68±1.90	11.30±0.26	92.09±1.65	17.63±0.58
	UVELRS	12.69±1.84	71.15±0.98	**73.10±1.30**	13.94±0.50	17.91±1.91	9.84±1.07	OM	OM
	CAL	16.87±2.00	24.59±3.31	61.34±3.32	24.94±2.06	7.88±1.17	8.08±0.44	OT	OT
	Ours	**23.06±1.57**	**72.19±0.59**	71.23±2.50	**39.57±1.88**	45.09±3.06	**11.92±0.31**	**94.26±2.18**	**19.47±0.41**



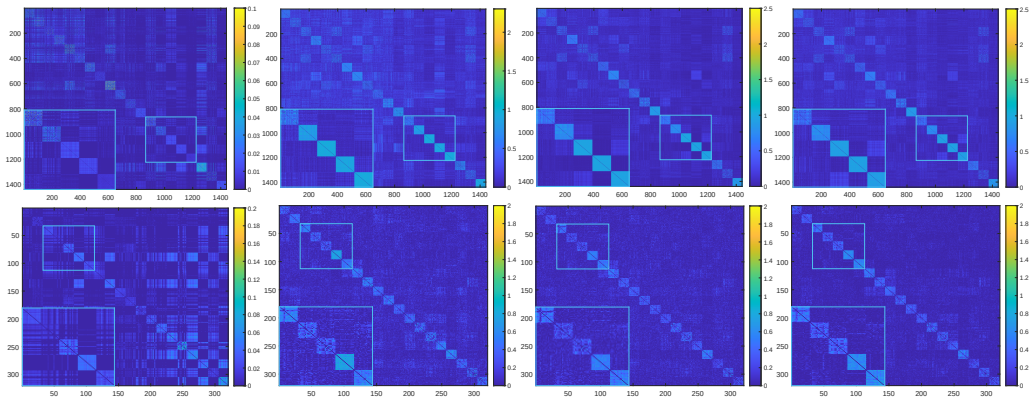
(a) LMVSC (b) ALPC (c) SE-FTMC (d) LMTC (e) Ours

Figure 5: Visualization results of five methods on COIL20 (the first row) and 100leaves (the second row) datasets.

A.5 Supplementary Experiments

Here, we present additional experimental results that were not included in the main text.

Clustering Results. Table 6 presents the F-score and Precision values on eight datasets. Consistent with the results reported in the main text, our method achieves the best or second-best F-score in most cases. It also successfully completes all experiments, whereas several competitors encounter OT or OM issues on large-scale datasets. These results further corroborate the effectiveness and scalability of our method across different evaluation metrics.



(a) S^2MVTC (b) SE-FTMC (c) LMTC (d) Ours

Figure 6: Comparison of affinity matrices of four methods on COIL20 dataset (the first row) and 100leaves dataset (the second row).

Table 7: Time consumptions (in seconds) of different methods on eight benchmark datasets.

Dataset	t-SVD-MSC	LMVSC	TLSpNM	Orth-NTF	ASR-ETR	NLRSC	S^2MVTC	STONE	ALPC	ESTMC	SE-FTMC	LMTC	UVELRS	CAL	Ours
ProteinFold	16.38	0.63	45.61	18.43	5.02	28.16	0.26	2.29	13.95	0.96	**0.13**	0.90	66.53	29.92	1.54
COIL20	91.52	6.13	509.81	7.23	30.71	612.80	**0.22**	1709.60	28.83	25.45	10.31	3.02	268.53	173.26	0.71
100leaves	29.59	3.14	52.82	30.26	14.93	205.78	**0.25**	3.32	25.45	3.51	0.75	6.34	168.06	703.79	7.33
Caltech101	28421.28	103.04	6515.91	2345.10	198.41	29460.00	**2.05**	284.75	242.00	70.39	8.32	140.93	25457.21	8753.09	162.69
ALOI	35301.23	25.09	6821.51	1290.90	182.00	*OT*	**3.18**	36.63	271.69	34.64	8.52	62.01	26902.64	4687.80	39.52
Animal	OT	93.83	8480.30	568.26	102.99	*OT*	**3.14**	1396.80	129.50	129.51	123.61	44.83	34825.94	911.37	31.96
CIFAR100	*OM*	184.40	*OM*	29011.00	596.37	*OM*	**9.77**	2321.50	1091.50	256.93	498.42	524.27	OM	*OT*	1152.70
YoutubeFace	*OM*	1345.30	*OM*	*OT*	972.57	*OM*	**32.07**	2685.50	2150.40	1305.20	371.38	1393.40	OM	*OT*	926.70

Visualization of Embedded Representation. The additional visualization results generated by five methods on COIL20 and 100leaves datasets are shown in Figure 5. It can be seen that the samples in Figure 5(a) - Figure 5(d) are disordered and scattered, while our method in Figure 5(e) produces clearly separable clusters. This also clearly demonstrates that our proposed method is capable of recovering coherent patterns characterized by separable cluster structures.

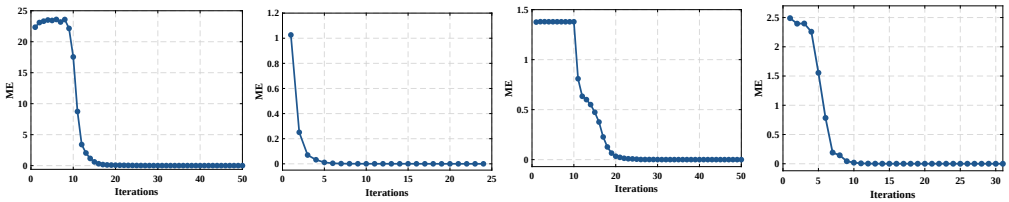
Affinity Matrices. The affinity graphs constructed by our method and three comparison methods on COIL20 and 100leaves datasets are shown in Figure 6. We can observe that for affinity matrices in Figure 6(a) - Figure 6(c), the block diagonal structures are indistinct, while our method exhibits a more pronounced block diagonal structure. This results indicates that our method effectively reveals more discriminative subspace information and the intrinsic clustering structure of the multiview data.

Running Time Comparison. We present the detail time consumption for fifteen methods on eight datasets in Table 7. we can see that the proposed method exhibits favorable computational efficiency across datasets of different scales. Compared with conventional tensor-based methods such as t-SVD-MSC, TLSpNM, and Orth-NTF, our method requires substantially less computation time on most datasets. The difference becomes particularly evident on relatively large datasets. Compared with recent large-scale approaches, the proposed method remains computationally competitive. In particular, our method achieves 0.71 seconds on COIL20 dataset and 31.96 seconds on Animal dataset. Compared with closely related LMTC, our method is faster on Animal and YoutubeFace datasets, with running times of 31.96, and 926.70 seconds, respectively. These observations are consistent with the theoretical complexity analysis and verify the scalability of the proposed method.

Parameter Settings. The specific parameter ranges of all compared methods are listed in Table 8.For fairness, the parameters of each baseline are tuned within the search ranges recommended in the corresponding papers or official implementations, while parameters related to the number

Table 8: Parameters selection range of comparison methods

Method	Parameter 1	Parameter 2	Parameter 3	Parameter 4	Parameter 5
t-SVD-MSC	[0.1, 2]	–	–	–	–
LMVSC	{c, 50, 100}	[0.001,10]	–	–	–
TLSpNM	[0.0001, 0.0009]	–	–	–	–
Orth-NTF	[0.5, 1]	[0.1, 1]	[100, 5000]	–	–
ASR-ETR	[2c, 8c]	$[10^{-6}, 1]$	$[10^{-4}, 1]$	–	–
NLRSC	$[10^{-6}, 10]$	$[10^{-6}, 10]$	–	–	–
S^{2}MVTC	$[10^{-4}, 10^{2}]$	$[10^{-4}, 10^{-2}]$	[500,3000]	[10, 40]	–
STONE	[0.1, 5]	[c, 7c]	$[10^{-5}, 10]$	$[10^{-5}, 10]$	$[10^{-5}, 10]$
ALPC	$[10^{-4}, 1]$	$[10^{-2}, 10^{2}]$	–	–	–
ESTMC	$[10^{-6}, 1]$	$[10^{-6}, 1]$	$[10^{-6}, 1]$	$[10^{-2}, 1]$	–
SE-FTMC	[0.001, 5]	[c, 8c]	–	–	–
LMTC	$[10^{-10}, 10^{5}]$	–	–	–	–
UVELRS	$[10^{-2}, 10^{2}]$	$[10^{-2}, 10^{2}]$	[0.1, 1]	–	–
CAL	$[10^{-4}, 10^{4}]$	$[10^{-4}, 10^{4}]$	$[10^{-4}, 10^{4}]$	–	–



(a) Caltech101 (b) Animal (c) CIFAR100 (d) YoutubeFace

Figure 7: The convergence curves on four datasets.

of clusters are adjusted accordingly based on c. The best results obtained within these predefined ranges are reported for comparison.

Convergence Analysis. To further validate the stability of the proposed model, we investigate the iterative behaviors of ME value on the remaining four datasets. As presented in Figure 7, it illustrates the empirical convergence behavior of the proposed optimization algorithm on the four remaining datasets. It can be observed that the ME rapidly decreases and eventually approaches zero on all datasets. These results empirically demonstrate the fast and stable convergence behavior of the proposed optimization algorithm.