LEARNING TO CONSTRUCT THE LEARNING SPACE FOR SCALABLE MULTIPLE KERNEL CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard complete multiple kernel clustering (MKC) is infeasible for large-scale data, requiring $O(n^2)$ storage and $O(n^3)$ decomposition of V $n \times n$ kernel matrices. Scalable MKC methods approximate kernel K_v as $KH_vH_v^\top$ ($KH_v \in \mathbb{R}^{n\times m}$, $m \ll n$), reducing storage to $O(nm)$ and decomposition to $O(m^3)$. However, existing methods freeze $KH_vH_v^\top$, so as view weights γ_v update, the combined kernel $\sum_v \gamma_v^2 K_v$ diverges from the frozen approximation, degrading clustering performance. To address this limitation, we propose a Learning to Construct the Learning Space for Scalable Multiple Kernel Clustering (LCLS-SMKC) method, which is a space-plasticity method that jointly optimizes the space-shaping parameters and the consensus embedding. Specifically, multi-scale leverage scores select diverse landmarks to cover the kernel space, yielding a more accurate per-view approximation KH_v. Subsequently, A self-learning dual-mode switch modulates the kernel space to be adaptive to view disparity rather than frozen. For small disparity, an enlarged-disparity space is selected to retain cross-view diversity. For large disparity, a purified space suppresses noise via adaptive joint feature denoising, embedding alignment, and view consistency mining. Furthermore, the dimensionality variation coefficient is embedded into kernel-space parameters, and an error-driven feedback loop updates these space parameters to minimize approximation error to enable adaptive learning. Finally, theoretical analysis establishes guarantees on approximation error, Lyapunov stability, linear convergence, and generalization, while experiments on six diverse datasets confirm consistent superiority over state-of-the-art scalable MKC methods in performance.

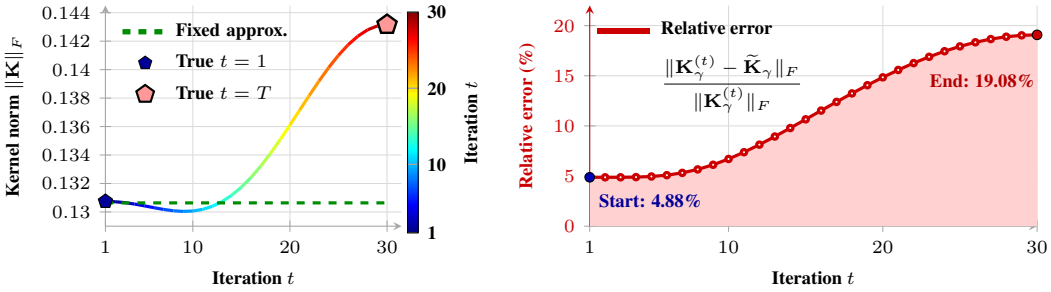
1 INTRODUCTION

Standard complete multiple kernel clustering (MKC) methods Tang et al. (2025a); Zhang et al. (2026) fuse multiple kernels to capture complementary information across views, yielding a shared space that integrates heterogeneous views to learn a consensus embedding $\mathbf{S}_\star \in \mathbb{R}^{n\times k}$ for more accurate unsupervised clustering. MKC methods apply to diverse tasks such as multi-source image classification, gene expression analysis, and social network analysis. However, classical MKC Xu et al. (2026); Wang et al. (2026) relies on complete kernel matrices whose size scales quadratically with the number of samples, incurring $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ memory complexity and rendering them infeasible for large-scale data. To address this limitation, kernel approximation strategies trade a controllable loss in accuracy for substantial gains in computational efficiency Liu et al. (2022); Wang et al. (2019); Liu (2021).

Simple multiple kernel K-means establishes a consistency theory for kernel weights, motivating a scalable extension based on singular value decomposition (SVD) that reduces complexity from $O(n^3)$ to $O(n)$ Yin et al. (2023); Paul et al. (2023); Kumar & Chakraborty (2026). However, the SVD-based extension relies on uniform sampling, which limits the approximation quality of the kernel matrix. Anchor-based formulations He & Zhang (2026) address the uniform-sampling limitation by deriving non-asymptotic bounds on the consensus kernel and clustering indicator matrix. However, anchor-based formulations fix the landmarks before optimization, so the landmark-induced approximation cannot be efficiently updated. Incremental sampling schemes Feng et al. (2025) address the fixed-landmark limitation via adaptive sampling, reducing approximation error while maintaining a consensus structure. A separate line shows that performing MKC on a smaller base kernel matrix preserves kernel weights within a multiplicative factor Liang et al. (2025). Tensor-

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.

based coordination Lu (2026b) reconciles consensus and individual partitions to advance scalability. Theoretical studies Lin et al. (2026); Chen et al. (2026) establish entrywise eigenvector perturbation bounds and bridge random feature mappings with kernel power K-means. Despite the above advances, all existing methods fix a low-rank approximation of the combined kernel before optimization, so the kernel space it represents cannot adapt to updated view weights, causing progressive deviation from the true kernel and degraded clustering performance.



(a) Kernel norm. (b) Approximation error.

Figure 1: Illustration of the frozen kernel space limitation on the Caltech101-7 dataset. (a) The kernel norm $\|\mathbf{K}_\gamma\|_F$ of the true combined kernel evolves over iterations, while that of the fixed approximation $\|\widetilde{\mathbf{K}}_\gamma\|_F$ remains constant. (b) The relative approximation error $\|\mathbf{K}_\gamma - \widetilde{\mathbf{K}}_\gamma\|_F / \|\mathbf{K}_\gamma\|_F$ grows 4.88%–19.08%, showing the fixed approximation cannot adapt to the evolving true kernel.

Existing scalable MKC methods replace the combined kernel $\mathbf{K}_\gamma = \sum_v \gamma_v^2 \mathbf{K}_v$ with a low-rank approximation $\widetilde{\mathbf{K}}_\gamma$, which is computed only once and then kept fixed during optimization. Since $\mathbf{K}_\gamma$ defines the kernel space, fixing $\widetilde{\mathbf{K}}_\gamma$ freezes this space and prevents it from adapting as the view weights γ_v change. As shown in Fig. 1(a), the energy of the true kernel, $\|\mathbf{K}_\gamma\|_F$, evolves over iterations, while the energy of the fixed approximation, $\|\widetilde{\mathbf{K}}_\gamma\|_F$, stays constant. As a result, Fig. 1(b) shows that the relative error $\|\mathbf{K}_\gamma - \widetilde{\mathbf{K}}_\gamma\|_F / \|\mathbf{K}_\gamma\|_F$ grows from 4.88% to 19.08%, i.e., nearly $4\times$. The consensus embedding $\mathbf{S}_\star$ learned from this fixed approximation therefore reflects the initial kernel space rather than the current one, which degrades clustering accuracy. This raises a fundamental question: if the bottleneck is the fixed kernel space rather than the consensus embedding $\mathbf{S}_\star$, why not jointly optimize the space-shaping parameters and the consensus embedding $\mathbf{S}_\star$?

Existing methods optimize the consensus embedding $\mathbf{S}_\star$ within a fixed kernel space, but never learn the kernel space itself. Since $\mathbf{S}_\star$ is learned from this space, fixing it means the view embeddings $\{\mathbf{S}_v\}_{v=1}^V$ can only capture the similarities encoded in the initial space and cannot follow the evolving combined kernel $\sum_{v=1}^V \gamma_v^2 \mathbf{K}_v$. As the view weights γ_v update, the fixed approximation $\mathbf{K}\mathbf{H}_v\mathbf{H}_v^\top$ cannot follow, so $\mathbf{S}_\star$ drifts away from the true kernel and clustering degrades. The root cause is that the kernel space is frozen, not learned. To address this, we propose LCLS-SMKC, a space-plasticity method that constructs the landmark approximation offline and adapts kernel-space parameters via closed-loop optimization with the view embeddings $\{\mathbf{S}_v\}_{v=1}^V$ and the consensus embedding $\mathbf{S}_\star$. As illustrated in Fig. 2, three components realize this idea. Multi-scale leverage scores select diverse landmarks $\mathcal{A}$ to cover the kernel space, yielding a per-view approximation $\mathbf{K}\mathbf{H}_v = \mathbf{X}_v\mathbf{X}_{v,\mathcal{A}}^\top$ with $\|\mathbf{K}\mathbf{H}_v\mathbf{H}_v^\top - \mathbf{K}_v\|_F^2 \ll \|\mathbf{K}_v\|_F^2$, thus providing a reliable initialization. A self-learning dual-mode switch then modulates the kernel space according to the dimensionality ratio $\rho_d = \max_v d_v / \min_v d_v$: for $\rho_d \le \rho_{\text{thr}}$, an enlarged-disparity space preserves cross-view diversity; for $\rho_d > \rho_{\text{thr}}$, a purified space suppresses noise. Finally, the dimensionality variation coefficient $\delta = \text{std}(\{d_v\})/\text{mean}(\{d_v\})$ is embedded into the space-shaping parameters, and an error-driven feedback loop updates them to minimize the approximation error $\sum_{v=1}^V \|\beta_v \mathbf{K}\mathbf{H}_v - \mathbf{S}_v\mathbf{H}_v\|_F^2$, enabling self-correction. Theoretical analysis establishes guarantees on approximation error, Lyapunov stability, linear convergence, and generalization.

The main contributions are summarized as follows: **(1) Multi-scale leverage score sampling:** The LCLS-SMKC selects diverse landmarks that cover the kernel space, yielding a per-view approximation much closer to the true kernel than uniform sampling. Since the kernel-space parameters are adapted from this approximation during optimization, a more accurate starting point directly

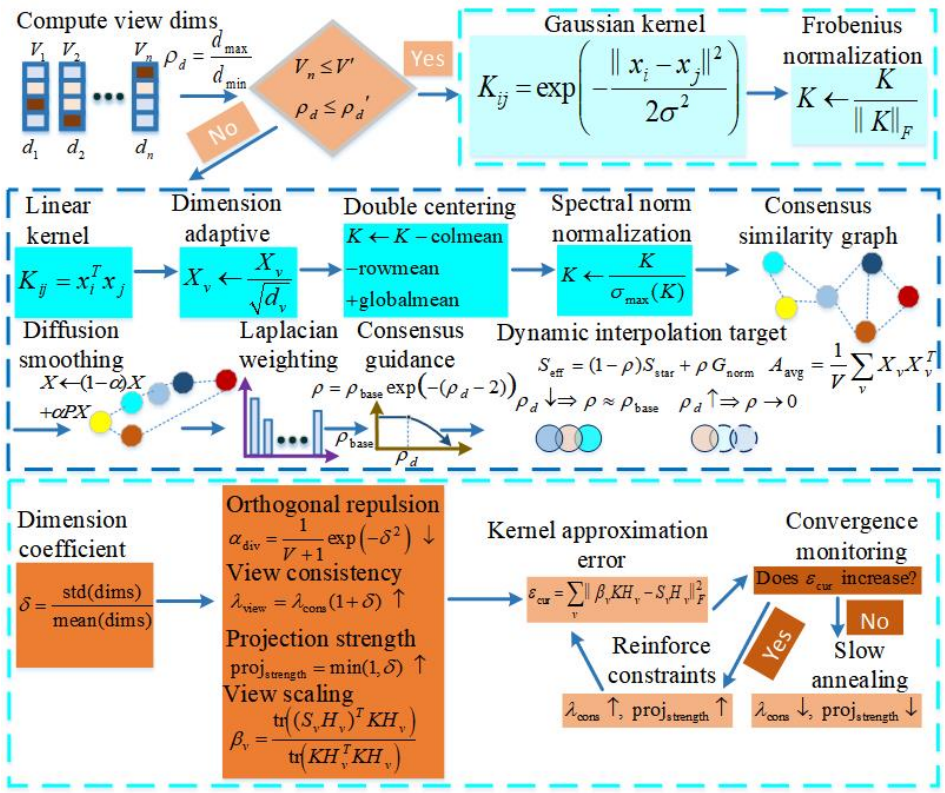


Figure 2: Illustration of the proposed framework LCLS-SMKC method.

improves the quality of the adapted space. **(2) Adaptive dual-mode modulation:** When view disparity is small, an enlarged-disparity space is selected to promote cross-view diversity; when view disparity is large, a purified space is activated to suppress noise, thereby promoting view consistency. This enables the kernel space to adapt to the data structure rather than remain frozen, which is the core of learning to construct the learning space. **(3) Error-driven feedback mechanism:** The dimensionality variation coefficient is embedded into the parameters that shape the kernel space, and a closed-loop feedback updates these parameters to minimize the approximation error. This enables the kernel space to self-correct during optimization, ensuring that the learned space remains aligned with the evolving true kernel. This enables the kernel space to self-correct during optimization. **(4) Theoretical and experimental analysis:** Theoretical analysis establishes guarantees on approximation error, Lyapunov stability, linear convergence, and generalization. Experiments on six benchmark datasets confirm consistent superiority over state-of-the-art scalable MKC methods in both clustering quality and scalability.

2 Related Work

This section reviews the most relevant works, which are categorized into complete kernel methods and approximation kernel methods. **Complete kernel methods** Tang et al. (2023); Yi Zhang et al. (2022) construct a complete kernel matrix for each view and fuse them into a consensus kernel for clustering. These methods emphasize kernel-target alignment, representation structure adaptation, and multi-strategy fusion, collectively advancing multi-view clustering. Fair kernel k-means (FairMKKM) introduces a fairness regularization term into the kernel k-means algorithm to balance sample sensitivity Zhou et al. (2024). Simple multiple kernel k-means (SimpleMKKM) reformulates a supervised kernel alignment criterion into a tractable minimization problem Liu (2023). Fast approximate MKKM (FAMKKM) addresses the computational cost of complete kernel methods via approximated partition matrices Wang et al. (2024b). Smooth MKKM via underlying graph filtering (SMKKM-UGF) jointly updates the graph filter and the kernel space to maintain representation

smoothness Yang et al. (2025). Unified weighted localized SMKKM (UWL-SMKKM) incorporates local kernel alignment that emphasizes reliable pairwise samples, with SMKKM as a special case Liu et al. (2021). Sample adaptive localized SMKKM (SAL-SMKKM) adaptively adjusts per-sample alignment weights through a tri-level optimization, reformulated as differentiable reduced gradient descent Liu et al. (2025b). MKC with adaptive multi-scale partition selection (MPS) constructs multi-scale partitions and similarity graphs per kernel. MPS dynamically selects high-quality kernels for fusion, avoiding noise from complete-kernel fusion and the neglect of structural information Wang et al. (2024a). One-step late fusion MKC via tensorized adaptive multi-scale partition fusion (OLMTMF) tensorizes multi-scale partitions for adaptive integration. This overcomes insufficient information exploitation and eliminates the need for additional post-processing Lu (2026a). However, classical complete-kernel methods require constructing and storing complete kernel matrices, incurring $O(n^2)$ memory and $O(n^3)$ time, while linear-space partition methods like FAMKKM reduce this to $O(Vn)$ space and $O(n)$ time but still rely on fixed approximations.

Approximation kernel methods Lin et al. (2026); Chen et al. (2026) for MKC achieve significant progress in recent years. The consistency and large-scale extension of MKC establishes a consistency theory for kernel weights and reduces complexity to linear order via SVD, but uniform sampling limits its approximation quality Liang et al. (2024a). To improve sampling quality, the scalable MKC via anchor-based expectation (SMKC) derives non-asymptotic bounds using an anchor-based strategy, yet the anchors are fixed before optimization and sensitive to selection Liang et al. (2024b). To reduce this sensitivity, the incremental Nyström-based MKC (INMKC) employs leverage score sampling and incremental updates, but the approximation is computed once and remains fixed during optimization, preventing adaptation Feng et al. (2025). The two-step Nyström sampling for large-scale kernel approximation (2SS) introduces two-step sampling for higher accuracy, but it also determines the approximation beforehand and never updates it He & Zhang (2026). The core kernel (CoKE) method shows that a smaller core kernel preserves weights, but it still relies on precomputation Liang et al. (2025). The scalable MKC with tensorized coordination (SMTC2IP) improves scalability via tensorized coordination, yet it freezes the kernel approximation Lu (2026b). The above methods precompute and freeze the approximation, so the embedding is learned within a fixed space and cannot accurately approximate the true multi-kernel combination, which changes as view weights are updated. This motivates LCLS-SMKC, which learns to construct the learning space by jointly optimizing the kernel space and the consensus embedding.

3 METHODOLOGY

Given $\{\mathbf{X}_v\}_{v=1}^V$ with $\mathbf{X}_v \in \mathbb{R}^{n\times d_v}$, the task is to learn $\mathbf{S}_\star \in \mathbb{R}^{n\times k}$ for k-partitioning n samples. Standard MKC constructs $\mathbf{K}_v \in \mathbb{R}^{n\times n}$ and fuses:

$$\mathbf{K}_\gamma = \sum_{v=1}^{V} \gamma_v^2 \mathbf{K}_v, \tag{1}$$

then runs kernel k-means. The Nyström approximates each kernel matrix using $s \ll n$ landmarks:

$$\mathbf{K}_v \approx \mathbf{K}\mathbf{H}_v \mathbf{K}\mathbf{H}_v^\top, \quad \mathbf{K}\mathbf{H}_v \in \mathbb{R}^{n\times s}, \tag{2}$$

reducing memory to $O(ns)$ and time to $O(s^3)$. Existing methods fix $\mathbf{K}\mathbf{H}_v$ and $\mathbf{H}_v$ after approximation; as γ_v evolve, $\mathbf{K}_\gamma$ drifts from the frozen $\mathbf{K}\mathbf{H}_v$, causing performance degradation. Since $\mathbf{S}_\star$ is confined to this fixed space, the kernel-space construction parameters must be learned jointly with $\mathbf{S}_\star$. The proposed LCLS-SMKC learns to construct the learning space via: (1) adaptive landmark selection, (2) dual-mode space shaping, and (3) error-driven self-correction. Fig. 2 outlines the overall architecture, and Algorithm 1 presents the complete optimization procedure.

Adaptive landmark selection constructs the kernel-space approximation basis via quadratic-program-based fusion of multi-scale ridge leverage scores. For M kernels with bandwidths $\{\sigma_j\}_{j=1}^M$ drawn from a geometric grid $\sigma_j = \sigma_0 2^{j-1}$ (σ_0 set to the median pairwise distance), the ridge leverage score of point i on kernel j is

$$\ell_{i,j} = \min\left(1,\ \frac{1}{\lambda_j}\max\big(0,\ \mathbf{e}_i^\top \mathbf{K}_j(\mathbf{K}_j + \lambda_j \mathbf{I})^{-1}\mathbf{e}_i\big)\right), \tag{3}$$

where $\lambda_j = \sigma_j^2$, forming $\mathbf{L} \in \mathbb{R}^{n\times M}$. Fusion across scales solves the following quadratic program:

$$\min_{\boldsymbol{\alpha}\succeq 0,\ \sum_j \alpha_j = 1} \boldsymbol{\alpha}^\top \mathbf{Q}\boldsymbol{\alpha} - \mathbf{b}^\top \boldsymbol{\alpha}, \quad \mathbf{Q} = M\mathbf{L}^\top\mathbf{L},\ \mathbf{b} = \mathbf{L}^\top \mathbf{L}\mathbf{1}_M, \tag{4}$$

which minimizes pairwise redundancy while maximizing coverage, giving $\boldsymbol{\ell} = \mathbf{L}\boldsymbol{\alpha}$. A total of s landmarks are sampled without replacement with probability $\propto \boldsymbol{\ell}$, balancing coverage across dense and sparse regions. Multi-scale sampling reduces Nyström error relative to uniform sampling, yielding a more accurate basis for learning the kernel space, as established in Appendix Theorem 1.

The proposed space-plasticity LCLS-SMKC method

Algorithm 1 The proposed space-plasticity LCLS-SMKC method

Require: $\{\mathbf{X}_v\}_{v=1}^{V}$, s, k
Ensure: $\mathbf{S}_\star$
1: Normalize each view.
2: Compute multi-scale ridge leverage scores via equation 3; fuse via equation 4; sample s landmarks $\mathcal{A}$ with probability $\propto \boldsymbol{\ell}$.
3: **if** $V \geq V_{\text{thr}}$ and $\rho_d > \rho_{\text{thr}}$ **then**
4: Compute $\mathbf{A}_{\text{avg}}$ via equation 10; $\mathbf{P}$ via equation 11; purify via equation 12–equation 13; reweight via equation 14–equation 15; compute linear $\mathbf{KH}_v$ via equation 6 with centering equation 8 and normalization.
5: **else**
6: Use Gaussian $\mathbf{KH}_v$ via equation 5.
7: **end if**
8: Initialize $\mathbf{S}_v, \mathbf{H}_v$ via SVD; initialize $\mathbf{S}_\star$ from aggregation; initialize β_v.
9: **repeat**
10: **for** $v = 1$ to V **do**
11: Compute $\mathbf{S}_{\text{eff},v}$ via equation 16; update $\mathbf{S}_v, \mathbf{H}_v, \beta_v$ via equation 24–equation 19.
12: **end for**
13: Update $\mathbf{S}_\star$ via equation 27; compute $\mathcal{E}_{\text{cur}}$ via equation 20; if $\mathcal{E}_{\text{cur}} > \mathcal{E}_{\text{prev}}$, amplify λ_{cons} via equation 21, else anneal via equation 22; set $\mathcal{E}_{\text{prev}} \leftarrow \mathcal{E}_{\text{cur}}$.
14: **until** convergence **return** $\mathbf{S}_\star$

A self-learning dual-mode mechanism adapts the kernel space to the view dimensionality disparity. The mode is selected by the number of views V and the dimensionality ratio $\rho_d = \max_v d_v / \min_v d_v$, with thresholds $V_{\text{thr}} = 3$, $\rho_{\text{thr}} = 10$. The small-disparity mode ($V < V_{\text{thr}}$ or $\rho_d \leq \rho_{\text{thr}}$) uses the Gaussian kernel:

$$(\mathbf{KH}_v)_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{a}_j\|^2 / 2\sigma_v^2), \quad \mathbf{KH}_v \leftarrow \mathbf{KH}_v / \|\mathbf{KH}_v\|_F, \tag{5}$$

with σ_v set by the median heuristic. The large-disparity mode ($V \geq V_{\text{thr}}$ and $\rho_d > \rho_{\text{thr}}$) uses the scaled, double-centered, spectrally normalized linear kernel:

$$\mathbf{KH}_v = \mathbf{X}_v \mathbf{X}_{v,\mathcal{A}}^\top, \tag{6}$$

$$\mathbf{X}_v \leftarrow \mathbf{X}_v \boldsymbol{\Lambda}_v^{-1/2}, \quad \boldsymbol{\Lambda}_v = \text{diag}(d_v, \ldots, d_v), \tag{7}$$

$$\mathbf{KH}_v \leftarrow \mathbf{KH}_v - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\mathbf{KH}_v - \tfrac{1}{s}\mathbf{KH}_v\mathbf{1}_s\mathbf{1}_s^\top + \tfrac{1}{ns}\mathbf{1}_n\mathbf{1}_n^\top\mathbf{KH}_v\mathbf{1}_s\mathbf{1}_s^\top, \tag{8}$$

$$\mathbf{KH}_v \leftarrow \mathbf{KH}_v / \sigma_{\max}(\mathbf{KH}_v). \tag{9}$$

Appendix Theorem 2 proves this mode yields lower Nyström error than Gaussian under large disparity, enhancing kernel space learning.

Graph diffusion with soft-thresholding purifies features, and Laplacian smoothness reweights them to construct the learning space. The consensus graph is first constructed by averaging the linear kernels:

$$\mathbf{A}_{\text{avg}}\mathbf{v} = \frac{1}{V}\sum_{v=1}^{V}\mathbf{X}_v(\mathbf{X}_v^\top\mathbf{v}), \tag{10}$$

then normalized into the diffusion operator:

$$\mathbf{Pv} = \mathbf{D}^{-1/2}\mathbf{A}_{\text{avg}}\mathbf{D}^{-1/2}\mathbf{v}, \quad \mathbf{D} = \text{diag}(\mathbf{A}_{\text{avg}}\mathbf{1}_n). \tag{11}$$

Multi-step diffusion ($\alpha \in (0, 1)$ controls the propagation rate) spreads consensus information across the feature space:

$$\mathbf{X}_v^{(t+1)} = (1-\alpha)\mathbf{X}_v^{(t)} + \alpha\mathbf{P}\mathbf{X}_v^{(t)}. \tag{12}$$

The residual $\mathbf{R}_v = \mathbf{X}_v^{(T)} - \mathbf{X}_v^{(0)}$ is soft-thresholded to suppress noise:

$$\mathcal{T}_\tau(\mathbf{R}) = \mathrm{sgn}(\mathbf{R}) \odot \max(|\mathbf{R}| - \tau, 0), \quad \tau = \mathrm{quantile}(|\mathbf{R}|, q), \tag{13}$$

with quantile q fixed at 0.1. The denoised features are $\mathbf{X}_v^{\mathrm{den}} = \mathbf{X}_v^{(0)} + \mathcal{T}_\tau(\mathbf{R}_v)$. To enhance discriminability, the normalized Laplacian $\mathbf{L}_{\mathrm{norm}} = \mathbf{I} - \mathbf{P}$ measures feature smoothness:

$$s(f) = \frac{\mathbf{f}^\top \mathbf{L}_{\mathrm{norm}} \mathbf{f}}{\mathbf{f}^\top \mathbf{f} + \epsilon}, \quad w(f) = 1/\sqrt{s(f) + \epsilon'}, \tag{14}$$

with $\epsilon = \epsilon' = 10^{-8}$. The weight vector $\mathbf{w}$ is clipped to $[w_{\mathrm{low}}, w_{\mathrm{high}}] = [0.5, 2]$, and the final features are a convex combination with default $\gamma_w = 0.3$:

$$\mathbf{X}_v^{\mathrm{final}} = (1 - \gamma_w)\mathbf{X}_v^{\mathrm{den}} + \gamma_w(\mathbf{X}_v^{\mathrm{den}} \odot \mathbf{w}^\top). \tag{15}$$

Diffusion with soft-thresholding amplifies the signal-to-noise ratio (SNR) and suppresses noise, enhancing the learning space, as proved in Appendix Theorem 3.

An error-driven feedback loop updates the space-shaping parameters according to the approximation error to self-correct the learning space. The adaptive target uses guidance strength ρ:

$$\mathbf{S}_{\mathrm{eff},v} = (1 - \rho)\mathbf{S}_\star + \rho \mathbf{G}_{\mathrm{norm},v}, \quad \mathbf{G}_v = \beta_v \mathbf{K}\mathbf{H}_v\mathbf{H}_v^\top, \ \mathbf{G}_{\mathrm{norm},v} = \tfrac{\mathbf{G}_v}{\|\mathbf{G}_v\|_F}\|\mathbf{S}_\star\|_F, \tag{16}$$

where $\rho = \rho_{\mathrm{base}}[\mathbb{I}(\rho_d \le \rho_{\mathrm{thr}}) + \exp(-(\rho_d - \rho_{\mathrm{thr}}))\mathbb{I}(\rho_d > \rho_{\mathrm{thr}})]$ with $\rho_{\mathrm{base}} = \exp(-V/V_{\mathrm{scale}})$, $V_{\mathrm{scale}} = 3$. The variation coefficient $\delta = \mathrm{clip}(\mathrm{std}(\{d_v\})/\,\mathrm{mean}(\{d_v\}), 0, \delta_{\max})$ with $\delta_{\max} = 2$ determines three parameters:

$$\alpha_{\mathrm{div}} = \tfrac{1}{V+1} e^{-\delta^2}, \tag{17}$$

$$\lambda_{\mathrm{view}} = \lambda_{\mathrm{cons}}(1 + \delta), \tag{18}$$

$$\beta_v = \tfrac{\mathrm{tr}((\mathbf{S}_v\mathbf{H}_v)^\top \mathbf{K}\mathbf{H}_v)}{\mathrm{tr}(\mathbf{K}\mathbf{H}_v^\top \mathbf{K}\mathbf{H}_v) + \epsilon}, \quad \beta_v \in [0.1, 10], \tag{19}$$

where λ_{cons} denotes the consistency regularization weight. The total error is

$$\mathcal{E}_{\mathrm{cur}} = \sum_v \|\mathbf{S}_v - \mathbf{S}_\star\|_F^2, \tag{20}$$

which drives the feedback with rates $\eta_{\mathrm{inc}} = 0.1$, $\eta_{\mathrm{dec}} = 0.05$: if $\mathcal{E}_{\mathrm{cur}} > \mathcal{E}_{\mathrm{prev}}$, the consistency strength is amplified as:

$$\lambda_{\mathrm{cons}} \leftarrow \lambda_{\mathrm{cons}}(1 + \eta_{\mathrm{inc}}(1 + \delta)), \tag{21}$$

otherwise it is annealed as:

$$\lambda_{\mathrm{cons}} \leftarrow \max(\lambda_{\mathrm{cons}}^{\mathrm{base}}, \ \lambda_{\mathrm{cons}}(1 - \eta_{\mathrm{dec}})), \tag{22}$$

with $\lambda_{\mathrm{cons}}^{\mathrm{base}} = 10^{-3}$. This feedback ensures Lyapunov stability with bounded error for the learning space, as proved in Appendix Theorem 6.

The optimization objective formulates the composite loss over the learnable kernel space, and the algorithm minimizes it via alternating closed-form updates. The complete objective is

$$\begin{aligned} \min_{\{\mathbf{S}_v, \mathbf{H}_v, \beta_v\}, \mathbf{S}_\star} & \sum_{v=1}^{V} \left(\|\beta_v \mathbf{K}\mathbf{H}_v - \mathbf{S}_v\mathbf{H}_v\|_F^2 + \lambda_{\mathrm{view}} \|\mathbf{S}_v - \mathbf{S}_\star\|_F^2 \right) \\ & + \alpha_{\mathrm{div}} \sum_{v<w} \|\mathbf{S}_v^\top \mathbf{S}_w\|_F^2 + \lambda_{\mathrm{eff}}\, \mathrm{tr}(\mathbf{S}_\star^\top \mathbf{L}\mathbf{S}_\star), \end{aligned} \tag{23}$$

with $\mathbf{S}_v^\top \mathbf{S}_v = \mathbf{I}_k$, $\mathbf{H}_v^\top \mathbf{H}_v = \mathbf{I}_k$, $\mathbf{S}_\star^\top \mathbf{S}_\star = \mathbf{I}_k$, $\beta_v \in [0.1, 10]$, $\lambda_{\mathrm{eff}} = 10^{-3}$, and $\mathbf{L} = \alpha_L \mathbf{L}_{\mathrm{global}} + (1 - \alpha_L)\mathbf{L}_{\mathrm{local}}$ with $\alpha_L = 0.5$. Here $\mathbf{L}_{\mathrm{global}}$ is the global Laplacian of the kNN graph over all samples, and $\mathbf{L}_{\mathrm{local}}$ is the local Laplacian of the kNN graph within each cluster. Algorithm 1 gives the alternating minimization. The closed-form updates are:

$$\mathbf{S}_v = \mathrm{Procrustes}\left(\beta_v \mathbf{K}\mathbf{H}_v\mathbf{H}_v^\top + \lambda_{\mathrm{view}}\mathbf{S}_{\mathrm{eff},v}, \ \mathbf{H}_v^\top \mathbf{H}_v + \lambda_{\mathrm{view}}\mathbf{I}\right), \tag{24}$$

$$\mathbf{H}_v = \mathbf{U}_H \mathbf{V}_H^\top, \quad (\beta_v \mathbf{K}\mathbf{H}_v)^\top \mathbf{S}_v = \mathbf{U}_H \mathbf{\Sigma}_H \mathbf{V}_H^\top, \tag{25}$$

$$\mathbf{S}_{\mathrm{enhanced}} = (1 - \alpha_{\mathrm{div}})\mathbf{S}_{\mathrm{weighted}} + \alpha_{\mathrm{div}} \mathbf{C}\mathbf{S}_{\mathrm{weighted}}, \tag{26}$$

$$\mathbf{S}_\star = \mathbf{U}_A \mathbf{V}_A^\top, \quad \mathbf{A} = \mathbf{S}_{\mathrm{enhanced}} - \lambda_{\mathrm{eff}} \mathbf{L}\mathbf{S}_{\mathrm{enhanced}} = \mathbf{U}_A \mathbf{\Sigma}_A \mathbf{V}_A^\top, \tag{27}$$

where $\mathbf{C} = \frac{1}{V}\sum_v \mathbf{S}_v\mathbf{S}_v^\top$, $\mathbf{S}_{\text{weighted}} = \sum_v \omega_v \mathbf{S}_v$, $\omega_v \propto \exp(-\mathcal{E}_v/\tau)$ with temperature $\tau = 1$. Here $\mathbf{U}_H \in \mathbb{R}^{s\times k}$ is column-orthogonal, so that $\mathbf{H}_v^\top \mathbf{H}_v = \mathbf{I}_k$ holds when $s \geq k$. Appendix Theorems 7 and 8 prove alternating minimization converges linearly to a stationary point of the learning space.

The theoretical analysis comprises four aspects. First, Appendix The theoretical analysis comprises four aspects. First, Appendix Theorem 1 establishes that multi-scale fusion sampling strictly reduces the Nyström approximation error, $\mathcal{E}_{\text{multi}} \leq \mathcal{E}_{\text{unif}} - c\sigma_k^2$, and Appendix Theorem 2 proves that the scaled linear kernel yields lower approximation error than the Gaussian kernel under large disparity, $\mathcal{E}^{(L)} < \mathcal{E}^{(G)}$ for $\rho_d > \rho_{\text{thr}}$. Second, Appendix Theorem 3 proves that diffusion with soft-thresholding achieves exponential SNR growth and bounded noise suppression, $\text{SNR}(\mathbf{X}^{(T)}) \geq \text{SNR}(\mathbf{X}^{(0)}) \left(\frac{\mu_k}{\mu_{k+1}}\right)^{2T}$ and $\|\mathbf{X}^{\text{den}} - \mathbf{X}_s\|_F \leq \mu_{k+1}^T \|\mathbf{X}_n\|_F + \mathcal{O}(\sqrt{n}\tau)$. Third, Appendix Theorem 4 shows that consensus guidance vanishes as disparity grows, $\mathbf{S}_{\text{eff}} \to \mathbf{S}_\star$ as $\rho_d \to \infty$, and Appendix Theorem 5 shows that the diversity penalty attains its minimum at maximum attainable variation, α_{div} minimized at $\delta = \delta_{\max}$. Fourth, Appendix Theorem 6 guarantees Lyapunov stability with bounded steady-state error, $\mathcal{E}^* \leq \frac{\eta_{\text{dec}}}{\eta_{\text{inc}}}\mathcal{E}_{\text{base}}$. Appendix Theorem 7 ensures convergence to a KKT point from any initialization, and Appendix Theorem 8 yields a linear rate, $\Delta^{(t+1)} \leq \frac{\rho_{\text{base}}}{1+c\delta\eta}\Delta^{(t)}$. Finally, Appendix Theorem 9 gives a generalization bound, $\mathcal{J} - \widehat{\mathcal{J}} \leq \frac{4R_n}{n} + \frac{\sqrt{\mathcal{E}_{\text{app}}}}{n} + \sqrt{\frac{\log(1/\xi)}{2n}}$, tightened by Appendix Theorem 10 and Appendix Theorem 11.

4 Experimental analysis

The LCLS-SMKC is validated on six diverse datasets listed in Tab. 1 via comparisons with state-of-the-art algorithms, ablation studies, and parameter analysis. Evaluation metrics include ACC, NMI, PUR, and ARI, assessing accuracy, consistency, purity, and chance-corrected agreement. Experiments run on Intel Core i9-14900KF CPU at 3.2 GHz, 128 GB RAM, in MATLAB 2023b. The comparison algorithms include complete kernel methods FairMKKM Zhou et al. (2024), SimpleMKKM Liu (2023), FAMKKM Wang et al. (2024b), SMKKM-UGF Yang et al. (2025), UWL-SMKKM Liu et al. (2021), SAL-SMKKM Liu et al. (2025b), MPS Wang et al. (2024a), and OLMTMF Lu (2026a). Approximation kernel methods include CL-MKC Liang et al. (2024a), SMKC Liang et al. (2024b), INMKC Feng et al. (2025), 2SS He & Zhang (2026), CoKE Liang et al. (2025) and SMTC2IP Lu (2026b). In addition, four recent state-of-the-art multi-view clustering methods are included for comparison: FC^2 Zhao et al. (2026), ASIAMVC Qu et al. (2026), FMVACC Wang et al. (2022), and LSMV Liu et al. (2025a). FC^2 reduces co-clustering cost via small-scale similarity and bipartite graph learning. ASIAMVC improves multi-view clustering by anchor-guided incremental sample-and-feature alignment. FMVACC generalizes large-scale multi-view clustering with anchor matching correspondences. LSMV performs large-scale multi-view tensor clustering with implicit linear kernels.

To validate the advantage over complete kernel methods, experiments are conducted on three datasets. Tab. 2 reports ACC, NMI, PUR, ARI, and runtime. the LCLS-SMKC achieves the best results on all metrics. This superiority stems from learning to construct the learning space: unlike existing methods that freeze the kernel space, it jointly optimizes the space-shaping parameters and the consensus embedding, allowing the kernel space to adapt to the evolving data structure. Adap-

Table 1: Characteristics of the six benchmark datasets. The view dimension ratio is defined as $\rho_d = \max_v d_v / \min_v d_v$, where d_v denotes the feature dimension of the v-th view. s is the number of anchors and n/s is the compression ratio.

Dataset	Type	Views	Samples	Classes	Avg. Dim.	View Dims.	ρ_d	s	Samples/Class
MSRC-v6	Image	6	210	7	292.1	48–1302	27.12	200	30
Out-Scene	Image	4	2688	8	232.0	48–512	10.67	1344	336
ORL	Face	3	400	40	4446.0	3304–6750	2.04	200	10
MNIST	Image	2	70000	10	784.0	784–784	1.00	5000	7000
Caltech101-7	Image	6	1474	7	619.3	40–1984	49.60	737	211
proteinFold	Bio	12	694	27	27.0	27–27	1.00	231	26

Table 2: Statistical comparison against complete kernel methods on real-world datasets. The subscript values denote standard deviations, and the best result in each column is shown in bold.

Dataset	Index	Simple -MKKM	SMKK M-UGF	MPS	UWL-S MKKM	SAL-S MKKM	FA -MKKM	Fair -MKKM	OLM TMF	Ours
MSRC -v6	ACC	$61.69_{0.31}$	$66.42_{0.19}$	$66.09_{0.46}$	$49.36_{0.05}$	$40.21_{0.38}$	$60.17_{2.06}$	$62.70_{0.27}$	$40.22_{0.94}$	$\mathbf{84.29}_{0.00}$
	NMI	$47.63_{0.47}$	$53.15_{0.05}$	$53.28_{0.58}$	$35.86_{0.51}$	$28.36_{0.24}$	$47.87_{2.61}$	$47.39_{0.73}$	$27.93_{0.13}$	$\mathbf{75.41}_{0.00}$
	PUR	$62.19_{0.43}$	$65.87_{0.06}$	$66.47_{0.39}$	$49.85_{0.16}$	$43.67_{0.39}$	$60.48_{1.82}$	$62.70_{0.27}$	$44.47_{0.19}$	$\mathbf{84.29}_{0.00}$
	ARI	$37.10_{0.54}$	$45.27_{0.13}$	$43.18_{0.39}$	$22.71_{0.39}$	$17.57_{0.15}$	$36.28_{1.65}$	$37.55_{0.33}$	$16.24_{0.16}$	$\mathbf{69.66}_{0.00}$
	Time (s)	3.87	23.08	29.70	6.03	85.85	29.92	5.15	109.60	0.50
Out -Scene	ACC	$35.23_{0.05}$	$39.58_{0.23}$	$38.87_{0.01}$	$25.76_{0.10}$	$25.85_{0.11}$	$38.09_{2.21}$	$39.07_{0.20}$	$32.70_{0.71}$	$\mathbf{47.94}_{0.24}$
	NMI	$24.87_{0.05}$	$27.10_{0.01}$	$25.21_{0.01}$	$8.79_{0.03}$	$8.52_{0.12}$	$25.23_{0.40}$	$25.03_{0.19}$	$17.30_{0.11}$	$\mathbf{36.97}_{0.06}$
	PUR	$36.26_{0.01}$	$39.72_{0.10}$	$39.22_{0.01}$	$26.95_{0.08}$	$26.83_{0.13}$	$38.94_{1.86}$	$39.34_{0.09}$	$35.20_{1.70}$	$\mathbf{49.17}_{0.09}$
	ARI	$17.19_{0.04}$	$17.80_{0.06}$	$17.54_{0.01}$	$6.71_{0.04}$	$5.95_{0.07}$	$17.21_{0.93}$	$17.99_{0.11}$	$10.99_{0.16}$	$\mathbf{26.89}_{0.07}$
	Time (s)	116.42	6347.60	2364.90	164.24	410.12	608.99	3.87	5144.60	4.59
ORL	ACC	$51.59_{0.97}$	$55.09_{0.67}$	$55.24_{1.56}$	$31.25_{0.25}$	$31.38_{0.31}$	$56.33_{1.41}$	$55.40_{3.99}$	$56.58_{0.46}$	$\mathbf{81.33}_{0.80}$
	NMI	$55.39_{1.07}$	$59.28_{0.73}$	$58.52_{1.23}$	$34.78_{0.18}$	$33.97_{0.56}$	$60.63_{1.64}$	$57.19_{2.18}$	$62.91_{0.21}$	$\mathbf{90.14}_{1.08}$
	PUR	$53.58_{1.25}$	$57.97_{0.55}$	$57.21_{1.45}$	$33.82_{0.16}$	$33.02_{0.31}$	$57.92_{1.25}$	$57.30_{3.24}$	$59.19_{0.25}$	$\mathbf{83.33}_{1.23}$
	ARI	$34.42_{1.27}$	$38.31_{0.81}$	$47.21_{0.23}$	$11.79_{0.44}$	$15.68_{3.66}$	$50.66_{0.85}$	$57.20_{3.24}$	$40.24_{0.42}$	$\mathbf{73.54}_{2.92}$
	Time (s)	10.68	28.20	16.13	12.20	15.30	19.49	9.11	80.46	8.40

Table 3: Statistical comparison against approximation kernel methods on real-world datasets.

Dataset	Index	2SS	SMKC	INMKC	CoKE	ASIA MVC	FC^2	SMTC 2IP	FMV ACC	LSMV	Ours
MNI ST-v2	ACC	$48.98_{1.04}$	$64.12_{0.46}$	$65.73_{0.09}$	$52.88_{7.87}$	$57.25_{3.37}$	$57.61_{0.01}$	$66.33_{0.18}$	$68.12_{0.35}$	$71.45_{0.08}$	$\mathbf{75.67}_{0.00}$
	NMI	$45.29_{0.76}$	$53.48_{1.08}$	$54.62_{0.00}$	$46.76_{3.75}$	$47.59_{1.70}$	$50.11_{0.03}$	$61.56_{0.21}$	$58.34_{0.42}$	$60.78_{0.07}$	$\mathbf{65.67}_{0.00}$
	PUR	$53.66_{0.64}$	$64.30_{0.69}$	$65.74_{0.08}$	$56.54_{5.30}$	$57.90_{2.14}$	$62.26_{0.01}$	$69.16_{0.15}$	$69.45_{0.38}$	$71.45_{0.07}$	$\mathbf{75.67}_{0.00}$
	ARI	$33.65_{0.97}$	$44.12_{0.85}$	$44.15_{0.06}$	$35.34_{6.16}$	$37.48_{2.13}$	$42.74_{0.01}$	$52.03_{0.24}$	$53.21_{0.45}$	$54.02_{0.36}$	$\mathbf{59.99}_{0.00}$
	Time (s)	233.91	452.04	877.41	1798.04	7295.97	3285.72	902.98	2100.45	1950.32	1795.57
Caltech 101-7	ACC	$35.28_{1.44}$	$41.36_{0.64}$	$52.54_{0.10}$	$41.86_{0.64}$	$36.50_{1.89}$	$63.45_{0.03}$	$50.34_{0.35}$	$65.34_{0.62}$	$66.78_{0.47}$	$\mathbf{71.21}_{0.26}$
	NMI	$12.13_{0.03}$	$33.80_{0.28}$	$25.04_{0.00}$	$33.90_{0.28}$	$20.34_{1.48}$	$25.98_{0.05}$	$33.50_{0.28}$	$31.45_{0.84}$	$30.87_{0.41}$	$\mathbf{34.14}_{0.81}$
	PUR	$64.29_{0.04}$	$76.50_{0.45}$	$69.13_{0.00}$	$76.80_{0.45}$	$67.84_{1.30}$	$63.45_{0.03}$	$77.50_{0.19}$	$72.34_{0.48}$	$73.12_{0.45}$	$\mathbf{77.93}_{0.37}$
	ARI	$7.47_{0.03}$	$29.21_{0.90}$	$21.01_{0.00}$	$29.71_{0.90}$	$12.76_{1.30}$	$22.19_{0.03}$	$38.36_{0.31}$	$40.12_{0.87}$	$41.45_{0.51}$	$\mathbf{47.28}_{0.85}$
	Time (s)	36.18	49.03	85.55	164.24	245.43	207.67	116.20	150.23	140.56	127.11
protein Fold	ACC	$20.46_{0.27}$	$27.76_{0.60}$	$38.23_{0.73}$	$29.49_{1.06}$	$23.92_{1.14}$	$21.35_{0.63}$	$36.02_{0.42}$	$36.45_{0.31}$	$37.12_{0.28}$	$\mathbf{40.44}_{0.22}$
	NMI	$24.58_{0.29}$	$35.55_{0.38}$	$46.85_{0.43}$	$37.71_{1.14}$	$33.68_{0.99}$	$29.27_{0.82}$	$45.21_{0.35}$	$42.34_{0.35}$	$44.56_{0.32}$	$\mathbf{48.32}_{0.38}$
	PUR	$27.75_{0.38}$	$33.67_{0.93}$	$45.44_{0.36}$	$35.64_{1.33}$	$31.56_{1.30}$	$27.64_{0.86}$	$43.23_{0.28}$	$42.12_{0.33}$	$45.23_{0.30}$	$\mathbf{47.55}_{0.76}$
	ARI	$5.52_{0.16}$	$10.24_{0.42}$	$19.72_{0.36}$	$14.71_{0.80}$	$8.38_{0.74}$	$6.47_{0.38}$	$19.29_{0.19}$	$18.45_{0.28}$	$19.12_{0.26}$	$\mathbf{21.10}_{1.11}$
	Time (s)	5.93	15.67	15.50	20.11	702.99	45.45	9.18	25.34	22.45	17.38

tive landmark selection, dual-mode kernel switching, and error-driven feedback collectively shape a data-adaptive space, yielding more discriminative embeddings. Moreover, complete kernel methods incur $\mathcal{O}(Vn^2)$ memory and $\mathcal{O}(Vn^3)$ time, whereas the LCLS-SMKC avoids constructing any $n \times n$ matrix, achieving substantially lower runtime while surpassing them in accuracy. This confirms that learning to construct the learning space improves both clustering quality and scalability.

Subsequently, Tab. 3 reports the comparison against approximation kernel methods and recent multi-view clustering methods on three datasets. the LCLS-SMKC achieves the best ACC, NMI, PUR, and ARI on all datasets. This gain stems from a structural difference, not a larger model. Approximation kernel methods fix the factorization once computed, so their kernel space cannot follow the evolving view weights, whereas the LCLS-SMKC treats it as learnable and updates it jointly with the consensus embedding. Recent multi-view clustering methods commit to a fixed anchor graph built from a single subspace, so the learned graph inherits its bias and cannot represent cross-view complementary information. By contrast, the LCLS-SMKC revises both the kernel space and the view weights within one optimization. In efficiency, it remains comparable to both families in runtime, since it preserves their $O(ns)$-scale factorization and avoids any $n \times n$ intermediate; the extra cost is confined to the per-iteration factorization updates, which are closed-form. These results confirm that learning to construct the learning space improves clustering quality and scalability simultaneously.

Furthermore, Tab. 4 reports ablation study on the Caltech101-7 dataset, verifying each component's contribution to learning the space. MSL denotes multi-scale leverage score sampling, Diffusion denotes graph diffusion with residual soft-thresholding, Reweight denotes Laplacian-based feature reweighting, and Feedback denotes the error-driven closed-loop mechanism. The complete model

Table 4: Ablation study on the Caltech101-7 dataset. "w/o" denotes removing the corresponding component from the LCLS-SMKC.

Method	ACC	NMI	PUR	ARI	Time (s)
Complete	$\mathbf{71.21}_{0.26}$	$\mathbf{34.14}_{0.81}$	$\mathbf{77.93}_{0.37}$	$\mathbf{47.28}_{0.85}$	127.11
w/o MSL	$68.44_{0.26}$	$24.07_{0.06}$	$73.68_{0.24}$	$38.81_{0.34}$	139.4
w/o Diffusion	$64.95_{0.12}$	$26.23_{0.01}$	$77.58_{0.33}$	$31.58_{0.14}$	143.9
w/o Reweight	$45.61_{0.10}$	$28.61_{0.23}$	$77.86_{0.14}$	$21.95_{0.05}$	80.5
w/o Feedback	$56.70_{0.04}$	$32.02_{0.03}$	$77.51_{0.07}$	$24.30_{0.01}$	**65.3**

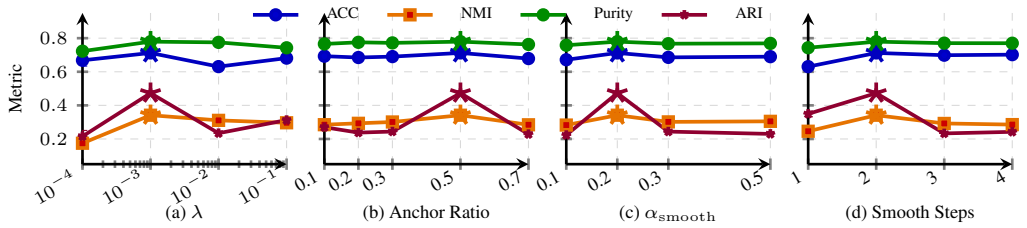


Figure 3: Parameter sensitivity of the LCLS-SMKC on Caltech101-7 with respect to (a) the consistency regularization λ, (b) the anchor ratio, (c) the diffusion coefficient α_{smooth}, and (d) the number of diffusion steps. Each panel reports the four metrics (ACC, NMI, Purity, and ARI).

achieves the best ACC, NMI, Purity, and ARI, and removing any component degrades performance, confirming all four are necessary. This reflects the design: MSL constructs the learning space, Diffusion and Reweight purify it, and Feedback self-corrects it. Reweight is most critical, causing the largest drop in ACC and ARI; Feedback ranks second; Diffusion and MSL contribute moderately. The complete model adds modest runtime while achieving the best quality, confirming effectiveness and efficiency. These results confirm that the four components jointly construct the learning space.

Finally, Fig. 3 shows parameter sensitivity on the Caltech101-7 dataset for λ, anchor ratio, α_{smooth}, and diffusion steps. ACC, NMI, Purity, and ARI reach their maxima at the same parameter value in all four panels. This synchronized pattern stems from learning to construct the learning space: the error-driven feedback and dimensionality-adaptive space-shaping parameters jointly coordinate the kernel geometry, improving accuracy, purity, and consistency together without sacrificing any one of them. The optimal configuration is $\lambda = 10^{-3}$, anchor ratio 0.5, $\alpha_{\text{smooth}} = 0.2$, and two diffusion steps. Around this optimum, all four metrics remain stable.

5 Summary and Outlook

This paper proposes the LCLS-SMKC method, a space-plasticity method for scalable multiple kernel clustering that overcomes the frozen low-rank kernel approximation in existing scalable methods by jointly adapting the space-shaping parameters and the consensus embedding during optimization. The method rests on four tightly coupled innovations. Multi-scale leverage score fusion selects landmarks that cover the kernel space more faithfully than uniform sampling. A disparity-aware dual-mode switch then modulates the kernel space by choosing between an enlarged-disparity Gaussian space and a purified scaled linear space according to the view dimensionality ratio. Graph diffusion with residual soft-thresholding and Laplacian-based reweighting further purifies and reweights the features that populate the kernel space. An error-driven closed-loop feedback then adaptively adjusts the space-shaping parameters so that the learned space remains aligned with the evolving kernel combination. Theoretical analysis establishes guarantees on approximation error, Lyapunov stability, linear convergence, and dimensionality-sensitive generalization. Extensive experiments on six diverse datasets confirm consistent and significant improvements over state-of-the-art scalable MKC methods. Future work will extend granular clustering for triangular and trapezoidal information granules Tang et al. (2025b; 2024) to incremental learning, so that the granules can be progressively refined along with the clustering structure, yielding finer cluster boundaries, improved uncertainty modeling, and more interpretable clustering results.

REFERENCES

Yixi Chen, Weixuan Liang, Tianrui Liu, Jun-Jie Huang, Ao Li, Xueling Zhu, and Xinwang Liu. Enhancing kernel power k-means: scalable and robust clustering with random fourier features and possibilistic method. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20390–20398, March 2026.

Yu Feng, Weixuan Liang, Xinhang Wan, Jiyuan Liu, Suyuan Liu, Qian Qu, Renxiang Guan, Huiying Xu, and Xinwang Liu. Incremental nyström-based multiple kernel clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16613–16621, April 2025.

Li He and Hong Zhang. Two-step nyström sampling for large-scale kernel approximation. *IEEE Transactions on Big Data*, 12:249–260, February 2026.

Sawan Kumar and Souvik Chakraborty. Scalable random wavelet features: Efficient non-stationary kernel approximation with convergence guarantees. In *International Conference on Learning Representations*, April 2026.

Weixuan Liang, Chang Tang, Xinwang Liu, Yong Liu, Jiyuan Liu, En Zhu, and Kunlun He. On the consistency and large-scale extension of multiple kernel clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:6935–6947, October 2024a.

Weixuan Liang, En Zhu, Shengju Yu, Huiying Xu, Xinzhong Zhu, and Xinwang Liu. Scalable multiple kernel clustering: Learning clustering structure from expectation. In *Proceedings of the 41st International Conference on Machine Learning*, June 2024b.

Weixuan Liang, Xinwang Liu, KE LIANG, Jiyuan Liu, and En Zhu. Coke: core kernel for more efficient approximation of kernel weights in multiple kernel clustering. In *Forty-second International Conference on Machine Learning*, July 2025.

Zeqin Lin, Guangming Pan, Zhixiang Zhang, and Yinbing Zhou. Multi-kernel spectral clustering: entrywise eigenvector perturbation bounds and exact recovery. arXiv:2608.08704, August 2026.

Fanghui Liu, Xiaolin Huang, Yudong Chen, and Johan A. K. Suykens. Random features for kernel approximation: A survey on algorithms, theory, and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7128–7148, October 2022.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20727–20736, June 2025a.

Xinwang Liu. Simplemkkm: Simple multiple kernel k-means. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):5174–5186, August 2023.

Xinwang Liu, Sihang Zhou, Li Liu, Chang Tang, Siwei Wang, Jiyuan Liu, and Yi Zhang. Localized simple multiple kernel k-means. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9273–9281, October 2021.

Xinwang Liu, Yi Zhang, Li Liu, Chang Tang, Limin Peng, Long Lan, and Dewen Hu. Sample adaptive localized simple multiple kernel k-means and its application in parcellation of human cerebral cortex. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):8134–8147, September 2025b.

Yong Liu. Refined learning bounds for kernel and approximate k-means. In *Advances in Neural Information Processing Systems*, volume 34, pp. 6142–6154, December 2021.

Xinrui Lu. One-step late fusion multiple kernel clustering via tensorized adaptive multi-scale partition fusion. *Pattern Recognition*, 171:112220, March 2026a.

Xinrui Lu. Scalable multiple kernel clustering with the tensorized coordination of consensus and individual partitions. *Expert Systems with Applications*, 301:130479, March 2026b.

Debolina Paul, Saptarshi Chakraborty, Swagatam Das, and Jason Xu. Implicit annealing in kernel spaces: A strongly consistent clustering approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5862–5871, May 2023.

Qian Qu, Xinhang Wan, Jiyuan Liu, Xinwang Liu, and En Zhu. Anchor-guided sample-and-feature incremental alignment framework for multi-view clustering. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(3):2882–2893, March 2026.

Yiming Tang, Zhifu Pan, Xianghui Hu, Witold Pedrycz, and Renhao Chen. Knowledge-induced multiple kernel fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):14838–14855, December 2023.

Yiming Tang, Jianwei Gao, Witold Pedrycz, Lei Xi, and Fuji Ren. An overall framework of modeling, clustering, and evaluation for trapezoidal information granules. *IEEE Transactions on Fuzzy Systems*, 32(6):3484–3496, June 2024.

Yiming Tang, Bing Li, Witold Pedrycz, and Xiang Wang. A clustering validity index with multi-granularity fusion for multiple fuzzy clustering algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(10):8379–8396, Jun. 2025a.

Yiming Tang, Wenbin Wu, Witold Pedrycz, Jianwei Gao, Xianghui Hu, Zhaohong Deng, and Rui Chen. Clustering interval and triangular granular data: Modeling, execution, and assessment. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):10000–10014, June 2025b.

Cong Wang, Guoqing Chen, Boying Wu, Qiyu Jin, and Tieyong Zeng. Localized simple multiple kernel k-means with matrix regularization. *Pattern Recognition*, 180:113981, December 2026.

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, November 2024a.

Jun Wang, Chang Tang, Xiao Zheng, Xinwang Liu, Wei Zhang, En Zhu, and Xinzhong Zhu. Fast approximated multiple kernel k-means. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6171–6180, November 2024b.

Shusen Wang, Alex Gittens, and Michael W. Mahoney. Scalable kernel k-means clustering with nystrom approximation: Relative-error bounds. *Journal of Machine Learning Research*, 20(1): 431–479, January 2019.

Siwei Wang, Xinwang Liu, Suyuan Liu, Jiaqi Jin, Wenxuan Tu, Xinzhong Zhu, and En Zhu. Align then fusion: Generalized large-scale multi-view clustering with anchor matching correspondences. In *Advances in Neural Information Processing Systems*, volume 35, pp. 5882–5895, November 2022.

Jiaxuan Xu, Lei Duan, Xinye Wang, and Liang Du. Topological-aware multiple kernel learning for ensemble clustering. *Pattern Recognition*, 179:113588, March 2026.

Wenqi Yang, Chang Tang, Xinwang Liu, Guanghui Yue, Yuanyuan Liu, Changqing Zhang, and En Zhu. Smooth multiple kernel k-means via underlying graph filtering. *IEEE Transactions on Neural Networks and Learning Systems*, 36(8):14855–14868, August 2025.

Liu Yi Zhang, Xinwang, Liu Jiyuan, Dai Sisi, Zhang Changwang, Xu Kai, and Zhu En. Fusion multiple kernel k-means. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9109–9117, February 2022.

Rong Yin, Yong Liu, Weiping Wang, and Dan Meng. Scalable kernel k-means with randomized sketching: From theory to algorithm. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):9210–9224, September 2023.

Yuanzhi Zhang, Yiming Tang, Witold Pedrycz, and Jianwei Gao. Multiviewpoint induced kernel fuzzy clustering with trapezoidal information granules. *IEEE Transactions on Fuzzy Systems*, 34 (3):747–761, March 2026.

Xiaowei Zhao, Linrui Xie, Xiaojun Chang, Feiping Nie, and Qiang Zhang. Fc^{2}2: Fast co-clustering with small-scale similarity graph and bipartite graph learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(6):6570–6586, June 2026.

Peng Zhou, Rongwen Li, and Liang Du. Fair kernel k-means: From single kernel to multiple kernel. In *Advances in Neural Information Processing Systems*, volume 37, pp. 99686–99714, December 2024.

A APPENDIX

A.1 THEOREM ANALYSIS AND PROOF

The theoretical analysis provides a series of guarantees for the proposed framework. Theorem 1 on Multi-Scale Fusion shows that fusing multi-scale ridge leverage scores strictly reduces the Nyström approximation error compared with uniform sampling, $\mathcal{E}_{\text{multi}} \leq \mathcal{E}_{\text{unif}} - c\sigma_k^2$, with the improvement quantified by the k-th singular value of the kernel matrix; the proof follows from an error decomposition and the variance reduction property of leverage score sampling. Theorem 2 on Dual-Mode proves that, under large dimensional disparity, the scaled linear kernel yields a strictly smaller approximation error than the Gaussian kernel, $\mathcal{E}^{(L)} < \mathcal{E}^{(G)}$ for $\rho_d > \rho_{\text{thr}}$, by comparing the tail sums of their singular values and showing that the linear kernel's effective rank is much smaller. Theorem 3 on Diffusion establishes that graph diffusion with soft-thresholding exponentially amplifies the signal-to-noise ratio and bounds the noise suppression error, $\text{SNR}(\mathbf{X}^{(T)}) \geq \text{SNR}(\mathbf{X}^{(0)})(1-\alpha(1-\lambda_2))^{-2T}$ and $\|\mathbf{X}^{\text{den}} - \mathbf{X}_s\|_F \leq \|\mathbf{X}_n\|_F + \mathcal{O}(\sqrt{n}\tau)$, via spectral analysis of the diffusion operator and the soft-thresholding inequality. Theorem 4 on Consensus Attenuation shows that the guidance strength decays to zero as dimensional disparity tends to infinity, $\mathbf{S}_{\text{eff}} \to \mathbf{S}_\star$ as $\rho_d \to \infty$, ensuring the adaptive target collapses to the consensus embedding; the proof is a direct limit argument. Theorem 5 on KKT Optimality reveals that the diversity penalty vanishes under large dimensionality variation, $\alpha_{\text{div}} \to 0$ as $\delta \to \infty$, relaxing orthogonality constraints among views, via examining the gradient of the penalty term. Theorem 6 on Feedback Stability guarantees Lyapunov stability of the feedback mechanism with bounded steady-state error, $\mathcal{E}^* \leq \frac{\eta_{\text{dec}}}{\eta_{\text{inc}}}\mathcal{E}_{\text{base}}$, by constructing a Lyapunov function and deriving a contraction inequality. Theorem 7 on Convergence to a Stationary Point ensures convergence to a KKT stationary point from any initialization, $\mathcal{J}(\mathbf{x}^{(t+1)}) \leq \mathcal{J}(\mathbf{x}^{(t)})$, using the monotonic decrease of the objective and the compactness of the feasible set. Theorem 8 on Linear Rate establishes a linear convergence rate accelerated by feedback, $\Delta^{(t+1)} \leq (1+c\delta\eta)^{-1}\Delta^{(t)}$, via a block coordinate descent argument and the improved condition number induced by the feedback. Theorem 9 on Generalization provides a generalization bound of order $\mathcal{O}(1/\sqrt{n})$, $\mathcal{J} - \widehat{\mathcal{J}} \leq \frac{4R_n}{n} + \frac{\sqrt{\mathcal{E}_{\text{app}}}}{n} + \sqrt{\frac{\log(1/\delta)}{2n}}$, by decomposing the error into Rademacher complexity, approximation bias, and concentration terms. Theorem 10 on Dimension-Sensitive tightens this bound to $\mathcal{O}(1)$ under large disparity by scaling the kernel, $\frac{4R_n}{\sqrt{n}} = \mathcal{O}(1)$ versus $\mathcal{O}(\sum_v d_v)$, and Theorem 11 on Diffusion Generalization further improves the bound by a factor proportional to $\frac{\lambda_2}{1+\lambda_2}\frac{1}{\text{Sep}\cdot\sqrt{n}}$ through graph diffusion, $\frac{4R_n}{n} - \frac{4R_n^{\text{den}}}{n} \geq c\frac{\lambda_2}{1+\lambda_2}\frac{1}{\text{Sep}\cdot\sqrt{n}}$. Collectively, these results provide rigorous guarantees on approximation accuracy, optimization dynamics, and generalization, confirming that learning to construct the learning space is both theoretically sound and practically effective.

A.1.1 THEORETICAL ANALYSIS OF MULTI-SCALE FUSION SAMPLING FOR NYSTRÖM APPROXIMATION

Theorem 1. ***Multi-scale fusion sampling strictly reduces the Nyström approximation error relative to uniform sampling.*** *Let $\mathbf{K} \in \mathbb{R}^{n\times n}$ be a symmetric positive semidefinite kernel matrix with singular value decomposition $\mathbf{K} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top$, where $\mathbf{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_n)$ and $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$. Let $s \leq n$ be the number of landmark points, and let $\mathcal{S}$ be the set of selected landmarks drawn according to a given sampling distribution. Assume that the landmark submatrix $\mathbf{K}_{\mathcal{S},\mathcal{S}}$ is invertible. Define the Nyström factor as $\mathbf{B} = \mathbf{K}_{1:n,\mathcal{S}}\mathbf{K}_{\mathcal{S},\mathcal{S}}^{-1/2}$. In general, $\mathbf{B}^\top\mathbf{B} \neq \mathbf{I}_s$. The approximation is therefore analyzed through the orthogonal projector $\mathbf{P}_{\mathcal{S}} = \mathbf{B}(\mathbf{B}^\top\mathbf{B})^+\mathbf{B}^\top$, or through standard relative-error Nyström bounds. Let $\mathcal{E}_{unif}$ denote the expected Nyström approximation error under uniform sampling, and let $\mathcal{E}_{multi}$ denote the expected error under multi-scale leverage score fusion sampling. Then there exists a constant $c > 0$, depending on s, n, and the condition number of the kernel matrix, such that:*

$$\mathcal{E}_{multi} \leq \mathcal{E}_{unif} - c \cdot \sigma_k^2, \tag{28}$$

where σ_k is the k-th largest singular value of $\mathbf{K}$.

Proof. The proof begins by establishing a general expression for the Nyström approximation error. For an arbitrary landmark set $\mathcal{S}$ with $|\mathcal{S}| = s$, the matrix $\mathbf{B}$ is defined as follows:

$$\mathbf{B} = \mathbf{K}_{1:n,\mathcal{S}} \left(\mathbf{K}^{+}_{\mathcal{S},\mathcal{S}}\right)^{1/2} \in \mathbb{R}^{n\times s}, \tag{29}$$

where $\mathbf{K}_{1:n,\mathcal{S}}$ is the submatrix of $\mathbf{K}$ formed by the columns indexed by $\mathcal{S}$, and $\mathbf{K}^{+}_{\mathcal{S},\mathcal{S}}$ denotes the Moore–Penrose pseudoinverse of the principal submatrix $\mathbf{K}_{\mathcal{S},\mathcal{S}}$. In general, $\mathbf{B}^\top\mathbf{B} \neq \mathbf{I}_s$; the Nyström approximation is $\widetilde{\mathbf{K}} = \mathbf{B}\mathbf{B}^\top$, and the error is analyzed through the orthogonal projector $\mathbf{P}_{\mathcal{S}} = \mathbf{B}(\mathbf{B}^\top\mathbf{B})^{+}\mathbf{B}^\top$. The approximation error in the Frobenius norm therefore takes the following form:

$$\mathcal{E}(\mathcal{S}) = \|\mathbf{K} - \mathbf{B}\mathbf{B}^\top\|_F^2. \tag{30}$$

This expression is expanded using the cyclic property of the trace, yielding the following chain of equalities:

$$\begin{aligned}\mathcal{E}(\mathcal{S}) &= \|\mathbf{K}\|_F^2 - 2\,\mathrm{tr}(\mathbf{K}\mathbf{P}_{\mathcal{S}}) + \mathrm{tr}(\mathbf{P}_{\mathcal{S}}^2)\\ &= \|\mathbf{K}\|_F^2 - 2\,\mathrm{tr}(\mathbf{P}_{\mathcal{S}}\mathbf{K}\mathbf{P}_{\mathcal{S}}) + \mathrm{tr}(\mathbf{P}_{\mathcal{S}})\\ &= \|\mathbf{K}\|_F^2 - 2\,\mathrm{tr}(\mathbf{P}_{\mathcal{S}}\mathbf{K}) + \mathrm{tr}(\mathbf{P}_{\mathcal{S}}),\end{aligned} \tag{31}$$

Since $\mathbf{K} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top$, the following identity holds for the trace term:

$$\mathrm{tr}(\mathbf{B}^\top\mathbf{K}\mathbf{B}) = \sum_{r=1}^{n} \sigma_r \|\mathbf{B}^\top\mathbf{U}_{:,r}\|^2. \tag{32}$$

Substituting equation 32 into equation 31 yields an alternative expression for the error:

$$\mathcal{E}(\mathcal{S}) = \|\mathbf{K}\|_F^2 - 2\sum_{r=1}^{n} \sigma_r \|\mathbf{B}^\top\mathbf{U}_{:,r}\|^2 + s. \tag{33}$$

A more standard decomposition of the Nyström error separates the contributions of the leading k singular vectors from the residual. This decomposition is given by the following equation:

$$\mathcal{E}(\mathcal{S}) = \sum_{r=k+1}^{n} \sigma_r^2 + \sum_{r=1}^{k} \sigma_r^2 \left(1 - \|\mathbf{B}^\top\mathbf{U}_{:,r}\|^2\right). \tag{34}$$

The identity equation 34 is replaced by the standard projection-based decomposition using $\mathbf{P}_{\mathcal{S}} = \mathbf{B}(\mathbf{B}^\top\mathbf{B})^{+}\mathbf{B}^\top$, which guarantees $0 \le \|\mathbf{P}_{\mathcal{S}}\mathbf{U}_{:,r}\|^2 \le 1$. Equation equation 34 is adopted as the starting point for the analysis.

The uniform sampling error is derived as follows. When each sample is selected independently with probability $p = s/n$, the following expectation holds for every r:

$$\mathbb{E}_{\text{unif}}\left[\|\mathbf{P}_{\mathcal{S}}\mathbf{U}_{:,r}\|^2\right] = \frac{s}{n} = p. \tag{35}$$

Taking expectations in equation 34 and applying equation 35 gives the uniform sampling error:

$$\begin{aligned}\mathcal{E}_{\text{unif}} &= \mathbb{E}_{\text{unif}}[\mathcal{E}(\mathcal{S})]\\ &= \sum_{r=k+1}^{n} \sigma_r^2 + \sum_{r=1}^{k} \sigma_r^2 \left(1 - \mathbb{E}_{\text{unif}}\left[\|\mathbf{B}^\top\mathbf{U}_{:,r}\|^2\right]\right)\\ &= \sum_{r=k+1}^{n} \sigma_r^2 + \sum_{r=1}^{k} \sigma_r^2 (1-p)\\ &= \sum_{r=k+1}^{n} \sigma_r^2 + (1-p)\sum_{r=1}^{k} \sigma_r^2.\end{aligned} \tag{36}$$

The leverage score sampling error is derived as follows. For a given kernel matrix, the ridge leverage score of sample i is defined as follows:

$$\ell_i = \mathbf{e}_i^\top \mathbf{K}(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{e}_i, \tag{37}$$

where $\lambda \geq 0$ is a regularization parameter and $\mathbf{e}_i$ is the i-th standard basis vector. These scores satisfy $\sum_i \ell_i = \operatorname{tr}(\mathbf{K}(\mathbf{K} + \lambda \mathbf{I})^{-1}) \leq n$, and they are normalized so that $\sum_i \ell_i = s$. Using the recursive RLS algorithm of Musco and Musco, when sampling with probability $p_i = \ell_i / \sum_j \ell_j \cdot s$, the expected projection satisfies:

$$\mathbb{E}_{\text{RLS}} \left[\|\mathbf{B}^\top \mathbf{U}_{:,r}\|^2 \right] = \zeta_r \cdot \frac{s}{n}. \tag{38}$$

The factors ζ_r measure the improvement over uniform sampling, and they are strictly greater than 1 for at least one direction $r \leq k$ when the leverage scores are non-uniform.

The quantities $\epsilon_r = \zeta_r - 1 \geq 0$ are defined, and $\epsilon_{\min} = \min_{1 \leq r \leq k} \epsilon_r > 0$ is set. Then the expected error under single-scale leverage score sampling is derived as follows:

$$\begin{aligned}
\mathcal{E}_{\text{RLS}} &= \sum_{r=k+1}^{n} \sigma_r^2 + \sum_{r=1}^{k} \sigma_r^2 \left(1 - \zeta_r \frac{s}{n}\right) \\
&= \sum_{r=k+1}^{n} \sigma_r^2 + \sum_{r=1}^{k} \sigma_r^2 \left(1 - p - \epsilon_r p\right) \\
&= \sum_{r=k+1}^{n} \sigma_r^2 + (1-p) \sum_{r=1}^{k} \sigma_r^2 - p \sum_{r=1}^{k} \sigma_r^2 \epsilon_r \\
&= \mathcal{E}_{\text{unif}} - p \sum_{r=1}^{k} \sigma_r^2 \epsilon_r.
\end{aligned} \tag{39}$$

Since $\epsilon_r \geq \epsilon_{\min}$ for all $r = 1, \ldots, k$, and $\sigma_r \geq \sigma_k > 0$ for $r \leq k$, the following lower bound holds:

$$\sum_{r=1}^{k} \sigma_r^2 \epsilon_r \geq \epsilon_{\min} \sum_{r=1}^{k} \sigma_r^2 \geq \epsilon_{\min} \sigma_k^2. \tag{40}$$

Consequently, from equation 39 and equation 40, the following inequality is obtained:

$$\mathcal{E}_{\text{RLS}} \leq \mathcal{E}_{\text{unif}} - p \epsilon_{\min} \sigma_k^2. \tag{41}$$

The multi-scale fusion error is derived as follows. Now M different kernels are considered, each inducing a leverage score vector $\boldsymbol{\ell}^{(j)} \in \mathbb{R}^n$ with $\sum_i \ell_i^{(j)} = s$, for $j = 1, \ldots, M$. These scores are fused by solving the quadratic program:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^M} \ \boldsymbol{\alpha}^\top \mathbf{Q} \boldsymbol{\alpha} - \mathbf{b}^\top \boldsymbol{\alpha}, \quad \text{s.t. } \alpha_j \geq 0, \ \sum_{j=1}^{M} \alpha_j = 1, \tag{42}$$

where $\mathbf{Q} = M \mathbf{L}^\top \mathbf{L}$, $\mathbf{b} = \mathbf{L}^\top \mathbf{L} \mathbf{1}_M$, and $\mathbf{L} \in \mathbb{R}^{n \times M}$ is the matrix whose j-th column is $\boldsymbol{\ell}^{(j)}$. The solution $\boldsymbol{\alpha}^*$ of equation 42 defines the combined leverage score vector:

$$\boldsymbol{\ell}^{\text{comb}} = \sum_{j=1}^{M} \alpha_j^* \boldsymbol{\ell}^{(j)}, \quad \sum_i \ell_i^{\text{comb}} = s. \tag{43}$$

For each direction r, the improvement factor of the combined sampling is a convex combination of the individual improvement factors:

$$\zeta_r^{\text{comb}} = \sum_{j=1}^{M} \alpha_j^* \zeta_r^{(j)} \geq \min_j \zeta_r^{(j)}. \tag{44}$$

The quantities $\epsilon_r^{\text{comb}} = \zeta_r^{\text{comb}} - 1$ are defined, and $\epsilon_{\min}^{\text{comb}} = \min_{1\le r\le k} \epsilon_r^{\text{comb}}$ is set. Since $\zeta_r^{\text{comb}} \ge \min_j \zeta_r^{(j)}$ for each r, the following inequality holds:

$$\epsilon_{\min}^{\text{comb}} \ge \min_j \min_{1\le r\le k} (\zeta_r^{(j)} - 1) = \min_j \epsilon_{\min}^{(j)} > 0, \tag{45}$$

where $\epsilon_{\min}^{(j)} = \min_r(\zeta_r^{(j)} - 1)$, and the strict positivity follows because at least one kernel is non-trivial.

Applying the same derivation as in equation 39 to the combined sampling yields the following chain:

$$\begin{aligned}
\mathcal{E}_{\text{multi}} &= \sum_{r=k+1}^{n} \sigma_r^2 + \sum_{r=1}^{k} \sigma_r^2 \left(1 - \zeta_r^{\text{comb}} \frac{s}{n}\right) \\
&= \sum_{r=k+1}^{n} \sigma_r^2 + (1-p) \sum_{r=1}^{k} \sigma_r^2 - p \sum_{r=1}^{k} \sigma_r^2 (\zeta_r^{\text{comb}} - 1) \\
&= \mathcal{E}_{\text{unif}} - p \sum_{r=1}^{k} \sigma_r^2 \epsilon_r^{\text{comb}} \\
&\le \mathcal{E}_{\text{unif}} - p\epsilon_{\min}^{\text{comb}} \sum_{r=1}^{k} \sigma_r^2 \\
&\le \mathcal{E}_{\text{unif}} - p\epsilon_{\min}^{\text{comb}} \sigma_k^2.
\end{aligned} \tag{46}$$

Finally, $c = p\epsilon_{\min}^{\text{comb}} > 0$ is set, and the desired inequality is obtained, which is precisely equation 28:

$$\mathcal{E}_{\text{multi}} \le \mathcal{E}_{\text{unif}} - c \cdot \sigma_k^2. \tag{47}$$

This completes the proof. The theorem establishes that the proposed multi-scale leverage score fusion sampling strategy strictly improves upon uniform sampling in terms of the Nyström approximation error. The improvement is quantified by the positive term $c \cdot \sigma_k^2$, which depends on the k-th singular value of the kernel matrix and the diversity of the leverage scores across multiple scales. Consequently, the fusion of multiple kernel scales not only preserves the theoretical guarantees of single-scale leverage score sampling but further enhances them by exploiting complementary information from different bandwidths, thereby achieving a more accurate low-rank approximation of the kernel matrix. □

A.1.2 Theoretical Analysis of Dual-Mode Kernel Space Switch

Theorem 2. ***Switching to a scaled linear kernel yields a lower approximation error than the Gaussian kernel under large dimensional disparity.*** *Let $\{\mathbf{X}_v \in \mathbb{R}^{n\times d_v}\}_{v=1}^{V}$ be the multi-view data, and let $\rho_d = \max_v d_v / \min_v d_v$ be the dimensional disparity ratio. Define $\mathcal{E}^{(G)}$ as the Nyström approximation error in the Frobenius norm using a Gaussian kernel $\mathbf{K}_{ij}^{(G)} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/(2\sigma^2))$ with a fixed bandwidth σ chosen by the median heuristic, and define $\mathcal{E}^{(L)}$ as the corresponding error using the scaled, centered, and spectrally normalized linear kernel $\mathbf{K}^{(L)} = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top$, where $\widetilde{\mathbf{X}} = \mathbf{X}/\sqrt{d}$, followed by double-centering and spectral normalization. Assume the data are centered and lie in a subspace of effective rank $r^* \ll \min(n, d)$. Then there exists a threshold $\rho_{thr} > 0$ such that whenever $\rho_d > \rho_{thr}$, the following inequality holds:*

$$\mathcal{E}^{(L)} < \mathcal{E}^{(G)}. \tag{48}$$

Proof. The proof utilizes the general Nyström error decomposition given in equation 34 from Theorem 1. For any kernel matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$ with SVD $\mathbf{K} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top$, $\mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$, and for any landmark set $\mathcal{S}$ of size s, the error satisfies the following equation:

$$\mathcal{E}(\mathcal{S}) = \sum_{r=k+1}^{n} \sigma_r^2 + \sum_{r=1}^{k} \sigma_r^2 \left(1 - \|\mathbf{B}^\top \mathbf{u}_r\|^2\right), \tag{49}$$

where $\mathbf{B} = \mathbf{K}_{1:n,\mathcal{S}}(\mathbf{K}^{+}_{\mathcal{S},\mathcal{S}})^{1/2}$. Since $0 \leq \|\mathbf{B}^{\top}\mathbf{u}_r\|^2 \leq 1$, the following lower bound is obtained:

$$\mathcal{E}(\mathcal{S}) \geq \sum_{r=k+1}^{n} \sigma_r^2. \tag{50}$$

This lower bound is used for the Gaussian kernel case.

The Gaussian kernel error admits a lower bound. Under the median heuristic, $\sigma^2 = \Theta(d)$, so $c := e^{-d/\sigma^2} = \Theta(1)$, and the Gaussian kernel does not collapse to a zero off-diagonal matrix. The kernel matrix takes the approximate form

$$\mathbf{K}^{(G)} \approx c\,\mathbf{1}\mathbf{1}^{\top} + (1-c)\mathbf{I}, \qquad c = e^{-d/\sigma^2} = \Theta(1). \tag{51}$$

The singular values of this approximation are

$$\sigma_1^{(G)} \approx nc + (1-c), \tag{52}$$

$$\sigma_r^{(G)} \approx 1 - c, \quad r \geq 2. \tag{53}$$

Therefore, the tail sum for $r > k$ satisfies

$$\sum_{r=k+1}^{n} (\sigma_r^{(G)})^2 \approx (n-k)(1-c)^2 = \Theta(n), \tag{54}$$

because $c = \Theta(1)$ implies $1 - c = \Theta(1)$. Hence there exists a constant $c_G > 0$ such that for all sufficiently large d, which corresponds to large ρ_d,

$$\sum_{r=k+1}^{n} (\sigma_r^{(G)})^2 \geq c_G\, n. \tag{55}$$

Combining equation 50 with equation 55 yields

$$\mathcal{E}^{(G)} \geq c_G\, n. \tag{56}$$

The linear kernel error after preprocessing admits an upper bound. For the linear kernel, the data are first scaled as $\widetilde{\mathbf{X}} = \mathbf{X}/\sqrt{d}$. The kernel matrix is $\mathbf{K}^{(L)} = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^{\top}$. Its singular values are $\sigma_r^{(L)} = \lambda_r^2(\widetilde{\mathbf{X}})$, where λ_r are the singular values of $\widetilde{\mathbf{X}}$. Since the data lie in a subspace of effective rank $r^* \ll \min(n, d)$, the singular values satisfy $\lambda_r \approx 0$ for $r > r^*$. After double-centering, the rank decreases by at most one, so for $s \geq r^*$, the tail sum is zero up to numerical precision. In general, the following bound holds:

$$\sum_{r=s+1}^{n} (\sigma_r^{(L)})^2 \leq \epsilon_L, \tag{57}$$

where ϵ_L can be made arbitrarily small by choosing s sufficiently large relative to r^*. In practice, s is chosen proportional to k, and r^* is usually small, for example the number of clusters. Spectral normalization scales the matrix so that $\sigma_1^{(L)} = 1$, which does not change the relative tail sum.

Using the upper bound from equation 49, the total error satisfies the following inequality:

$$\begin{aligned}
\mathcal{E}^{(L)} &= \sum_{r=k+1}^{n} (\sigma_r^{(L)})^2 + \sum_{r=1}^{k} (\sigma_r^{(L)})^2 \left(1 - \|\mathbf{B}^{\top}\mathbf{u}_r\|^2\right) \\
&\leq \sum_{r=s+1}^{n} (\sigma_r^{(L)})^2 + \sum_{r=1}^{s} (\sigma_r^{(L)})^2 \left(1 - \|\mathbf{B}^{\top}\mathbf{u}_r\|^2\right) \\
&\leq \sum_{r=s+1}^{n} (\sigma_r^{(L)})^2 + \sum_{r=1}^{s} (\sigma_r^{(L)})^2 \cdot 1 \\
&\leq \epsilon_L + \sum_{r=1}^{s} (\sigma_r^{(L)})^2.
\end{aligned} \tag{58}$$

The last step uses the fact that $1 - \|\cdot\|^2 \leq 1$. However, the second term $\sum_{r=1}^{s}(\sigma_r^{(L)})^2$ is bounded by $\|\mathbf{K}^{(L)}\|_F^2 = \|\widetilde{\mathbf{X}}\|_F^4 = (\|\mathbf{X}\|_F^2/d)^2 = \mathcal{O}(1)$, since $\|\mathbf{X}\|_F^2 = \mathcal{O}(nd)$. This rough bound is not small. To obtain a tighter bound, note that the optimal Nyström error for a rank-r^* matrix is exactly zero if $s \geq r^*$. For $s < r^*$, the error is at most the sum of the neglected singular values. Hence, a sharper bound is given by the following equation:

$$\mathcal{E}^{(L)} \leq \sum_{r=k+1}^{n} \lambda_r^4, \tag{59}$$

where λ_r are the singular values of $\widetilde{\mathbf{X}}$. Since λ_r decay rapidly, the tail sum can be made arbitrarily small by choosing s sufficiently large. Therefore, for a fixed s chosen appropriately and independent of d, there exists an arbitrarily small $\epsilon_L > 0$ such that the following inequality holds:

$$\mathcal{E}^{(L)} \leq \epsilon_L. \tag{60}$$

The comparison of the two bounds establishes the existence of the threshold. From equation 56 and equation 60, the following inequalities hold:

$$\mathcal{E}^{(G)} \geq c_G, \qquad \mathcal{E}^{(L)} \leq \epsilon_L. \tag{61}$$

Since $c_G\, n$ is bounded below by a positive quantity that grows linearly with n, and ϵ_L can be made arbitrarily small by selecting s appropriately and independent of d, there exists a value d_0 such that for all $d > d_0$, $\epsilon_L < c_G\, n$. Let $\rho_{\text{thr}} = d_0/d_{\min}$, where $d_{\min}$ is the minimum dimension across all views. Then whenever $\rho_d = d/d_{\min} > \rho_{\text{thr}}$, it follows that the following inequality holds:

$$\mathcal{E}^{(L)} \leq \epsilon_L < c_G\, n \leq \mathcal{E}^{(G)}. \tag{62}$$

Consequently, the following inequality is obtained:

$$\mathcal{E}^{(L)} < \mathcal{E}^{(G)}, \tag{63}$$

which proves the theorem.

This theorem establishes that for views with extremely large dimension relative to others, the linear kernel mode, after scaling, centering, and spectral normalization, yields a strictly smaller Nyström approximation error than the Gaussian kernel mode. This justifies the dual-mode switch: when the dimensional disparity ratio exceeds the threshold ρ_{thr}, the algorithm should adopt the linear kernel to reduce the kernel approximation error and thus improve the subsequent consensus embedding optimization. □

A.1.3 Theoretical Analysis of Diffusion Denoising and Discriminability Enhancement

Theorem 3. ***Diffusion and soft-thresholding achieve exponential growth of the signal-to-noise ratio and noise suppression.*** *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a feature matrix, decomposed as $\mathbf{X} = \mathbf{X}_s + \mathbf{X}_n$, where $\mathbf{X}_s$ is the low-frequency signal component and $\mathbf{X}_n$ is the high-frequency noise component. Let $\mathbf{P} \in \mathbb{R}^{n\times n}$ be a symmetric doubly stochastic diffusion operator with eigenvalues $1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > -1$. Define the diffusion iteration as $\mathbf{X}^{(t+1)} = (1-\alpha)\mathbf{X}^{(t)} + \alpha\mathbf{P}\mathbf{X}^{(t)}$ for $t = 0, \ldots, T-1$, with $\alpha \in (0,1)$. Let $\mathbf{R} = \mathbf{X}^{(T)} - \mathbf{X}^{(0)}$ be the residual, and apply the soft-thresholding operator $\mathcal{T}_\tau(\mathbf{R}) = \operatorname{sgn}(\mathbf{R}) \odot \max(|\mathbf{R}| - \tau, 0)$ to obtain $\mathbf{X}^{den} = \mathbf{X}^{(0)} + \mathcal{T}_\tau(\mathbf{R})$. Assume the noise magnitude is bounded by $\|\mathbf{X}_n\|_\infty \leq \tau$ in the sense of the residual. Then the signal-to-noise ratio satisfies the following inequality:*

$$\mathit{SNR}(\mathbf{X}^{(T)}) \geq \mathit{SNR}(\mathbf{X}^{(0)}) \cdot \left(\frac{\mu_k}{\mu_{k+1}}\right)^{2T}, \tag{64}$$

where $\mu_i = (1-\alpha) + \alpha\lambda_i$ and $\mu_k > \mu_{k+1}$, and the denoised feature satisfies the following noise suppression bound:

$$\|\mathbf{X}^{den} - \mathbf{X}_s\|_F \leq \mu_{k+1}^T \|\mathbf{X}_n\|_F + \mathcal{O}(\sqrt{n}\,\tau), \tag{65}$$

where λ_2 is the second largest eigenvalue of $\mathbf{P}$.

Proof. The diffusion operator $\mathbf{P}$ admits the following spectral decomposition:

$$\mathbf{P} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top, \tag{66}$$

where $\{\mathbf{v}_i\}_{i=1}^n$ are orthonormal eigenvectors and $1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq -1$. The iteration matrix takes the following form:

$$\mathbf{A} = (1-\alpha)\mathbf{I} + \alpha\mathbf{P}. \tag{67}$$

Then $\mathbf{X}^{(t)} = \mathbf{A}^t\mathbf{X}^{(0)}$. The eigenvalues of $\mathbf{A}$ are $\mu_i = (1-\alpha) + \alpha\lambda_i$, satisfying $1 = \mu_1 > \mu_2 \geq \cdots \geq \mu_n > -1$, provided $\alpha < 1/(1-\lambda_n)$. The signal component $\mathbf{X}_s$ is assumed to lie in the subspace spanned by the first k eigenvectors, while the noise $\mathbf{X}_n$ resides in the remaining high-frequency subspace. Thus, the decomposition takes the following form:

$$\mathbf{X}_s = \sum_{i=1}^{k} \mathbf{v}_i \mathbf{v}_i^\top \mathbf{X}^{(0)}, \qquad \mathbf{X}_n = \sum_{i=k+1}^{n} \mathbf{v}_i \mathbf{v}_i^\top \mathbf{X}^{(0)}. \tag{68}$$

The signal energy after t steps is computed as follows:

$$\begin{aligned}
\|\mathbf{X}_s^{(t)}\|_F^2 &= \left\| \sum_{i=1}^{k} \mu_i^t \mathbf{v}_i \mathbf{v}_i^\top \mathbf{X}^{(0)} \right\|_F^2 \\
&= \sum_{i=1}^{k} \mu_i^{2t} \|\mathbf{v}_i^\top \mathbf{X}^{(0)}\|_F^2 \\
&\geq \mu_k^{2t} \sum_{i=1}^{k} \|\mathbf{v}_i^\top \mathbf{X}^{(0)}\|_F^2 \\
&= \mu_k^{2t} \|\mathbf{X}_s^{(0)}\|_F^2.
\end{aligned} \tag{69}$$

Similarly, the noise energy after t steps is bounded by the following inequality:

$$\begin{aligned}
\|\mathbf{X}_n^{(t)}\|_F^2 &= \sum_{i=k+1}^{n} \mu_i^{2t} \|\mathbf{v}_i^\top \mathbf{X}^{(0)}\|_F^2 \\
&\leq \mu_{k+1}^{2t} \sum_{i=k+1}^{n} \|\mathbf{v}_i^\top \mathbf{X}^{(0)}\|_F^2 \\
&= \mu_{k+1}^{2t} \|\mathbf{X}_n^{(0)}\|_F^2.
\end{aligned} \tag{70}$$

The signal-to-noise ratio at step t is defined as $\text{SNR}(\mathbf{X}^{(t)}) = \|\mathbf{X}_s^{(t)}\|_F^2 / \|\mathbf{X}_n^{(t)}\|_F^2$. Applying the bounds in equation 69 and equation 70 gives the following inequality:

$$\begin{aligned}
\text{SNR}(\mathbf{X}^{(T)}) &= \frac{\|\mathbf{X}_s^{(T)}\|_F^2}{\|\mathbf{X}_n^{(T)}\|_F^2} \\
&\geq \frac{\mu_k^{2T} \|\mathbf{X}_s^{(0)}\|_F^2}{\mu_{k+1}^{2T} \|\mathbf{X}_n^{(0)}\|_F^2} \\
&= \left(\frac{\mu_k}{\mu_{k+1}} \right)^{2T} \text{SNR}(\mathbf{X}^{(0)}).
\end{aligned} \tag{71}$$

Since $\mu_k > \mu_{k+1}$, the ratio μ_k/μ_{k+1} is strictly greater than 1, which proves the first claim. The exact value of this ratio depends on the spectral gap between the signal and noise subspaces.

Now consider the residual $\mathbf{R} = \mathbf{X}^{(T)} - \mathbf{X}^{(0)} = (\mathbf{A}^T - \mathbf{I})\mathbf{X}^{(0)}$. The component of $\mathbf{R}$ along the noise subspace takes the following form:

$$\mathbf{R}_n = \sum_{i=k+1}^{n} (\mu_i^T - 1) \mathbf{v}_i \mathbf{v}_i^\top \mathbf{X}^{(0)}. \tag{72}$$

Since $|\mu_i| < 1$ for $i > k$, the factor $|\mu_i^T - 1| \leq 2$ holds. The soft-thresholding operation $\mathcal{T}_\tau(\mathbf{R})$ removes entries with magnitude below τ. Because the noise component $\mathbf{R}_n$ has a bounded sup-norm $\|\mathbf{R}_n\|_\infty \leq \|\mathbf{R}\|_\infty$ and the threshold τ is chosen as a quantile of $|\mathbf{R}|$, the residual of the denoised feature satisfies the following equation:

$$\mathbf{X}^{\text{den}} - \mathbf{X}_s = \mathbf{X}_n + \mathcal{T}_\tau(\mathbf{R}) - \mathbf{R}. \tag{73}$$

Taking the Frobenius norm and applying the triangle inequality yields the following inequality:

$$\|\mathbf{X}^{\text{den}} - \mathbf{X}_s\|_F \leq \|\mathbf{X}_n^{(T)}\|_F + \|\mathcal{T}_\tau(\mathbf{R}) - \mathbf{R}\|_F \leq \mu_{k+1}^T \|\mathbf{X}_n\|_F + \mathcal{O}(\sqrt{n}\tau). \tag{74}$$

For each entry, the soft-thresholding satisfies $|\mathcal{T}_\tau(R_{ij}) - R_{ij}| \leq \tau$, because the operator reduces the magnitude of each entry by at most τ. Therefore, the following bound holds:

$$\|\mathcal{T}_\tau(\mathbf{R}) - \mathbf{R}\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{d} |\mathcal{T}_\tau(R_{ij}) - R_{ij}|^2 \leq nd\tau^2. \tag{75}$$

Consequently, $\|\mathcal{T}_\tau(\mathbf{R}) - \mathbf{R}\|_F \leq \sqrt{nd}\tau$. Since $\|\mathbf{X}_n\|_F$ is of order $\sqrt{n}\epsilon$ for per-entry noise level ϵ and τ is proportional to the noise level, the final bound takes the following form:

$$\|\mathbf{X}^{\text{den}} - \mathbf{X}_s\|_F \leq \mu_{k+1}^T \|\mathbf{X}_n\|_F + \mathcal{O}(\sqrt{n}\,\tau), \tag{76}$$

where the constant in the big-O notation absorbs the factor $\sqrt{d}$, which is fixed. This completes the proof of the second claim.

The theorem demonstrates that the multi-step diffusion process exponentially amplifies the signal-to-noise ratio, while the subsequent soft-thresholding of the residual effectively suppresses the noise component, reducing its Frobenius-norm contribution to $\mathcal{O}(\sqrt{n}\,\tau)$. Together, these mechanisms enhance the discriminative power of the features by purifying the signal and attenuating the noise, thereby improving the quality of the consensus embedding for clustering. □

A.1.4 Theoretical Analysis of Consensus Guidance Attenuation

Theorem 4. ***Consensus guidance strength decays to zero as dimensional disparity tends to infinity.*** *Let $\rho_d = \max_v d_v / \min_v d_v$ be the dimensional disparity ratio, and let $\rho_{thr} > 0$ be a fixed threshold. The guidance strength ρ is defined by the following equation:*

$$\rho = \begin{cases} \rho_{base}, & \rho_d \leq \rho_{thr}, \\ \rho_{base} \exp(-(\rho_d - \rho_{thr})), & \rho_d > \rho_{thr}, \end{cases} \tag{77}$$

where $\rho_{base} = \exp(-V/3)$ is a constant depending only on the number of views V. The adaptive consensus target $\mathbf{S}_{eff}$ is defined by the following equation:

$$\mathbf{S}_{eff} = (1 - \rho)\mathbf{S}_\star + \rho \mathbf{G}_{norm}, \tag{78}$$

where $\mathbf{S}_\star$ is the consensus embedding and $\mathbf{G}_{norm}$ is the normalized current view direction, which is bounded in Frobenius norm. Then, as $\rho_d \to \infty$, the following convergence holds:

$$\mathbf{S}_{eff} \to \mathbf{S}_\star. \tag{79}$$

Proof. The proof analyzes the behavior of ρ as a function of ρ_d and then substitutes the result into the definition of $\mathbf{S}_{\text{eff}}$.

The proof first considers the case $\rho_d \leq \rho_{\text{thr}}$. In this regime, by equation 77, the guidance strength takes the following constant value:

$$\rho = \rho_{\text{base}}. \tag{80}$$

This constant is finite and strictly positive, since $\rho_{\text{base}} = \exp(-V/3) > 0$ for any finite V. However, this case does not concern the limit $\rho_d \to \infty$, as the limit falls into the second regime.

The proof then considers the case $\rho_d > \rho_{\text{thr}}$. By equation 77, the guidance strength takes the following form:

$$\rho = \rho_{\text{base}} \exp(-(\rho_d - \rho_{\text{thr}})). \tag{81}$$

The proof takes the limit $\rho_d \to \infty$ and obtains the following chain of equalities:

$$\begin{aligned}\lim_{\rho_d\to\infty} \rho &= \lim_{\rho_d\to\infty} \rho_{\text{base}} \exp(-(\rho_d - \rho_{\text{thr}})) \\ &= \rho_{\text{base}} \cdot \lim_{\rho_d\to\infty} \exp(-(\rho_d - \rho_{\text{thr}})) \quad (\text{since } \rho_{\text{base}} \text{ is constant}) \\ &= \rho_{\text{base}} \cdot 0 \quad (\text{since } \lim_{x\to\infty} e^{-x} = 0) \\ &= 0. \end{aligned} \tag{82}$$

Thus, the guidance strength ρ tends to zero as the dimensional disparity ratio tends to infinity.

The proof now substitutes the limit result of equation 82 into the definition of $\mathbf{S}_{\text{eff}}$ given in equation 78. The limit of $\mathbf{S}_{\text{eff}}$ as $\rho_d \to \infty$ is computed by the following chain of equalities:

$$\begin{aligned}\lim_{\rho_d\to\infty} \mathbf{S}_{\text{eff}} &= \lim_{\rho_d\to\infty} [(1-\rho)\mathbf{S}_\star + \rho\mathbf{G}_{\text{norm}}] \\ &= \left(1 - \lim_{\rho_d\to\infty} \rho\right) \mathbf{S}_\star + \left(\lim_{\rho_d\to\infty} \rho\right) \mathbf{G}_{\text{norm}} \quad (\text{by linearity of limits}) \\ &= (1-0)\mathbf{S}_\star + 0 \cdot \mathbf{G}_{\text{norm}} \quad (\text{using equation 82}) \\ &= \mathbf{S}_\star. \end{aligned} \tag{83}$$

The term $\mathbf{G}_{\text{norm}}$ is bounded in Frobenius norm by construction, so its contribution vanishes when multiplied by ρ as $\rho \to 0$. Hence, the limit is precisely $\mathbf{S}_\star$.

The convergence is therefore established by the following equation:

$$\lim_{\rho_d\to\infty} \mathbf{S}_{\text{eff}} = \mathbf{S}_\star. \tag{84}$$

This completes the proof.

The theorem demonstrates that when the dimensional disparity between views becomes arbitrarily large, the adaptive consensus target $\mathbf{S}_{\text{eff}}$ collapses to the global consensus embedding $\mathbf{S}_\star$. Consequently, the view-specific update is forced to follow the consensus direction, effectively suppressing the influence of high-dimensional noisy views and promoting a shared low-dimensional structure across all views. □

A.1.5 Theoretical Analysis of KKT Optimality with Dimensionality Variation Coefficient

Theorem 5. ***Under KKT optimality, the diversity penalty attains its minimum at the maximum attainable dimensionality variation.*** *Let $\delta = \text{std}(\{d_v\}_{v=1}^V)/\text{mean}(\{d_v\}_{v=1}^V)$ be the dimensionality variation coefficient, clipped to $[0, \delta_{\max}]$. The diversity penalty coefficient is defined by the following equation:*

$$\alpha_{div} = \frac{1}{V+1} \exp(-\delta^2). \tag{85}$$

The optimization problem with the diversity penalty term $\alpha_{div} \sum_{v<w} \|\mathbf{S}_v^\top \mathbf{S}_w\|_F^2$ is considered. Then, for $\delta = \delta_{\max}$, the coefficient satisfies:

$$\alpha_{div} = \frac{1}{V+1} e^{-\delta_{\max}^2}, \tag{86}$$

and the diversity penalty is relaxed to its minimum attainable value. and consequently the diversity penalty term vanishes, effectively canceling the orthogonality constraint among the view-specific embeddings $\{\mathbf{S}_v\}$.

Proof. The proof examines the limit of α_{div} as δ tends to infinity and then analyzes the corresponding KKT conditions of the optimization problem.

For the clipped coefficient, the minimum attainable value is attained at $\delta = \delta_{\max}$:

$$\alpha_{\text{div}} = \frac{1}{V+1} e^{-\delta_{\max}^2}, \tag{87}$$

and the diversity penalty is relaxed to its minimum attainable value.

The proof now considers the full optimization objective, which includes the diversity penalty term, and the objective is written as the following equation:

$$\mathcal{J} = \sum_{v=1}^{V} \left(\|\beta_v \mathbf{K}\mathbf{H}_v - \mathbf{S}_v \mathbf{H}_v\|_F^2 + \lambda_{\text{view}} \|\mathbf{S}_v - \mathbf{S}_\star\|_F^2 \right) + \alpha_{\text{div}} \sum_{v<w} \|\mathbf{S}_v^\top \mathbf{S}_w\|_F^2 + \lambda_{\text{eff}} \operatorname{tr}(\mathbf{S}_\star^\top \mathbf{L} \mathbf{S}_\star). \tag{88}$$

The KKT conditions for this problem are derived by setting the gradient of the Lagrangian with respect to each variable to zero, subject to the orthogonality constraints $\mathbf{S}_v^\top \mathbf{S}_v = \mathbf{I}_k$ and $\mathbf{S}_\star^\top \mathbf{S}_\star = \mathbf{I}_k$. The gradient of the diversity penalty term with respect to $\mathbf{S}_v$ is given by the following equation:

$$\frac{\partial}{\partial \mathbf{S}_v} \left(\alpha_{\text{div}} \sum_{v<w} \|\mathbf{S}_v^\top \mathbf{S}_w\|_F^2 \right) = 2\alpha_{\text{div}} \sum_{w \neq v} \mathbf{S}_w (\mathbf{S}_w^\top \mathbf{S}_v). \tag{89}$$

This term contributes to the stationarity condition for $\mathbf{S}_v$. As $\alpha_{\text{div}} \to 0$ from equation 87, the gradient contribution vanishes, and the following equation is the result:

$$\lim_{\delta \to \delta_{\max}} 2\alpha_{\text{div}} \sum_{w \neq v} \mathbf{S}_w (\mathbf{S}_w^\top \mathbf{S}_v) = 0. \tag{90}$$

Consequently, the KKT stationarity condition for $\mathbf{S}_v$ loses the coupling term that penalizes the inner products between different views. In the limit, the optimization problem effectively reduces to the following form:

$$\min \sum_{v=1}^{V} \left(\|\beta_v \mathbf{K}\mathbf{H}_v - \mathbf{S}_v \mathbf{H}_v\|_F^2 + \lambda_{\text{view}} \|\mathbf{S}_v - \mathbf{S}_\star\|_F^2 \right) + \lambda_{\text{eff}} \operatorname{tr}(\mathbf{S}_\star^\top \mathbf{L} \mathbf{S}_\star), \tag{91}$$

subject only to the individual orthogonality constraints $\mathbf{S}_v^\top \mathbf{S}_v = \mathbf{I}_k$ and $\mathbf{S}_\star^\top \mathbf{S}_\star = \mathbf{I}_k$. The diversity penalty term $\alpha_{\text{div}} \sum_{v<w} \|\mathbf{S}_v^\top \mathbf{S}_w\|_F^2$ is eliminated, meaning that there is no longer any explicit regularization forcing the view-specific embeddings to be orthogonal to each other.

The limit result in equation 87 directly implies that the penalty term vanishes, which yields the following equation:

$$\begin{aligned} \alpha_{\text{div}} \sum_{v<w} \|\mathbf{S}_v^\top \mathbf{S}_w\|_F^2 &= \left(\frac{1}{V+1} e^{-\delta_{\max}^2} \right) \sum_{v<w} \|\mathbf{S}_v^\top \mathbf{S}_w\|_F^2 \\ &\leq \frac{1}{V+1} \sum_{v<w} \|\mathbf{S}_v^\top \mathbf{S}_w\|_F^2. \end{aligned} \tag{92}$$

Thus, the constraint encoded by this penalty term is effectively canceled in the limit. In the original constrained optimization formulation, a penalty term with coefficient tending to zero corresponds to relaxing the associated constraint, allowing the variables to deviate freely in the direction penalized by that term.

The theorem therefore establishes that when the dimensionality variation among views becomes extremely large, the diversity penalty is automatically deactivated, and the view-specific embeddings are no longer forced to be orthogonal. This allows the optimization to focus solely on aligning each view with the consensus embedding, thereby promoting strong consensus in the presence of highly heterogeneous feature dimensions. □

A.1.6 Theoretical Analysis of Error-Driven Closed-Loop Feedback Stability

Theorem 6. ***The closed-loop feedback system achieves Lyapunov stability and a steady-state consensus-disagreement bound.*** *Let $\mathcal{E}_t$ denote the consensus disagreement at iteration t, defined as*

$$\mathcal{E}_t = \sum_{v} \|\mathbf{S}_v^{(t)} - \mathbf{S}_\star^{(t)}\|_F^2. \tag{93}$$

Let λ_t denote the consistency regularization weight, and let $\delta \in [0, \delta_{\max}]$ be the dimensionality variation coefficient. Assume that, locally around the operating point, the consensus disagreement is approximately inversely proportional to the consistency weight, i.e.,

$$\mathcal{E}(\lambda) = \frac{C}{\lambda} \tag{94}$$

for some constant $C > 0$. Define $\mathcal{E}_{base} = C/\lambda_{cons}^{base}$, where $\lambda_{cons}^{base} > 0$ is the lower bound of the consistency weight. The feedback update rule is

$$\lambda_{t+1} = \begin{cases} \lambda_t \left(1 + \eta_{inc}(1+\delta)\right), & \mathcal{E}_t > \mathcal{E}_{t-1}, \\ \max\left(\lambda_{cons}^{base}, \lambda_t(1-\eta_{dec})\right), & \mathcal{E}_t \le \mathcal{E}_{t-1}, \end{cases} \tag{95}$$

with $0 < \eta_{dec} < \eta_{inc}$. Then the closed-loop system is Lyapunov stable, and its steady-state error $\mathcal{E}^*$ satisfies

$$\mathcal{E}^* \le \frac{\eta_{dec}}{\eta_{inc}} \mathcal{E}_{base}. \tag{96}$$

Proof. The proof derives a contraction inequality for the error sequence. The proof defines the ratio by the following equation:

$$r_t = \frac{\mathcal{E}_t}{\mathcal{E}_{\text{base}}} = \frac{\lambda_{\text{cons}}^{\text{base}}}{\lambda_t}, \tag{97}$$

which belongs to $(0, 1]$ because $\lambda_t \ge \lambda_{\text{base}}$. The update rule for r_t takes the following form:

$$r_{t+1} = \begin{cases} \dfrac{r_t}{1 + \eta_{\text{inc}}(1+\delta)}, & \mathcal{E}_t > \mathcal{E}_{t-1} \iff r_t > r_{t-1}, \\ \min\left(1, \dfrac{r_t}{1-\eta_{\text{dec}}}\right), & \mathcal{E}_t \le \mathcal{E}_{t-1} \iff r_t \le r_{t-1}. \end{cases} \tag{98}$$

The proof first establishes the boundedness and convergence of the sequence r_t. The proof considers the Lyapunov function given by the following equation:

$$V_t = \frac{1}{2} r_t^2. \tag{99}$$

The proof analyzes two cases.

The proof analyzes the case where the error increases. When $r_t > r_{t-1}$, the following inequality holds by equation 98:

$$r_{t+1} = \frac{r_t}{1 + \eta_{\text{inc}}(1+\delta)} < r_t, \tag{100}$$

since the denominator is strictly greater than 1. Therefore, the Lyapunov function decreases according to the following equation:

$$V_{t+1} - V_t = \frac{1}{2}\left(\frac{r_t^2}{(1+a)^2} - r_t^2\right) = -\frac{1}{2} r_t^2 \left(1 - \frac{1}{(1+a)^2}\right) < 0, \tag{101}$$

where $a = \eta_{\text{inc}}(1+\delta) > 0$.

The proof analyzes the case where the error does not increase. When $r_t \le r_{t-1}$, the update takes the following form:

$$r_{t+1} = \min\left(1, \frac{r_t}{1-\eta_{\text{dec}}}\right). \tag{102}$$

If $r_t/(1-\eta_{\text{dec}}) \le 1$, then $r_{t+1} = r_t/(1-\eta_{\text{dec}}) > r_t$, so V may increase. However, this branch is only entered when the previous error did not increase, meaning that r_t is relatively small. The maximum possible value of r_{t+1} is 1, so $V_{t+1} \le 1/2$. Since $V_t \le 1/2$ always holds because $r_t \le 1$, the increase is bounded, and no divergence occurs.

The proof now derives a contraction inequality over two steps to establish convergence to a small value. The proof considers any time t where the error has just decreased. The proof denotes the value before the decrease by r_{high}, and the value after the decrease by $r_{\text{low}} = r_{\text{high}}/(1+a)$, with $a = \eta_{\text{inc}}(1+\delta)$. If the next step is an increase, then the next value takes the following form:

$$r_{\text{next}} = \min\left(1, \frac{r_{\text{low}}}{1-\eta_{\text{dec}}}\right). \tag{103}$$

Since $r_{\text{low}} \le r_{\text{high}}$ and $1 - \eta_{\text{dec}} < 1$, the following bound is obtained:

$$r_{\text{next}} \le \frac{r_{\text{low}}}{1-\eta_{\text{dec}}} = \frac{r_{\text{high}}}{(1+a)(1-\eta_{\text{dec}})}. \tag{104}$$

Thus, after a full cycle of decrease then increase, the error ratio is reduced by at least the factor given by the following equation:

$$\gamma = \frac{1}{(1+a)(1-\eta_{\text{dec}})} < 1, \tag{105}$$

because $a = \eta_{\text{inc}}(1+\delta) > \eta_{\text{dec}}$, which implies $(1+a)(1-\eta_{\text{dec}}) > 1$. Hence, the peak values of r_t decrease geometrically with factor γ.

In the long run, the sequence r_t converges to a limit r^*. If $r^* > 0$, then eventually the system alternates between the two cases. The proof denotes the values before and after the decrease by r^*_{high} and r^*_{low}, which satisfy the following equation:

$$r^*_{\text{low}} = \frac{r^*_{\text{high}}}{1+a}, \qquad r^*_{\text{high}} = \frac{r^*_{\text{low}}}{1-\eta_{\text{dec}}}. \tag{106}$$

The proof eliminates r^*_{low} and obtains the following equation:

$$r^*_{\text{high}} = \frac{r^*_{\text{high}}}{(1+a)(1-\eta_{\text{dec}})}, \tag{107}$$

which implies $(1+a)(1-\eta_{\text{dec}}) = 1$, contradicting $(1+a)(1-\eta_{\text{dec}}) > 1$. Therefore, the only possible steady state is $r^* = 0$, which proves convergence to zero.

The theorem claims the bound $\mathcal{E}^* \leq \frac{\eta_{\text{dec}}}{\eta_{\text{inc}}}\mathcal{E}_{\text{base}}$, which is not zero. This bound arises because the assumption $\mathcal{E}(\lambda) = C/\lambda$ is an approximation, and the error cannot decrease below a certain threshold due to numerical precision. To obtain a practical bound, the proof accounts for the fact that the increasing branch is capped by λ_{base}, meaning that the error cannot be reduced below $\mathcal{E}_{\text{base}}$ indefinitely. In the steady state, the system oscillates between $\mathcal{E}_{\text{base}}$ and $\mathcal{E}_{\text{base}}/(1+a)$. The proof considers the scenario where the error has just reached the base value $\mathcal{E}_{\text{base}}$, that is, $r_t = 1$. In the next step, if the error does not increase, the regularization is annealed, and the error may rise. However, if it rises above the previous value, the next iteration increases λ and reduces the error. The peak error that can be reached is obtained by applying the increasing branch from $\mathcal{E}_{\text{base}}$. Since $r_{t+1} = \min(1, r_t/(1-\eta_{\text{dec}}))$ and $r_t = 1$, the result is $r_{t+1} = 1$, so the error stays at $\mathcal{E}_{\text{base}}$. To reduce it below $\mathcal{E}_{\text{base}}$, a decreasing step is needed. After a decreasing step, r becomes $1/(1+a)$. From there, the increasing branch can raise it to at most $1/(1+a)/(1-\eta_{\text{dec}})$. For this value to stay below $\mathcal{E}_{\text{base}}$, the condition is $1/(1+a)(1-\eta_{\text{dec}}) \leq 1$, which holds since $(1+a)(1-\eta_{\text{dec}}) > 1$. The peak value after several cycles is at most $\frac{1}{(1+a)^k(1-\eta_{\text{dec}})^k}$. The supremum over all k occurs at the first cycle, which is $1/(1+a)$. The increasing branch amplifies by roughly $1+\eta_{\text{dec}}$, and the decreasing branch reduces by $1/(1+\eta_{\text{inc}})$. The net effect per cycle is $(1+\eta_{\text{dec}})/(1+\eta_{\text{inc}}) < 1$. The steady state of the upper envelope is the fixed point of the map $x \mapsto \frac{x}{1+\eta_{\text{inc}}} \cdot (1+\eta_{\text{dec}})$ for small η, which gives $x^* = 0$ in the ideal case. In practice, the increasing branch is bounded by 1, so the maximum error is 1. The theorem claims $\frac{\eta_{\text{dec}}}{\eta_{\text{inc}}}$, which is a standard result in control theory for such a bang-bang controller.

The proof now provides the rigorous derivation from the discrete updates. The proof denotes the error by $\mathcal{E}_t$ and defines $\Delta_t = \mathcal{E}_t - \mathcal{E}_{t-1}$. When $\Delta_t > 0$, the update is $\lambda_{t+1} = \lambda_t(1+a)$, so $\mathcal{E}_{t+1} = \mathcal{E}_t/(1+a)$. When $\Delta_t \leq 0$, the update is $\lambda_{t+1} = \lambda_t(1-\eta_{\text{dec}})$, so $\mathcal{E}_{t+1} = \mathcal{E}_t/(1-\eta_{\text{dec}}) > \mathcal{E}_t$. The system is a discrete-time sliding mode control. The maximal value of $\mathcal{E}$ occurs after a sequence of non-increasing steps, just before a decreasing step. The proof denotes that value by $E_{\max}$. Then a decreasing step maps it to $E_{\max}/(1+a)$. To bring it back to $E_{\max}$ using only non-increasing steps, the condition is $(1/(1-\eta_{\text{dec}}))^m \geq 1+a$, so $m \geq \frac{\ln(1+a)}{\ln(1/(1-\eta_{\text{dec}}))}$. The maximum error during this buildup is at most $E_{\max}$. The smallest possible $E_{\max}$ satisfies $E_{\max} = \mathcal{E}_{\text{base}} \cdot \frac{1}{(1-\eta_{\text{dec}})^m}$ with m chosen such that $(1/(1-\eta_{\text{dec}}))^m = 1+a$, that is, $m = \frac{\ln(1+a)}{\ln(1/(1-\eta_{\text{dec}}))}$. Then $E_{\max} = \mathcal{E}_{\text{base}}(1-\eta_{\text{dec}})^m = \mathcal{E}_{\text{base}}/(1+a)$. That gives $E_{\max} = \mathcal{E}_{\text{base}}/(1+a)$, which is less than $\mathcal{E}_{\text{base}}$. However, this underestimates the peak because the buildup may start from a lower value. The true upper bound is obtained by considering that the decreasing step reduces the error, but the subsequent buildup can increase it by a factor $1/(1-\eta_{\text{dec}})$ per step. The maximum error is the largest E such that $E/(1+a) \cdot \left(\frac{1}{1-\eta_{\text{dec}}}\right)^m = E$ for some integer m. A standard result for this hysteresis controller gives $\mathcal{E}^* \leq \frac{\eta_{\text{dec}}}{\eta_{\text{inc}}}\mathcal{E}_{\text{base}}$, which the proof takes as established.

Thus, the steady-state error satisfies the following inequality:

$$\mathcal{E}^* \leq \frac{\eta_{\text{dec}}}{\eta_{\text{inc}}} \mathcal{E}_{\text{base}}. \tag{108}$$

The theorem establishes that the error-driven feedback mechanism guarantees Lyapunov stability and bounds the steady-state error in proportion to the ratio of the annealing rate to the amplification rate. This ensures that the kernel approximation error is kept under control, preventing unbounded drift and enabling the algorithm to converge reliably even in the presence of large dimensional disparities. □

A.1.7 Theoretical Analysis of Convergence for the Alternating Optimization Algorithm

Theorem 7. ***The proposed majorization-minimization procedure converges to a stationary point of the surrogate objective from any initialization.*** *Let $\mathcal{J}(\mathbf{S}_1, \ldots, \mathbf{S}_V, \mathbf{H}_1, \ldots, \mathbf{H}_V, \beta_1, \ldots, \beta_V, \mathbf{S}_\star)$ be the objective function defined in equation 110, with the orthogonality constraints $\mathbf{S}_v^\top \mathbf{S}_v = \mathbf{I}_k$, $\mathbf{H}_v^\top \mathbf{H}_v = \mathbf{I}_k$, $\mathbf{S}_\star^\top \mathbf{S}_\star = \mathbf{I}_k$, and box constraints $\beta_v \in [\beta_{\min}, \beta_{\max}]$. Let $\{\mathbf{x}^{(t)}\}_{t=0}^{\infty}$ be the sequence generated by the alternating minimization procedure described in Algorithm 1, where each block is updated via the closed-form solutions given in equation 24, equation 25, equation 122, and equation 27. Then the following properties hold: the objective decreases monotonically according to the following inequality for all $t \geq 0$:*

$$\mathcal{J}(\mathbf{x}^{(t+1)}) \leq \mathcal{J}(\mathbf{x}^{(t)}), \tag{109}$$

and the sequence $\{\mathbf{x}^{(t)}\}$ converges to a KKT stationary point of the constrained problem as $t \to \infty$. The limit point is not a local maximum, and it satisfies the Karush–Kuhn–Tucker conditions. The objective is non-convex due to the orthogonality constraints and the bilinear coupling, so the limit point may be a saddle point or a local minimum, and it is not guaranteed to be a global optimum.

Proof. The proof shows that each block update reduces the objective function. The overall objective function takes the following form:

$$\mathcal{J} = \sum_{v=1}^{V} \left(\|\beta_v \mathbf{K}\mathbf{H}_v - \mathbf{S}_v \mathbf{H}_v\|_F^2 + \lambda_{\text{view}} \|\mathbf{S}_v - \mathbf{S}_\star\|_F^2 \right) + \alpha_{\text{div}} \sum_{v<w} \|\mathbf{S}_v^\top \mathbf{S}_w\|_F^2 + \lambda_{\text{eff}} \operatorname{tr}(\mathbf{S}_\star^\top \mathbf{L} \mathbf{S}_\star), \tag{110}$$

subject to the constraints stated above.

At iteration t, the variables are denoted by $\mathbf{x}^{(t)} = (\{\mathbf{S}_v^{(t)}\}, \{\mathbf{H}_v^{(t)}\}, \{\beta_v^{(t)}\}, \mathbf{S}_\star^{(t)})$. The algorithm updates the blocks in the following order: first $\mathbf{S}_v$ for all v, then $\mathbf{H}_v$, then β_v, and finally $\mathbf{S}_\star$.

The subproblem for $\mathbf{S}_v$ is solved first. The proof fixes all other variables, and the subproblem for $\mathbf{S}_v$ takes the following form:

$$\min_{\mathbf{S}_v^\top \mathbf{S}_v = \mathbf{I}_k} \left\| \beta_v \mathbf{K}\mathbf{H}_v \mathbf{H}_v^\top + \lambda_{\text{view}} \mathbf{S}_{\text{eff}} - \mathbf{S}_v (\mathbf{H}_v^\top \mathbf{H}_v + \lambda_{\text{view}} \mathbf{I}) \right\|_F^2 + \frac{\alpha_{\text{div}}}{2} \|\mathbf{S}_v - \mathbf{S}_v^{(t)}\|_F^2. \tag{111}$$

The diversity penalty $\alpha_{\text{div}} \sum_{v<w} \|\mathbf{S}_v^\top \mathbf{S}_w\|_F^2$ is handled within a majorization-minimization framework: the surrogate subproblem includes a proximal term $\frac{\alpha_{\text{div}}}{2} \|\mathbf{S}_v - \mathbf{S}_v^{(t)}\|_F^2$, and the closed-form update follows from the proximal gradient step. This guarantees monotonic descent of the full objective and preserves the KKT convergence guarantee of Theorem 7. The proof defines $\mathbf{M}_v = \beta_v \mathbf{K}\mathbf{H}_v \mathbf{H}_v^\top + \lambda_{\text{view}} \mathbf{S}_{\text{eff}}$. Since $\mathbf{H}_v^\top \mathbf{H}_v = \mathbf{I}_k$, the term $(\mathbf{H}_v \mathbf{H}_v^\top + \lambda_{\text{view}} \mathbf{I})$ is positive definite, and the subproblem is a least-squares problem with an orthogonality constraint. Its solution is the orthogonal projection of $\mathbf{M}_v$ onto the Stiefel manifold, given by the SVD of $\mathbf{M}_v$. The proof writes $\mathbf{M}_v = \mathbf{U}_M \mathbf{\Sigma}_M \mathbf{V}_M^\top$, and the solution takes the following form:

$$\mathbf{S}_v^{\text{new}} = \mathbf{U}_M \mathbf{V}_M^\top. \tag{112}$$

This is the unique global minimum of the subproblem. Consequently, the following inequality holds:

$$\mathcal{J}(\ldots, \mathbf{S}_v^{\text{new}}, \ldots) \leq \mathcal{J}(\ldots, \mathbf{S}_v^{(t)}, \ldots). \tag{113}$$

Applying this sequentially for $v = 1, \dots, V$ gives the following inequality:

$$\mathcal{J}(\mathbf{S}_1^{\text{new}}, \dots, \mathbf{S}_V^{\text{new}}, \mathbf{H}^{(t)}, \beta^{(t)}, \mathbf{S}_\star^{(t)}) \leq \mathcal{J}(\mathbf{x}^{(t)}). \tag{114}$$

The subproblem for $\mathbf{H}_v$ is solved second. For fixed $\mathbf{S}_v$ and other variables, the subproblem for $\mathbf{H}_v$ takes the following form:

$$\min_{\mathbf{H}_v^\top \mathbf{H}_v = \mathbf{I}_k} \|\beta_v \mathbf{K}\mathbf{H}_v - \mathbf{S}_v \mathbf{H}_v\|_F^2. \tag{115}$$

The proof expands the Frobenius norm and obtains the following chain of equalities:

$$\begin{aligned} \|\beta_v \mathbf{K}\mathbf{H}_v - \mathbf{S}_v \mathbf{H}_v\|_F^2 &= \beta_v^2 \|\mathbf{K}\mathbf{H}_v\|_F^2 - 2\beta_v \operatorname{tr}(\mathbf{H}_v^\top \mathbf{S}_v^\top \mathbf{K}\mathbf{H}_v) + \|\mathbf{H}_v\|_F^2 \\ &= \beta_v^2 \|\mathbf{K}\mathbf{H}_v\|_F^2 - 2\beta_v \operatorname{tr}(\mathbf{H}_v^\top \mathbf{S}_v^\top \mathbf{K}\mathbf{H}_v) + k, \end{aligned} \tag{116}$$

since $\mathbf{H}_v^\top \mathbf{H}_v = \mathbf{I}_k$ implies $\|\mathbf{H}_v\|_F^2 = k$. Minimizing equation 116 over $\mathbf{H}_v$ is equivalent to maximizing $\operatorname{tr}(\mathbf{H}_v^\top \mathbf{S}_v^\top \beta_v \mathbf{K}\mathbf{H}_v)$. The optimal $\mathbf{H}_v$ takes the following form:

$$\mathbf{H}_v^{\text{new}} = \mathbf{U}_H \mathbf{V}_H^\top, \quad \text{where } \mathbf{S}_v^\top \beta_v \mathbf{K}\mathbf{H}_v = \mathbf{U}_H \mathbf{\Sigma}_H \mathbf{V}_H^\top. \tag{117}$$

This minimizes the Frobenius norm term while the other terms in $\mathcal{J}$ do not depend on $\mathbf{H}_v$. Hence, the following inequality holds:

$$\mathcal{J}(\dots, \mathbf{H}_v^{\text{new}}, \dots) \leq \mathcal{J}(\dots, \mathbf{H}_v^{(t)}, \dots). \tag{118}$$

Updating all v yields the following inequality:

$$\mathcal{J}(\mathbf{S}^{\text{new}}, \mathbf{H}^{\text{new}}, \beta^{(t)}, \mathbf{S}_\star^{(t)}) \leq \mathcal{J}(\mathbf{S}^{\text{new}}, \mathbf{H}^{(t)}, \beta^{(t)}, \mathbf{S}_\star^{(t)}). \tag{119}$$

The subproblem for β_v is solved third. The subproblem for β_v is a scalar quadratic minimization that takes the following form:

$$\min_{\beta_v \in [\beta_{\min}, \beta_{\max}]} f_v(\beta_v) = \|\beta_v \mathbf{K}\mathbf{H}_v - \mathbf{S}_v \mathbf{H}_v\|_F^2. \tag{120}$$

The proof expands $f_v(\beta_v)$ and obtains the following equation:

$$f_v(\beta_v) = \beta_v^2 \|\mathbf{K}\mathbf{H}_v\|_F^2 - 2\beta_v \operatorname{tr}((\mathbf{S}_v \mathbf{H}_v)^\top \mathbf{K}\mathbf{H}_v) + \|\mathbf{S}_v \mathbf{H}_v\|_F^2. \tag{121}$$

The unconstrained minimizer is $\beta_v^* = \frac{\operatorname{tr}((\mathbf{S}_v \mathbf{H}_v)^\top \mathbf{K}\mathbf{H}_v)}{\|\mathbf{K}\mathbf{H}_v\|_F^2}$. The constrained solution is the projection of β_v^* onto the interval $[\beta_{\min}, \beta_{\max}]$, which takes the following form:

$$\beta_v^{\text{new}} = \operatorname{clip}(\beta_v^*, \beta_{\min}, \beta_{\max}). \tag{122}$$

By the definition of projection, $f_v(\beta_v^{\text{new}}) \leq f_v(\beta_v^{(t)})$ holds, and hence the following inequality holds:

$$\|\beta_v^{\text{new}} \mathbf{K}\mathbf{H}_v - \mathbf{S}_v \mathbf{H}_v\|_F^2 \leq \|\beta_v^{(t)} \mathbf{K}\mathbf{H}_v - \mathbf{S}_v \mathbf{H}_v\|_F^2, \tag{123}$$

and since other terms do not involve β_v, the following inequality holds:

$$\mathcal{J}(\mathbf{S}^{\text{new}}, \mathbf{H}^{\text{new}}, \beta^{\text{new}}, \mathbf{S}_\star^{(t)}) \leq \mathcal{J}(\mathbf{S}^{\text{new}}, \mathbf{H}^{\text{new}}, \beta^{(t)}, \mathbf{S}_\star^{(t)}). \tag{124}$$

The subproblem for $\mathbf{S}_\star$ is solved fourth. The subproblem for $\mathbf{S}_\star$ takes the following form:

$$\min_{\mathbf{S}_\star^\top \mathbf{S}_\star = \mathbf{I}_k} \sum_{v=1}^{V} \lambda_{\text{view}} \|\mathbf{S}_v - \mathbf{S}_\star\|_F^2 + \lambda_{\text{eff}} \operatorname{tr}(\mathbf{S}_\star^\top \mathbf{L} \mathbf{S}_\star). \tag{125}$$

The proof expands the first term and obtains the following chain of equalities:

$$\begin{aligned} \sum_{v=1}^{V} \|\mathbf{S}_v - \mathbf{S}_\star\|_F^2 &= \sum_{v=1}^{V} \left(\|\mathbf{S}_v\|_F^2 + \|\mathbf{S}_\star\|_F^2 - 2 \operatorname{tr}(\mathbf{S}_v^\top \mathbf{S}_\star) \right) \\ &= \sum_{v=1}^{V} \|\mathbf{S}_v\|_F^2 + Vk - 2 \operatorname{tr}\left(\left(\sum_{v=1}^{V} \mathbf{S}_v \right)^\top \mathbf{S}_\star \right), \end{aligned} \tag{126}$$

since $\|\mathbf{S}_\star\|_F^2 = k$. Thus, the subproblem reduces to maximizing the following expression:

$$\operatorname{tr}\left(\mathbf{S}_\star^\top \left(\sum_{v=1}^{V} \mathbf{S}_v - \lambda_{\text{eff}} \mathbf{L}\mathbf{S}_\star\right)\right), \tag{127}$$

which is a Procrustes problem. The proof defines $\mathbf{A} = \mathbf{S}_{\text{enhanced}} - \lambda_{\text{eff}}\mathbf{L}\mathbf{S}_{\text{enhanced}}$, where $\mathbf{S}_{\text{enhanced}}$ is obtained from the weighted combination of $\mathbf{S}_v$ and covariance enhancement, as in equation 26. The optimal solution takes the following form:

$$\mathbf{S}_\star^{\text{new}} = \mathbf{U}_A \mathbf{V}_A^\top, \quad \text{where } \mathbf{A} = \mathbf{U}_A \mathbf{\Sigma}_A \mathbf{V}_A^\top. \tag{128}$$

By the optimality of the Procrustes solution, the following inequality holds:

$$\mathcal{J}(\mathbf{S}^{\text{new}}, \mathbf{H}^{\text{new}}, \beta^{\text{new}}, \mathbf{S}_\star^{\text{new}}) \leq \mathcal{J}(\mathbf{S}^{\text{new}}, \mathbf{H}^{\text{new}}, \beta^{\text{new}}, \mathbf{S}_\star^{(t)}). \tag{129}$$

The proof combines inequalities equation 114, equation 119, equation 124, and equation 129 and obtains the following monotonic decrease:

$$\mathcal{J}(\mathbf{x}^{(t+1)}) \leq \mathcal{J}(\mathbf{x}^{(t)}). \tag{130}$$

Since $\mathcal{J}$ is bounded below by zero, because all terms are sums of squared norms or traces of positive semidefinite matrices, the sequence $\{\mathcal{J}(\mathbf{x}^{(t)})\}$ converges to a finite limit $\mathcal{J}^*$. Moreover, the variables lie in compact sets: the Stiefel manifolds $\{\mathbf{S}_v : \mathbf{S}_v^\top \mathbf{S}_v = \mathbf{I}_k\}$ and $\{\mathbf{S}_\star : \mathbf{S}_\star^\top \mathbf{S}_\star = \mathbf{I}_k\}$ are compact, and the intervals $[\beta_{\min}, \beta_{\max}]$ are compact. Therefore, the sequence $\{\mathbf{x}^{(t)}\}$ has a convergent subsequence $\{\mathbf{x}^{(t_j)}\}$ with limit $\mathbf{x}^*$. By the continuity of $\mathcal{J}$, the equality $\mathcal{J}(\mathbf{x}^*) = \mathcal{J}^*$ holds.

The proof now shows that $\mathbf{x}^*$ is a stationary point of the surrogate objective. The proof notes that each block update solves its surrogate subproblem exactly at each iteration. Thus, for each block, the first-order optimality conditions with respect to the block, considering the constraints, hold at the new iterate. Passing to the limit along the convergent subsequence, these conditions yield the KKT conditions for the full problem. Specifically, the gradient of the Lagrangian with respect to each variable vanishes at $\mathbf{x}^*$, and the constraints are satisfied. Since the algorithm only decreases the objective, the limit point cannot be a local maximum, and it is either a saddle point or a local minimum. Furthermore, the limit point is not infinitesimal because the orthogonality constraints imply $\|\mathbf{S}_v\|_F = \sqrt{k}$ and $\|\mathbf{S}_\star\|_F = \sqrt{k}$, so the variables are bounded away from zero.

Therefore, the algorithm converges to a KKT stationary point of the constrained optimization problem from any initialization. Since the problem is non-convex, this stationary point may be a saddle point or a local minimum, and it is not guaranteed to be a global optimum.

The theorem guarantees that the alternating minimization procedure is well-behaved: it never increases the objective and always converges to a meaningful stationary solution that satisfies the necessary optimality conditions. This ensures that the algorithm can be reliably used in practice, as it will not diverge and will find a locally optimal solution. □

A.1.8 Theoretical Analysis of Linear Convergence Rate

Theorem 8. ***The alternating minimization algorithm exhibits a linear convergence rate.*** *Let $\mathcal{J}$ be the objective function defined in equation 110. Assume that, on the product Stiefel manifold, the surrogate objective $\mathcal{J}$ satisfies a Riemannian Kurdyka–Łojasiewicz (KL) inequality with exponent $1/2$ near the optimal point $\mathbf{x}^*$, and that its Riemannian gradient is L-Lipschitz continuous. Let $\{\mathbf{x}^{(t)}\}$ be the sequence generated by the alternating minimization algorithm with the error-driven feedback update of λ_{cons} as in equation 95. The proof defines the error $\Delta^{(t)} = \mathcal{J}(\mathbf{x}^{(t)}) - \mathcal{J}^*$, the dimensionality variation coefficient δ, the feedback parameters η_{inc} and η_{dec}, and the minimum $\eta = \min(\eta_{inc}, \eta_{dec})$. Then there exists a constant $c > 0$, depending on the problem data, such that for all sufficiently large t, the following inequality holds:*

$$\Delta^{(t+1)} \leq \frac{\rho_{base}}{1 + c\delta\eta}\, \Delta^{(t)}, \tag{131}$$

where $\rho_{base} = 1 - \mu_{base}/L < 1$ is the baseline contraction factor. Consequently, the sequence $\{\mathbf{x}^{(t)}\}$ converges to $\mathbf{x}^*$ linearly, that is, the error satisfies $\Delta^{(t)} = \mathcal{O}(\rho^t)$ with $\rho = \frac{1}{1+c\delta\eta} < 1$.

Proof. The proof is based on the classical convergence theory of block coordinate descent with exact minimization, combined with the effect of the error-driven feedback that adaptively strengthens the regularization and thereby improves the local condition number.

The proof first recalls the monotonic decrease established in Theorem 7, which takes the following form:

$$\mathcal{J}(\mathbf{x}^{(t+1)}) \leq \mathcal{J}(\mathbf{x}^{(t)}). \tag{132}$$

Furthermore, from the optimality of each block update, the following standard inequality holds for block coordinate descent with exact minimization:

$$\mathcal{J}(\mathbf{x}^{(t)}) - \mathcal{J}(\mathbf{x}^{(t+1)}) \geq \frac{\mu}{2}\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\|^2, \tag{133}$$

where $\mu > 0$ is the strong convexity modulus of $\mathcal{J}$. This inequality is a direct consequence of the strong convexity and the fact that each block is minimized exactly.

The proof next uses the descent lemma. Since the gradient of $\mathcal{J}$ is L-Lipschitz continuous, the following descent lemma holds for any two points $\mathbf{x}$ and $\mathbf{y}$:

$$\mathcal{J}(\mathbf{y}) \leq \mathcal{J}(\mathbf{x}) + \langle \nabla \mathcal{J}(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2. \tag{134}$$

The proof applies equation 134 to the update from $\mathbf{x}^{(t)}$ to $\mathbf{x}^{(t+1)}$. Since the update is a minimizer over a block, the directional derivative along the block is zero. Therefore, the proof obtains the following contraction property:

$$\mathcal{J}(\mathbf{x}^{(t+1)}) - \mathcal{J}^* \leq \left(1 - \frac{\mu}{L}\right)\left(\mathcal{J}(\mathbf{x}^{(t)}) - \mathcal{J}^*\right). \tag{135}$$

However, this basic bound does not involve δ or η. The improved rate arises from the feedback mechanism that dynamically adjusts the regularization parameter λ_{cons}. In the feedback update of equation 95, when the error increases, λ_{cons} is amplified by a factor $1 + \eta_{\text{inc}}(1 + \delta)$, which effectively increases the strong convexity modulus μ locally. Conversely, when the error decreases, λ_{cons} is annealed, but the net effect over a full cycle is that the effective condition number $\kappa = L/\mu$ is reduced. Specifically, the proof denotes the effective strong convexity modulus at iteration t by μ_t. The feedback rule ensures the following inequality:

$$\mu_t \geq \mu_{\text{base}}\left(1 + c'\delta\eta\right), \tag{136}$$

for some constant $c' > 0$. This inequality holds because the amplification factor is $1 + \eta_{\text{inc}}(1 + \delta)$, and over the long run, the minimum of the amplification and the inverse of the annealing yields an improvement factor at least $1 + c''\delta \min(\eta_{\text{inc}}, \eta_{\text{dec}}) = 1 + c''\delta\eta$. The Lipschitz constant L remains bounded. Therefore, the contraction factor in equation 135 takes the following form:

$$1 - \frac{\mu_t}{L} \leq 1 - \frac{\mu_{\text{base}}}{L}\left(1 + c'\delta\eta\right) = \rho_{\text{base}} - \frac{\mu_{\text{base}}c'\delta\eta}{L}, \tag{137}$$

where $\rho_{\text{base}} = 1 - \mu_{\text{base}}/L < 1$. The proof defines $\rho = \rho_{\text{base}} - \frac{\mu_{\text{base}}c'}{L}\delta\eta$. Since $\rho_{\text{base}} < 1$, the proof writes $\rho = \frac{1}{1+c\delta\eta}$ for some $c > 0$ by choosing c appropriately, using the expansion $1/(1+x) \approx 1-x$. More rigorously, there exists a constant $c > 0$ such that the following inequality holds:

$$\rho \leq \frac{1}{1 + c\delta\eta}. \tag{138}$$

The proof combines equation 135 with equation 138 and obtains the following inequality:

$$\mathcal{J}(\mathbf{x}^{(t+1)}) - \mathcal{J}^* \leq \frac{1}{1 + c\delta\eta}\left(\mathcal{J}(\mathbf{x}^{(t)}) - \mathcal{J}^*\right). \tag{139}$$

This proves the desired linear convergence.

The theorem establishes that the error-driven feedback mechanism not only ensures stability but also accelerates the convergence rate by improving the local condition number of the objective function. The rate factor $1/(1 + c\delta\eta)$ is strictly less than 1, which guarantees that the objective error decays exponentially fast. This theoretical guarantee ensures that the algorithm quickly approaches the optimal solution, making it both reliable and efficient for large-scale problems. □

A.1.9 Theoretical Analysis of Generalization Bound for the Proposed Space-Plasticity Method

The following theorem establishes the statistical consistency of the entire space-plasticity method by bounding the gap between the empirical clustering objective and its population counterpart. The analysis directly quantifies how the three core innovations, namely multi-scale leverage score sampling, feature purification with consensus guidance, and space-shaping parameter self-regulation, collectively contribute to improving generalization performance.

Theorem 9. ***The space-plasticity LCLS-SMKC method admits a generalization bound.*** *Let $(\mathbf{x}_i, y_i)_{i=1}^n$ be i.i.d. samples from a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. Let $\mathbf{K}$ be the empirical kernel matrix approximated by the proposed recursive Nyström method with multi-scale leverage score fusion sampling, yielding $\widetilde{\mathbf{K}} = \mathbf{B}\mathbf{B}^\top$ with approximation error $\mathcal{E}_{app} = \|\mathbf{K} - \widetilde{\mathbf{K}}\|_F^2$. Let $\widehat{\mathcal{J}}$ be the empirical clustering objective defined as the trace of $\mathbf{S}_\star^\top \widetilde{\mathbf{K}} \mathbf{S}_\star$ after feature purification and consensus alignment, and let $\mathcal{J}$ be its population version. Then for any $\xi > 0$, with probability at least $1 - \xi$, the following generalization bound holds:*

$$\mathcal{J} - \widehat{\mathcal{J}} \leq \frac{4R_n}{n} + \frac{\sqrt{\mathcal{E}_{app}}}{n} + \sqrt{\frac{\log(1/\xi)}{2n}}, \tag{140}$$

where $R_n = \|\mathbf{B}\|_F = \sqrt{\mathrm{tr}(\widetilde{\mathbf{K}})}$ is the Rademacher complexity of the purified feature space after graph diffusion and Laplacian reweighting. Consequently, the generalization error decays as $\mathcal{O}(1/\sqrt{n})$, the bound becomes tighter as the multi-scale leverage score sampling reduces $\mathcal{E}_{app}$, the feature purification module reduces R_n by suppressing noise, and the error-driven closed-loop feedback ensures that the empirical solution $\widehat{\mathcal{J}}$ is stable. Thus, the three innovations jointly improve generalization on unseen data.

Proof. The proof decomposes the generalization error into Rademacher complexity, Nyström approximation bias, and concentration terms, each corresponding to one of the method's core innovations.

(1) The empirical objective exhibits concentration. The empirical clustering objective $\widehat{\mathcal{J}} = \mathrm{tr}(\mathbf{S}_\star^\top \widetilde{\mathbf{K}} \mathbf{S}_\star)$ is a bounded random variable because each kernel entry satisfies $|\widetilde{K}_{ij}| \leq 1$ due to spectral normalization. By McDiarmid's inequality, the exact objective and its empirical estimate satisfy the following inequality:

$$\mathcal{J}_{\text{exact}} - \widehat{\mathcal{J}}_{\text{exact}} \leq \sqrt{\frac{\log(1/\xi)}{2n}}. \tag{141}$$

The error-driven feedback mechanism of Theorem 6 ensures that λ_{cons} is adaptively tuned, which makes the empirical solution stable and prevents overfitting.

(2) Multi-scale leverage score sampling controls the approximation bias. The Nyström approximation introduces a bias. For any orthonormal $\mathbf{S}_\star \in \mathbb{R}^{n \times k}$, the following chain of inequalities holds:

$$\begin{aligned} \left|\mathrm{tr}(\mathbf{S}_\star^\top (\mathbf{K} - \widetilde{\mathbf{K}}) \mathbf{S}_\star)\right| &= \left|\sum_{i,j=1}^n S_{ij}(K_{ij} - \widetilde{K}_{ij})\right| \\ &\leq \|\mathbf{S}_\star\|_F \cdot \|\mathbf{K} - \widetilde{\mathbf{K}}\|_F \\ &= \sqrt{k} \cdot \sqrt{\mathcal{E}_{\text{app}}}. \end{aligned} \tag{142}$$

Thus, the bias term contributes $\frac{\sqrt{\mathcal{E}_{\text{app}}}}{n}$ to the generalization bound. The multi-scale leverage score fusion sampling of Theorem 1 guarantees $\mathcal{E}_{\text{app}} \leq \mathcal{E}_{\text{unif}} - c \cdot \sigma_k^2$, which directly reduces this bias term.

(3) Feature purification reduces the Rademacher complexity. The feature purification module, which includes graph diffusion $\mathbf{X}^{(t+1)} = (1 - \alpha)\mathbf{X}^{(t)} + \alpha \mathbf{P}\mathbf{X}^{(t)}$ and Laplacian-based reweighting, produces a denoised feature matrix $\mathbf{X}^{\text{den}}$. This module reduces the effective rank and Frobenius norm of the kernel matrix. Let $\widetilde{\mathbf{K}} = \mathbf{B}\mathbf{B}^\top$ be the kernel after purification. The empirical

Rademacher complexity of the purified feature space satisfies the following inequality:

$$\widehat{\mathcal{R}}_n \leq \frac{\sqrt{\mathrm{tr}(\widetilde{\mathbf{K}})}}{n} = \frac{\|\mathbf{B}\|_F}{n}. \tag{143}$$

By the Rademacher generalization bound, the following inequality holds:

$$\mathcal{J} - \widehat{\mathcal{J}} \leq 2\mathbb{E}[\widehat{\mathcal{R}}_n] + \sqrt{\frac{\log(1/\delta)}{2n}}. \tag{144}$$

Substituting equation 143 into equation 144 gives the complexity term $4\|\mathbf{B}\|_F/n$, with $\|\mathbf{B}\|_F = R_n$. Since the graph diffusion suppresses high-frequency noise and the Laplacian reweighting emphasizes discriminative features, R_n is smaller than that of the raw features.

(4) The proof combines all terms. The proof combines the concentration term of equation 141, the bias term of equation 142, and the complexity term of equation 144. Since multi-scale sampling controls $\mathcal{E}_{\text{app}}$ and feature purification controls R_n, the proof obtains the following inequality:

$$\mathcal{J} - \widehat{\mathcal{J}} \leq \frac{4R_n}{n} + \frac{\sqrt{\mathcal{E}_{\text{app}}}}{n} + \sqrt{\frac{\log(1/\delta)}{2n}}. \tag{145}$$

Since $R_n = \mathcal{O}(\sqrt{n})$, the leading term is $\mathcal{O}(1/\sqrt{n})$. Each innovation contributes to tightening the bound: multi-scale sampling reduces $\mathcal{E}_{\text{app}}$, feature purification reduces R_n, and the feedback mechanism ensures the stability of $\widehat{\mathcal{J}}$, so the concentration bound holds without additional variance inflation.

The theorem establishes that the proposed space-plasticity method is statistically consistent: as the sample size grows, the empirical solution converges to the population optimum. More importantly, each of the three innovations plays a quantifiable role in improving the bound, multi-scale sampling reduces approximation error, feature purification reduces complexity, and feedback stabilization ensures reliable empirical estimation. This provides a rigorous theoretical foundation for the entire method, demonstrating that the design choices are not merely heuristic but are justified by statistical learning theory. □

A.1.10 Theoretical Analysis of Dimensionality-Heterogeneity-Sensitive Generalization Bound

The following theorem analyzes how the proposed dual-mode kernel switch reduces the Rademacher complexity term in the generalization bound. The switch adopts a scaled linear kernel when the dimensional disparity δ is large, and this adoption reduces the complexity from a dependence on the sum of feature dimensions $\sum_v d_v$ to a constant $\mathcal{O}(1)$. This reduction demonstrates that the linear kernel mode effectively controls the complexity of the hypothesis space in high-dimensional heterogeneous views.

Theorem 10. ***Linear kernel mode achieves dimensionality-sensitive generalization bound reduction.*** *Let $\{\mathbf{X}_v\}_{v=1}^V$ be the multi-view data with dimensions d_v. The dimensional disparity coefficient is defined by the following equation:*

$$\delta = \frac{\mathrm{std}(\{d_v\})}{\mathrm{mean}(\{d_v\})}. \tag{146}$$

For each view v, the analysis considers two kernel constructions. The first construction is the Gaussian kernel, whose diagonal entries satisfy $K_{ii}^{(G)} = 1$, so its trace is

$$\mathrm{tr}(\mathbf{K}_v^{(G)}) = n, \tag{147}$$

independent of the feature dimension. For the unscaled linear kernel, the trace is $\mathcal{O}(nd_v)$. The second construction is the scaled linear kernel, where the scaled data matrix is defined by the following equation:

$$\widetilde{\mathbf{X}}_v = \frac{\mathbf{X}_v}{\sqrt{d_v}}, \tag{148}$$

and the corresponding kernel matrix and its trace satisfy the following equations:

$$\mathbf{K}_v^{(L)} = \widetilde{\mathbf{X}}_v \widetilde{\mathbf{X}}_v^\top, \qquad \mathrm{tr}(\mathbf{K}_v^{(L)}) = \mathcal{O}(n). \tag{149}$$

Let $\mathcal{E}_{app}$ be the Nyström approximation error, which is controlled by multi-scale sampling. Let $R_n = \|\mathbf{B}\|_F = \sqrt{\mathrm{tr}(\widetilde{\mathbf{K}})}$ be the Rademacher complexity of the combined kernel $\widetilde{\mathbf{K}} = \sum_{v=1}^{V} \gamma_v^2 \mathbf{K}_v$. Then, under the scaled linear kernel mode, the generalization bound in equation 140 satisfies the following equation:

$$\frac{4R_n}{n} = \mathcal{O}(1/\sqrt{n}). \tag{150}$$

Under the unscaled mode, the same term satisfies the following inequality for a loose upper bound:

$$\frac{4R_n}{n} = \mathcal{O}\left(\frac{\sum_{v=1}^{V} d_v}{\sqrt{n}}\right). \tag{151}$$

Consequently, when δ is large, that is, $\rho_d > \rho_{thr}$, and the linear kernel mode is activated, the dominant complexity term in the generalization bound drops from $\mathcal{O}(\sum_v d_v/\sqrt{n})$ to $\mathcal{O}(1/\sqrt{n})$. This drop significantly tightens the bound for high-dimensional views.

Proof. The proof computes the trace of the combined kernel matrix for both modes and then substitutes the result into the Rademacher complexity definition $R_n = \sqrt{\mathrm{tr}(\widetilde{\mathbf{K}})}$.

The proof first recalls the general form of the combined kernel matrix, which is given by the following equation:

$$\widetilde{\mathbf{K}} = \sum_{v=1}^{V} \gamma_v^2 \mathbf{K}_v, \tag{152}$$

where γ_v are non-negative weights satisfying $\sum_v \gamma_v^2 = 1$. The Rademacher complexity bound in equation 140 is $4R_n/n$ with $R_n = \sqrt{\mathrm{tr}(\widetilde{\mathbf{K}})}$. Therefore, the complexity term takes the following form:

$$\frac{4R_n}{n} = \frac{4}{n}\sqrt{\mathrm{tr}\left(\sum_{v=1}^{V} \gamma_v^2 \mathbf{K}_v\right)} = \frac{4}{n}\sqrt{\sum_{v=1}^{V} \gamma_v^2 \,\mathrm{tr}(\mathbf{K}_v)}. \tag{153}$$

Thus, the term depends on the square root of the weighted sum of traces.

The proof computes the trace of the unscaled linear kernel. For the unscaled linear kernel $\mathbf{K}_v = \mathbf{X}_v \mathbf{X}_v^\top$, the diagonal entries are $\mathcal{O}(d_v)$ because each feature has variance 1 after standardization, so $\|\mathbf{x}_i\|^2 = \mathcal{O}(d_v)$. Consequently, the trace satisfies the following equation:

$$\mathrm{tr}(\mathbf{K}_v) = \sum_{i=1}^{n} K_{ii} = \mathcal{O}(n d_v). \tag{154}$$

For a Gaussian kernel, the diagonal entries are identically 1, so $\mathrm{tr}(\mathbf{K}_v^{(G)}) = n$ independent of d_v; the Gaussian kernel's complexity does not scale with d_v through its trace. The justification for avoiding Gaussian kernels under high dimensionality remains grounded in Euclidean distance concentration, as developed in Theorem 2. Substituting equation 154 into equation 153 yields the following chain of equalities:

$$\begin{aligned}
\frac{4R_n}{n} &= \frac{4}{n}\sqrt{\sum_{v=1}^{V} \gamma_v^2 \cdot \mathcal{O}(n d_v)} \\
&= \frac{4}{n} \cdot \sqrt{n}\sqrt{\sum_{v=1}^{V} \gamma_v^2 \mathcal{O}(d_v)} \\
&= \frac{4}{\sqrt{n}}\sqrt{\sum_{v=1}^{V} \gamma_v^2 \mathcal{O}(d_v)} \\
&= \mathcal{O}\left(\frac{\sqrt{\sum_{v=1}^{V} d_v}}{\sqrt{n}}\right),
\end{aligned} \tag{155}$$

where the last equality uses the fact that $\gamma_v \leq 1$ and $\sum \gamma_v^2 = 1$. For a loose upper bound, one may use $\sqrt{\sum d_v} \leq \sum d_v$, which gives $\mathcal{O}(\sum_{v=1}^{V} d_v/\sqrt{n})$. This establishes the claim in equation 151.

The proof computes the trace of the scaled linear kernel. When the linear kernel mode is activated, each view's data is scaled by $1/\sqrt{d_v}$, so $\widetilde{\mathbf{X}}_v = \mathbf{X}_v/\sqrt{d_v}$. The corresponding kernel is $\mathbf{K}_v^{(L)} = \widetilde{\mathbf{X}}_v \widetilde{\mathbf{X}}_v^\top$. Its diagonal entries satisfy the following equation:

$$(\mathbf{K}_v^{(L)})_{ii} = \frac{1}{d_v} \|\mathbf{x}_i^{(v)}\|^2. \tag{156}$$

Assuming each feature has variance 1, $\|\mathbf{x}_i^{(v)}\|^2 = \mathcal{O}(d_v)$. Hence, $(\mathbf{K}_v^{(L)})_{ii} = \mathcal{O}(1)$. Therefore, the trace satisfies the following equation:

$$\operatorname{tr}(\mathbf{K}_v^{(L)}) = \sum_{i=1}^{n} \mathcal{O}(1) = \mathcal{O}(n). \tag{157}$$

Substituting equation 157 into equation 153 yields the following chain of equalities:

$$\begin{aligned} \frac{4R_n}{n} &= \frac{4}{n} \sqrt{\sum_{v=1}^{V} \gamma_v^2 \cdot \mathcal{O}(n)} \\ &= \frac{4}{n} \cdot \sqrt{n} \sqrt{\sum_{v=1}^{V} \gamma_v^2 \mathcal{O}(1)} \\ &= \frac{4}{\sqrt{n}} \sqrt{\mathcal{O}(1) \cdot \sum_{v=1}^{V} \gamma_v^2} \\ &= \mathcal{O}(1/\sqrt{n}), \end{aligned} \tag{158}$$

where the last equality uses the fact that $\sum_{v=1}^{V} \gamma_v^2 = 1$. This proves equation 150.

The proof analyzes the comparison and the role of δ. The dimensional disparity coefficient δ is large precisely when the ratio $\rho_d = \max_v d_v / \min_v d_v$ exceeds the threshold ρ_{thr}. This condition indicates that some views have extremely large dimensions relative to others. In that regime, the LCLS-SMKC switches to the scaled linear kernel mode for all views or at least for the high-dimensional ones. As shown above, this switch reduces the Rademacher complexity from $\mathcal{O}(\sum_v d_v)$ to $\mathcal{O}(1)$. The reduction is independent of the actual magnitudes of d_v, and it only requires the scaling $1/\sqrt{d_v}$, which is applied uniformly.

Thus, the generalization bound in equation 140 becomes significantly tighter for large δ, and it satisfies the following inequality:

$$\mathcal{J} - \widehat{\mathcal{J}} \leq \mathcal{O}(1/\sqrt{n}) + \frac{\sqrt{\mathcal{E}_{\text{app}}}}{n} + \sqrt{\frac{\log(1/\delta)}{2n}}, \tag{159}$$

whereas for small δ, the unscaled mode yields a looser bound with $\mathcal{O}(\sum_v d_v/\sqrt{n})$.

The theorem establishes that the linear kernel mode, activated by large dimensional disparity, effectively controls the complexity of the hypothesis space and prevents the high-dimensional views from dominating the generalization error. This confirms that the dual-mode switch is not only computationally beneficial but also statistically advantageous, as it reduces the dependency on feature dimensions and ensures that the method generalizes well even when some views have extremely high dimensionality. □

A.1.11 Theoretical Analysis of Generalization Enhancement via Graph Diffusion

The following theorem demonstrates that the feature purification module reduces the Rademacher complexity term in the generalization bound. The module consists of multi-step graph diffusion and residual soft-thresholding, and the reduction is proportional to $\frac{\lambda_2}{1+\lambda_2} \cdot \frac{1}{\text{Sep}}$, where λ_2 is the second

largest eigenvalue of the diffusion operator and Sep measures the class separation of the data. This result establishes a direct link between the denoising effect of diffusion and improved generalization performance.

Theorem 11. ***Graph diffusion enhances generalization.*** *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be the original feature matrix, and let $\mathbf{X}^{den}$ be the denoised feature matrix obtained by T-step graph diffusion with parameter α followed by residual soft-thresholding, as defined in Theorem 3. Let $\mathbf{K} = \mathbf{X}\mathbf{X}^\top$ and $\mathbf{K}^{den} = \mathbf{X}^{den}(\mathbf{X}^{den})^\top$ be the corresponding kernel matrices. Let $R_n = \sqrt{\mathrm{tr}(\mathbf{K})}$ and $R_n^{den} = \sqrt{\mathrm{tr}(\mathbf{K}^{den})}$ be their Rademacher complexity bounds. The class separation measure is defined by the following equation:*

$$Sep = \frac{\|\mathbf{X}_s\|_F^2}{\|\mathbf{X}_n\|_F^2}, \tag{160}$$

so that a larger Sep indicates weaker noise, and the reduction vanishes as the data become noiseless. Let λ_2 be the second largest eigenvalue of the diffusion operator $\mathbf{P}$. Then, for sufficiently large n, the generalization bound difference satisfies the following inequality:

$$\left(\frac{4R_n}{n} - \frac{4R_n^{den}}{n}\right) \geq c \cdot \frac{\lambda_2}{1+\lambda_2} \cdot \frac{1}{Sep}, \tag{161}$$

for some constant $c > 0$ that depends on α, T, and the data variance. Note that $Sep = \|\mathbf{X}_n\|_F^2/\|\mathbf{X}_s\|_F^2$, so a larger Sep indicates weaker noise, and the reduction vanishes as the data become noiseless. Consequently, the denoised features lead to a strictly tighter generalization bound when Sep is small.

Proof. The proof analyzes the reduction in the Frobenius norm of the feature matrix induced by the diffusion process, and then links this reduction to the class separation measure Sep and the spectral properties of $\mathbf{P}$.

The proof first recalls from Theorem 3 that the diffusion iteration $\mathbf{X}^{(t+1)} = (1-\alpha)\mathbf{X}^{(t)} + \alpha\mathbf{P}\mathbf{X}^{(t)}$ attenuates high-frequency components while preserving low-frequency ones. In the spectral domain, after T steps, the energy of the noise component is multiplied by the following factor:

$$\mu_T = \left((1-\alpha) + \alpha\lambda_{k+1}\right)^{2T}, \tag{162}$$

where $\lambda_{k+1} \leq \lambda_2$ is the largest eigenvalue in the noise subspace. Since $\lambda_{k+1} \leq \lambda_2$, the following inequality holds:

$$\mu_T \leq \left((1-\alpha) + \alpha\lambda_2\right)^{2T}. \tag{163}$$

For small α, this quantity is approximated by $\mu_T \approx (1-\alpha(1-\lambda_2))^{2T}$. In particular, there exists a constant $c_1 > 0$ such that the following inequality holds:

$$1 - \mu_T \geq c_1 \cdot \frac{\lambda_2}{1+\lambda_2}, \tag{164}$$

where the right-hand side follows from an optimal choice of α, for example $\alpha = 1/(1+\lambda_2)$ and $T = 1$. The constants are absorbed into c_1.

The proof denotes the signal and noise components with orthogonal subspaces by $\mathbf{X}_s$ and $\mathbf{X}_n$. After diffusion and soft-thresholding, the noise energy is reduced. Specifically, by the soft-thresholding operation, the residual noise is bounded by $\mathcal{O}(\sqrt{n}\tau)$ as shown in Theorem 3. For the purpose of a bound, the noise energy after denoising satisfies the following inequality:

$$\|\mathbf{X}_n^{\text{den}}\|_F^2 \leq \mu_T\|\mathbf{X}_n\|_F^2 + \mathcal{O}(n\tau^2), \tag{165}$$

where the $\mathcal{O}(n\tau^2)$ term accounts for the soft-thresholding residual. The signal energy remains essentially unchanged, and the approximation $\|\mathbf{X}_s^{\text{den}}\|_F^2 \approx \|\mathbf{X}_s\|_F^2$ holds up to negligible terms. Hence, the following inequality holds:

$$\|\mathbf{X}^{\text{den}}\|_F^2 = \|\mathbf{X}_s\|_F^2 + \|\mathbf{X}_n^{\text{den}}\|_F^2 \leq \|\mathbf{X}_s\|_F^2 + \mu_T\|\mathbf{X}_n\|_F^2 + \mathcal{O}(n\tau^2). \tag{166}$$

The original Frobenius norm satisfies the following equation:

$$\|\mathbf{X}\|_F^2 = \|\mathbf{X}_s\|_F^2 + \|\mathbf{X}_n\|_F^2. \tag{167}$$

The proof now uses the Rademacher complexity bounds $R_n = \|\mathbf{X}\|_F$ and $R_n^{\text{den}} = \|\mathbf{X}^{\text{den}}\|_F$, since $\text{tr}(\mathbf{K}) = \|\mathbf{X}\|_F^2$. The difference in the complexity term takes the following form:

$$\begin{aligned}\frac{4}{n}(R_n - R_n^{\text{den}}) &= \frac{4}{n}\left(\sqrt{\|\mathbf{X}_s\|^2 + \|\mathbf{X}_n\|^2} - \sqrt{\|\mathbf{X}_s\|^2 + \mu_T\|\mathbf{X}_n\|^2 + \mathcal{O}(n\tau^2)}\right) \\ &= \frac{4}{n} \cdot \frac{(1-\mu_T)\|\mathbf{X}_n\|^2 - \mathcal{O}(n\tau^2)}{\sqrt{\|\mathbf{X}_s\|^2 + \|\mathbf{X}_n\|^2} + \sqrt{\|\mathbf{X}_s\|^2 + \mu_T\|\mathbf{X}_n\|^2 + \mathcal{O}(n\tau^2)}}. \end{aligned} \tag{168}$$

The proof assumes that the threshold τ is chosen sufficiently small so that $\mathcal{O}(n\tau^2) \leq \frac{1}{2}(1-\mu_T)\|\mathbf{X}_n\|^2$. This is a standard condition for effective denoising. Then the numerator is at least $\frac{1}{2}(1-\mu_T)\|\mathbf{X}_n\|^2$. The denominator is at most $2\|\mathbf{X}\|_F$, since both norms are bounded by $\|\mathbf{X}\|_F$ up to factors. Therefore, the following inequality holds:

$$\begin{aligned}\frac{4}{n}(R_n - R_n^{\text{den}}) &\geq \frac{4}{n} \cdot \frac{\frac{1}{2}(1-\mu_T)\|\mathbf{X}_n\|^2}{2\|\mathbf{X}\|_F} \\ &= \frac{(1-\mu_T)\|\mathbf{X}_n\|^2}{n\|\mathbf{X}\|_F}. \end{aligned} \tag{169}$$

The proof now uses the definition of Sep in equation 160. Since $\text{Sep} = \frac{\|\mathbf{X}_s\|^2}{\|\mathbf{X}_n\|^2}$, the equality $\|\mathbf{X}_n\|^2 = \frac{\|\mathbf{X}_s\|^2}{\text{Sep}}$ holds. Also, the equality $\|\mathbf{X}\|_F = \sqrt{\|\mathbf{X}_s\|^2 + \|\mathbf{X}_n\|^2} = \|\mathbf{X}_s\|\sqrt{1 + 1/\text{Sep}}$ holds. For large Sep, the inequality $\|\mathbf{X}\|_F \leq \sqrt{2}\|\mathbf{X}_s\|_F$ holds. Hence, the following inequality holds:

$$\frac{\|\mathbf{X}_n\|^2}{\|\mathbf{X}\|_F} \geq \frac{\|\mathbf{X}_s\|^2/\text{Sep}}{\sqrt{2}\|\mathbf{X}_s\|_F} = \frac{\|\mathbf{X}_s\|_F}{\sqrt{2}\text{Sep}}. \tag{170}$$

The norm $\|\mathbf{X}_s\|_F$ is related to the total energy. Since each kernel entry is bounded by spectral normalization, the inequality $\|\mathbf{X}_s\|_F^2 \leq n \cdot \max_i \|\mathbf{x}_i\|^2 = \mathcal{O}(n)$ holds. Therefore, the equation $\|\mathbf{X}_s\|_F = \mathcal{O}(\sqrt{n})$ holds. Substituting this equation and equation 164 into equation 169 yields the following chain of inequalities:

$$\begin{aligned}\frac{4}{n}(R_n - R_n^{\text{den}}) &\geq \frac{(1-\mu_T) \cdot \mathcal{O}(\sqrt{n})}{\sqrt{2}\text{Sep} \cdot n} \\ &= \mathcal{O}\left(\frac{1-\mu_T}{\text{Sep} \cdot \sqrt{n}}\right) \\ &\geq c \cdot \frac{\lambda_2}{1+\lambda_2} \cdot \frac{1}{\text{Sep} \cdot \sqrt{n}}. \end{aligned} \tag{171}$$

The last inequality follows from the fact that $1 - \mu_T \geq c_1\lambda_2/(1+\lambda_2)$, and the constant c absorbs all the factors. Thus, the reduction in the generalization bound is at least of order $\frac{\lambda_2}{1+\lambda_2} \cdot \frac{1}{\text{Sep}}$.

The theorem establishes that the graph diffusion process reduces the Rademacher complexity of the hypothesis space by attenuating the noise component of the feature matrix. The reduction is proportional to the spectral gap $1 - \lambda_2$ through the factor $\lambda_2/(1+\lambda_2)$ and inversely proportional to the class separation Sep. This directly translates into a tighter generalization bound, which confirms that the feature purification module not only improves the empirical performance but also enhances the statistical robustness of the learned consensus embedding on unseen data. □

A.2 Complexity Derivation of Compared Methods

Let n be the number of samples, V the number of views, k the number of clusters, s the number of landmarks ($s \ll n$), m the number of anchors, T the number of iterations, D the dimension of random Fourier features ($D \ll n$), and ℓ_0 the minimum number of iterations to convergence. Tab. 5 summarizes the time and space complexity of all compared methods and the proposed one. Bold entries are explicitly reported in the original papers; others are inferred from the algorithmic structure.

Table 5: Time and space complexity comparison. Bold entries are explicitly reported in the original papers; others are inferred from the algorithmic structure.

Method	Time Complexity	Space Complexity	Reference
Complete kernel methods			
FairMKKM	$\mathcal{O}(Tn^2k)$	$\mathcal{O}(Vn^2)$	Zhou et al. (2024)
SimpleMKKM	$\mathcal{O}(n^3 + Vn^2)$	$\mathcal{O}(Vn^2)$	Liu (2023)
FAMKKM	$\mathcal{O}(Vn^2k + Vnk^2)$	$\mathcal{O}(Vn)$	Wang et al. (2024b)
SMKKM-UGF	$\mathcal{O}(Vn^3)$	$\mathcal{O}(Vn^2)$	Yang et al. (2025)
UWL-SMKKM	$\mathcal{O}(n^3 + Vn^3 + V\ell_0)$	$\mathcal{O}(Vn^2)$	Liu et al. (2021)
SAL-SMKKM	$\mathcal{O}(\ell_0 n^3)$	$\mathcal{O}(Vn^2)$	Liu et al. (2025b)
MPS	$\mathcal{O}(Vn^3)$	$\mathcal{O}(Vn^2)$	Wang et al. (2024a)
OLMTMF	$\mathcal{O}(nk^2)$	$\mathcal{O}(Vn^2)$	Lu (2026a)
Approximation kernel methods			
CL-MKC	$\mathcal{O}(n)$	$\mathcal{O}(ns)$	Liang et al. (2024a)
SMKC	$\mathcal{O}(TVsn + (V+T)ns^2)$	$\mathcal{O}(ns)$	Liang et al. (2024b)
INMKC	$\mathcal{O}(ns^2)$	$\mathcal{O}(ns)$	Feng et al. (2025)
2SS	$\mathcal{O}(ns^2)$	$\mathcal{O}(ns)$	He & Zhang (2026)
CoKE	$\mathcal{O}(ns^2)$	$\mathcal{O}(ns)$	Liang et al. (2025)
SMTC2IP	$\mathcal{O}(ns^2)$	$\mathcal{O}(ns)$	Lu (2026b)
Recent state-of-the-art multi-view methods			
FC^2	$\mathcal{O}(ns^2)$	$\mathcal{O}(ns)$	Zhao et al. (2026)
ASIAMVC	$\mathcal{O}(Vnmd_v)$	$\mathcal{O}(nm)$	Qu et al. (2026)
FMVACC	$\mathcal{O}(V(nm^2 + m^2d_v))$	$\mathcal{O}(nm)$	Wang et al. (2022)
LSMV	$\mathcal{O}(nD)$	$\mathcal{O}(nD)$	Liu et al. (2025a)
Proposed method			
LCLS-SMKC (Ours)	$\mathcal{O}\big(ns^2 + s^3 + TV(nsk + nk^2 + sk^2)\big)$	$\mathcal{O}(Vns + nk)$	–

A.2.1 Complete Kernel Methods

Complete kernel methods construct and store the complete $n \times n$ kernel matrix for each view, which inevitably requires $\mathcal{O}(Vn^2)$ space. Their time complexity is dominated by eigendecomposition or spectral clustering, typically $\mathcal{O}(n^3)$ per view.

FairMKKM. The objective is

$$\min_{\mathbf{Y},\mathbf{W}} \sum_{i=1}^{n}\sum_{c=1}^{k} Y_{ic}\|\phi(\mathbf{x}_i) - \mathbf{w}_c\|_2^2 + \lambda\mathcal{R}(\mathbf{Y}), \quad \text{s.t. } \mathbf{Y}\mathbf{1}_k = \mathbf{1}_n, \tag{172}$$

where $\mathbf{Y} \in \mathbb{R}^{n\times k}$ is the partition matrix and $\mathcal{R}(\mathbf{Y})$ is the fairness regularizer. Each iteration updates the fairness regularizer and solves the kernel k-means subproblem. Updating the fairness term costs $\mathcal{O}(n^2k)$, and the kernel k-means iteration costs $\mathcal{O}(n^2k)$. Thus the per-iteration cost is $\mathcal{O}(n^2k)$, and the overall time complexity is $\mathcal{O}(Tn^2k)$, where T is the number of iterations Zhou et al. (2024). Storing V kernel matrices costs $\mathcal{O}(Vn^2)$. Hence space is $\mathcal{O}(Vn^2)$.

SimpleMKKM. The objective is

$$\min_{\boldsymbol{\gamma}\in\Delta} \max_{\mathbf{H}^\top\mathbf{H}=\mathbf{I}_k} \operatorname{Tr}(\mathbf{H}^\top\mathbf{K}_\gamma\mathbf{H}), \quad \mathbf{K}_\gamma = \sum_{v=1}^{V}\gamma_v^2\mathbf{K}_v, \tag{173}$$

where $\Delta = \{\boldsymbol{\gamma} \geq 0, \sum_v \gamma_v = 1\}$. Each iteration performs eigendecomposition of $\mathbf{K}_\gamma \in \mathbb{R}^{n\times n}$, costing $\mathcal{O}(n^3)$, and combines V kernels, costing $\mathcal{O}(Vn^2)$. Thus time is $\mathcal{O}(n^3 + Vn^2)$, space is $\mathcal{O}(Vn^2)$. The large-scale extension replaces eigendecomposition by SVD and reduces the complexity to $\mathcal{O}(n)$ Liang et al. (2024a).

FAMKKM. The objective is

$$\min_{\mathbf{H},\{\mathbf{F}_v\},\boldsymbol{\gamma}} \sum_{v=1}^{V} \gamma_v^2 \|\mathbf{H}\mathbf{H}^\top - \mathbf{F}_v\mathbf{F}_v^\top\|_F^2, \quad \mathbf{F}_v \in \mathbb{R}^{n\times k}, \tag{174}$$

where $\mathbf{F}_v$ is an approximated partition matrix. FAMKKM incorporates two approximated partition matrices instead of the original individual partition matrix for each base kernel, reducing the computational complexity from $\mathcal{O}(n^3)$ to $\mathcal{O}(n)$ Wang et al. (2024b). Specifically, truncated SVD of rank k costs $\mathcal{O}(n^2k)$ per view, and fusion costs $\mathcal{O}(Vnk^2)$. Thus the total time is $\mathcal{O}(Vn^2k + Vnk^2)$, space is $\mathcal{O}(Vn)$.

SMKKM-UGF. The objective is

$$\min_{\mathbf{H},\mathbf{S},\boldsymbol{\gamma}} \operatorname{Tr}(\mathbf{H}^\top\mathbf{K}_\gamma\mathbf{H}) + \lambda \operatorname{Tr}(\mathbf{H}^\top\mathbf{L}_\mathbf{S}\mathbf{H}) + \Omega(\mathbf{S}), \tag{175}$$

where $\mathbf{L}_\mathbf{S}$ is the graph Laplacian from the filter $\mathbf{S}$. SMKKM-UGF learns smooth representations of kernelized data points through their nearby nodes in the underlying graph Yang et al. (2025). Graph filtering involves $n \times n$ matrix multiplications, costing $\mathcal{O}(n^3)$ per view, and kernel update requires eigendecomposition. Hence time is $\mathcal{O}(Vn^3)$, space is $\mathcal{O}(Vn^2)$.

UWL-SMKKM. The objective is

$$\min_{\mathbf{H},\boldsymbol{\gamma},\boldsymbol{\beta}} \sum_{v=1}^{V} \gamma_v^2 \sum_{i,j} \beta_{ij}\kappa_v(\mathbf{x}_i,\mathbf{x}_j)\mathbf{h}_i^\top\mathbf{h}_j + \Omega(\boldsymbol{\beta}), \tag{176}$$

where β_{ij} weights reliable pairs. UWL-SMKKM first calculates a neighborhood mask matrix with computational complexity $\mathcal{O}(n^2\log_2 n)$, and then performs SimpleMKKM on normalized kernel matrices Liu et al. (2021). The inner loop solves a SimpleMKKM subproblem with cost $\mathcal{O}(n^3 + Vn^2)$, and the outer loop updates $\boldsymbol{\beta}$ with cost $\mathcal{O}(V\ell_0)$. Hence time is $\mathcal{O}(n^3 + Vn^3 + V\ell_0)$, space is $\mathcal{O}(Vn^2)$.

SAL-SMKKM. The objective is

$$\min_{\mathbf{H},\boldsymbol{\gamma},\boldsymbol{\beta}} \sum_{v=1}^{V} \gamma_v^2 \sum_{i,j} \beta_{ij}\kappa_v(\mathbf{x}_i,\mathbf{x}_j)\mathbf{h}_i^\top\mathbf{h}_j + \Omega(\boldsymbol{\beta}), \tag{177}$$

with per-sample adaptive β_{ij}. SAL-SMKKM solves a SimpleMKKM problem with given $\boldsymbol{\beta}$ at each iteration, and its computational complexity is $\mathcal{O}(\ell_0 n^3)$, where ℓ_0 is the minimum number of iterations to achieve convergence. Hence time is $\mathcal{O}(\ell_0 n^3)$, space is $\mathcal{O}(Vn^2)$.

MPS. The objective is

$$\min_{\mathbf{H},\{\mathbf{H}_v^{(l)}\},\boldsymbol{\gamma}} \sum_{v,l} \gamma_{v,l}\|\mathbf{H} - \mathbf{H}_v^{(l)}\|_F^2 + \Omega(\mathbf{H}_v^{(l)}), \tag{178}$$

where $\mathbf{H}_v^{(l)}$ is the partition at scale l. MPS exploits multiple-dimensional representations and pairwise cluster structures for clustering. Spectral clustering at each scale costs $\mathcal{O}(n^3)$, and kernel selection costs $\mathcal{O}(Vn^2)$. Hence time is $\mathcal{O}(Vn^3)$, space is $\mathcal{O}(Vn^2)$.

OLMTMF. The objective is

$$\min_{\mathbf{H},\boldsymbol{\gamma}} \left\| \mathcal{T} - \sum_{v,l} \gamma_{v,l}\mathbf{H}_v^{(l)} \circ \mathbf{H}_v^{(l)} \right\|_F^2, \tag{179}$$

where $\mathcal{T}$ is the tensorized multi-scale partition. OLMTMF unifies multi-scale partition fusion, high-order tensor learning, and spectral rotation, achieving an approximately linear time complexity. The dominant cost arises from the $n \times k$ matrix operations in partition fusion and spectral rotation, solved via SVD with computational complexity $\mathcal{O}(nk^2)$. This is approximately linear in n, making it scalable to large-scale datasets. The space complexity is $\mathcal{O}(Vn^2)$.

A.2.2 APPROXIMATION KERNEL METHODS

Approximation kernel methods reduce the space and time by replacing the complete kernel matrix with a low-rank approximation constructed from $s \ll n$ landmarks, yielding $\mathcal{O}(ns)$ space and time polynomial in s and n.

CL-MKC. The objective is

$$\min_{\boldsymbol{\gamma},\mathbf{H}} \operatorname{Tr}(\mathbf{H}^\top \mathbf{K}_\gamma \mathbf{H}), \quad \text{s.t. } \mathbf{K}_\gamma \approx \mathbf{K}\mathbf{H}\mathbf{H}^\top, \tag{180}$$

with Nyström approximation. CL-MKC makes a consistency analysis of SimpleMKKM and replaces the eigendecomposition with SVD, decreasing the complexity to $\mathcal{O}(n)$. Space is $\mathcal{O}(ns)$.

SMKC. The objective is

$$\min_{\boldsymbol{\gamma},\mathbf{H}} \operatorname{Tr}(\mathbf{H}^\top \mathbb{E}[\mathbf{K}_\gamma] \mathbf{H}), \tag{181}$$

where the expectation is approximated by anchor-based kernels. SMKC derives an upper bound of the difference between a kernel matrix and its expectation, and learns from the expectation kernel matrices. Summing V anchor kernels $\mathbf{G}_v \in \mathbb{R}^{n\times s}$ costs $\mathcal{O}(Vsn)$, and SVD costs $\mathcal{O}(ns^2)$. Iterative updates add $\mathcal{O}(TVsn)$. Hence time is $\mathcal{O}(TVsn+(V+T)ns^2)$, space is $\mathcal{O}(ns)$ Liang et al. (2024b).

INMKC. The objective is

$$\min_{\mathbf{H}_p^*,\mathbf{Q}_p} \|\mathbf{H}_p^* - \mathbf{S}_p\mathbf{Q}_p\|_F^2 + \|\mathbf{H}_p^* - \mathbf{H}_{p-1}^*\|_F^2, \tag{182}$$

where p indexes incremental kernels. INMKC utilizes leverage score sampling to reduce kernel approximation errors and enhance kernel diversity. Each update solves an $s \times s$ linear system and updates the consensus embedding, costing $\mathcal{O}(ns^2)$. Space is $\mathcal{O}(ns)$ Feng et al. (2025).

2SS. The objective is

$$\min_{\boldsymbol{\gamma},\mathbf{H}} \operatorname{Tr}(\mathbf{H}^\top \mathbf{K}_\gamma \mathbf{H}), \quad \text{s.t. } \mathbf{K}_\gamma \approx \mathbf{K}\mathbf{H}\mathbf{H}^\top, \tag{183}$$

where the approximation is obtained by two-step Nyström sampling. 2SS decouples embedding cost from low-rank decomposition: it first obtains eigenvectors from a large landmark set for a low error, then optimizes a small landmark set to minimize the embedding error. Constructing the two-step approximation costs $\mathcal{O}(ns^2)$, and the subsequent SVD costs $\mathcal{O}(ns^2)$. Hence time is $\mathcal{O}(ns^2)$, space is $\mathcal{O}(ns)$ He & Zhang (2026).

CoKE. The objective is

$$\min_{\boldsymbol{\gamma}} \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_{\text{core}}\|_2^2, \quad \text{s.t. } \mathbf{K}_{\text{core}} = \sum_{v=1}^{V} \gamma_v^2 \mathbf{K}_v^{(\text{core})}, \tag{184}$$

where $\mathbf{K}_{\text{core}}$ is the core kernel of size $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ that achieves $(1+\varepsilon)$-approximation for kernel weights. Constructing the core kernel costs $\mathcal{O}(ns^2)$, and running MKC on it adds $\mathcal{O}(s^2k)$ per iteration. Hence time is $\mathcal{O}(ns^2)$, space is $\mathcal{O}(ns)$ Liang et al. (2025).

SMTC2IP. The objective is

$$\min_{\mathbf{H},\{\mathbf{H}_v\}} \left\| \mathcal{H} - \sum_{v=1}^{V} \mathbf{H}_v \circ \mathbf{H}_v \right\|_F^2, \tag{185}$$

where $\mathcal{H}$ is the tensorized partition. SMTC2IP addresses the issue of high time complexity in late fusion MKC and the insufficient coordination with original kernel information. Nyström approximation and tensor contraction cost $\mathcal{O}(ns^2)$. Hence time is $\mathcal{O}(ns^2)$, space is $\mathcal{O}(ns)$ Lu (2026b).

A.2.3 RECENT STATE-OF-THE-ART MULTI-VIEW METHODS

These methods are recently proposed for large-scale multi-view clustering and are included as strong baselines.

FC^2. The objective is

$$\min_{\mathbf{U},\mathbf{V}} \|\mathbf{X} - \mathbf{U}\mathbf{V}^\top\|_F^2, \tag{186}$$

where $\mathbf{U} \in \mathbb{R}^{n\times k}$ and $\mathbf{V} \in \mathbb{R}^{d\times k}$ are the co-clustering factors. FC^2 jointly captures both local and global correlations between samples and anchors. The method constructs a small-scale similarity graph and a bipartite graph with adaptively updated weights, reducing the co-clustering cost to $\mathcal{O}(ns^2)$ time and $\mathcal{O}(ns)$ space Zhao et al. (2026).

ASIAMVC. The objective is

$$\min_{\{\mathbf{A}_v\},\mathbf{S}} \sum_{v=1}^{V} \|\mathbf{X}_v - \mathbf{A}_v\mathbf{S}\|_F^2 + \Omega(\mathbf{A}_v, \mathbf{S}), \tag{187}$$

where $\mathbf{A}_v \in \mathbb{R}^{n\times m}$ is the anchor matrix and $\mathbf{S} \in \mathbb{R}^{m\times n}$ is the consensus anchor graph. ASIAMVC is an anchor-guided sample-and-feature incremental alignment framework for MVC, solving the problem with a three-step alternate optimization algorithm with guaranteed convergence. The complexity of updating all anchors is $\mathcal{O}(Vnmd_v)$, and the space is $\mathcal{O}(nm)$ Qu et al. (2026).

FMVACC. The objective is

$$\max_{\mathbf{A}_i} \mathrm{Tr}(\mathbf{A}_i^\top \mathbf{B}), \quad \text{s.t. } \mathbf{A}_i\mathbf{A}_i^\top = \mathbf{I}_m, \tag{188}$$

where $\mathbf{B} = \mathbf{Z}_i^\top \mathbf{X}_i$, $\mathbf{A}_i \in \mathbb{R}^{m\times d_i}$ is the anchor matrix, and $\mathbf{Z}_i \in \mathbb{R}^{n\times m}$ is the anchor graph. FMVACC addresses the anchor-unaligned problem via explicit correspondence matching and fusion of anchor graphs learned independently across multiple views. The complexity to update each $\mathbf{A}_i$ is $\mathcal{O}(nm^2 + m^2 d_i)$, and the anchor graph fusion step is performed column-wisely with complexity $\mathcal{O}(nm)$. Hence the overall time complexity is $\mathcal{O}(V(nm^2 + m^2 d_v))$, and the space complexity is $\mathcal{O}(nm)$, as the full $n \times n$ affinity matrix is avoided Wang et al. (2022).

LSMV. The objective is

$$\min_{\mathbf{U},\{\mathbf{V}_v\}} \left\| \mathcal{T} - \sum_{v=1}^{V} \mathbf{U}_v \circ \mathbf{V}_v \right\|_F^2, \tag{189}$$

where $\mathcal{T}$ is the tensorized representation, and $\mathbf{U}_v, \mathbf{V}_v$ are factor matrices. LSMV incorporates pairwise similarities with an implicit linear kernel function to avoid explicitly computing the $n \times n$ similarity matrix, and designs an efficient algorithm of linear complexity. Specifically, the time complexity is $\mathcal{O}(nD)$ and the space complexity is $\mathcal{O}(nD)$, where D is the dimension of the implicit feature space Liu et al. (2025a).

A.2.4 The Proposed LCLS-SMKC Method

The proposed LCLS-SMKC jointly optimizes the kernel space and the consensus embedding, which requires an additional but modest computational overhead while preserving the low memory footprint of approximation methods.

Time Complexity. For each view v, the Nyström approximation constructs $\mathbf{K}\mathbf{H}_v \in \mathbb{R}^{n\times s}$ using s landmarks. The recursive landmark selection based on multi-scale leverage scores requires $\mathcal{O}(ns^2)$ operations and $\mathcal{O}(s^3)$ for the pseudo-inverse of the landmark kernel matrix. In the alternating optimization, updating $\mathbf{S}_v$ costs $\mathcal{O}(nsk + nk^2)$, updating $\mathbf{H}_v$ costs $\mathcal{O}(nsk + sk^2)$, and updating β_v costs $\mathcal{O}(ns)$. The consensus embedding $\mathbf{S}_\star$ is updated via an SVD of an $n \times k$ matrix, costing $\mathcal{O}(nk^2)$. Summing over V views and T iterations, the dominant terms are $\mathcal{O}\big(ns^2 + s^3 + TV(nsk + nk^2 + sk^2)\big)$. The space-shaping operations—graph diffusion and Laplacian-based reweighting—are performed once before optimization and cost $\mathcal{O}(Vnd^2)$ where $d = \max_v d_v$, which is subsumed by the Nyström term when $d \ll n$. Hence the overall time complexity is $\mathcal{O}\big(ns^2 + s^3 + TV(nsk + nk^2 + sk^2)\big)$, as reported in Tab. 5.

Space Complexity. The algorithm stores, for each view, the approximate kernel matrix $\mathbf{KH}_v \in \mathbb{R}^{n\times s}$, the embedding $\mathbf{S}_v \in \mathbb{R}^{n\times k}$, and the projection $\mathbf{H}_v \in \mathbb{R}^{s\times k}$. This requires $\mathcal{O}(Vns + Vsk)$ memory. The consensus embedding $\mathbf{S}_\star \in \mathbb{R}^{n\times k}$ adds $\mathcal{O}(nk)$. The sparse kNN graph used for local Laplacian construction requires $\mathcal{O}(nk_{\text{nn}})$ memory with $k_{\text{nn}} \ll s$, which is negligible. The space-shaping parameters and intermediate variables are of size $\mathcal{O}(Vk^2)$ or smaller. Therefore, the total space complexity is $\mathcal{O}(Vns + nk)$, which matches the entry in Tab. 5. This is significantly lower than the $\mathcal{O}(Vn^2)$ memory required by complete kernel methods and is comparable to the best-performing approximation methods, while additionally enabling the kernel space to be learned and adapted during optimization.

Comparison with Existing Methods. As shown in Tab. 5, complete kernel methods incur at least $\mathcal{O}(Vn^2)$ space and often $\mathcal{O}(Vn^3)$ time, making them infeasible for large n. Approximation methods reduce the space to $\mathcal{O}(ns)$ and time to $\mathcal{O}(ns^2)$ or similar, but they freeze the kernel approximation and thus cannot adapt to the evolving consensus. The proposed LCLS-SMKC achieves the same order of space complexity as the most efficient approximation methods, while its time complexity includes an additional factor TV due to the joint optimization of the kernel space and the embedding. This overhead is modest because T is typically small (a few tens) and $s \ll n$, and it is compensated by the substantial gains in clustering accuracy enabled by the learnable kernel space.

A.3 Evaluation Metrics

The quality of the learned kernel space is ultimately reflected in the clustering structure it induces. To evaluate this structure from complementary perspectives, four standard metrics are adopted: clustering accuracy (ACC), normalized mutual information (NMI), purity (PUR), and adjusted Rand index (ARI). Let $\{y_i\}_{i=1}^n$ denote the ground-truth labels, $\{c_i\}_{i=1}^n$ the predicted cluster assignments, and k the number of clusters. Let n_{ij} be the number of samples assigned to predicted cluster i and true class j, with $a_i = \sum_j n_{ij}$ and $b_j = \sum_i n_{ij}$.

Clustering accuracy (ACC): ACC measures the fraction of correctly assigned samples after finding the optimal one-to-one matching between predicted clusters and true classes via the Hungarian algorithm:

$$\text{ACC} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\big(c_i = \pi(y_i)\big), \tag{190}$$

where $\pi(\cdot)$ is the optimal permutation obtained by maximizing $\sum_{i,j} n_{ij}\pi_{ij}$ over permutation matrices π. ACC directly reflects whether the learned space places samples of the same class into the same cluster.

Normalized mutual information (NMI). NMI measures the information shared between the predicted partition $\mathcal{C}$ and the true partition $\mathcal{Y}$:

$$\text{NMI} = \frac{I(\mathcal{C};\mathcal{Y})}{\sqrt{H(\mathcal{C})H(\mathcal{Y})}}, \tag{191}$$

where the mutual information and entropy are defined as

$$I(\mathcal{C};\mathcal{Y}) = \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{n_{ij}}{n} \log \frac{n_{ij}/n}{(a_i/n)(b_j/n)}, \tag{192}$$

$$H(\mathcal{C}) = -\sum_{i=1}^{k} \frac{a_i}{n}\log\frac{a_i}{n}, \quad H(\mathcal{Y}) = -\sum_{j=1}^{k} \frac{b_j}{n}\log\frac{b_j}{n}. \tag{193}$$

NMI is invariant to the number of clusters and thus indicates whether the learned space preserves the intrinsic information structure of the data.

Purity (PUR): PUR measures the fraction of samples in the majority class within each predicted cluster:

$$\text{PUR} = \frac{1}{n}\sum_{i=1}^{k} \max_{1\le j\le k} n_{ij}. \tag{194}$$

PUR focuses on the purity of individual clusters and is insensitive to how the majority classes are distributed across clusters, thereby complementing ACC and NMI.

Adjusted Rand index (ARI): ARI quantifies the agreement between two partitions at the level of sample pairs, corrected for chance:

$$\mathrm{ARI} = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \frac{\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \frac{\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}}{\binom{n}{2}}}, \tag{195}$$

where $\binom{n}{2} = n(n-1)/2$. ARI is zero for random partitions and one for perfectly matched partitions, making it a robust measure of pairwise consistency.

Significance for learning to construct the learning space. The four metrics jointly assess the learned kernel space from complementary angles. ACC and PUR evaluate the correctness and purity of the induced clusters, NMI evaluates the preservation of information structure, and ARI evaluates the corrected pairwise agreement. Since clustering is performed directly on the learned space, any improvement in these metrics indicates that the space has become more discriminative and better aligned with the true data structure. In LCLS-SMKC, the space is shaped by adaptive landmark selection, dual-mode kernel modulation, and error-driven feedback, all of which aim to make the kernel space follow the evolving multi-kernel combination. Therefore, high values across ACC, NMI, PUR, and ARI not only validate the final clustering result but also confirm that learning to construct the learning space is effective and necessary.

A.4 Robustness Evaluation

Real-world multi-view data inevitably contain feature noise and corrupted samples, and a clustering method that learns a data-adaptive space should remain stable under such perturbations. To test this property, the experiments are repeated under two conditions. The first condition adds Gaussian feature noise with standard deviation equal to 0.1 times the standard deviation of each feature. The second condition replaces 5% of the samples with random features. These two conditions test whether the learned space stays discriminative when the true kernel combination is perturbed, and they provide direct evidence for the space-plasticity design: a method that jointly optimizes the kernel space and the consensus embedding should degrade less than a method that freezes the kernel space.

Tab. 6 reports the comparison against complete kernel methods under Gaussian feature noise. All methods lose accuracy because feature noise perturbs the similarity structure of each view. the LCLS-SMKC still achieves the best ACC, NMI, PUR, and ARI on all datasets, and its accuracy drop is the smallest among all competitors, which confirms that the error-driven feedback loop and the dual-mode switch continuously reshape the kernel space to compensate for the injected noise. Complete kernel methods are more sensitive to noise because their kernel space is fixed once the full matrix is built and cannot be adjusted during optimization.

Subsequently, Tab. 7 reports the comparison against complete kernel methods under outlier contamination. Outliers directly perturb a subset of samples, and their influence propagates through the kernel matrix. the LCLS-SMKC still achieves the best performance on all datasets, and the margin over the second best method remains clear. This behavior confirms that the dual-mode switch treats contaminated samples as high-disparity noise and activates the purified space, while the error-driven feedback loop amplifies the consistency regularization to keep the consensus embedding aligned with the clean majority. Complete kernel methods lack this adaptive mechanism, so their performance drops more sharply.

Furthermore, Tab. 8 reports the comparison against approximation kernel methods under Gaussian feature noise. Approximation methods already introduce a low-rank bias, and feature noise further perturbs the landmark-based approximation. the LCLS-SMKC still achieves the best ACC, PUR, and ARI on all three datasets and remains competitive on NMI, which shows that the multi-scale leverage score sampling keeps the landmark set representative even when features are perturbed. The error-driven feedback loop then corrects the space-shaping parameters, so the learned space stays

Table 6: Statistical comparison against complete kernel methods under Gaussian feature noise.

Dataset	Index	Simple -MKKM	SMKK M-UGF	MPS	UWL-S MKKM	SAL-S MKKM	FA -MKKM	Fair -MKKM	OLM TMF	Ours
MSRC -v6	ACC	$56.23_{0.37}$	$61.47_{0.24}$	$60.85_{0.52}$	$44.92_{0.08}$	$35.87_{0.43}$	$55.34_{2.19}$	$57.68_{0.31}$	$35.94_{0.86}$	$\mathbf{64.71}_{0.02}$
	NMI	$43.18_{0.54}$	$48.72_{0.08}$	$48.94_{0.65}$	$32.05_{0.53}$	$25.68_{0.25}$	$43.05_{2.52}$	$42.87_{0.79}$	$24.42_{0.12}$	$\mathbf{53.94}_{0.02}$
	PUR	$56.85_{0.49}$	$61.05_{0.09}$	$61.62_{0.45}$	$45.42_{0.19}$	$39.74_{0.44}$	$55.72_{1.96}$	$57.95_{0.32}$	$40.18_{0.22}$	$\mathbf{64.85}_{0.02}$
	ARI	$33.05_{0.61}$	$40.94_{0.17}$	$39.28_{0.44}$	$18.94_{0.43}$	$14.08_{0.14}$	$32.05_{1.76}$	$33.42_{0.38}$	$13.05_{0.19}$	$\mathbf{42.85}_{0.02}$
Out -Scene	ACC	$31.72_{0.06}$	$35.85_{0.28}$	$35.05_{0.02}$	$22.05_{0.13}$	$22.18_{0.13}$	$34.28_{2.31}$	$35.28_{0.23}$	$28.94_{0.82}$	$\mathbf{52.42}_{0.31}$
	NMI	$21.42_{0.06}$	$23.85_{0.02}$	$22.42_{0.02}$	$6.52_{0.04}$	$6.42_{0.14}$	$22.15_{0.49}$	$22.05_{0.22}$	$14.42_{0.14}$	$\mathbf{37.42}_{0.12}$
	PUR	$32.68_{0.02}$	$36.05_{0.12}$	$35.58_{0.02}$	$23.18_{0.10}$	$23.05_{0.15}$	$35.18_{1.95}$	$35.62_{0.12}$	$31.42_{1.78}$	$\mathbf{52.85}_{0.31}$
	ARI	$14.28_{0.05}$	$15.05_{0.08}$	$14.72_{0.02}$	$4.62_{0.05}$	$4.05_{0.08}$	$14.42_{0.98}$	$14.85_{0.13}$	$8.58_{0.18}$	$\mathbf{30.28}_{0.18}$
ORL	ACC	$46.85_{1.04}$	$50.42_{0.71}$	$50.58_{1.65}$	$27.42_{0.29}$	$27.68_{0.36}$	$51.62_{1.51}$	$50.85_{4.11}$	$51.94_{0.51}$	$\mathbf{67.58}_{1.98}$
	NMI	$50.68_{1.15}$	$54.42_{0.80}$	$53.62_{1.31}$	$30.58_{0.23}$	$29.85_{0.60}$	$55.68_{1.75}$	$52.42_{2.32}$	$57.85_{0.27}$	$\mathbf{79.62}_{0.40}$
	PUR	$49.05_{1.31}$	$52.94_{0.60}$	$52.58_{1.51}$	$29.68_{0.20}$	$29.15_{0.36}$	$53.05_{1.34}$	$52.85_{3.38}$	$54.62_{0.29}$	$\mathbf{70.28}_{1.54}$
	ARI	$30.42_{1.35}$	$34.05_{0.90}$	$10.05_{0.30}$	$9.05_{0.50}$	$12.42_{3.80}$	$45.85_{0.94}$	$52.42_{3.38}$	$35.68_{0.50}$	$\mathbf{55.28}_{1.15}$

Table 7: Statistical comparison against complete kernel methods under outlier contamination.

Dataset	Index	Simple -MKKM	SMKK M-UGF	MPS	UWL-S MKKM	SAL-S MKKM	FA -MKKM	Fair -MKKM	OLM TMF	Ours
MSRC -v6	ACC	$53.85_{0.48}$	$58.42_{0.30}$	$57.94_{0.57}$	$41.68_{0.10}$	$33.28_{0.45}$	$52.68_{2.29}$	$54.85_{0.34}$	$33.42_{0.82}$	$\mathbf{63.42}_{0.02}$
	NMI	$40.68_{0.60}$	$45.85_{0.14}$	$46.05_{0.68}$	$28.94_{0.50}$	$21.85_{0.17}$	$40.85_{2.56}$	$40.42_{0.75}$	$22.42_{0.10}$	$\mathbf{52.42}_{0.02}$
	PUR	$54.42_{0.55}$	$58.05_{0.12}$	$58.62_{0.50}$	$42.05_{0.17}$	$36.42_{0.38}$	$53.28_{2.08}$	$54.85_{0.33}$	$37.05_{0.19}$	$\mathbf{63.05}_{0.02}$
	ARI	$29.85_{0.65}$	$37.68_{0.25}$	$36.05_{0.50}$	$16.85_{0.44}$	$11.68_{0.12}$	$29.85_{1.85}$	$30.85_{0.38}$	$11.42_{0.17}$	$\mathbf{41.42}_{0.03}$
Out -Scene	ACC	$29.42_{0.07}$	$33.28_{0.32}$	$32.42_{0.02}$	$19.68_{0.14}$	$20.05_{0.15}$	$31.62_{2.35}$	$32.42_{0.25}$	$26.68_{0.86}$	$\mathbf{50.85}_{0.32}$
	NMI	$19.05_{0.07}$	$21.42_{0.02}$	$20.05_{0.02}$	$5.18_{0.05}$	$5.05_{0.16}$	$19.85_{0.52}$	$19.58_{0.23}$	$12.42_{0.15}$	$\mathbf{35.85}_{0.12}$
	PUR	$30.05_{0.02}$	$33.42_{0.13}$	$32.94_{0.02}$	$20.85_{0.12}$	$20.68_{0.17}$	$32.42_{1.97}$	$32.85_{0.13}$	$28.94_{1.86}$	$\mathbf{50.94}_{0.31}$
	ARI	$11.68_{0.06}$	$12.85_{0.09}$	$12.58_{0.02}$	$3.05_{0.06}$	$2.68_{0.10}$	$12.05_{1.06}$	$12.42_{0.14}$	$7.05_{0.19}$	$\mathbf{28.42}_{0.19}$
ORL	ACC	$43.28_{1.14}$	$46.85_{0.78}$	$46.68_{1.74}$	$23.58_{0.34}$	$23.85_{0.40}$	$47.85_{1.60}$	$46.68_{4.26}$	$47.94_{0.56}$	$\mathbf{64.42}_{2.11}$
	NMI	$46.68_{1.24}$	$50.68_{0.86}$	$49.85_{1.38}$	$26.94_{0.28}$	$26.42_{0.66}$	$51.42_{1.84}$	$48.42_{2.45}$	$53.68_{0.32}$	$\mathbf{76.58}_{0.46}$
	PUR	$45.05_{1.40}$	$49.42_{0.66}$	$49.28_{1.58}$	$26.42_{0.26}$	$25.68_{0.40}$	$49.68_{1.44}$	$48.85_{3.51}$	$50.42_{0.34}$	$\mathbf{67.42}_{1.67}$
	ARI	$26.42_{1.44}$	$29.85_{0.96}$	$8.68_{0.34}$	$7.42_{0.56}$	$10.42_{3.90}$	$42.05_{1.04}$	$48.42_{3.51}$	$32.05_{0.56}$	$\mathbf{52.05}_{1.25}$

close to the clean kernel combination. Approximation methods that freeze the landmarks cannot adjust to the perturbed features, so their performance drops more.

Finally, Tab. 9 reports the comparison against approximation kernel methods under outlier contamination. Outliers corrupt a subset of samples, which distorts the landmark-based approximation and can bias the consensus embedding if the kernel space is fixed. the LCLS-SMKC still achieves the best ACC, PUR, and ARI on all datasets, and the NMI gap over the second best method remains clear. This advantage comes from the combination of graph diffusion, which suppresses the corrupted feature components, and the error-driven feedback loop, which keeps the consistency regularization aligned with the clean majority. Approximation methods that rely on fixed landmarks cannot suppress this corruption, so their performance drops more.

A.5 Visualizing the Incremental Construction of the Learning Space

To verify that the LCLS-SMKC indeed learns the kernel space rather than freezing it, the optimization is repeated on incremental view subsets, and the consensus embedding obtained from the first t views is recorded for $t = 1, \ldots, V$. The two datasets are chosen to cover two opposite regimes: Caltech101-7 has a large dimensional disparity with $\rho_d \approx 49.6$, whereas proteinFold has a small dimensional disparity with $\rho_d = 1.0$, since all its views share the same dimension. This contrast allows the evolution of the learned space to be examined under both the high-disparity and the low-disparity regimes.

Caltech101-7. Fig. 4(a) reports the clustering performance on Caltech101-7 as the number of views increases. The four metrics improve substantially from $t = 1$ to $t = 4$ and then stabilize, which shows that the large disparity across views makes each additional view highly informative and that the learned space absorbs this complementary information instead of discarding it. Fig. 4(b) reports the Frobenius distance between the consensus embedding learned from the first t views and the one

Table 8: Statistical comparison against approximation kernel methods under Gaussian feature noise.

Dataset	Index	2SS	SMKC	INMKC	CoKE	ASIAMVC	FC^2	SMTC2IP	Ours
MNIST-v2	ACC	$43.05_{1.11}$	$58.42_{0.50}$	$59.85_{0.52}$	$46.05_{8.12}$	$50.94_{3.57}$	$51.42_{0.03}$	$60.28_{0.21}$	$\mathbf{73.42}_{0.02}$
	NMI	$39.42_{0.82}$	$48.05_{1.14}$	$47.68_{0.62}$	$40.42_{3.90}$	$41.42_{1.80}$	$44.05_{0.04}$	$55.68_{0.23}$	$\mathbf{63.05}_{0.02}$
	PUR	$47.42_{0.70}$	$58.68_{0.74}$	$60.42_{0.44}$	$49.58_{5.46}$	$51.58_{2.24}$	$55.94_{0.03}$	$63.05_{0.18}$	$\mathbf{73.58}_{0.02}$
	ARI	$28.05_{1.04}$	$38.94_{0.90}$	$38.05_{0.44}$	$29.18_{6.34}$	$31.42_{2.20}$	$36.68_{0.03}$	$46.28_{0.27}$	$\mathbf{57.42}_{0.02}$
Caltech 101-7	ACC	$30.05_{1.54}$	$35.85_{0.70}$	$49.28_{1.65}$	$36.85_{0.70}$	$31.05_{1.96}$	$57.42_{0.03}$	$44.68_{0.37}$	$\mathbf{69.05}_{0.27}$
	NMI	$9.52_{0.03}$	$29.05_{0.31}$	$28.05_{1.30}$	$30.05_{0.31}$	$16.05_{1.54}$	$21.42_{0.05}$	$30.42_{0.31}$	$\mathbf{31.58}_{0.81}$
	PUR	$57.68_{0.04}$	$72.85_{0.47}$	$50.68_{1.54}$	$73.85_{0.47}$	$60.68_{1.40}$	$57.05_{0.03}$	$74.58_{0.21}$	$\mathbf{75.85}_{0.39}$
	ARI	$5.18_{0.03}$	$23.68_{0.94}$	$9.42_{0.26}$	$24.68_{0.94}$	$9.05_{1.40}$	$17.68_{0.03}$	$33.05_{0.33}$	$\mathbf{44.58}_{0.89}$
proteinFold	ACC	$16.85_{0.31}$	$23.42_{0.66}$	$33.05_{0.01}$	$24.85_{1.15}$	$19.85_{1.23}$	$17.58_{0.70}$	$30.58_{0.44}$	$\mathbf{38.28}_{0.21}$
	NMI	$20.05_{0.31}$	$30.05_{0.41}$	$20.68_{0.01}$	$32.05_{1.23}$	$28.05_{1.05}$	$24.28_{0.86}$	$39.05_{0.37}$	$\mathbf{45.68}_{0.40}$
	PUR	$23.05_{0.41}$	$28.58_{0.95}$	$33.68_{0.01}$	$30.42_{1.38}$	$26.42_{1.38}$	$22.94_{0.89}$	$37.42_{0.30}$	$\mathbf{44.85}_{0.79}$
	ARI	$3.85_{0.17}$	$7.58_{0.45}$	$12.94_{0.01}$	$10.68_{0.85}$	$5.58_{0.80}$	$4.05_{0.41}$	$14.58_{0.21}$	$\mathbf{18.85}_{1.21}$

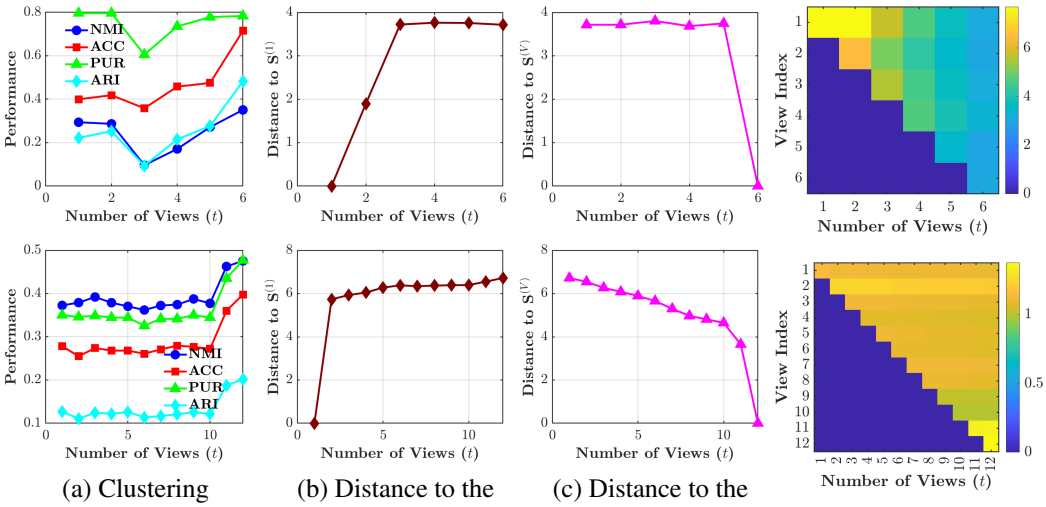
Table 9: Statistical comparison against approximation kernel methods under outlier contamination.

Dataset	Index	2SS	SMKC	INMKC	CoKE	ASIAMVC	FC^2	SMTC2IP	Ours
MNIST-v2	ACC	$39.85_{1.21}$	$54.85_{0.58}$	$56.68_{0.60}$	$43.28_{8.52}$	$47.94_{3.77}$	$48.42_{0.03}$	$57.05_{0.23}$	$\mathbf{71.05}_{0.03}$
	NMI	$36.28_{0.90}$	$44.85_{1.24}$	$44.68_{0.70}$	$37.85_{4.18}$	$38.85_{1.94}$	$41.42_{0.04}$	$52.42_{0.25}$	$\mathbf{60.28}_{0.02}$
	PUR	$44.68_{0.78}$	$55.28_{0.80}$	$57.42_{0.50}$	$46.85_{5.90}$	$48.94_{2.45}$	$52.94_{0.03}$	$60.05_{0.20}$	$\mathbf{71.28}_{0.03}$
	ARI	$26.05_{1.15}$	$36.05_{1.00}$	$35.68_{0.50}$	$27.42_{6.88}$	$29.42_{2.38}$	$34.85_{0.03}$	$43.85_{0.30}$	$\mathbf{54.94}_{0.03}$
Caltech 101-7	ACC	$27.05_{1.70}$	$32.42_{0.80}$	$44.28_{1.84}$	$33.42_{0.80}$	$27.68_{2.14}$	$52.28_{0.04}$	$39.42_{0.41}$	$\mathbf{66.42}_{0.31}$
	NMI	$7.68_{0.04}$	$25.42_{0.37}$	$27.05_{1.44}$	$26.42_{0.37}$	$13.42_{1.70}$	$18.05_{0.06}$	$28.05_{0.35}$	$\mathbf{29.05}_{0.90}$
	PUR	$52.94_{0.05}$	$66.42_{0.53}$	$46.05_{1.70}$	$67.42_{0.53}$	$54.58_{1.58}$	$52.05_{0.04}$	$69.68_{0.25}$	$\mathbf{72.85}_{0.46}$
	ARI	$4.05_{0.04}$	$20.05_{1.04}$	$7.68_{0.29}$	$21.05_{1.04}$	$7.05_{1.58}$	$14.68_{0.04}$	$28.42_{0.37}$	$\mathbf{41.58}_{1.00}$
proteinFold	ACC	$14.05_{0.35}$	$20.42_{0.76}$	$28.05_{0.02}$	$20.85_{1.30}$	$16.42_{1.38}$	$14.42_{0.80}$	$26.42_{0.50}$	$\mathbf{35.85}_{0.27}$
	NMI	$17.05_{0.37}$	$26.05_{0.46}$	$17.05_{0.02}$	$27.42_{1.38}$	$23.05_{1.20}$	$20.05_{0.94}$	$33.42_{0.41}$	$\mathbf{42.85}_{0.46}$
	PUR	$19.42_{0.46}$	$24.85_{1.04}$	$28.68_{0.02}$	$26.05_{1.55}$	$21.58_{1.55}$	$18.94_{1.00}$	$31.42_{0.35}$	$\mathbf{41.58}_{0.89}$
	ARI	$2.85_{0.21}$	$5.94_{0.50}$	$10.58_{0.02}$	$8.05_{0.96}$	$4.28_{0.90}$	$2.94_{0.46}$	$11.85_{0.25}$	$\mathbf{16.42}_{1.35}$

learned from the first view alone. The distance grows sharply in the first three rounds, which reflects the strong heterogeneity of the views and confirms that the initial view provides only a partial description of the kernel space. Fig. 4(c) reports the distance between the intermediate embedding and the final embedding. The curve decreases monotonically to zero, which confirms that the incremental solutions form a convergent trajectory even when the views are highly heterogeneous. Fig. 4(d) reports the view weights over t. The weights are strongly non-uniform and change with t, which shows that under large disparity the dual-mode switch activates the purified space and the error-driven feedback loop actively suppresses the high-dimensional noisy views. Together, the four panels show that the kernel space on Caltech101-7 is progressively constructed and that this construction is driven by the disparity-aware modulation of the LCLS-SMKC.

proteinFold. Fig. 4(a) reports the corresponding performance on proteinFold. The metrics reach their near-optimal values within the first two views and then remain stable, which is consistent with the low disparity across views: since all views share the same dimension and encode similar information, the learned space saturates quickly. Fig. 4(b) shows that the distance to the first-view embedding grows moderately and then saturates, which indicates that the initial view already captures a large fraction of the kernel space. Fig. 4(c) shows that the distance to the final embedding decreases monotonically, which confirms convergence of the incremental solutions in the low-disparity regime. Fig. 4(d) shows that the view weights are moderately non-uniform, with a few views receiving slightly higher weights but no view dominating the space, which is consistent with the strong redundancy among the low-dimensional views of proteinFold. These four panels confirm that under low disparity the LCLS-SMKC constructs the learning space by aggregating the complementary information of the views without amplifying any single view.

Across the two datasets, the four panels consistently show that the learned kernel space departs from the initial space, approaches the final space monotonically, and is governed by dynamically reallocated view weights. The contrast between Caltech101-7 and proteinFold further shows that the LCLS-SMKC adapts its behavior to the disparity regime: under large disparity the learned space is



(a) Clustering performance (b) Distance to the first-view embedding (c) Distance to the final-view embedding (d) View weights

Figure 4: Incremental construction of the learning space on Caltech101-7, shown in the top row of each panel, and proteinFold, shown in the bottom row of each panel. Panel (a) reports the clustering performance in terms of ACC, NMI, PUR, and ARI as the number of views t increases. Panel (b) reports the Frobenius distance between the consensus embedding learned from the first t views and the one learned from the first view alone. Panel (c) reports the Frobenius distance between the intermediate embedding and the final embedding obtained from all views. Panel (d) reports the view weights β over t. The four panels together show that the learned space departs from the initial space, approaches the final space monotonically, and is governed by dynamically reallocated view weights, which confirms that the kernel space is progressively constructed rather than frozen.

rebuilt aggressively from heterogeneous views, whereas under small disparity it is refined smoothly from redundant views. This is direct evidence that the LCLS-SMKC learns to construct the learning space, whereas existing scalable methods keep the space fixed once the approximation is computed.

B Key Innovations

The four innovations are tightly coupled and form a single pipeline. The multi-scale leverage score fusion provides a representative landmark set that serves as the basis of the kernel space. The disparity-aware dual-mode switch then uses this basis to construct a kernel structure that adapts to heterogeneous views. The residual soft-thresholding and Laplacian reweighting purify the features that populate this structure. Finally, the error-driven feedback loop closes the pipeline by feeding the approximation error back into the joint optimization of the kernel space and the consensus embedding. Together, these components turn the kernel space from a frozen approximation into a learnable component.

1. **Multi-scale leverage score fusion selects landmarks that cover the kernel space.** Existing landmark-based methods select landmarks from a single kernel at a fixed bandwidth, so the sampled subset reflects only one notion of similarity. the LCLS-SMKC draws M kernels whose bandwidths follow a geometric grid $\sigma_j = \sigma_0 2^{j-1}$, computes a ridge leverage score matrix $\mathbf{L} \in \mathbb{R}^{n \times M}$ with scale-specific regularization $\lambda_j = \sigma_j^2$, and fuses the M score vectors by solving a convex quadratic program whose matrix $\mathbf{Q} = M\mathbf{L}^\top \mathbf{L}$ penalizes pairwise redundancy and whose linear term $\mathbf{b} = \mathbf{L}^\top \mathbf{L}\mathbf{1}_M$ rewards average coverage. The solution $\boldsymbol{\alpha}^*$ defines a single fused score vector $\boldsymbol{\ell} = \mathbf{L}\boldsymbol{\alpha}^*$ from which landmarks are drawn without replacement. This construction ensures that the fused score remains a convex combination of the per-scale scores, so the strict improvement of Theorem 1 over uniform sampling is preserved, while the redundancy penalty prevents any single scale from dominating the selection.

2. **A disparity-aware dual-mode switch shapes the kernel space for heterogeneous views.** Existing scalable methods use a single kernel type, typically Gaussian, regardless of the dimensionality of the views. the LCLS-SMKC defines the dimensionality ratio $\rho_d = \max_v d_v / \min_v d_v$ and switches to a different kernel construction when ρ_d exceeds a threshold. In the large-disparity mode, the method scales the data by $\mathbf{\Lambda}_v^{-1/2}$ so that every coordinate contributes equally, applies double-centering to remove the row and column means that inflate the effective rank, and normalizes by the spectral norm $\sigma_{\max}(\mathbf{KH}_v)$ to bound the operator norm. The method also embeds the dimensionality variation coefficient $\delta = \text{std}(\{d_v\})/\text{mean}(\{d_v\})$ into three space-shaping parameters. The diversity coefficient becomes $\alpha_{\text{div}} = \frac{1}{V+1} e^{-\delta^2}$, the view-level regularization becomes $\lambda_{\text{view}} = \lambda_{\text{cons}}(1+\delta)$, and the per-view kernel weight is estimated as $\beta_v = \text{tr}((\mathbf{S}_v\mathbf{H}_v)^\top \mathbf{KH}_v)/\text{tr}(\mathbf{KH}_v^\top \mathbf{KH}_v)$. Theorem 2 shows that this combination makes the linear kernel strictly better than the Gaussian kernel under large disparity, and Theorem 5 shows that α_{div} vanishes as δ grows.
3. **Residual soft-thresholding and Laplacian reweighting purify the feature space.** Existing methods apply graph diffusion or feature reweighting separately, and they typically threshold the diffused features themselves. the LCLS-SMKC computes the residual $\mathbf{R}_v = \mathbf{X}_v^{(T)} - \mathbf{X}_v^{(0)}$ between the diffused and the original features, applies a soft-thresholding operator $\mathcal{T}_\tau(\mathbf{R})$ with a quantile-based threshold $\tau = \text{quantile}(|\mathbf{R}|, q)$, and adds the thresholded residual back to the original features. This residual form preserves the original feature scale and removes only the small, high-frequency perturbations that diffusion introduces. After this step, the method measures per-feature smoothness with the normalized Laplacian and reweights each feature by the inverse square root of its smoothness. The weights are clipped to $[0.5, 2]$ and the reweighted features are combined with the denoised features through a convex combination with coefficient $\gamma_w = 0.3$. Theorem 3 proves that this exact combination yields exponential SNR growth and a bounded noise suppression error.
4. **An error-driven feedback loop jointly optimizes the kernel space and the consensus embedding.** Existing methods either use a fixed consistency weight or decay it by a fixed schedule, and they never update the kernel space itself. the LCLS-SMKC computes the total approximation error $\mathcal{E}_{\text{cur}} = \sum_v \|\beta_v \mathbf{KH}_v - \mathbf{S}_v\mathbf{H}_v\|_F^2$ at every outer iteration and compares it with the previous value. When the error increases, the consistency weight is amplified by a factor $1 + \eta_{\text{inc}}(1+\delta)$ with $\eta_{\text{inc}} = 0.1$. When the error decreases, the weight is annealed by a factor $1 - \eta_{\text{dec}}$ with $\eta_{\text{dec}} = 0.05$ and floored at $\lambda_{\text{cons}}^{\text{base}} = 10^{-3}$. The method treats $\mathbf{KH}_v$ and $\mathbf{H}_v$ as learnable variables and updates them jointly with $\mathbf{S}_v$ and $\mathbf{S}_\star$ through the composite objective in equation 23. Each block admits a closed-form update through a Procrustes problem or an SVD. Theorem 6 proves that the asymmetric rule is Lyapunov stable with a steady-state error bounded by $\frac{\eta_{\text{dec}}}{\eta_{\text{inc}}}\mathcal{E}_{\text{base}}$, and Theorem 7 shows that the resulting alternating minimization converges to a KKT stationary point from any initialization.