A NONLINEAR KERNEL SPECTRAL PERSPECTIVE ON MULTI-VIEW CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-view contrastive learning has become a powerful paradigm for exploiting cross-view consistency, yet most existing methods treat contrastive objectives primarily as representation-level regularizers within larger learning pipelines. The spectral structure induced by these objectives remains less understood. We present a nonlinear kernel spectral perspective on multi-view contrastive learning and derive **MVCL-KS**, a direct spectral realization of this perspective. Starting from bidirectional InfoNCE, we show that its exact dot-product gradient induces a signed cross-view relation operator; aggregating all view pairs yields a unified multi-view block spectral system, which is further lifted to view-specific reproducing kernel Hilbert spaces. Because the contrastive probabilities depend on the current representation, the resulting formulation is a nonlinear self-consistent spectral system. This formulation leads to dominant-eigenspace spectral refinement with basis-invariant cycle-aware stabilization, together with low-rank and anchor-based implementations for scalability. Under the same evaluation setting, MVCL-KS is $1.7\times$ to $22.1\times$ faster than the fastest competing multi-view baseline across all benchmarks, while maintaining competitive or superior clustering quality. Experiments with multiple kernel families further show that the same spectral formulation works with different nonlinear view geometries.

1 INTRODUCTION

Multi-view data arise whenever the same objects are described by heterogeneous feature sets, sensors, modalities, or semantic descriptors. A central goal of multi-view learning is to exploit information shared across views while preserving complementary structures that are useful for downstream tasks such as clustering. Classical approaches address this problem through subspace learning, graph fusion, matrix/tensor factorization, and spectral embedding (Gao et al., 2015; Zhang et al., 2017; Xia et al., 2022). More recent methods increasingly rely on deep representation learning to extract nonlinear view-specific features and align them in a shared latent space (Shaham et al., 2018; Huang et al., 2019; Eisenberg et al., 2025). Contrastive learning is well suited to multi-view data because corresponding observations naturally provide cross-view positive relations, while competition against other instances promotes discriminative structure (Chen et al., 2020b; Zhang et al., 2026).

Along this line, a broad family of contrastive multi-view clustering methods has emerged. Representative approaches align instance-level or cluster-level representations (Pan & Kang, 2021; Xu et al., 2022; Chen et al., 2023), improve the reliability of positive and negative relations (Lu et al., 2024; Guo et al., 2024; Sun et al., 2024; 2025), or account for heterogeneous view quality through self-weighted and reliability-aware objectives (Wu et al., 2024; Lou et al., 2025; He et al., 2026). Kernel-based variants further combine contrastive objectives with multi-kernel clustering (Liu et al., 2023; Su et al., 2024). Although these methods have substantially improved clustering performance, the contrastive term is typically embedded inside a larger representation-learning pipeline. This makes it difficult to isolate the geometric transformation induced by the contrastive objective itself and, in particular, to understand how pairwise interactions among multiple views organize a common spectral representation.

Recent theoretical work has begun to reveal the spectral and algebraic structure underlying contrastive learning. Tan et al. connect InfoNCE-style learning with spectral clustering on a similarity graph (Tan et al., 2024); Nakada et al. relate multimodal contrastive learning to a contrastive cross-

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.

covariance structure and an SVD characterization (Nakada et al., 2023); and UniCon studies efficient contrastive alignment through a kernel-based gradient-equivalence view (Sui et al., 2026). These results provide important foundations, but they do not directly answer a multi-view question: *can the attractive and repulsive relations induced by all pairwise contrastive objectives be represented by one unified spectral system and lifted to nonlinear view-specific geometries in a direct kernel formulation?*

We answer this question from a signed-operator perspective. We show that bidirectional dot-product InfoNCE induces a signed cross-view relation whose positive correspondence term and softmax competition term jointly encode attraction and repulsion. Stacking all pairwise relations produces a symmetric multi-view block operator, converting separate pairwise contrastive interactions into a single spectral system. We then lift this system to heterogeneous nonlinear geometries through view-specific kernels. Since the contrastive probabilities themselves depend on the current spectral representation, the resulting formulation becomes a nonlinear self-consistent eigensystem rather than an ordinary one-shot spectral decomposition.

This perspective leads to a simple method, termed **MVCL-KS**. It alternates between constructing the contrastive signed operator and extracting the dominant algebraic eigenspace of its kernelized form. We interpret this procedure as *self-consistent spectral refinement*. To improve robustness to non-contractive dynamics, we detect two-cycle behavior using basis-invariant subspace distances and stabilize it through projector averaging. A low-rank reproducing kernel Hilbert space (RKHS) formulation and an anchor-based out-of-sample extension make the method applicable to larger datasets.

The resulting framework also reduces the computational burden of multi-view contrastive learning. Deep competitors typically require repeated forward/backward passes through view-specific encoders and many gradient updates, whereas MVCL-KS operates directly on kernel and spectral quantities. In our experiments, the fixed radial basis function (RBF) setting of MVCL-KS obtains the best ACC on all six main benchmarks among the compared methods. Under the common GPU setting, MVCL-KS is $1.7\times$ to $22.1\times$ faster than the fastest competing multi-view method across all eight benchmarks. An additional kernel comparison shows that the proposed spectral formulation is not tied to RBF and can be instantiated with linear, cosine, polynomial, and Laplacian geometries.

Our main contributions are summarized as follows:

- We derive a signed-operator representation of bidirectional dot-product InfoNCE for multi-view data, exposing the attractive and repulsive cross-view structure hidden in its gradient.
- We assemble all view-pair interactions into a unified symmetric block spectral system and lift it to view-specific RKHS geometries, yielding a nonlinear self-consistent spectral formulation for multi-view contrastive learning.
- We develop MVCL-KS, a kernel-spectral method with cycle-aware subspace stabilization and low-rank and anchor-based extensions. The theoretical development distinguishes conditional spectral optimality from global convergence of the nonlinear refinement.

2 From Multi-View InfoNCE to a Signed Spectral System

2.1 Problem Setup and Spectral Perspective

Let $\mathcal{X} = \{X^{(v)}\}_{v=1}^{V}$ be a multi-view dataset with V views, where $X^{(v)} = [\mathbf{x}_1^{(v)}, \ldots, \mathbf{x}_N^{(v)}]^\top \in \mathbb{R}^{N\times d_v}$ contains N observations from view v. The goal is to partition the N objects into C clusters without label supervision. For each view, let $F^{(v)} = [\mathbf{f}_1^{(v)}, \ldots, \mathbf{f}_N^{(v)}]^\top \in \mathbb{R}^{N\times C}$ denote its current C-dimensional representation, and define the stacked representation

$$F = \begin{bmatrix} F^{(1)\top} & \cdots & F^{(V)\top} \end{bmatrix}^\top \in \mathbb{R}^{VN\times C}. \tag{1}$$

Our starting point is that the pairwise InfoNCE objectives used in multi-view contrastive clustering are more than sample-wise alignment losses. Their gradients induce *signed cross-view linear operators*. By assembling these operators across all view pairs and then lifting them to view-specific reproducing kernel Hilbert spaces (RKHSs), multi-view contrastive clustering becomes a nonlinear

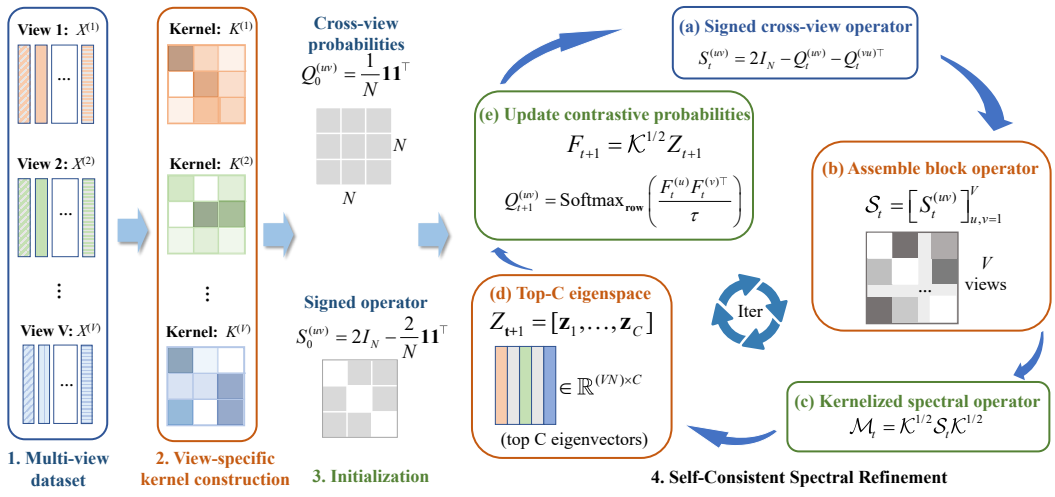

Figure 1: Overview of MVCL-KS in the full sample space. Cross-view signed operators are assembled into a V-partite block operator, lifted by view-specific kernels, and coupled to the dominant algebraic eigenspace of the kernelized operator.

self-consistent spectral system:

$$Z \longrightarrow \mathcal{S}(Z) \longrightarrow \mathcal{M}(Z) \longrightarrow \mathrm{Eig}_C\big(\mathcal{M}(Z)\big), \tag{2}$$

Here $\mathcal{S}(Z)$ denotes the representation-dependent signed block operator induced by the current cross-view contrastive probabilities, and $\mathcal{M}(Z)$ denotes its kernelized form in the joint RKHS. Their exact definitions are introduced in Sections 2.3 and 3.2, and the procedure is illustrated in Figure 1. The operator $\mathrm{Eig}_C(\cdot)$ returns the eigenspace associated with the C largest *algebraic* eigenvalues. Algebraic rather than magnitude ordering is required because the signed operators can be indefinite.

2.2 From Pairwise InfoNCE to a Signed Cross-View Operator

For a pair of views $u \neq v$, we use the dot-product similarity

$$s_{ij}^{(uv)} = \mathbf{f}_i^{(u)\top} \mathbf{f}_j^{(v)}. \tag{3}$$

With temperature $\tau > 0$, the directional contrastive probability from view u to view v is

$$q_{ij}^{(uv)} = \frac{\exp\left(s_{ij}^{(uv)}/\tau\right)}{\sum_{\ell=1}^{N} \exp\left(s_{i\ell}^{(uv)}/\tau\right)}, \qquad Q^{(uv)} = [q_{ij}^{(uv)}]_{i,j=1}^{N}. \tag{4}$$

For multi-view data, the positive target for sample i is its corresponding sample in the other view. The directional loss is

$$\mathcal{L}_{u\to v} = -\frac{1}{N}\sum_{i=1}^{N} \log q_{ii}^{(uv)}, \tag{5}$$

and we symmetrize the two directions as

$$\mathcal{L}_{uv} = \frac{1}{2}\left(\mathcal{L}_{u\to v} + \mathcal{L}_{v\to u}\right). \tag{6}$$

The difference between the one-to-one correspondence target and the current softmax association yields the signed cross-view operator

$$S^{(uv)} = 2I_N - Q^{(uv)} - Q^{(vu)\top}. \tag{7}$$

Unlike a conventional nonnegative affinity matrix, $S^{(uv)}$ can contain both positive and negative entries. It therefore represents attractive and repulsive corrections induced by contrastive learning.

Proposition 1 (Exact pairwise gradient representation)**.** *For the dot-product similarity in equation 3, the bidirectional loss in equation 6 satisfies*

$$-\frac{\partial \mathcal{L}_{uv}}{\partial F^{(u)}} = \frac{1}{2N\tau} S^{(uv)} F^{(v)}, \qquad -\frac{\partial \mathcal{L}_{uv}}{\partial F^{(v)}} = \frac{1}{2N\tau} S^{(uv)\top} F^{(u)}. \tag{8}$$

A proof is provided in Appendix E.1. Proposition 1 exposes the central spectral object hidden inside pairwise InfoNCE: once the current probabilities are fixed, the local update direction is fully characterized by the action of $S^{(uv)}$.

2.3 Unified Multi-View Block Spectral System

We aggregate all pairwise losses with equal weight,

$$\mathcal{L}_{\mathrm{MVC}} = \sum_{1 \le u < v \le V} \mathcal{L}_{uv}, \tag{9}$$

without introducing learnable view or view-pair weights. The pairwise signed operators are assembled into the symmetric block matrix

$$\mathcal{S}(F) = \begin{bmatrix} 0 & S^{(12)} & \cdots & S^{(1V)} \\ S^{(12)\top} & 0 & \cdots & S^{(2V)} \\ \vdots & \vdots & \ddots & \vdots \\ S^{(1V)\top} & S^{(2V)\top} & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{VN \times VN}. \tag{10}$$

The zero diagonal blocks isolate the structure induced by *cross-view* contrastive interactions without adding a separate intra-view graph objective. Thus, $\mathcal{S}$ can be interpreted as the adjacency operator of a signed V-partite graph.

By stacking the pairwise gradient identities, we obtain

$$-\nabla_F \mathcal{L}_{\mathrm{MVC}} = \frac{1}{2N\tau} \mathcal{S}(F) F. \tag{11}$$

For a fixed operator $\mathcal{S}$, the corresponding conditional spectral problem is

$$\max_{Z \in \mathbb{R}^{VN \times C}} \operatorname{Tr}\left(Z^\top \mathcal{S} Z\right) \quad \text{s.t.} \quad Z^\top Z = I_C. \tag{12}$$

Theorem 2 (Conditional multi-view spectral characterization)**.** *Let $\lambda_1(\mathcal{S}) \ge \cdots \ge \lambda_{VN}(\mathcal{S})$ be the eigenvalues of a fixed $\mathcal{S}$, and let U_C contain eigenvectors associated with its C largest algebraic eigenvalues. Then*

$$Z^\star = U_C R, \qquad R^\top R = I_C, \tag{13}$$

is a global solution of equation 12, with optimal value $\sum_{c=1}^{C} \lambda_c(\mathcal{S})$. Every global solution spans a maximizing C-dimensional invariant subspace; if $\lambda_C(\mathcal{S}) > \lambda_{C+1}(\mathcal{S})$, this subspace is unique.

The proof follows from the Rayleigh–Ritz theorem and is given in Appendix E.2. This result converts the collection of $\binom{V}{2}$ pairwise contrastive interactions into a single joint spectral partitioning system.

3 Nonlinear Kernel Spectral Learning

3.1 Kernelization: Nonlinear Contrastive Spectral Geometry

The signed operator captures cross-view attraction and repulsion, while heterogeneous views may exhibit strongly nonlinear geometry. For each view v, let $k_v(\cdot,\cdot)$ be a positive semidefinite kernel with Gram matrix

$$K^{(v)} = \left[k_v\left(\mathbf{x}_i^{(v)}, \mathbf{x}_j^{(v)}\right)\right]_{i,j=1}^{N}. \tag{14}$$

We center and scale each kernel by

$$\bar{K}^{(v)} = J K^{(v)} J, \qquad J = I_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top, \qquad K^{(v)} \leftarrow \frac{N\bar{K}^{(v)}}{\operatorname{Tr}(\bar{K}^{(v)})}, \tag{15}$$

which removes arbitrary offsets and places view-specific kernels on a comparable scale without learnable fusion weights. Define the block-diagonal kernel

$$\mathcal{K} = \mathrm{blkdiag}\left(K^{(1)}, \ldots, K^{(V)}\right). \tag{16}$$

Let $A \in \mathbb{R}^{VN \times C}$ denote RKHS coefficients and let $F = \mathcal{K}A$ be the stacked sample representation. We impose the joint RKHS orthogonality constraint

$$A^\top \mathcal{K} A = I_C. \tag{17}$$

With the coordinate transformation

$$Z = \mathcal{K}^{1/2} A, \qquad F = \mathcal{K}^{1/2} Z, \tag{18}$$

the kernelized conditional problem becomes

$$\max_{Z^\top Z = I_C} \mathrm{Tr}\left(Z^\top \mathcal{M} Z\right), \qquad \mathcal{M} = \mathcal{K}^{1/2} \mathcal{S} \mathcal{K}^{1/2}. \tag{19}$$

We refer to $\mathcal{M}$ as the *kernelized multi-view contrastive operator*. Its (u, v)-th off-diagonal block is

$$M^{(uv)} = K^{(u)1/2} S^{(uv)} K^{(v)1/2}, \tag{20}$$

which explicitly couples view-specific nonlinear geometry with contrastively induced signed relations.

Theorem 3 (Kernel spectral characterization)**.** *For a fixed $\mathcal{S}$, the solution of equation 19 is the dominant C-dimensional algebraic eigenspace of $\mathcal{M}$. If $\mathcal{K} \succ 0$ and U_C denotes this eigenspace, the corresponding RKHS coefficients and sample representation satisfy*

$$A^\star = \mathcal{K}^{-1/2} U_C R, \qquad F^\star = \mathcal{K}^{1/2} U_C R, \qquad R^\top R = I_C. \tag{21}$$

For singular $\mathcal{K}$, the inverse is restricted to its effective positive-eigenvalue subspace.

The proof is deferred to Appendix E.3.

Special cases. When $V = 2$, $\mathcal{M}$ reduces to the bipartite operator

$$\mathcal{M} = \begin{bmatrix} 0 & C^{(12)} \\ C^{(12)\top} & 0 \end{bmatrix}, \qquad C^{(12)} = K^{(1)1/2} S^{(12)} K^{(2)1/2}, \tag{22}$$

whose positive eigenvectors are obtained from the left and right singular vectors of $C^{(12)}$. With a linear kernel $K^{(v)} = X^{(v)} X^{(v)\top}$, the formulation reduces, up to the nonzero spectrum, to feature-space signed cross-covariance operators of the form $X^{(u)\top} S^{(uv)} X^{(v)}$. Hence, the proposed formulation simultaneously generalizes two-view spectral alignment to arbitrary V and linear contrastive structure to nonlinear RKHS geometry.

3.2 Self-Consistent Spectral Refinement

A key distinction from ordinary spectral clustering is that $\mathcal{S}$ is representation-dependent: $Q^{(uv)}$, and therefore $S^{(uv)}$, changes with the current embedding. Using the kernelized operator in equation 19 and the relation $F = \mathcal{K}^{1/2} Z$ in equation 18, the nonlinear operator is

$$\mathcal{M}(Z) = \mathcal{K}^{1/2} \mathcal{S}(Z) \mathcal{K}^{1/2}. \tag{23}$$

Proposition 4 (First-order self-consistent stationarity)**.** *Under the joint orthogonality constraint $Z^\top Z = I_C$, every first-order stationary point of the dot-product multi-view InfoNCE objective satisfies*

$$\mathcal{M}(Z^\star) Z^\star = Z^\star \Lambda \tag{24}$$

for some symmetric $\Lambda \in \mathbb{R}^{C \times C}$. Consequently, any fixed point whose span equals $\mathrm{Eig}_C(\mathcal{M}(Z^\star))$ is a first-order stationary point.

The proof, based on equation 11 and the Stiefel constraint, is provided in Appendix E.4.

Motivated by Proposition 4, we use a *self-consistent spectral refinement*:

$$F_t = \mathcal{K}^{1/2} Z_t, \tag{25}$$

$$Q_t^{(uv)} = \mathrm{Softmax}_{\mathrm{row}}\left(\frac{F_t^{(u)} F_t^{(v)\top}}{\tau}\right), \tag{26}$$

$$S_t^{(uv)} = 2I_N - Q_t^{(uv)} - Q_t^{(vu)\top}, \tag{27}$$

$$\mathcal{M}_t = \mathcal{K}^{1/2} \mathcal{S}_t \mathcal{K}^{1/2}, \tag{28}$$

$$Z_{t+1} = \mathrm{Eig}_C(\mathcal{M}_t). \tag{29}$$

For a fixed $\mathcal{S}_t$, equation 29 exactly solves the conditional Rayleigh problem in equation 19. However, because $\mathcal{S}_t$ itself depends on Z_t, we do *not* interpret the eigenspace replacement as a monotone descent step for the original InfoNCE objective. Its role is instead to enforce spectral self-consistency between the representation and the contrastive operator that it induces.

Parameter-free initialization. We initialize every cross-view probability matrix with the uninformative distribution

$$Q_0^{(uv)} = \frac{1}{N}\mathbf{1}\mathbf{1}^\top, \qquad u \neq v, \tag{30}$$

which yields

$$S_0^{(uv)} = 2I_N - \frac{2}{N}\mathbf{1}\mathbf{1}^\top. \tag{31}$$

The first spectral subspace is therefore obtained directly from the normalized kernels and the uninformative contrastive initialization.

3.3 Cycle-Aware Stabilization

Because the map in equation 2 is nonlinear, repeated dominant-eigenspace replacement need not be contractive. We therefore monitor convergence on the Grassmann manifold rather than comparing basis matrices directly. For two orthonormal C-dimensional bases Z_a and Z_b, define

$$d(Z_a, Z_b) = \sqrt{1 - \frac{1}{C}\left\|Z_a^\top Z_b\right\|_F^2}. \tag{32}$$

This distance is invariant to sign changes and orthogonal rotations of the basis.

At iteration t, we monitor the one-step and two-step distances

$$d_1(t) = d(Z_t, Z_{t-1}), \qquad d_2(t) = d(Z_t, Z_{t-2}). \tag{33}$$

If $d_1(t) \leq \varepsilon$, the refinement has stabilized. If instead

$$d_2(t) \leq \varepsilon_{\mathrm{cyc}} \quad \text{and} \quad d_1(t) > \varepsilon_{\mathrm{sep}}, \tag{34}$$

we regard the iteration as entering a two-cycle. Rather than selecting either state, we average the corresponding orthogonal projectors,

$$\bar{P}_t = \frac{1}{2}\left(Z_t Z_t^\top + Z_{t-1} Z_{t-1}^\top\right), \tag{35}$$

and use its dominant C-dimensional eigenspace as the stabilized output:

$$Z_{\mathrm{out}} = \mathrm{Eig}_C(\bar{P}_t). \tag{36}$$

This safeguard is basis-invariant, label-free, and leaves the underlying contrastive operator unchanged. In implementation, $\bar{P}_t$ need not be explicitly materialized; its dominant eigenspace can be recovered through a $2C \times 2C$ Gram matrix. Unless otherwise specified, all datasets use the fixed stopping configuration $T = 15$, $\varepsilon = 10^{-4}$, $\varepsilon_{\mathrm{cyc}} = 5 \times 10^{-4}$, and $\varepsilon_{\mathrm{sep}} = 10^{-4}$.

4 PRACTICAL MVCL-KS

4.1 CONSENSUS EMBEDDING AND CLUSTERING

Let the representation used for the final clustering results be partitioned by views as

$$R_{\text{out}} = \begin{bmatrix} R^{(1)\top} & \cdots & R^{(V)\top} \end{bmatrix}^\top, \qquad R^{(v)} \in \mathbb{R}^{N\times C}. \tag{37}$$

The implementation uses $R^{(v)} = Z^{(v)}$ in exact mode and the Nyström-extended RKHS representation $R^{(v)} = F^{(v)}$ in anchor mode. In either mode, the view blocks inherit the common orientation of the learned joint reduced eigenspace. We therefore use the parameter-free mean

$$R_{\text{con}} = \frac{1}{V}\sum_{v=1}^{V} R^{(v)}. \tag{38}$$

Each row is normalized,

$$\widetilde{\mathbf{r}}_i = \frac{\mathbf{r}_{\text{con},i}}{\|\mathbf{r}_{\text{con},i}\|_2}, \tag{39}$$

and K-means with C clusters produces the final assignments.

4.2 LOW-RANK KERNEL FORM

The exact formulation uses $N \times N$ kernels. To reduce the spectral cost, we retain the leading positive kernel components

$$K^{(v)} \approx U_r^{(v)} \Lambda_r^{(v)} U_r^{(v)\top}, \qquad U_r^{(v)} \in \mathbb{R}^{N\times r}. \tag{40}$$

Define

$$\mathcal{U} = \text{blkdiag}\left(U_r^{(1)}, \ldots, U_r^{(V)}\right), \qquad \mathbf{\Lambda} = \text{blkdiag}\left(\Lambda_r^{(1)}, \ldots, \Lambda_r^{(V)}\right). \tag{41}$$

Then the kernelized operator can be solved in the Vr-dimensional effective RKHS subspace through

$$\widehat{\mathcal{M}}_t = \mathbf{\Lambda}^{1/2}\mathcal{U}^\top \mathcal{S}_t \mathcal{U}\mathbf{\Lambda}^{1/2} \in \mathbb{R}^{Vr\times Vr}. \tag{42}$$

If W_{t+1} is its dominant C-dimensional eigenspace, the full spectral coordinate is recovered as

$$Z_{t+1} \approx \mathcal{U} W_{t+1}. \tag{43}$$

The representation used to update the contrastive probabilities is computed directly in reduced coordinates as

$$F_t = \mathcal{U}\mathbf{\Lambda}^{1/2} W_t. \tag{44}$$

Moreover, because $\mathcal{U}^\top\mathcal{U} = I$, the subspace distance is preserved: $d(\mathcal{U}W_a, \mathcal{U}W_b) = d(W_a, W_b)$. Hence, both convergence detection and cycle stabilization can be carried out entirely in the Vr-dimensional space. Exact mode forms $Z_{\text{out}} = \mathcal{U}W_{\text{out}}$ only for the final results; anchor mode instead extends the corresponding RKHS representation as described below. The cross-view probability matrices need not be stored simultaneously; their products with the retained kernel bases can be computed in chunks. For larger N, we use an anchor/Nyström extension based on the same reduced-RKHS formulation; the exact out-of-sample map and its consistency on anchor samples are given in Appendix E.10.

Anchor mode for large N. When the full $N \times N$ spectral system is too expensive, we randomly select $m \ll N$ samples without replacement as anchors and denote their index set by L. The kernels, signed cross-view operator, and self-consistent refinement are then constructed only on these anchors. For each view, the centered and scaled anchor kernel is factorized as $K_{LL}^{(v)} = U^{(v)}\Lambda^{(v)}U^{(v)\top}$. After learning the reduced spectral block $W^{(v)}$, the implementation extends the corresponding RKHS representation as

$$\mathbf{f}^{(v)}(x) = \bar{\mathbf{k}}_{xL}^{(v)} U^{(v)} \Lambda^{(v)-1/2} W^{(v)}, \tag{45}$$

using only the retained positive eigenvalues and the same centering and scaling as the anchor kernel. For every anchor x_i, this map recovers $\mathbf{f}^{(v)}(x_i) = \mathbf{e}_i^\top U^{(v)}\Lambda^{(v)1/2}W^{(v)}$. The extended $F^{(v)}$ blocks are combined by the consensus rule in Section 4.1. Anchor mode should therefore be interpreted as a scalable RKHS variant of the same reduced spectral training system.

Table 1: Clustering performance on 3Sources, BBCSport, and Caltech.

Method	3Sources				BBCSport				Caltech			
	ACC ↑	NMI ↑	ARI ↑	F1 ↑	ACC ↑	NMI ↑	ARI ↑	F1 ↑	ACC ↑	NMI ↑	ARI ↑	F1 ↑
SPC (Shi & Malik, 2000)	48.52	30.99	14.09	37.86	74.82	59.74	55.01	67.04	55.36	62.66	46.10	46.06
LPGMVC (Huang et al., 2023)	53.49	43.34	34.89	48.94	44.12	13.51	11.57	32.09	41.00	27.71	17.92	30.53
OMVCDR (Wan et al., 2024)	52.07	41.33	33.71	47.42	51.84	34.10	25.24	47.97	64.21	52.45	44.38	52.35
MFLVC (Xu et al., 2022)	63.31	48.41	43.87	62.87	40.81	7.59	7.23	37.80	72.36	65.54	55.38	62.39
CVCL (Chen et al., 2023)	62.72	56.23	**50.49**	**63.59**	60.57	45.31	49.53	62.65	40.93	34.53	21.97	37.55
CMK (Liu et al., 2023)	62.13	60.43	46.54	57.72	79.04	**80.18**	75.03	80.81	73.00	60.22	53.33	60.00
DCMVSC (Yu et al., 2024)	56.21	57.23	38.49	53.18	69.60	63.91	51.47	62.65	56.43	48.57	41.03	49.94
RMCNC (Sun et al., 2024)	39.64	22.87	34.55	12.04	37.13	7.88	33.92	5.69	31.64	20.53	27.48	12.80
COPER (Eisenberg et al., 2025)	47.34	39.77	25.53	40.55	69.67	54.13	47.69	59.34	56.86	46.66	36.21	46.02
ROLL (Sun et al., 2025)	44.38	31.51	41.13	15.64	39.71	17.29	32.31	9.17	37.43	22.60	31.30	16.00
BMvC (Li et al., 2026)	59.14	55.67	43.64	57.27	82.64	64.46	62.95	73.18	80.95	**77.07**	**71.26**	**76.36**
MVCL-KS	**65.09**	**63.20**	49.98	62.31	**88.24**	78.45	**82.10**	**86.55**	**83.43**	71.37	69.13	73.82

Best results are in **bold** and second-best results are underlined. All values are percentages.

Table 2: Clustering performance on 100leaves, Notting-Hill, and NUS-WIDE.

Method	100leaves				Notting-Hill				NUS-WIDE			
	ACC ↑	NMI ↑	ARI ↑	F1 ↑	ACC ↑	NMI ↑	ARI ↑	F1 ↑	ACC ↑	NMI ↑	ARI ↑	F1 ↑
SPC (Shi & Malik, 2000)	69.13	83.90	56.90	57.33	49.55	39.06	29.28	38.21	30.38	15.15	8.97	6.09
LPGMVC (Huang et al., 2023)	33.19	61.95	13.71	14.80	84.51	78.62	79.41	84.03	26.21	11.98	7.59	20.47
OMVCDR (Wan et al., 2024)	89.81	93.27	82.05	82.22	80.39	74.04	69.89	76.56	36.44	**24.88**	18.65	**28.80**
MFLVC (Xu et al., 2022)	30.63	64.88	18.95	25.61	60.84	56.88	43.04	55.23	37.94	20.77	15.55	26.84
CVCL (Chen et al., 2023)	51.81	73.12	36.19	42.89	61.72	62.01	47.43	62.45	26.56	13.11	8.58	25.75
CMK (Liu et al., 2023)	88.94	95.21	85.02	85.20	82.66	74.07	74.99	80.43	36.38	22.51	15.81	26.43
DCMVSC (Yu et al., 2024)	38.51	66.39	24.28	32.76	58.22	38.11	35.40	49.04	32.00	16.52	11.32	23.16
RMCNC (Sun et al., 2024)	57.31	78.18	56.18	42.92	67.73	59.00	60.92	48.84	28.75	11.12	**27.65**	8.77
COPER (Eisenberg et al., 2025)	21.44	58.21	13.70	21.41	87.23	77.25	78.95	83.63	29.94	13.91	9.68	23.64
ROLL (Sun et al., 2025)	75.25	88.63	74.02	65.50	37.43	22.60	31.30	16.00	25.56	10.01	23.54	6.26
BMvC (Li et al., 2026)	87.98	95.27	83.94	87.23	85.31	81.42	79.45	82.92	36.14	21.45	15.00	27.04
MVCL-KS	**91.38**	**95.94**	**87.59**	**89.13**	**89.25**	**84.08**	**83.51**	**87.14**	**39.88**	21.21	17.54	28.72

Best results are in **bold** and second-best results are underlined. All values are percentages.

5 EXPERIMENTS

Detailed dataset statistics, baseline descriptions, and parameter settings are provided in Appendix B. The six main benchmarks use a fixed RBF kernel to avoid dataset-specific kernel selection; additional kernel comparisons are reported in Appendix C.

5.1 COMPARISON WITH STATE-OF-THE-ART METHODS

Tables 1 and 2 compare MVCL-KS with existing methods on the six main benchmarks. Notably, using an RBF kernel across all six datasets, MVCL-KS obtains the highest ACC. Relative to the best competing ACC on each dataset, the gains are $+1.78$, $+5.60$, $+2.48$, $+1.57$, $+2.02$, and $+1.94$ percentage points on 3Sources, BBCSport, Caltech, 100leaves, Notting-Hill, and NUS-WIDE, respectively.

MVCL-KS is particularly strong on BBCSport, 100leaves, and Notting-Hill, where the spectral representation simultaneously yields high ACC, NMI, ARI, and F1. On Caltech and NUS-WIDE, some competitors remain stronger on NMI, ARI, and F1, indicating that the proposed spectral perspective does not merely optimize a single external clustering criterion. Consequently, the RBF-only results support the main claim that the signed kernel spectral system itself provides a competitive clustering representation before exploiting kernel-specific adaptation.

Table 3: Wall-clock runtime (seconds) on the six main benchmarks. Only multi-view methods executed on a single NVIDIA H20 GPU under the same experimental platform are included.

Method	3Sources	BBCSport	Caltech	100leaves	Notting-Hill	NUS-WIDE
MFLVC	8.99	12.28	51.91	48.18	93.58	174.16
CVCL	2.67	23.07	45.00	<u>14.20</u>	85.39	60.66
CMK	<u>1.38</u>	6.42	122.23	103.06	1546.90	94.62
DCMVSC	5.89	6.98	45.75	30.52	161.92	23.19
RMCNC	1.48	<u>4.60</u>	<u>7.93</u>	21.28	<u>32.16</u>	<u>7.91</u>
COPER	46.22	125.03	562.26	2527.66	1884.08	372.24
ROLL	8.85	36.73	74.34	54.79	74.34	55.71
BMvC	17.62	18.78	18.38	43.05	761.55	53.52
MVCL-KS	**0.83**	**0.35**	**0.79**	**1.65**	**3.67**	**0.98**

Table 4: Evaluation on Cifar100 (50,000 samples) and YouTubeVideo (101,499 samples).

	Cifar100					YouTubeVideo				
Method	ACC	NMI	ARI	F1	Time	ACC	NMI	ARI	F1	Time
SPC (Shi & Malik, 2000)	94.64	98.02	94.07	94.25	63.64	16.52	<u>17.04</u>	2.89	2.54	14.60
MFLVC (Xu et al., 2022)	76.61	95.68	79.22	79.53	3038.22	**21.68**	16.99	2.77	<u>13.23</u>	5187.84
CVCL (Chen et al., 2023)	77.74	82.98	65.32	65.80	1710.32	13.67	8.22	1.34	10.27	7169.98
CMK (Liu et al., 2023)	<u>98.49</u>	<u>99.73</u>	<u>98.69</u>	<u>98.70</u>	584.63	20.09	6.58	0.41	**21.52**	2567.33
RMCNC (Sun et al., 2024)	95.60	99.26	96.22	95.08	102.43	15.59	13.07	**16.56**	2.10	1370.12
CMSCU (Zhao et al., 2025)	97.82	98.52	97.01	97.40	4849.05	14.96	10.98	2.07	10.32	4501.61
ROLL (Sun et al., 2025)	95.68	99.10	95.12	96.04	2543.57	14.39	12.46	<u>14.68</u>	1.88	3286.38
BMvC (Li et al., 2026)	96.70	99.45	97.12	98.24	3913.80	10.46	8.39	1.07	7.91	6955.19
DMVC-CSG (Cui et al., 2026)	94.56	99.00	94.19	94.32	2587.24	11.17	11.54	1.54	7.26	6207.88
MVCL-KS	**99.97**	**99.96**	**99.94**	**99.94**	17.96	<u>21.38</u>	**20.68**	4.11	8.80	62.10

5.2 Efficiency

Table 3 reports wall-clock time only for multi-view methods executed in the same NVIDIA H20 GPU environment. MVCL-KS requires only 0.35–3.67 seconds on the six main benchmarks. Relative to the fastest competing GPU baseline on each dataset, the speedup ranges from approximately $1.7\times$ to $13.1\times$. The fastest competing GPU baselines require 1.38, 4.60, 7.93, 14.20, 32.16, and 7.91 seconds on 3Sources, BBCSport, Caltech, 100leaves, Notting-Hill, and NUS-WIDE, whereas MVCL-KS requires 0.83, 0.35, 0.79, 1.65, 3.67, and 0.98 seconds, respectively.

5.3 Large-Scale Evaluation

We further evaluate scalability on Cifar100 with $50{,}000$ samples and YouTubeVideo with $101{,}499$ samples using the anchor mode introduced in Section 4.2. For each dataset, we randomly select $m = 512$ sample indices with the fixed experimental seed. On Cifar100, MVCL-KS achieves 99.97% ACC while requiring only 17.96 seconds. Among the competing multi-view methods executed on the same GPU, the fastest is RMCNC at 102.43 seconds, so MVCL-KS is approximately $5.7\times$ faster. On YouTubeVideo, MVCL-KS remains competitive with MFLVC in clustering quality while requiring only 62.10 seconds, compared with $1{,}370.12$ seconds for the fastest competing multi-view GPU baseline (RMCNC), corresponding to a $22.1\times$ speedup.

6 Conclusion

We introduced MVCL-KS, which recasts bidirectional multi-view InfoNCE as a unified signed block spectral system in heterogeneous RKHS geometries. The resulting self-consistent refinement incorporates a basis-invariant cycle-aware safeguard and scalable low-rank and anchor implementations. Across the main benchmarks, the fixed RBF setting is competitive or leading while being substantially faster than GPU-based baselines; additional results demonstrate compatibility with multiple kernel geometries.

Reproducibility Statement

Reproducibility is supported by the mathematical formulation and fixed stopping criteria in the main paper, the reduced-space procedure in Algorithm 1, the proofs in Appendix E, and the dataset and parameter specifications in Appendix B. The complete code will be released publicly upon acceptance.

AI Use Statement

AI tools were used to aid and polish writing and for retrieval and discovery of related work. All mathematical derivations, experimental results, technical claims, citations, and final manuscript content were reviewed and verified by the authors.

References

Jie Chen, Hua Mao, Wai Lok Woo, and Xi Peng. Deep multiview clustering by contrasting cluster assignments. In *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 16752–16761, 2023.

Man-Sheng Chen, Ling Huang, Chang-Dong Wang, and Dong Huang. Multi-view clustering in latent embedding space. In *Proc. AAAI Conf. Artif. Intell.*, volume 34, pp. 3513–3520, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. Int. Conf. Mach. Learn.*, pp. 1597–1607, 2020b.

Yongyong Chen, Xiaolin Xiao, and Yicong Zhou. Jointly learning kernel representation tensor and affinity matrix for multi-view clustering. *IEEE Trans. Multimedia*, 22(8):1985–1997, 2019.

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

Ran Eisenberg, Jonathan Svirsky, and Ofir Lindenbaum. COPER: Correlation-based permutations for multi-view clustering. In *Proc. Int. Conf. Learn. Represent.*, pp. 10464–10489, 2025.

Hongchang Gao, Feiping Nie, Xuelong Li, and Heng Huang. Multi-view subspace clustering. In *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 4238–4246, 2015.

Ruiming Guo, Mouxing Yang, Yijie Lin, Xi Peng, and Peng Hu. Robust contrastive multi-view clustering against dual noisy correspondence. In *Adv. Neural Inf. Proces. Syst.*, volume 37, pp. 121401–121421, 2024.

Hongqing He, Jie Xu, Wenyuan Yang, Yonghua Zhu, Guoqiu Wen, and Xiaofeng Zhu. Global-graph guided and local-graph weighted contrastive learning for unified clustering on incomplete and noise multi-view data. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 24813–24822, 2026.

Shudong Huang, Ivor W Tsang, Zenglin Xu, and Jiancheng Lv. Latent representation guided multi-view clustering. *IEEE Trans. Knowl. Data Eng.*, 2023.

Zhenyu Huang, Joey Tianyi Zhou, Xi Peng, Changqing Zhang, Hongyuan Zhu, and Jiancheng Lv. Multi-view spectral clustering network. In *Proc. Int. Joint Conf. Artif. Intell.*, volume 2, pp. 4, 2019.

Zhenglai Li, Jun Wang, Chang Tang, Xinzhong Zhu, Wei Zhang, and Xinwang Liu. Balanced multi-view clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2026.

Jiyuan Liu, Xinwang Liu, Yuexiang Yang, Qing Liao, and Yuanqing Xia. Contrastive multi-view kernel learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(8):9552–9566, 2023.

Zhengzheng Lou, Hang Xue, Yanzheng Wang, Chaoyang Zhang, Xin Yang, and Shizhe Hu. Parameter-free deep multi-modal clustering with reliable contrastive learning. *IEEE Trans. Image Process.*, 2025.

Yiding Lu, Yijie Lin, Mouxing Yang, Dezhong Peng, Peng Hu, and Xi Peng. Decoupled contrastive multi-view clustering with high-order random walks. In *Proc. AAAI Conf. Artif. Intell.*, volume 38, pp. 14193–14201, 2024.

Ryumei Nakada, Halil Ibrahim Gulluk, Zhun Deng, Wenlong Ji, James Zou, and Linjun Zhang. Understanding multimodal contrastive learning and incorporating unpaired data. In *ACM Int. Conf. Artif. Intell. Stat.*, volume 206, pp. 4348–4380, 2023.

Erlin Pan and Zhao Kang. Multi-view contrastive graph clustering. In *Adv. Neural Inf. Proces. Syst.*, volume 34, pp. 2148–2159, 2021.

Uri Shaham, Kelly Stanton, Henry Li, Boaz Nadler, Ronen Basri, and Yuval Kluger. Spectralnet: Spectral clustering using deep neural networks. In *Proc. Int. Conf. Learn. Represent.*, pp. 1668–1687, 2018.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

Peng Su, Yixi Liu, Shujian Li, Shudong Huang, and Jiancheng Lv. Robust contrastive multi-view kernel clustering. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 4938–4945, 2024.

Hangke Sui, Yuqing Wang, and Minh Do. UniCon: Unified framework for efficient contrastive alignment via kernels. In *Proc. Int. Conf. Learn. Represent.*, volume 2026, pp. 85142–85174, 2026.

Yuan Sun, Yang Qin, Yongxiang Li, Dezhong Peng, Xi Peng, and Peng Hu. Robust multi-view clustering with noisy correspondence. *IEEE Trans. Knowl. Data Eng.*, 2024.

Yuan Sun, Yongxiang Li, Zhenwen Ren, Guiduo Duan, Dezhong Peng, and Peng Hu. Roll: Robust noisy pseudo-label learning for multi-view clustering with noisy correspondence. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 30732–30741, 2025.

Zhiquan Tan, Yifan Zhang, Jingqin Yang, and Yang Yuan. Contrastive learning is spectral clustering on similarity graph. In *Proc. Int. Conf. Learn. Represent.*, pp. 7190–7206, 2024.

Xinhang Wan, Jiyuan Liu, Xinbiao Gan, Xinwang Liu, Siwei Wang, Yi Wen, Tianjiao Wan, and En Zhu. One-step multi-view clustering with diverse representation. *IEEE Trans. Neural Netw. Learn. Syst.*, 2024.

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

Wei Xia, Quanxue Gao, Qianqian Wang, Xinbo Gao, C. Ding, and Dacheng Tao. Tensorized bipartite graph learning for multi-view clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45: 5187–5202, 2022.

Yuan Xie, Jinyan Liu, Yanyun Qu, Dacheng Tao, Wensheng Zhang, Longquan Dai, and Lizhuang Ma. Robust kernelized multiview self-representation for subspace clustering. *IEEE Trans. Neural Netw. Learn. Syst.*, 32(2):868–881, 2020.

Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 16051–16060, 2022.

Xuejiao Yu, Yi Jiang, Guoqing Chao, and Dianhui Chu. Deep contrastive multi-view subspace clustering with representation and cluster interactive learning. *IEEE Trans. Knowl. Data Eng.*, 2024.

Changqing Zhang, Qinghua Hu, Huazhu Fu, Pengfei Zhu, and Xiaochun Cao. Latent multi-view subspace clustering. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 4279–4287, 2017.

Pei Zhang, Xinwang Liu, Jian Xiong, Sihang Zhou, Wentao Zhao, En Zhu, and Zhiping Cai. Consensus one-step multi-view subspace clustering. *IEEE Trans. Knowl. Data Eng.*, 34(10):4676–4689, 2020.

Yi-Ge Zhang, Jingyi Cui, Qiran Li, and Yisen Wang. Difficult examples hurt unsupervised contrastive learning: A theoretical perspective. In *Proc. Int. Conf. Learn. Represent.*, volume 2026, pp. 101184–101214, 2026.

Yang Zhao, Daidai Zhu, Aihong Yuan, and Xuelong Li. Consensus multi-view spectral clustering network with unified similarity. *Neural Netw.*, pp. 107647, 2025.

A Related Work

Traditional multi-view clustering exploits cross-view consistency through subspace representations, graph fusion, spectral embeddings, and matrix/tensor models. MVSC couples view-specific self-representations through a common spectral partition (Gao et al., 2015), LMSC maps heterogeneous views into a shared latent space before self-representation (Zhang et al., 2017), and tensorized methods model higher-order dependencies across views (Xia et al., 2022). Spectral and graph-based approaches further learn or fuse multiple affinities, often with partition fusion, spectral rotation, or anchor approximations (Zhang et al., 2020; Chen et al., 2020a), while LPGMVC and OMVCDR integrate latent or diverse representations with graph/spectral structures (Huang et al., 2023; Wan et al., 2024). Kernelized methods instead model nonlinear similarities or self-representations in RKHSs (Chen et al., 2019; Xie et al., 2020). These methods establish the value of spectral and kernel geometry, but the graph or kernel is generally introduced as an explicit clustering component rather than derived from the contrastive objective itself.

Contrastive learning has become a major mechanism for multi-view alignment. MCGC constructs graph-based positive relations in a shallow contrastive framework (Pan & Kang, 2021), whereas MFLVC and CVCL learn deep view-specific representations and impose contrastive consistency at the feature or cluster level (Xu et al., 2022; Chen et al., 2023). DCMVSC combines deep representation learning, self-expression, and clustering interaction (Yu et al., 2024), while DIVIDE, CANDY, RMCNC, and ROLL improve robustness to false relations, noisy correspondence, or noisy pseudo-labels (Lu et al., 2024; Guo et al., 2024; Sun et al., 2024; 2025). Self-weighted and reliability-aware methods further address heterogeneous view quality (Wu et al., 2024; Lou et al., 2025). Contrastive multi-view kernel learning also combines kernels with contrastive objectives, as in CMK and R-CMK (Liu et al., 2023; Su et al., 2024). In contrast, MVCL-KS keeps the instance-level contrastive target simple and asks which spectral operator the objective itself induces before introducing kernel geometry.

The closest theoretical works study contrastive learning through spectral or algebraic structure. Tan et al. show that contrastive learning can be interpreted through spectral clustering on a similarity graph (Tan et al., 2024), Nakada et al. connect multimodal contrastive optimization with cross-covariance and SVD structure (Nakada et al., 2023), and UniCon develops a kernel-based gradient-equivalence perspective for efficient contrastive alignment (Sui et al., 2026). MVCL-KS builds on this direction but asks a complementary multi-view question: it derives pairwise signed relations from the exact contrastive gradient, assembles all view pairs into one symmetric block spectral system, and then realizes that representation-dependent system in heterogeneous RKHS geometries. This distinction leads directly to the nonlinear self-consistent refinement studied in the following sections.

B Algorithm Analysis and Experimental Configuration

B.1 Algorithm Analysis

Algorithm 1 summarizes the common steps of exact and anchor modes. All iterative updates are performed in the Vr-dimensional reduced space. Exact mode lifts the final reduced eigenspace to Z, whereas anchor mode follows the implementation and extends the RKHS representation F using the Nyström map in Appendix E.10.

Complexity. Let $n = N$ in exact mode and $n = m$ in anchor mode. Dense kernel construction costs $O(n^2 \sum_v d_v)$. The current implementation uses dense symmetric eigendecomposition before retaining rank r, which costs $O(Vn^3)$; replacing it with a partial eigensolver would reduce this stage to approximately $O(Vn^2 r)$. At each refinement, forming all cross-view logits and applying the row-wise softmax costs $O(V^2 n^2 C)$. A direct evaluation of the reduced blocks $U_r^{(u)\top} S_t^{(uv)} U_r^{(v)}$ costs $O(V^2(n^2 r + nr^2))$, while extracting the leading C eigenvectors of the $Vr \times Vr$ reduced operator costs $O((Vr)^2 C)$ with an iterative eigensolver. The total refinement cost is therefore

$$O\big(T\left[V^2 n^2 (C + r) + V^2 n r^2 + (Vr)^2 C\right]\big). \tag{46}$$

In anchor mode, $n = m \ll N$, and extending F to all samples costs $O(Nm \sum_v d_v + VNmC + VmrC)$ for dense sample-to-anchor kernels, coefficient construction, and RKHS eval-

Algorithm 1 MVCL-KS with cycle-aware spectral refinement

Require: Multi-view data $\{X^{(v)}\}_{v=1}^{V}$, cluster numbers C, temperature τ, retained rank r
Ensure: Cluster labels $\mathbf{y}$

Select the training set: all N samples (exact) or m anchors (anchor)
for $v = 1, \ldots, V$ **do**
    Build and normalize $K^{(v)}$ via equation 15
    Retain rank-r factors $U_r^{(v)}, \Lambda_r^{(v)}$ via equation 40
end for
Form $\mathcal{U}$ and $\mathbf{\Lambda}$ via equation 41
Initialize $Q_{\text{init}}^{(uv)}$ and $S_{\text{init}}^{(uv)}$ via equation 30–equation 31
Form $\widehat{\mathcal{M}}_{\text{init}}$ from S_{init}; set $W_0 \leftarrow \text{Eig}_C(\widehat{\mathcal{M}}_{\text{init}})$
for $t = 0, \ldots, T-1$ **do**
    Compute $F_t \leftarrow \mathcal{U}\mathbf{\Lambda}^{1/2} W_t$ via equation 44
    Update $Q_t^{(uv)}$ and $S_t^{(uv)}$ via equation 26–equation 27
    Form $\widehat{\mathcal{M}}_t$ by equation 42; set $W_{t+1} \leftarrow \text{Eig}_C(\widehat{\mathcal{M}}_t)$
    Compute $d_1 = d(W_{t+1}, W_t)$ and, if available, $d_2 = d(W_{t+1}, W_{t-1})$
    if $d_1 \le \varepsilon$ **then**
        $W_{\text{out}} \leftarrow W_{t+1}$; **break**
    else if $t \ge 1$, $d_2 \le \varepsilon_{\text{cyc}}$, and $d_1 > \varepsilon_{\text{sep}}$ **then**
        $W_{\text{out}} \leftarrow \text{Eig}_C\big((W_{t+1}W_{t+1}^\top + W_t W_t^\top)/2\big)$; **break**
    end if
end for
if no earlier stopping condition is met **then**
    $W_{\text{out}} \leftarrow W_T$
end if
if exact mode **then**
    Set $R_{\text{out}} = Z_{\text{out}} \leftarrow \mathcal{U} W_{\text{out}}$
else
    Split W_{out} into $\{W^{(v)}\}_{v=1}^{V}$ and extend $R^{(v)} = F^{(v)}$ via equation 45
end if
Average $\{R^{(v)}\}_{v=1}^{V}$ by equation 38; row-normalize; apply K-means
return $\mathbf{y}$

uation. Chunking avoids storing all $V(V-1)$ probability matrices simultaneously. The training kernels, reduced factors, embeddings, and reduced operator require $O(Vn^2 + Vnr + VnC + V^2r^2)$ memory. The low-rank perturbation bound and the anchor out-of-sample formula are given in Appendices E.9 and E.10, respectively.

B.2 Experimental Setup

Datasets. We evaluate MVCL-KS on eight multi-view benchmarks covering text, object, face, and video data. Six datasets—3Sources, BBCSport, Caltech, 100leaves, NUS-WIDE, and Notting-Hill—are used for the main comparison, while Cifar100 and YouTubeVideo are used to further evaluate scalability. The dataset statistics are summarized in Table 5. The benchmark sizes range from 169 to 101,499 samples, with 2–6 views and up to 100 clusters.

Compared methods. We compare against a diverse set of spectral, shallow multi-view, kernel-contrastive, and deep contrastive clustering approaches. SPC denotes classical normalized-cut spectral clustering (Shi & Malik, 2000). LPGMVC and OMVCDR are representative shallow multi-view methods based on latent/adaptive graph learning and one-step diverse representation learning, respectively (Huang et al., 2023; Wan et al., 2024). CMK is the closest kernel-contrastive baseline and learns a contrastive multi-view kernel representation (Liu et al., 2023). The deep contrastive baselines include MFLVC (Xu et al., 2022) and CVCL (Chen et al., 2023); DCMVSC combines deep representation learning, self-expression, and clustering interaction (Yu et al., 2024); RMCNC focuses on robust cross-view correspondence (Sun et al., 2024); ROLL performs robust pseudo-label refinement (Sun et al., 2025); and COPER is an ICLR multi-view clustering method based on

Table 5: Statistics of the datasets used in the experiments.

Dataset	Type	N	C	V	Dimensionality per view
3Sources	Text	169	6	3	3560 / 3631 / 3068
BBCSport	Text	544	5	2	3183 / 3203
Caltech	Object	1,400	7	6	48 / 40 / 254 / 1984 / 512 / 928
100leaves	Object	1,600	100	3	64 / 64 / 64
NUS-WIDE	Object	1,600	8	6	64 / 144 / 73 / 128 / 225 / 500
Notting-Hill	Face	4,660	5	3	6750 / 3304 / 2000
Cifar100	Object	50,000	100	3	512 / 2048 / 1024
YouTubeVideo	Video	101,499	31	3	512 / 647 / 838

correlation-driven permutations (Eisenberg et al., 2025). We also include BMvC (Li et al., 2026) in the numerical comparison. On the large-scale benchmarks, we additionally include CMSCU (Zhao et al., 2025) and DMVC-CSG (Cui et al., 2026).

Evaluation metrics. We report clustering accuracy (ACC), normalized mutual information (NMI), adjusted Rand index (ARI), and pairwise F-score (F1). All metrics are reported as percentages. Since runtime is a central practical advantage of MVCL-KS, we also report end-to-end wall-clock time in seconds.

Implementation details. All GPU-based methods included in the runtime comparison were executed in the same hardware environment. We use the RBF kernel family throughout the six-dataset primary comparison because it is positive semidefinite, broadly applicable to heterogeneous feature views, and efficient under dense matrix operations. Here, "fixed RBF" refers to the kernel family rather than to identical numerical hyperparameters: the dataset-specific temperatures and retained ranks are reported in Table 6. Table 7 separately evaluates alternative kernel families and is not used to replace the primary RBF results. We use the dot-product logits in equation 3 without additional feature normalization. Runtime comparisons include only methods executed in the same GPU environment; LPGMVC and OMVCDR are MATLAB implementations and are therefore excluded from runtime rankings and speedup calculations. SPC is retained as a classical single-view baseline and is not used in the GPU runtime comparison.

B.3 Dataset-Specific Parameters

Table 6 reports the kernel family, temperature, and retained kernel rank used in the reported experiments. The six main benchmarks use exact mode with RBF kernels. The two large-scale datasets use anchor mode with $m = 512$ anchors sampled uniformly without replacement with seed 42; Cifar100 uses a linear kernel, whereas YouTubeVideo uses a cosine kernel. All datasets use $T = 15$, $\varepsilon = 10^{-4}$, $\varepsilon_{\text{cyc}} = 5 \times 10^{-4}$, and $\varepsilon_{\text{sep}} = 10^{-4}$.

Table 6: Dataset-specific MVCL-KS parameters used in the reported experiments.

Dataset	Mode	Kernel	Temperature τ	Kernel rank r
3Sources	exact	RBF	0.2	64
BBCSport	exact	RBF	0.2	64
Caltech	exact	RBF	1.0	256
100leaves	exact	RBF	0.5	512
Notting-Hill	exact	RBF	0.2	256
NUS-WIDE	exact	RBF	0.2	64
Cifar100	anchor	Linear	2.0	128
YouTubeVideo	anchor	Cosine	0.2	256

Table 7: Kernel adaptability of MVCL-KS. Each entry reports ACC (%) for the indicated kernel using the best tested configuration for each kernel in this diagnostic comparison. These results characterize kernel flexibility and are not used to select the primary comparison results in Tables 1–2.

Kernel	3Sources	BBCSport	Caltech	100leaves	Notting-Hill	NUS-WIDE
RBF	65.09	**88.24**	**83.43**	91.38	89.25	39.88
Linear	64.50	**88.24**	82.64	76.25	90.79	38.50
Cosine	**68.05**	79.60	**83.43**	76.81	88.67	37.12
Polynomial	64.50	**88.24**	82.71	90.81	90.64	38.38
Laplacian	63.31	86.21	82.86	**95.38**	**93.43**	**42.69**

C Additional Experimental Analysis

This section provides additional analyses of kernel choice, refinement dynamics, ablation studies, and parameter sensitivity.

C.1 Adaptability to Different Kernels

The main comparison uses RBF as a general-purpose and computationally efficient nonlinear kernel, avoiding dataset-specific kernel selection. We next test whether the proposed spectral formulation is tied to this particular geometry. Table 7 reports ACC obtained with linear, cosine, RBF, polynomial, and Laplacian kernels; their exact definitions and the common kernel normalization procedure are provided in Appendix D.

The preferred kernel varies substantially across datasets. Cosine gives the highest ACC on 3Sources; RBF is among the best choices on BBCSport and Caltech; and Laplacian is particularly effective on 100leaves, Notting-Hill, and NUS-WIDE. For example, the Laplacian kernel reaches 95.38% ACC on 100leaves and 93.43% on Notting-Hill, compared with 91.38% and 89.25%, respectively, for the corresponding RBF settings. This behavior is consistent with the role of the kernel in MVCL-KS: the signed contrastive operator remains unchanged, whereas the kernel determines the nonlinear geometry in which the operator is spectrally realized.

C.2 Spectral Refinement Dynamics and Three Stopping Behaviors

Figure 2 shows $d_1(t) = d(W_t, W_{t-1})$ and $d_2(t) = d(W_t, W_{t-2})$ on the six main datasets. Because $\mathcal{U}^\top \mathcal{U} = I$, these reduced-space distances equal their full-space counterparts. The figure uses three stop markers corresponding to the stopping rules in Section 3.3: a **circle** for convergence with $d_1(t) \le \varepsilon$, a **star** for a two-cycle with $d_2(t) \le \varepsilon_{\mathrm{cyc}}$ and $d_1(t) > \varepsilon_{\mathrm{sep}}$, resolved by projector averaging, and a **square** for reaching the maximum iteration number T. BBCSport reaches the fixed-subspace criterion in two refinements. In contrast, 3Sources and Caltech terminate through the two-cycle rule; their star markers identify the iteration at which the alternating-subspace behavior is resolved by projector averaging. The remaining datasets—100leaves, Notting-Hill, and NUS-WIDE—reach the maximum iteration number and are therefore marked by squares. In particular, 100leaves shows a steadily decreasing two-step distance while its one-step distance remains nonzero, which is consistent with an emerging period-two regime even though the cycle threshold is not reached within the allotted iterations. Notting-Hill and NUS-WIDE retain larger one-step distances and exhibit persistent nonlinear oscillation. These observations explain why MVCL-KS uses three distinct stopping outcomes and support our interpretation of the procedure as self-consistent spectral refinement rather than a globally contractive optimizer.

Additional analyses. Appendix C reports three complementary analyses that are useful for interpreting MVCL-KS but are not required for the main comparison: adaptability to different kernel families, spectral refinement dynamics under the three stopping outcomes, and sensitivity to kernel rank and temperature. These results confirm that the method is not tied to one kernel family and clarify how the refinement behaves across datasets and parameter settings.

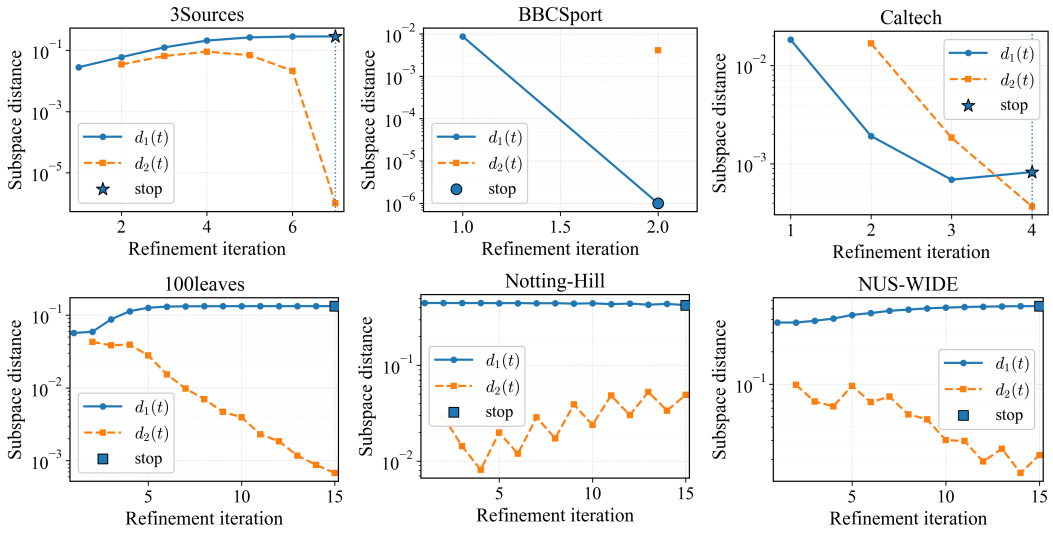

Figure 2: Spectral refinement dynamics of MVCL-KS on six datasets. Blue solid curves show the one-step subspace distance $d_1(t)$, and orange dashed curves show the two-step distance $d_2(t)$. Stop markers correspond to the three termination rules: a **circle** marks convergence with $d_1(t) \leq \varepsilon$, a **star** marks a detected two-cycle with $d_2(t) \leq \varepsilon_{\text{cyc}}$ and $d_1(t) > \varepsilon_{\text{sep}}$, which is resolved by projector averaging, and a **square** marks reaching the maximum number of iterations T.

Table 8: Ablation study of MVCL-KS using ACC (%). Each variant removes or disables a specific component of the formulation while keeping the same RBF kernel setting.

Variant	3Sources	BBCSport	Caltech	100leaves	Notting-Hill	NUS-WIDE
w/o signed contrastive operator	62.13	84.38	69.07	81.12	86.18	32.62
w/o self-consistent updates	64.50	88.05	82.36	90.00	87.51	38.50
MVCL-KS (full)	65.09	88.24	83.43	91.38	89.25	39.88

C.3 Ablation of the Signed Operator and Self-Consistent Updates

Table 8 evaluates the two structural components removed from the full model. **w/o signed contrastive operator** removes equation 7 and the resulting block operator in equation 10, leaving only the fixed multi-view kernel geometry for spectral clustering. **w/o self-consistent updates** keeps the signed initialization in equation 30–equation 31 but disables the representation-dependent updates in equation 26–equation 29. **MVCL-KS (full)** retains both components together with the cycle-aware stopping rule. The signed operator improves all six datasets over the fixed-kernel variant, and self-consistent updates provide a further gain. The effect of cycle-aware stopping is examined separately through the refinement dynamics in Appendix C, because it changes the numerical termination behavior rather than the contrastive objective.

C.4 Hyperparameter Sensitivity: Kernel Rank and Temperature

Kernel rank and contrastive temperature control complementary aspects of MVCL-KS. The retained rank determines the capacity of the effective RKHS subspace and the size of the reduced spectral system, whereas the temperature controls the concentration of cross-view contrastive probabilities. Figure 3 varies each factor independently while keeping the remaining configuration fixed. Across the tested kernel ranks, BBCSport, Caltech, and Notting-Hill are comparatively stable, whereas 100leaves benefits substantially from a richer retained kernel subspace; 3Sources and NUS-WIDE are mildly better at lower ranks. Across the tested temperatures, most datasets remain stable over a broad interval, while Caltech is more sensitive to an overly small temperature and improves

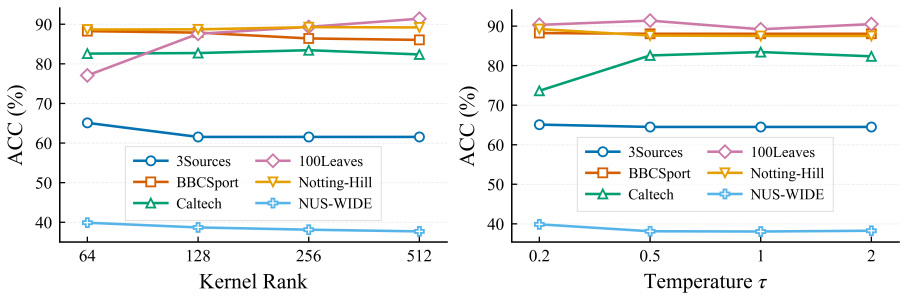

Figure 3: Sensitivity of MVCL-KS to kernel rank (left) and temperature (right) on the six main datasets. Each factor is varied independently while all other settings are held fixed.

markedly once the contrastive distribution is less sharp. These trends indicate that MVCL-KS does not rely on a single narrow operating point.

The two sensitivity analyses complement the kernel-family study in Section C.1: kernel choice changes the geometry supplied to the formulation, kernel rank controls how much of that geometry is retained computationally, and temperature controls the sharpness of contrastive feedback.

D Kernel Definitions Used in the Adaptability Study

For completeness, we summarize the five kernels used in Table 7. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{d_v}$ be two samples from view v. The linear kernel is

$$k_{\mathrm{lin}}(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}. \tag{47}$$

The cosine kernel is

$$k_{\cos}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^\top \mathbf{y}}{\max(\|\mathbf{x}\|_2 \|\mathbf{y}\|_2, \delta_{\mathrm{num}})}, \qquad \delta_{\mathrm{num}} > 0. \tag{48}$$

The radial basis function (RBF) kernel is

$$k_{\mathrm{rbf}}(\mathbf{x}, \mathbf{y}) = \exp\left(-\gamma \|\mathbf{x} - \mathbf{y}\|_2^2\right). \tag{49}$$

The polynomial kernel is

$$k_{\mathrm{poly}}(\mathbf{x}, \mathbf{y}) = \left(\gamma \mathbf{x}^\top \mathbf{y} + c_0\right)^p. \tag{50}$$

The Laplacian kernel is

$$k_{\mathrm{lap}}(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|_1). \tag{51}$$

For RBF, polynomial, and Laplacian kernels, when γ is not explicitly specified, we use the deterministic view-wise default $\gamma = 1/d_v$. For the polynomial kernel, $c_0 \geq 0$ is the offset and $p \geq 1$ is an integer degree. After construction, every view-specific Gram matrix is centered in sample space and trace-normalized before entering the spectral formulation; thus, the comparison changes only the kernel geometry while retaining the remainder of the MVCL-KS pipeline.

E Proofs and Additional Theoretical Analysis

This appendix provides the proofs of the propositions and theorems developed in Sections 2–4, together with several auxiliary results that clarify the initialization, cycle-aware stabilization, low-rank reduction, and anchor extension used by the implementation. Unless stated otherwise, all kernels and operators are real-valued, and eigenspaces are selected by descending *algebraic* eigenvalue.

E.1 Proof of Proposition 1

We first derive the gradient of the directional loss $\mathcal{L}_{u \to v}$ with respect to $F^{(u)}$. For a fixed anchor i, define

$$\ell_i^{(u \to v)} = -\log q_{ii}^{(uv)} = -\frac{1}{\tau} \mathbf{f}_i^{(u)\top} \mathbf{f}_i^{(v)} + \log \sum_{j=1}^{N} \exp\left(\frac{\mathbf{f}_i^{(u)\top} \mathbf{f}_j^{(v)}}{\tau}\right). \tag{52}$$

Differentiating with respect to the anchor representation $\mathbf{f}_i^{(u)}$ gives

$$\frac{\partial \ell_i^{(u\to v)}}{\partial \mathbf{f}_i^{(u)}} = -\frac{1}{\tau}\mathbf{f}_i^{(v)} + \frac{1}{\tau}\sum_{j=1}^{N} q_{ij}^{(uv)}\mathbf{f}_j^{(v)} = \frac{1}{\tau}\left(\sum_{j=1}^{N} q_{ij}^{(uv)}\mathbf{f}_j^{(v)} - \mathbf{f}_i^{(v)}\right). \tag{53}$$

Stacking all anchors and including the factor $1/N$ in equation 5 yields

$$\frac{\partial \mathcal{L}_{u\to v}}{\partial F^{(u)}} = \frac{1}{N\tau}\left(Q^{(uv)} - I_N\right)F^{(v)}. \tag{54}$$

The reverse loss $\mathcal{L}_{v\to u}$ contains $F^{(u)}$ as the *candidate* representation. Differentiating all anchors in view v with respect to the u-view candidates gives

$$\frac{\partial \mathcal{L}_{v\to u}}{\partial F^{(u)}} = \frac{1}{N\tau}\left(Q^{(vu)\top} - I_N\right)F^{(v)}. \tag{55}$$

Using the symmetric loss $\mathcal{L}_{uv} = \frac{1}{2}(\mathcal{L}_{u\to v} + \mathcal{L}_{v\to u})$, we therefore obtain

$$\frac{\partial \mathcal{L}_{uv}}{\partial F^{(u)}} = \frac{1}{2N\tau}\left(Q^{(uv)} + Q^{(vu)\top} - 2I_N\right)F^{(v)} = -\frac{1}{2N\tau}S^{(uv)}F^{(v)}. \tag{56}$$

The expression for $F^{(v)}$ follows analogously:

$$\frac{\partial \mathcal{L}_{uv}}{\partial F^{(v)}} = -\frac{1}{2N\tau}S^{(uv)\top}F^{(u)}. \tag{57}$$

This proves equation 8. □

Block form. Summing equation 56 over all pairs incident to each view and stacking the resulting gradients gives

$$-\nabla_F \mathcal{L}_{\mathrm{MVC}} = \frac{1}{2N\tau}\mathcal{S}(F)F, \tag{58}$$

which establishes equation 11. Importantly, equation 58 is an *exact gradient identity* for dot-product logits; $\mathcal{S}(F)$ is shorthand for the representation-dependent gradient operator and is not treated as an independent optimization variable.

E.2 Proof of Theorem 2

Because $\mathcal{S}$ is symmetric, let

$$\mathcal{S} = U\Lambda U^\top, \qquad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_{VN}), \qquad \lambda_1 \geq \cdots \geq \lambda_{VN}, \tag{59}$$

where U is orthogonal. For any feasible Z with $Z^\top Z = I_C$, write $Y = U^\top Z$. Then $Y^\top Y = I_C$, and

$$\mathrm{Tr}(Z^\top \mathcal{S} Z) = \mathrm{Tr}(Y^\top \Lambda Y) = \sum_{i=1}^{VN} \lambda_i \|Y_{i,:}\|_2^2. \tag{60}$$

Define $\alpha_i = \|Y_{i,:}\|_2^2$. Because Y has C orthonormal columns,

$$0 \leq \alpha_i \leq 1, \qquad \sum_{i=1}^{VN} \alpha_i = C. \tag{61}$$

The linear expression in equation 60 is maximized by placing unit mass on the C largest eigenvalues, giving

$$\mathrm{Tr}(Z^\top \mathcal{S} Z) \leq \sum_{c=1}^{C} \lambda_c. \tag{62}$$

Equality is attained by taking $Z = U_C R$, where $U_C = [\mathbf{u}_1, \ldots, \mathbf{u}_C]$ and R is any $C \times C$ orthogonal matrix. Conversely, equality requires the column space of Z to lie in a maximizing C-dimensional invariant subspace. If $\lambda_C > \lambda_{C+1}$, the maximizing invariant subspace is exactly $\mathrm{span}(U_C)$, hence it is unique although its orthonormal basis is not. This proves Theorem 2. □

E.3 Proof of Theorem 3

We first consider $\mathcal{K} \succ 0$. The RKHS coefficient formulation associated with the signed operator is

$$\max_A \mathrm{Tr}\left(A^\top \mathcal{K}\mathcal{S}\mathcal{K} A\right) \quad \text{s.t.} \quad A^\top \mathcal{K} A = I_C. \tag{63}$$

Introduce $Z = \mathcal{K}^{1/2} A$, or equivalently $A = \mathcal{K}^{-1/2} Z$. The constraint becomes

$$A^\top \mathcal{K} A = Z^\top Z = I_C, \tag{64}$$

and the objective becomes

$$\begin{aligned} \mathrm{Tr}\left(A^\top \mathcal{K}\mathcal{S}\mathcal{K} A\right) &= \mathrm{Tr}\left(Z^\top \mathcal{K}^{1/2}\mathcal{S}\mathcal{K}^{1/2} Z\right) \\ &= \mathrm{Tr}\left(Z^\top \mathcal{M} Z\right). \end{aligned} \tag{65}$$

Thus equation 63 is exactly equivalent to equation 19. Theorem 2, applied to the symmetric matrix $\mathcal{M}$, then implies that

$$Z^\star = U_C R, \qquad R^\top R = I_C, \tag{66}$$

where U_C is the dominant algebraic eigenspace of $\mathcal{M}$. Transforming back gives

$$A^\star = \mathcal{K}^{-1/2} U_C R, \qquad F^\star = \mathcal{K} A^\star = \mathcal{K}^{1/2} U_C R, \tag{67}$$

which is equation 21.

If $\mathcal{K}$ is positive semidefinite but singular, let $\mathcal{K} = U_+ \Lambda_+ U_+^\top$ be its eigendecomposition restricted to strictly positive eigenvalues. All feasible RKHS coordinates can be restricted without loss to $\mathrm{range}(\mathcal{K})$, and the same argument applies with $\mathcal{K}^{-1/2} = U_+ \Lambda_+^{-1/2} U_+^\top$ interpreted as the Moore–Penrose inverse square root on that effective subspace. □

E.4 Proof of Proposition 4

Let

$$\Phi(Z) = \mathcal{L}_{\mathrm{MVC}}(F), \qquad F = \mathcal{K}^{1/2} Z. \tag{68}$$

By the chain rule and equation 11,

$$\begin{aligned} \nabla_Z \Phi(Z) &= \mathcal{K}^{1/2} \nabla_F \mathcal{L}_{\mathrm{MVC}} \\ &= -\frac{1}{2N\tau} \mathcal{K}^{1/2} \mathcal{S}(F) F \\ &= -\frac{1}{2N\tau} \mathcal{K}^{1/2} \mathcal{S}(Z) \mathcal{K}^{1/2} Z \\ &= -\frac{1}{2N\tau} \mathcal{M}(Z) Z. \end{aligned} \tag{69}$$

The dependence of $\mathcal{S}$ on Z does not require an additional "derivative of $\mathcal{S}$" term in equation 69: $\mathcal{S}(Z)$ is the exact representation of the already-differentiated InfoNCE gradient established in Proposition 1.

Consider the orthogonality constraint $Z^\top Z = I_C$. The Lagrangian can be written as

$$\mathscr{L}(Z, \Gamma) = \Phi(Z) + \frac{1}{2} \mathrm{Tr}\left[\Gamma\left(Z^\top Z - I_C\right)\right], \tag{70}$$

where $\Gamma = \Gamma^\top$. First-order stationarity gives

$$-\frac{1}{2N\tau}\mathcal{M}(Z^\star)Z^\star + Z^\star\Gamma = 0. \tag{71}$$

Multiplying by $2N\tau$ and setting $\Lambda = 2N\tau\,\Gamma$ gives

$$\mathcal{M}(Z^\star)Z^\star = Z^\star\Lambda, \qquad \Lambda = \Lambda^\top, \tag{72}$$

which proves the first statement.

Now suppose $Z^\star$ is a fixed point whose span equals the dominant C-dimensional eigenspace of $\mathcal{M}(Z^\star)$. Because this span is invariant,

$$\mathcal{M}(Z^\star)Z^\star = Z^\star\Lambda_C, \qquad \Lambda_C = {Z^\star}^\top\mathcal{M}(Z^\star)Z^\star = \Lambda_C^\top. \tag{73}$$

Equation 72 is therefore satisfied, so the fixed point is first-order stationary. The converse need not hold: a stationary invariant subspace need not be the dominant eigenspace. This is why the method is described as a self-consistent *spectral refinement*, rather than a general-purpose monotone optimizer of InfoNCE. □

E.5 Additional Result: Structure of the Uniform Initialization

The parameter-free initialization in equation 30–equation 31 admits a simple spectral interpretation.

Proposition 5 (Centered-kernel form of the initial operator)**.** *Assume the theoretical kernels are exactly centered, so that $K^{(v)}\mathbf{1} = 0$ for every view v. Then $K^{(v)1/2}\mathbf{1} = 0$, and the (u, v)-th kernelized cross-view block at initialization is*

$$M_0^{(uv)} = 2K^{(u)1/2}K^{(v)1/2}. \tag{74}$$

Proof. Because $K^{(v)} \succeq 0$ and $K^{(v)}\mathbf{1} = 0$, the vector $\mathbf{1}$ lies in the null space of $K^{(v)}$. The positive semidefinite square root has the same null space, hence $K^{(v)1/2}\mathbf{1} = 0$. Using equation 31,

$$\begin{aligned} M_0^{(uv)} &= K^{(u)1/2}\left(2I_N - \frac{2}{N}\mathbf{1}\mathbf{1}^\top\right)K^{(v)1/2} \\ &= 2K^{(u)1/2}K^{(v)1/2} - \frac{2}{N}\left(K^{(u)1/2}\mathbf{1}\right)\left(K^{(v)1/2}\mathbf{1}\right)^\top \\ &= 2K^{(u)1/2}K^{(v)1/2}. \end{aligned} \tag{75}$$

□

Thus the first refinement is a purely kernel-spectral cross-view alignment. A tiny diagonal numerical ridge, if used in implementation, perturbs this identity only for numerical stabilization and is not part of the exact centered-kernel statement.

E.6 Additional Result: Two-View SVD Correspondence

Proposition 6 (Bipartite eigenspace–SVD correspondence)**.** *Let*

$$\mathcal{B} = \begin{bmatrix} 0 & C \\ C^\top & 0 \end{bmatrix}, \tag{76}$$

and let $C = U\Sigma V^\top$ be a singular value decomposition. For every singular triplet $(\mathbf{u}_i, \sigma_i, \mathbf{v}_i)$,

$$\frac{1}{\sqrt{2}}\begin{bmatrix}\mathbf{u}_i \\ \mathbf{v}_i\end{bmatrix} \quad \textit{and} \quad \frac{1}{\sqrt{2}}\begin{bmatrix}\mathbf{u}_i \\ -\mathbf{v}_i\end{bmatrix} \tag{77}$$

are eigenvectors of $\mathcal{B}$ with eigenvalues σ_i and $-\sigma_i$, respectively.

Proof. Using $C\mathbf{v}_i = \sigma_i\mathbf{u}_i$ and $C^\top\mathbf{u}_i = \sigma_i\mathbf{v}_i$,

$$\mathcal{B}\begin{bmatrix}\mathbf{u}_i \\ \mathbf{v}_i\end{bmatrix} = \begin{bmatrix}C\mathbf{v}_i \\ C^\top\mathbf{u}_i\end{bmatrix} = \sigma_i\begin{bmatrix}\mathbf{u}_i \\ \mathbf{v}_i\end{bmatrix}. \tag{78}$$

Replacing $\mathbf{v}_i$ by $-\mathbf{v}_i$ gives the negative eigenvalue. □

Applying Proposition 6 to $C = C^{(12)}$ in equation 22 proves the special-case statement in Section 3.1.

E.7 Additional Result: Cycle-Aware Projector Averaging

The stabilization in equation 35 is not an arbitrary average of two bases. It is naturally defined on subspaces.

Proposition 7 (Basis invariance and chordal mean)**.** *Let $Z_a, Z_b \in \mathbb{R}^{m\times C}$ have orthonormal columns and define $P_a = Z_a Z_a^\top$, $P_b = Z_b Z_b^\top$, and $\bar{P} = (P_a + P_b)/2$. Then:*

1. *$P_a, P_b, \bar{P}$ are invariant to arbitrary right orthogonal changes of basis $Z_a \mapsto Z_a R_a$, $Z_b \mapsto Z_b R_b$;*
2. *the rank-C projector $P_\star = Z_\star Z_\star^\top$ whose range is the dominant C-dimensional eigenspace of $\bar{P}$ solves*
$$\min_{\substack{P=P^\top=P^2\\ \mathrm{rank}(P)=C}} \left(\|P - P_a\|_F^2 + \|P - P_b\|_F^2\right); \tag{79}$$
3. *the same dominant subspace can be recovered from a $2C \times 2C$ Gram matrix without explicitly forming $\bar{P}$.*

Proof. For any orthogonal R_a,
$$(Z_a R_a)(Z_a R_a)^\top = Z_a R_a R_a^\top Z_a^\top = P_a,$$
and similarly for P_b, proving basis invariance.

For any rank-C orthogonal projector P, $\|P\|_F^2 = \mathrm{Tr}(P) = C$ and $\|P_a\|_F^2 = \|P_b\|_F^2 = C$. Hence
$$\begin{aligned}\|P - P_a\|_F^2 + \|P - P_b\|_F^2 &= 4C - 2\,\mathrm{Tr}(P(P_a + P_b))\\ &= 4C - 4\,\mathrm{Tr}(P\bar{P}).\end{aligned} \tag{80}$$
Minimizing equation 80 is therefore equivalent to maximizing $\mathrm{Tr}(P\bar{P})$ over rank-C orthogonal projectors. By Theorem 2, the solution is the projector onto the dominant C-dimensional eigenspace of $\bar{P}$.

For the low-dimensional computation, define
$$B = \frac{1}{\sqrt{2}}\,[Z_a \quad Z_b] \in \mathbb{R}^{m\times 2C}. \tag{81}$$
Then $\bar{P} = BB^\top$. The nonzero eigenvalues of $BB^\top$ equal those of $G = B^\top B \in \mathbb{R}^{2C\times 2C}$. If $G = R\Omega R^\top$ and Ω_C contains the largest C positive eigenvalues, then
$$Z_\star = B R_C \Omega_C^{-1/2} \tag{82}$$
has orthonormal columns and spans the dominant eigenspace of $\bar{P}$. Thus only a $2C \times 2C$ eigenproblem is required. □

Relation to the subspace distance. For orthonormal Z_a, Z_b,
$$\|P_a - P_b\|_F^2 = 2C - 2\|Z_a^\top Z_b\|_F^2 = 2C\, d(Z_a, Z_b)^2, \tag{83}$$
so equation 32 is the normalized chordal distance between the corresponding rank-C projectors. This explains why both the two-cycle detector and the projector mean are invariant to arbitrary eigenvector signs and basis rotations.

E.8 Additional Result: Exact Reduced-RKHS Eigensystem

The reduced operator in equation 42 is exact whenever the retained eigendecomposition exactly represents the kernels, and is the natural projected approximation otherwise.

Proposition 8 (Reduced-RKHS equivalence)**.** *Suppose*
$$\mathcal{K} = \mathcal{U}\mathbf{\Lambda}\mathcal{U}^\top, \qquad \mathcal{U}^\top\mathcal{U} = I_R, \tag{84}$$

where $R = \sum_v r_v$ collects all retained positive kernel components. Define

$$\widehat{\mathcal{M}} = \mathbf{\Lambda}^{1/2}\mathcal{U}^\top \mathcal{S}\mathcal{U}\mathbf{\Lambda}^{1/2}. \tag{85}$$

Then

$$\mathcal{M} = \mathcal{U}\widehat{\mathcal{M}}\mathcal{U}^\top. \tag{86}$$

Consequently, every nonzero eigenpair $\widehat{\mathcal{M}}\mathbf{w} = \lambda\mathbf{w}$ induces the eigenpair

$$\mathcal{M}(\mathcal{U}\mathbf{w}) = \lambda(\mathcal{U}\mathbf{w}), \tag{87}$$

and the nonzero spectra coincide.

Proof. From equation 84, $\mathcal{K}^{1/2} = \mathcal{U}\mathbf{\Lambda}^{1/2}\mathcal{U}^\top$ on the effective kernel range. Substitution gives

$$\begin{aligned}\mathcal{M} &= \mathcal{K}^{1/2}\mathcal{S}\mathcal{K}^{1/2} \\ &= \mathcal{U}\mathbf{\Lambda}^{1/2}\mathcal{U}^\top \mathcal{S}\mathcal{U}\mathbf{\Lambda}^{1/2}\mathcal{U}^\top \\ &= \mathcal{U}\widehat{\mathcal{M}}\mathcal{U}^\top.\end{aligned} \tag{88}$$

If $\widehat{\mathcal{M}}\mathbf{w} = \lambda\mathbf{w}$, then

$$\mathcal{M}\mathcal{U}\mathbf{w} = \mathcal{U}\widehat{\mathcal{M}}\mathbf{w} = \lambda\mathcal{U}\mathbf{w}.$$

Since $\mathcal{U}$ has orthonormal columns, the lifting is norm-preserving and establishes the spectral equivalence on the effective range. □

E.9 Spectral Perturbation Bound for Kernel Approximation

The following result quantifies how a same-dimensional low-rank kernel approximation affects the dominant spectral subspace.

Theorem 9 (Subspace stability under operator perturbation)**.** *Let $\mathcal{M}$ and $\widetilde{\mathcal{M}} = \mathcal{M} + E$ be symmetric. Let Z and $\widetilde{Z}$ contain their respective dominant C-dimensional algebraic eigenspaces, and suppose*

$$\delta = \lambda_C(\mathcal{M}) - \lambda_{C+1}(\mathcal{M}) > 0. \tag{89}$$

If $\|E\|_2 = \eta < \delta$, then

$$\|\sin\Theta(\widetilde{Z}, Z)\|_2 \le \frac{\eta}{\delta - \eta}. \tag{90}$$

In particular, if $\eta \le \delta/2$,

$$\|\sin\Theta(\widetilde{Z}, Z)\|_2 \le \frac{2\eta}{\delta}. \tag{91}$$

Proof. Let $[Z, Z_\perp]$ be an orthogonal eigenbasis of $\mathcal{M}$, with

$$\mathcal{M}Z = Z\Lambda_1, \qquad \mathcal{M}Z_\perp = Z_\perp\Lambda_2,$$

where every eigenvalue in Λ_1 is at least $\lambda_C(\mathcal{M})$ and every eigenvalue in Λ_2 is at most $\lambda_{C+1}(\mathcal{M})$. Let $\widetilde{\mathcal{M}}\widetilde{Z} = \widetilde{Z}\widetilde{\Lambda}_1$. Premultiplying by $Z_\perp^\top$ and using $\widetilde{\mathcal{M}} = \mathcal{M} + E$ gives the Sylvester equation

$$\Lambda_2\left(Z_\perp^\top\widetilde{Z}\right) - \left(Z_\perp^\top\widetilde{Z}\right)\widetilde{\Lambda}_1 = -Z_\perp^\top E\widetilde{Z}. \tag{92}$$

By Weyl's inequality,

$$\lambda_C(\widetilde{\mathcal{M}}) \ge \lambda_C(\mathcal{M}) - \eta. \tag{93}$$

Hence the spectral separation between Λ_2 and $\widetilde{\Lambda}_1$ is at least

$$(\lambda_C(\mathcal{M}) - \eta) - \lambda_{C+1}(\mathcal{M}) = \delta - \eta. \tag{94}$$

The standard norm bound for the symmetric Sylvester equation therefore gives

$$\left\|Z_\perp^\top\widetilde{Z}\right\|_2 \le \frac{\|Z_\perp^\top E\widetilde{Z}\|_2}{\delta - \eta} \le \frac{\eta}{\delta - \eta}. \tag{95}$$

Finally, $\|Z_\perp^\top\widetilde{Z}\|_2 = \|\sin\Theta(\widetilde{Z}, Z)\|_2$, which proves equation 90. Equation 91 follows from $\delta - \eta \ge \delta/2$. □

Theorem 9 shows that a low-rank approximation is reliable when its lifted operator error is small relative to the target eigengap. It is a *conditional* stability result: the theorem does not assume that the nonlinear self-consistent iteration is globally contractive. Because anchor mode changes the training dimension from VN to Vm, this theorem does not by itself provide an anchor approximation guarantee; such a result would additionally require an explicit Nyström lifting and a bound on the resulting full-space operator error.

E.10 Anchor/Nyström Extension

For large N, the anchor implementation in Section 4.2 samples $m \ll N$ observations uniformly without replacement as anchors, uses the same anchor indices across all views, and learns the same self-consistent spectral system on that subset. We now give the exact out-of-sample extension for one view and omit the superscript v.

Let the centered and scaled anchor kernel be

$$K_{LL} = U\Lambda U^\top, \tag{96}$$

where $U \in \mathbb{R}^{m\times r}$ contains the retained eigenvectors and $\Lambda \succ 0$ contains the retained positive eigenvalues. Let $W \in \mathbb{R}^{r\times C}$ denote the learned reduced spectral block for that view. On the anchors, the RKHS representation used in anchor mode is

$$F_L = U\Lambda^{1/2}W. \tag{97}$$

Define its RKHS coefficient matrix as

$$A_L = U\Lambda^{-1/2}W. \tag{98}$$

For a new sample x, let $\bar{\mathbf{k}}_{xL} \in \mathbb{R}^{1\times m}$ be its kernel evaluations against the anchors after applying the same centering and scale used for K_{LL}. Its out-of-sample RKHS representation is

$$\mathbf{f}(x) = \bar{\mathbf{k}}_{xL}A_L = \bar{\mathbf{k}}_{xL}U\Lambda^{-1/2}W. \tag{99}$$

Proposition 10 (Consistency on anchor samples)**.** *For an anchor sample x_i, equation 99 exactly recovers the corresponding row of F_L on the retained positive kernel subspace.*

Proof. For an anchor x_i, $\bar{\mathbf{k}}_{x_iL}$ is the i-th row of K_{LL}. Therefore,

$$\begin{aligned}\bar{\mathbf{k}}_{x_iL}U\Lambda^{-1/2}W &= \mathbf{e}_i^\top U\Lambda U^\top U\Lambda^{-1/2}W \\ &= \mathbf{e}_i^\top U\Lambda^{1/2}W,\end{aligned} \tag{100}$$

which is exactly the i-th row of equation 97. □

This extension preserves the learned anchor RKHS functions while avoiding a full $N \times N$ spectral training problem. It also matches the implementation: exact mode clusters the available spectral blocks $Z^{(v)}$, whereas anchor mode clusters the Nyström-extended RKHS blocks $F^{(v)}$. The two modes share the same reduced self-consistent training system but use distinct, explicitly stated representations.