BILEVEL CERTIFIED MULTIPLE KERNEL CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multiple kernel clustering (MKC) integrates heterogeneous data views through dynamic kernel weighting, yet its robustness is typically evaluated only against empirical attacks. Crucially, existing studies overlook weight re-optimization, discrete assignment stability, and numerical approximation errors. To address this, we propose **B**ilevel **C**ertified **MKC** (BC-MKC). **First**, we employ a bilevel optimization framework during training: the lower level computes a self-consistent spectral response via a strongly convex fixed-subspace weight formulation, while the upper level optimizes a smooth proxy—targeting alignment, clustering margin, and eigengap—using guarded implicit differentiation. **Second**, we incorporate a trace-normalized Nyström bank to stabilize the approximation geometry. **Then**, for post-training evaluation, we introduce a fixed-core verifier that propagates admissible base-kernel perturbations through weight re-optimization, kernel fusion, and spectral projection. This verifier either rigorously certifies the full partition and selected co-assignments up to label permutation, or safely abstains. **Finally**, we formally bound fixed-point and eigensolver residuals against the certification margin, seamlessly accommodating uniform full-kernel approximation bounds when available. Extensive experiments on ten benchmarks validate our guarantees, demonstrating state-of-the-art clustering accuracy while reliably certifying partition invariance.

Code: `https://anonymous.4open.science/r/anonymous27`

1 INTRODUCTION

Clustering heterogeneous data requires reconciling multiple similarity structures. Kernel k-means and spectral clustering provide a natural connection between nonlinear geometry and partition learning, while multiple-kernel clustering (MKC) extends this principle by combining candidate kernels that encode complementary views (Dhillon et al., 2004; Ng et al., 2001; von Luxburg, 2007; Gönen & Alpaydın, 2011). Since views can differ substantially in informativeness, modern MKC methods move beyond uniform fusion by learning kernel weights, neighborhood graphs, subspace representations, or partition-level consensus (Liu et al., 2016; 2017; Wang et al., 2017b; Zhou et al., 2020; Liu, 2022; Zhang et al., 2022b).

This adaptive-fusion paradigm has subsequently been extended to hierarchical structure, corruption, imbalance, fairness, and large-scale settings (Wang et al., 2019; Liu et al., 2021a; Zhang et al., 2025; Chen et al., 2025; Zhou et al., 2024; Liang et al., 2024), with further advances based on contrastive refinement and granular-ball representations (Su et al., 2024; Xia et al., 2025). Scalability, however, often requires replacing full kernels with Nyström approximations, core kernels, incremental constructions, or random features (Williams & Seeger, 2000; Musco & Musco, 2017; Feng et al., 2025; Liang et al., 2025; Rahimi & Recht, 2007). Although these approximations reduce computation, the resulting geometry can depend strongly on landmarks, initialization, processing order, and bandwidth, while finite sampling and rank truncation introduce unavoidable approximation errors (Bach, 2013; Rudi et al., 2015; Gittens & Mahoney, 2013). Consequently, scalable adaptive fusion alone does not guarantee stability of the spectral structure ultimately used for clustering.

These limitations introduce three challenges: (1) *configuration approximations inherently destabilize the spectral subspace*; (2) *kernel and solver perturbations unpredictably shift lower-level responses*; and (3) *existing bounds (Huang et al., 2008; Cortes et al., 2010) cannot certify the robustness of discrete assignments*. As illustrated in Fig. 1, our framework overcomes these issues by providing transparent diagnostics for lower-response stability, view screening, and certified assignment coverage.

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.

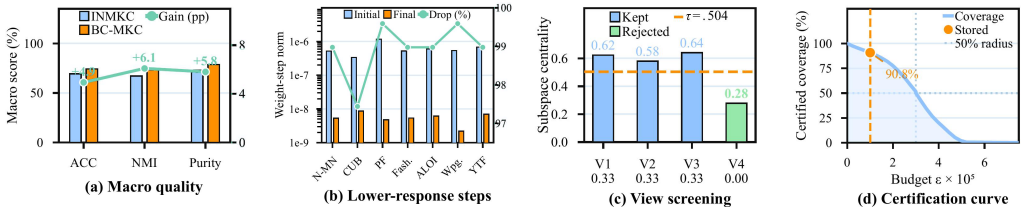


Figure 1: BC-MKC diagnostics. **(a)** Seven-dataset macro scores against INMKC. **(b)** Lower-response step norms and reduction. **(c)** YTF10 view screening. **(d)** YTF10 certified coverage; orange marks $\varepsilon_j = 10^{-5}$ and gray marks the 50% radius. Panels (a)–(c) summarize one saved run.

BC-MKC addresses these gaps through a unified framework. **First**, we stabilize the approximation geometry using trace-normalized Nyström factors and a reliability-weighted projector barycenter. **Second**, we formulate a self-consistent simplex response that robustly guides kernel weighting via guarded numerical implicit updates (Colson et al., 2007; Franceschi et al., 2018). **Finally**, we introduce a separate frozen-core verifier that rigorously propagates kernel perturbations and measured solver residuals to assignment margins using spectral bounds (Davis & Kahan, 1970; Yu et al., 2015). This mechanism allows the framework to explicitly report global, partial, or abstaining certification outcomes for the frozen approximate-kernel computation.

The contributions are:

- **Reliability-stabilized kernel geometry.** We construct a rotation-invariant Nyström kernel bank that aggregates multiple approximation configurations through label-free reliability weighting and trace normalization, reducing configuration-induced geometric variation.
- **Self-consistent bilevel kernel weighting.** We formulate adaptive kernel weighting as a bilevel response problem, coupling a strongly regularized simplex fixed point with a clustering-oriented upper objective and guarded implicit differentiation.
- **Certified assignment stability.** We derive a residual-aware verifier that propagates admissible kernel perturbations through weight re-optimization and spectral projection to discrete clustering margins, certifying global or partial partition invariance while safely abstaining when the conditions are insufficient.

2 Related Work

2.1 Multiple Kernel k-Means Clustering (MKKM)

Kernel k-means minimizes within-cluster scatter in a reproducing kernel Hilbert space and is closely related to spectral clustering (Dhillon et al., 2004). Following multiple kernel learning (Rakotomamonjy et al., 2008; Kloft et al., 2011), MKKM replaces a single similarity matrix with a simplex-weighted combination of M positive-semidefinite base kernels:

$$\min_{\boldsymbol{\beta}\in\Delta_M,\ H^\top H=I_k} \operatorname{Tr}\big[K(\boldsymbol{\beta})(I_n - HH^\top)\big], \qquad K(\boldsymbol{\beta}) = \sum_{j=1}^{M} \beta_j K_j. \tag{1}$$

Here, $H \in \mathbb{R}^{n\times k}$ is a normalized relaxation and $\Delta_M = \{\boldsymbol{\beta} \geq 0 : \mathbf{1}^\top\boldsymbol{\beta} = 1\}$. Spectral co-regularization and low-rank recovery model agreement across views (Kumar et al., 2011; Xia et al., 2014). Full-kernel methods reduce redundancy or regularize weights and partitions (Liu et al., 2016; Liu, 2022; Chen et al., 2025); partition methods align or reconstruct compact partitions (Wang et al., 2019; Liu et al., 2021b; Zhang et al., 2022b; 2025); and low-rank methods use landmarks, anchors, incremental alignment, or core kernels (Musco & Musco, 2017; Liang et al., 2024; Feng et al., 2025; Liang et al., 2025). These approaches address fusion, scale, balance, or empirical resistance (Yu et al., 2025), but rarely combine response re-optimization, numerical residuals, and discrete-assignment verification. Further comparisons appear in Appendix C.

2.2 Bilevel Certified Multiple Kernel Clustering

BC-MKC combines stabilized approximation, bilevel reweighting, and assignment verification. For predetermined configurations, let U_ℓ span the retained Nyström subspace and ω_ℓ be its label-free

weight. The projector barycenter and trace-one meta-kernel are

$$\bar{P} = \sum_{\ell=1}^{Q} \omega_\ell U_\ell U_\ell^\top, \qquad U_s = \mathrm{TopEig}_r(\bar{P}), \qquad K_s = r^{-1} U_s U_s^\top. \tag{2}$$

The bank contains the V direct-view kernels and K_s; projector aggregation removes basis rotations and keeps r distinct from k.

Given $\boldsymbol{\beta}_0 = \mathrm{softmax}(\boldsymbol{\theta})$, the lower level applies $\mathcal{R} = \Pi_{\Delta_M}[\boldsymbol{\beta}_0 + \mu^{-1}\boldsymbol{s}(U(\boldsymbol{\beta}))]$, where $s_j(U) = \mathrm{Tr}(U^\top K_j U)$. For a stable active set and simplex tangent basis T,

$$\boldsymbol{\beta}^\star = \mathcal{R}(\boldsymbol{\beta}^\star, \boldsymbol{\theta}), \qquad \widehat{D}_\theta \boldsymbol{\beta}^\star = T(I - \widehat{J}_\beta)^{-1} \widehat{J}_\theta T^\top. \tag{3}$$

The upper level scores alignment, margin, eigengap, contraction, balance, and entropy. Approximate implicit updates (Franceschi et al., 2018; Lorraine et al., 2020; Pedregosa, 2016; Salzo, 2020; Blondel et al., 2022) require active-set, conditioning, and descent checks here. This proxy is not a certificate.

After training, the approximate kernels and decoding choices are frozen. Let $D_\varepsilon = \|\widetilde{\boldsymbol{\varepsilon}}\|_\infty$, $A_\varepsilon = r\|\widetilde{\boldsymbol{\varepsilon}}\|_2/\mu$, $B_K = \|(\|K_j\|_F)_j\|_2/\mu$, $\kappa = \|(\|K_j\|_2)_j\|_2$, $L = 2\sqrt{2r}/\gamma_r$, and $\chi = L\kappa B_K < 1$. With residuals $\eta_{\mathrm{eig}}, \eta_{\mathrm{fp}}$ and safety factor $\sigma \geq 1$,

$$R_P = \frac{L(D_\varepsilon + \kappa A_\varepsilon) + 2\sqrt{2}\,\sigma\eta_{\mathrm{eig}}/\gamma_r + \sigma L\kappa\eta_{\mathrm{fp}}}{1-\chi}, \qquad m_i(R_P) > 0 \Longrightarrow \widetilde{y}_i = y_i. \tag{4}$$

Here $m_i(R_P)$ is the residual-adjusted fixed-core margin; eigengap and normalization checks determine global, partial, or reason-specific abstaining states.

2.3 Preliminaries

Simplex and spectral projectors. $\Delta_M = \{\boldsymbol{\beta} \geq 0 : \mathbf{1}^\top\boldsymbol{\beta} = 1\}$ and Π_{Δ_M} is its Euclidean projection. For $U^\top U = I_r$, $P_U = UU^\top$ is basis-invariant; spectral change is therefore measured through projectors, with k distinct from r. Notation is summarized in Appendix A.

Perturbation model and certification states. The eigengap $\gamma_r = \lambda_r(K) - \lambda_{r+1}(K)$ controls subspace sensitivity (Davis & Kahan, 1970). For frozen approximate kernels, $\|\Delta K_j\|_2 \leq \varepsilon_j$ and an available Nyström bound gives $\widetilde{\varepsilon}_j = \varepsilon_j + 2\varrho_j$. Certification concerns this frozen computation, excluding semantic correctness, parameter reselection, unrestricted feature changes, and a new discretization.

3 Proposed

Fig. 2 illustrates BC-MKC; pseudocode is provided in Appendix B.

3.1 Formulation

Let $\{X^{(v)}\}_{v=1}^V$ be V views of the same n objects, with requested cluster number k and retained rank $r \geq k$. Building on the fused kernel in Eq. (1) and the response in Eq. (3), BC-MKC learns kernel weights without ground-truth labels and later tests which fixed-core assignments satisfy its stated perturbation conditions.

Stabilized kernel bank. Let b denote the landmark count. For view v, let $C_v \in \mathbb{R}^{n\times b}$ be its Nyström cross-kernel and let the regularized landmark block $W_v \in \mathbb{R}^{b\times b}$ satisfy $W_v = Q_v\Lambda_v Q_v^\top$. We set $\widetilde{G}_v = C_v Q_v \Lambda_v^{-1/2}$, $G_v = \widetilde{G}_v/\|\widetilde{G}_v\|_F$, and $K_v = G_v G_v^\top$. For each of Q predetermined configurations, U_ℓ spans the retained subspace. Its label-free reliability a_ℓ combines median projector affinity with spectral separation, and $\omega_\ell = (1-\rho)/Q + \rho a_\ell/\sum_q a_q$; a degenerate reliability vector falls back to uniform weights. Since U_ℓ and $U_\ell R$ represent the same subspace, the resulting trace-one bank is

$$\begin{aligned} &\bar{P} = \sum_{\ell=1}^{Q} \omega_\ell U_\ell U_\ell^\top, \quad U_s = \mathrm{TopEig}_r(\bar{P}), \quad K_s = r^{-1}U_s U_s^\top, \quad \mathcal{K} = \{K_1, \ldots, K_V, K_s\}, \\ &s_b = (b/b_0)_{\mathrm{req}}, \quad b = \min\{n, \lfloor s_b b_0 \rceil\}, \qquad s_r = (r/k)_{\mathrm{req}}, \quad r = \min\{r_{\max}, \max\{k, \lfloor s_r k \rceil\}\}. \end{aligned} \tag{5}$$

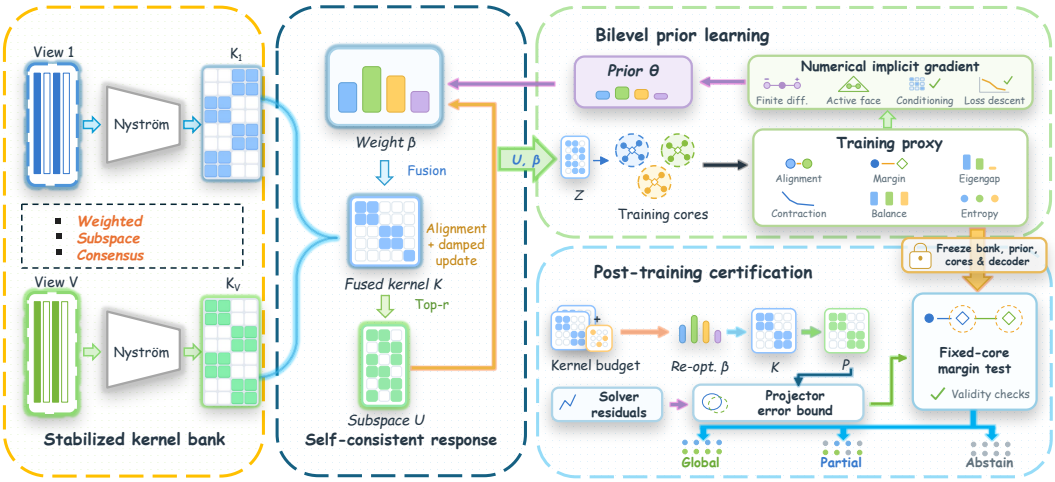


Figure 2: Overview of BC-MKC: stabilized kernel construction, self-consistent response, and bilevel prior learning, followed by fixed-core certification with global, partial, or abstaining outcomes.

Here $\lfloor\cdot\rceil$ denotes nearest-integer rounding and b_0 is the effective baseline landmark budget; s_b, s_r are the requested ratios reported in Tables 5–6, before rounding and clipping. The feasible cap $r_{\max}$ accounts for sample size, configured rank, and factor dimensions. We have $\mathrm{Tr}(K_j) = 1$, $\boldsymbol{\omega} \in \Delta_Q$, and $M = V + 1$. This realizes the stabilized meta-kernel in Eq. (2) and keeps r independent of k.

Lower response and upper objective. For $\boldsymbol{\beta} \in \Delta_M$, set $K(\boldsymbol{\beta}) = \sum_j \beta_j K_j$, $U(\boldsymbol{\beta}) = \mathrm{TopEig}_r(K(\boldsymbol{\beta}))$, and $s_j(U) = \mathrm{Tr}(U^\top K_j U)$. With $\boldsymbol{\beta}_0(\boldsymbol{\theta}) = \mathrm{softmax}(\boldsymbol{\theta})$, the fixed-$U$ weight problem and its self-consistent response are

$$\begin{aligned} &\min_{\boldsymbol{\beta}\in\Delta_M} \frac{s_\mu \mu_0}{2}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2^2 - \boldsymbol{s}(U)^\top\boldsymbol{\beta}, \qquad s_\mu = \mu/\mu_0, \\ &\mathcal{R}_{\boldsymbol{\theta}}(\boldsymbol{\beta}) = \Pi_{\Delta_M}\left[\boldsymbol{\beta}_0 + (s_\mu\mu_0)^{-1}\boldsymbol{s}(U(\boldsymbol{\beta}))\right], \qquad \boldsymbol{\beta}^\star = \mathcal{R}_{\boldsymbol{\theta}}(\boldsymbol{\beta}^\star). \end{aligned} \tag{6}$$

The reference scale is $\mu_0 = 5 \times 10^5$. The quadratic term is μ-strongly convex only when U is fixed; the resulting system is solved with damping. We retain $\eta_{\mathrm{fp}} = \|\widehat{\boldsymbol{\beta}} - \mathcal{R}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\beta}})\|_2$. The prior parameter $\boldsymbol{\theta}$, response $\boldsymbol{\beta}$, and retained subspace U are collected into one constrained training loss in Section 3.2. To form a label-free signal, a deterministic initial partition selects $A^{\mathrm{tr}} \in \mathbb{R}^{n\times k}$ with $\mathbf{1}^\top A^{\mathrm{tr}}_{:c} = 1$; it is held fixed during outer steps. With $Z = \mathrm{Diag}(UU^\top)^{-1/2}U$, define $g^{\mathrm{tr}}_c = Z^\top A^{\mathrm{tr}}_{:c}$, $D^{\mathrm{tr}}_{ic} = \|z_i - g^{\mathrm{tr}}_c\|_2^2$, and score $S_{ic} = -D^{\mathrm{tr}}_{ic}$. The implemented upper loss is $\mathcal{L}_{\mathrm{up}} = -\Phi$, where

$$\Phi = w_a\boldsymbol{\beta}^\top\boldsymbol{s}(U) + w_m\mathcal{M}_\tau + w_g\gamma_r - w_c[\chi-\chi_0]_+^2 - w_b\mathcal{B}_\tau + w_h\mathcal{H}(\boldsymbol{\beta}_0), \tag{7}$$

For fixed reference label q_i, the implemented margin, balance, and entropy terms are

$$\begin{aligned} &m_i^\tau = S_{iq_i} - \tau\log\sum_{c\neq q_i} e^{S_{ic}/\tau}, \quad \mathcal{M}_\tau = -\tau\log\left[n^{-1}\sum_i e^{-m_i^\tau/\tau}\right], \\ &p_{ic} = \mathrm{softmax}_c(S_i/\tau_a), \qquad \mathcal{B}_\tau = \sum_c (n^{-1}\sum_i p_{ic} - 1/k)^2, \quad \mathcal{H} = -\sum_j \beta_{0j}\log\beta_{0j}. \end{aligned}$$

Thus Eq. (7) specifies the training proxy, while the non-smooth certificate remains separate.

3.2 Optimization and Certification

Total training loss. Following a blockwise strategy, we optimize one state while the others are induced by the current block. The training objective is

$$\begin{aligned} &\min_{\boldsymbol{\theta},\boldsymbol{\beta},U} \mathcal{J}(\boldsymbol{\theta},\boldsymbol{\beta},U) := -w_a\boldsymbol{\beta}^\top\boldsymbol{s}(U) - w_m\mathcal{M}_\tau - w_g\gamma_r + w_c[\chi-\chi_0]_+^2 + w_b\mathcal{B}_\tau - w_h\mathcal{H}(\boldsymbol{\beta}_0), \\ &\text{s.t.} \quad \boldsymbol{\beta} = \mathcal{R}_{\boldsymbol{\theta}}(\boldsymbol{\beta}), \quad U = \mathrm{TopEig}_r(K(\boldsymbol{\beta})), \quad \boldsymbol{\beta}\in\Delta_M, \quad U^\top U = I_r. \end{aligned} \tag{8}$$

The coefficients in Eq. (7) are fixed hyperparameters. Hence $\boldsymbol{\theta}$ is the independent outer variable, whereas $\boldsymbol{\beta}$ and U are recomputed responses rather than free unconstrained parameters.

Update U and $\boldsymbol{\beta}$. For fixed $\boldsymbol{\theta}$, let $\boldsymbol{\beta}_0 = \mathrm{softmax}(\boldsymbol{\theta})$. Starting from the current response, the lower block alternates the spectral update and a damped simplex update,

$$U^{(s)} = \mathrm{TopEig}_r(K(\boldsymbol{\beta}^{(s)})), \qquad \boldsymbol{\beta}^{(s+1)} = (1-\alpha)\boldsymbol{\beta}^{(s)} + \alpha \mathcal{R}_{\boldsymbol{\theta}}(\boldsymbol{\beta}^{(s)}), \quad 0 < \alpha \le 1. \tag{9}$$

This step targets the minimizer of the fixed-U strongly convex weight subproblem in Eq. (6) and then refreshes the retained eigenspace. Iteration stops when the weight and projector changes satisfy their numerical tolerances.

Assumption 1 (Regular lower response)**.** *There exists a nonempty compact convex set $\mathcal{B} \subset \mathrm{relint}(\Delta_M)$ that contains the lower iterates and satisfies $\mathcal{R}_{\boldsymbol{\theta}}(\mathcal{B}) \subseteq \mathcal{B}$. On $\mathcal{B}$, the top-r eigengap is at least $\underline{\gamma} > 0$, the simplex face is unchanged, and $\mathcal{R}_{\boldsymbol{\theta}}$ is continuously differentiable. With $\bar{L} = 2\sqrt{2r}/\underline{\gamma}$, $\kappa = \|(\|K_j\|_2)_j\|_2$, and $B_K = \|(\|K_j\|_F)_j\|_2/\mu$, assume $\chi = \bar{L}\kappa B_K < 1$.*

Theorem 1 (Unique and differentiable lower response)**.** *Under Assumption 1, $\mathcal{R}_{\boldsymbol{\theta}}$ has a unique fixed point in $\mathcal{B}$. The damped iteration in Eq. (9) contracts by at most $1-\alpha(1-\chi)$. If $T^\top T = I$ and $T^\top \mathbf{1} = 0$, define $J_\beta = T^\top D_\beta \mathcal{R} T$ and $J_\theta = T^\top D_\theta \mathcal{R} T$. Then*

$$D_\theta \boldsymbol{\beta}^\star = T(I - J_\beta)^{-1} J_\theta T^\top, \qquad \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 \le \frac{\eta_{\mathrm{fp}}}{1-\chi}. \tag{10}$$

Theorem 1 supplies two quantities used by the optimizer: its contraction bound controls the lower stopping error, and its implicit derivative carries the effect of $\boldsymbol{\theta}$ through the re-optimized response. The proof is in Appendix D.

Update $\boldsymbol{\theta}$. Let $\bar{\Phi}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \Phi(\boldsymbol{\beta}, \boldsymbol{\theta}; U(\boldsymbol{\beta}))$, so each weight perturbation recomputes the spectral response. On the simplex tangent space, centered finite differences estimate J_β, J_θ, and the partial derivatives of $\bar{\Phi}$. Theorem 1 gives a proxy-ascent, or equivalently loss-descent, direction

$$g_{\theta,T} = \partial_{\theta_T} \bar{\Phi} + J_\theta^\top (I - J_\beta)^{-\top} \partial_{\beta_T} \bar{\Phi}, \qquad \boldsymbol{g}_\theta = T g_{\theta,T}. \tag{11}$$

The factor bank and A^{tr} remain fixed during this update. A direction is used only if the perturbed responses retain the same active face and $\mathrm{rcond}(I - \widehat{J}_\beta)$ exceeds a preset threshold. Let $(\boldsymbol{\beta}^+, U^+)$ denote the lower response recomputed at the candidate prior. Backtracking accepts

$$\boldsymbol{\theta}^+ = \boldsymbol{\theta} + \eta \widehat{\boldsymbol{g}}_\theta / \max(\|\widehat{\boldsymbol{g}}_\theta\|_2, 1) \quad \text{only if} \quad \mathcal{J}(\boldsymbol{\theta}^+, \boldsymbol{\beta}^+, U^+) < \mathcal{J}(\boldsymbol{\theta}, \boldsymbol{\beta}, U). \tag{12}$$

If a check fails, the step is rejected and the previous model is retained. The acceptance rule requires a lower recomputed training loss up to the prescribed numerical tolerance, but neither the finite-difference gate nor Eq. (7) is treated as a robustness certificate.

Residual-aware verification. Training ends before certification. From the final nominal embedding, the deterministic discretizer reselects $A \in \mathbb{R}^{n \times k}$ with $\mathbf{1}^\top A_{:c} = 1$; this matrix, the row-normalized decoder, factor bank, landmarks, bandwidths, k, r, and $\boldsymbol{\theta}$ are then frozen. Let $g_c = Z^\top A_{:c}$ and $D_{ic} = \|z_i - g_c\|_2^2$. The measured eigensolver residual is $\eta_{\mathrm{eig}} = \|KU - U(U^\top K U)\|_F$. The threat model is posed on the final kernel operators:

$$\widetilde{K}_j = K_j + E_j, \quad \|E_j\|_2 \le \widetilde{\varepsilon}_j, \quad \widetilde{\varepsilon}_j = \varepsilon_j + 2\varrho_j, \qquad D_\varepsilon = \|\widetilde{\boldsymbol{\varepsilon}}\|_\infty, \quad A_\varepsilon = \frac{r\|\widetilde{\boldsymbol{\varepsilon}}\|_2}{\mu}. \tag{13}$$

Here ϱ_j is an optional uniform approximation bound, charged once at each endpoint. If it is unavailable, $\varrho_j = 0$ and the scope is the fixed-landmark approximate kernels. Let $L = 2\sqrt{2r}/\gamma_r$, retain B_K and κ from Assumption 1, and set $\chi = L\kappa B_K$ and $R_K = (D_\varepsilon + \kappa A_\varepsilon)/(1-\chi)$.

Assumption 2 (Frozen scope and valid residual bounds)**.** *The clean and perturbed responses remain in the regular spectral region, $\chi < 1$, and $R_K < \gamma_r/2$. The reported $\sigma\eta_{\mathrm{fp}}$ and $\sigma\eta_{\mathrm{eig}}$, with $\sigma \ge 1$, are valid endpoint bounds for the lower fixed-point and top-r eigensolver errors, respectively. The eigensolver targets the same separated spectral cluster, and the core matrix and listed decoding choices in the preceding paragraph remain fixed.*

Theorem 2 (Residual-aware fixed-core assignment invariance)**.** *Under Assumptions 1–2, the numerical nominal projector $\widehat{P}$ and the re-optimized perturbed projector $\widetilde{P}$ satisfy*

$$\|\widetilde{P} - \widehat{P}\|_F \le R_P := \frac{L(D_\varepsilon + \kappa A_\varepsilon) + 2\sqrt{2}\,\sigma\eta_{\mathrm{eig}}/\gamma_r + \sigma L\kappa\eta_{\mathrm{fp}}}{1-\chi}. \tag{14}$$

Let $\nu_i = \|\widehat{u}_i\|_2$, $c_i = \arg\min_c D_{ic}$, *and, for the fixed core matrix* A, *define*

$$h_{ic} = \nu_i^{-2} + \sum_{\ell=1}^{n} A_{\ell c}^2 \nu_\ell^{-2} - 2A_{ic}\nu_i^{-2}, \qquad \delta_{ic} = 2R_P\sqrt{h_{ic}}, \qquad e_{ic} = 2\sqrt{D_{ic}}\,\delta_{ic} + \delta_{ic}^2. \quad (15)$$

If $R_P < \min_i \nu_i$ *and*

$$D_{id} - D_{ic_i} > e_{ic_i} + e_{id} \quad \textit{for every } d \neq c_i, \quad (16)$$

then sample i *keeps label* c_i *under every admissible perturbation. If all samples pass, the co-assignment matrix* $C(\boldsymbol{y})_{ii'} = \mathbf{1}\{y_i = y_{i'}\}$ *is invariant; if only* $S \subset [n]$ *passes, the restriction* $C(\boldsymbol{y})_{S,S}$ *is invariant.*

Theorem 2 is applied after the final training update. For each tested budget, it converts kernel and solver errors into R_P, then into the samplewise distance errors used by Eq. (16); it does not update $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, or U. The proof is in Appendix D.

4 Experiments

4.1 Experiment Setting

Metrics. Accuracy (ACC) is the proportion of label-aligned samples assigned to their reference classes. Normalized Mutual Information (NMI) measures shared information between the predicted and reference partitions; F1 and Adjusted Rand Index (ARI) measures pairwise agreement after chance correction.

Datasets. We evaluate BC-MKC on ten datasets, whose statistics are summarized in Table 1.

Dataset	#Samples	#kernel	#clusters
BBCSport	544	2	5
CUB	600	2	10
ProteinFold	694	12	27
Reuters	1,200	5	6
Scene15	4,485	3	15
Fashion	10,000	3	10
ALOI	10,800	4	100
YTF10	38,654	4	10
NoisyMNIST	70,000	2	10
Winnipeg	325,834	2	7

Table 1: Statistics of ten benchmarks.

Baselines. The 22 MKC baselines are LMKKM (Gönen & Margolin, 2014), RMKKM (Du et al., 2015), MKKM-MR (Liu et al., 2016), LKAM (Li et al., 2016), ONKC (Liu et al., 2017), NKSS (Zhou et al., 2019), SS-MKC (Zhou et al., 2020), SimpleMKKM (Liu, 2022), MKKM-SR (Lu et al., 2022), MVC-LFA (Wang et al., 2019), OPLFMVC (Liu et al., 2021b), HMKC (Liu et al., 2021a), FMKKM (Zhang et al., 2022b), DLEFT-MKC (Zhang et al., 2025), Approximate MKKM (Wang et al., 2017a), SMKC (Liang et al., 2024), INMKC (Feng et al., 2025), COKE (Liang et al., 2025), SWMKKM (Zhang et al., 2022a), HFLSMKKM (Liu, 2023), sLGm (Wu et al., 2024), and BMKC (Chen et al., 2025). For a rigorous comparison, we evaluate all runnable baselines using the supplied views, the ground-truth cluster number k, and the authors' recommended settings.

Implementation. The MATLAB code RMS-standardizes features and sets RBF bandwidths by median distance. Landmark budgets are 300, except 150 for YTF10 and 100 for Winnipeg; their ranks are 10 and 7, while the remaining ranks use feasible k or $2k$ settings. We use $\mu = 5 \times 10^5$, nominal $\varepsilon_j = 10^{-5}$, and standard algorithmic defaults.

4.2 Experimental Results

Clustering Performance. Table 2 reports reproduced ACC for BC-MKC and 22 MKC baselines on ten benchmarks; entries are mean ± standard deviation. The best three values are marked in red, blue, and orange, respectively. NMI, F1, and ARI results are in Appendix E. Three observations follow.

(1) **Higher overall clustering accuracy.** The leading reproduced baseline varies by dataset. Under the common protocol, BC-MKC attains the largest reported mean ACC in the ten reproduced columns and improves the macro-average ACC by 29.2% relative to INMKC, the strongest baseline by this aggregate.

(2) **Gains over dataset-specific leaders.** Relative to the strongest reproduced baseline for each dataset, the improvement ranges from 5.1% on P-Fold to 30.2% on CUB; the corresponding gains on Reuters, Scene15, and YTF10 are 19.0%, 17.8%, and 26.5%, respectively.

(3) **Advantage over representative MKC methods.** In macro-average ACC, BC-MKC exceeds DLEFT-MKC, INMKC, and BMKC by 30.6%, 29.2%, and 34.1%, respectively. These substantial margins highlight the superiority of our framework. By integrating stabilized geometry with self-consistent weight optimization, BC-MKC robustly achieves state-of-the-art clustering performance.

Algorithms	BBCSport	CUB	P-Fold	Reuters	Scene15	Fashion	ALOI	YTF10	N-MNIST	Winnipeg
LMKKM	63.9 ± 1.4	31.9 ± 3.3	32.4 ± 0.7	34.1 ± 1.5	30.9 ± 0.1	49.2 ± 0.9	47.4 ± 0.6	58.5 ± 1.5	39.6 ± 2.4	40.7 ± 3.3
RMKKM	71.4 ± 0.7	43.6 ± 1.2	34.7 ± 2.1	35.8 ± 3.0	36.4 ± 0.3	57.8 ± 1.3	39.1 ± 2.1	50.2 ± 3.0	41.3 ± 0.3	42.4 ± 1.2
MKKM-MR	63.2 ± 1.5	35.3 ± 2.7	24.5 ± 1.8	46.2 ± 1.4	38.1 ± 1.1	30.3 ± 1.1	40.8 ± 0.0	41.9 ± 0.9	33.0 ± 1.8	44.1 ± 2.7
LKAM	73.9 ± 0.5	37.0 ± 0.6	27.7 ± 1.2	45.5 ± 0.0	31.7 ± 0.5	64.1 ± 0.5	44.5 ± 1.5	47.6 ± 2.4	44.7 ± 3.3	45.8 ± 0.6
ONKC	63.4 ± 1.4	38.7 ± 2.1	26.3 ± 1.5	41.8 ± 1.2	39.9 ± 1.4	55.5 ± 0.7	44.2 ± 3.0	55.3 ± 0.3	56.4 ± 1.2	47.5 ± 2.1
NKSS	64.1 ± 1.2	50.4 ± 1.0	25.4 ± 0.7	37.7 ± 1.4	40.4 ± 0.3	46.7 ± 1.4	35.9 ± 0.9	47.0 ± 1.8	48.1 ± 2.7	49.2 ± 0.5
SS-MKC	45.3 ± 2.1	42.1 ± 1.5	31.2 ± 2.4	44.3 ± 3.3	36.3 ± 0.6	53.6 ± 2.2	47.6 ± 2.4	48.7 ± 3.3	49.8 ± 0.6	50.9 ± 1.5
SimpleMKKM	69.6 ± 0.5	43.8 ± 3.0	29.6 ± 0.3	46.0 ± 1.2	29.1 ± 2.1	61.0 ± 1.7	49.3 ± 0.3	50.4 ± 1.2	61.5 ± 2.1	52.6 ± 3.0
MKKM-SR	57.8 ± 0.9	55.5 ± 0.9	21.3 ± 1.8	47.7 ± 2.7	36.6 ± 0.5	53.4 ± 0.8	61.0 ± 1.8	42.1 ± 2.7	53.2 ± 1.3	54.3 ± 0.9
MVC-LFA	62.1 ± 1.1	47.2 ± 2.4	37.2 ± 3.3	49.4 ± 0.6	32.6 ± 1.5	45.9 ± 1.2	52.7 ± 3.3	53.8 ± 0.6	54.9 ± 1.5	56.0 ± 2.4
OPLFMVC	79.2 ± 3.2	48.9 ± 0.3	31.1 ± 2.6	43.9 ± 1.0	33.9 ± 1.8	52.8 ± 1.8	54.4 ± 1.2	55.5 ± 2.1	46.6 ± 3.0	57.7 ± 0.3
HMKC	91.1 ± 3.7	50.6 ± 1.8	25.3 ± 1.5	46.8 ± 0.3	40.5 ± 0.1	67.1 ± 0.5	56.1 ± 2.7	67.2 ± 0.3	58.3 ± 0.9	54.4 ± 1.8
FMKKM	78.2 ± 1.6	52.3 ± 3.3	30.2 ± 0.6	34.5 ± 1.5	37.6 ± 2.4	58.5 ± 3.4	48.8 ± 0.6	58.9 ± 1.5	60.0 ± 2.4	49.7 ± 3.3
DLEFT-MKC	80.2 ± 0.1	54.0 ± 1.2	28.5 ± 2.9	47.0 ± 4.0	35.2 ± 0.1	66.2 ± 3.1	59.5 ± 2.1	70.6 ± 3.0	61.7 ± 0.3	60.3 ± 1.2
Approx. MKKM	81.5 ± 1.8	45.7 ± 2.7	24.6 ± 0.0	41.9 ± 0.9	28.8 ± 1.8	43.6 ± 0.2	61.2 ± 0.6	62.3 ± 0.9	56.4 ± 1.8	60.8 ± 2.7
SMKC	69.3 ± 0.4	57.4 ± 0.6	27.9 ± 1.5	39.6 ± 2.4	33.4 ± 3.3	75.7 ± 1.1	62.9 ± 1.5	64.0 ± 2.4	55.1 ± 3.3	57.6 ± 0.6
INMKC	77.1 ± 1.6	59.1 ± 2.1	23.2 ± 2.0	51.3 ± 0.3	39.8 ± 1.2	60.9 ± 1.7	74.6 ± 3.0	65.7 ± 0.3	63.8 ± 1.2	53.2 ± 2.1
COKE	75.4 ± 2.1	60.8 ± 0.6	30.1 ± 0.9	46.7 ± 1.8	32.6 ± 2.7	74.2 ± 1.3	66.3 ± 0.9	57.4 ± 1.8	61.5 ± 2.7	58.4 ± 0.7
SWMKKM	68.3 ± 0.9	62.5 ± 1.5	26.4 ± 2.4	49.2 ± 3.3	28.7 ± 0.6	58.6 ± 0.8	58.0 ± 2.4	69.1 ± 3.3	58.2 ± 0.6	39.3 ± 1.5
HFLSMKKM	51.6 ± 1.3	64.2 ± 3.0	31.8 ± 1.1	37.5 ± 0.8	35.0 ± 0.4	49.5 ± 2.1	69.7 ± 0.3	68.8 ± 1.2	49.9 ± 2.1	31.0 ± 3.0
sLGm	77.2 ± 0.2	55.9 ± 0.9	25.2 ± 1.8	44.3 ± 2.7	37.6 ± 0.9	73.4 ± 0.6	59.4 ± 1.8	70.5 ± 2.7	51.6 ± 0.0	42.7 ± 0.9
BMKC	79.5 ± 1.1	67.6 ± 2.4	28.4 ± 3.3	52.2 ± 0.6	28.9 ± 1.5	70.1 ± 0.9	61.1 ± 3.3	62.2 ± 0.6	63.3 ± 1.5	34.4 ± 2.4
BC-MKC	$\mathbf{96.6 \pm 0.4}$	$\mathbf{88.0 \pm 0.5}$	$\mathbf{39.1 \pm 2.0}$	$\mathbf{62.1 \pm 0.1}$	$\mathbf{47.7 \pm 0.4}$	$\mathbf{86.0 \pm 0.8}$	$\mathbf{84.6 \pm 0.1}$	$\mathbf{89.3 \pm 0.8}$	$\mathbf{74.5 \pm 0.0}$	$\mathbf{66.9 \pm 0.6}$
Δ (%)	$\mathbf{\Delta + 5.5}$	$\mathbf{\Delta + 20.4}$	$\mathbf{\Delta + 1.9}$	$\mathbf{\Delta + 9.9}$	$\mathbf{\Delta + 7.2}$	$\mathbf{\Delta + 10.3}$	$\mathbf{\Delta + 10.0}$	$\mathbf{\Delta + 18.7}$	$\mathbf{\Delta + 10.7}$	$\mathbf{\Delta + 6.1}$

Table 2: ACC comparison of BC-MKC with 22 reproduced MKC baselines on ten benchmarks (%). Δ denotes the gain over the best baseline in percentage points.

Stability Across Random Seeds. Fig. 3 compares the mean ACC and standard deviation of eight MKC methods on BBCSport, CUB, and Fashion. Relative to the highest baseline means, BC-MKC improves ACC by 20.4%, 44.6%, and 13.8%, respectively (annotated in blue). Beyond these accuracy gains, the significantly tighter error bars reflect exceptional algorithmic reliability. This reduced variance demonstrates that BC-MKC is highly robust to initialization shifts, ensuring predictable convergence and structural stability.

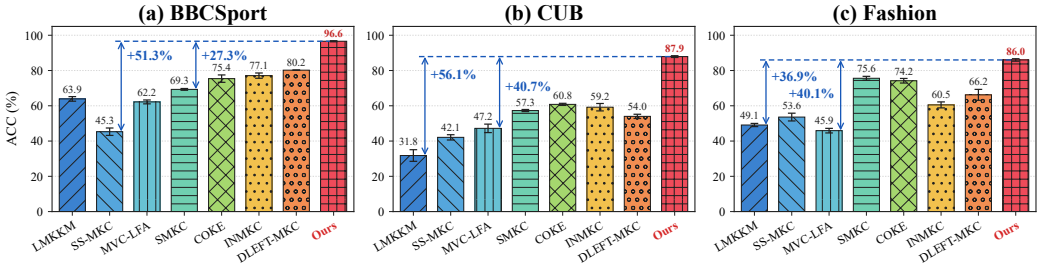


Figure 3: Random-seed ACC comparison of eight MKC methods on BBCSport, CUB, and Fashion.

4.3 Ablation Study

Stabilized kernel bank. Fig. 4 compares Full with NoNorm and Uniform-ω at training batches 10, 20, 30, 40, and 50. Relative to these variants, Full improves mean ACC across the four datasets by significant margins ranging from 7.7% to 37.4%. These substantial gains demonstrate that by combining trace normalization with adaptive reliability weighting, BC-MKC effectively mitigates approximation-induced geometric variations, establishing a highly stable spectral foundation for clustering.

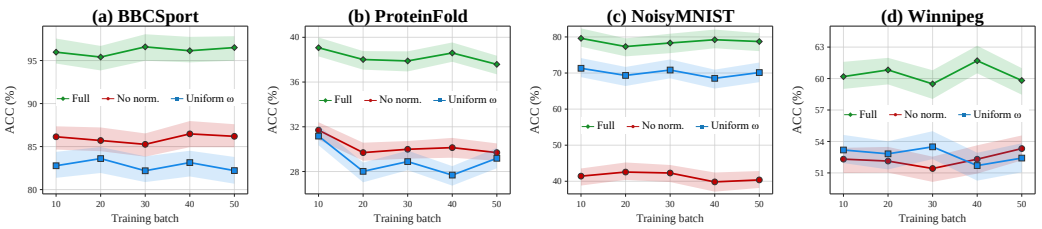


Figure 4: Training-batch ACC of Full, NoNorm, and Uniform-ω on (a) BBCSport, (b) ProteinFold, (c) NoisyMNIST, and (d) Winnipeg.

Self-consistent lower response. Table 3 compares Full-FP with Static-β, One-step, and Uniform-β on Fashion. Across ACC, NMI, F1, and ARI, Full-FP improves upon these variants by significant margins ranging from 9.1% to 16.2%. These comprehensive gains highlight the practical benefit of our adaptive design. By dynamically aligning the spectral subspace with the kernel weights,

BC-MKC effectively avoids the representational biases of static or heuristic assignments, leading to more reliable clustering structures.

Variant	Method	ACC	NMI	F1	ARI
#1	**Full-FP (Ours)**	**86.0 ± 0.8**	**82.1 ± 0.6**	**87.0 ± 0.5**	**77.5 ± 0.7**
#2	Static-β	75.4 ± 1.7	70.2 ± 1.2	72.3 ± 1.2	66.2 ± 0.8
#3	One-step	71.5 ± 1.1	71.3 ± 0.5	73.4 ± 1.5	68.4 ± 1.0
#4	Uniform-β	72.7 ± 0.9	70.8 ± 0.7	70.8 ± 0.9	67.8 ± 1.1

Table 3: Lower-response ablation on the Fashion dataset.

Variant	Method	ACC	ARI	NMI
#1	**Full (Ours)**	**86.0 ± 0.8**	**77.5 ± 0.7**	**82.1 ± 0.6**
#2	NoMargin	70.2 ± 1.3	64.5 ± 1.3	69.2 ± 0.7
#3	NoEigengap	71.4 ± 0.8	64.2 ± 0.7	71.5 ± 1.2
#4	NoContraction	69.2 ± 1.6	65.3 ± 1.2	70.3 ± 0.4
#5	NoBal./Ent.	68.8 ± 1.7	63.8 ± 1.4	69.7 ± 1.6
#6	StopGradient	70.1 ± 1.3	64.6 ± 0.9	71.8 ± 1.1
#7	Ungated	71.7 ± 0.9	63.1 ± 1.7	69.6 ± 0.8

Table 4: Upper-objective and update-control ablation on the Fashion dataset.

Upper objective and guarded update. Table 4 compares Full with NoMargin, NoEigengap, NoContraction, NoBal./Ent., StopGradient, and Ungated on Fashion. Across ACC, NMI, and ARI, Full consistently outperforms these variants by substantial margins ranging from 10.3% to 17.2%. These comprehensive improvements confirm that integrating the full upper objective with guarded updates effectively steers the learning trajectory. By explicitly enforcing structural margins while preventing unstable numerical deviations, BC-MKC ensures a safely controlled optimization process that reliably extracts discriminative partitions. Further ablations appear in Appendix F.3.

4.4 Parameter Sensitivity Analysis

Nyström landmark budget. Table 5 varies $s_b = (b/b_0)_{\text{req}}$ in Eq. (5) from 0.50 to 1.50, controlling the landmark count. Macro ACC ranges from 76.0% to 77.4%, with the highest mean at the default budget. The small variation suggests limited sensitivity of mean ACC to b/b_0 over the tested range.

b/b_0	BBC	CUB	P-Fold	Fashion
0.50	95.7±0.7	86.2±0.4	36.8±0.9	85.4±1.6
0.75	95.5±0.7	86.0±0.7	36.9±1.1	85.5±1.1
1.00	**96.6±0.4**	**88.0±0.5**	**39.1±2.0**	**86.0±0.8**
1.25	95.4±1.1	86.1±1.8	36.7±1.0	85.7±0.5
1.50	95.8±0.6	86.0±1.0	37.0±1.4	85.7±1.2

Table 5: ACC under different Nyström landmark budgets on four benchmarks.

r/k	BBC	CUB	P-Fold	Fashion
1.0	96.3±0.5	**88.0±1.5**	39.0±1.6	85.8±1.1
1.5	96.4±0.3	87.9±1.3	38.9±1.1	85.9±2.2
2.0	**96.6±0.4**	**88.0±0.5**	**39.1±2.0**	**86.0±0.8**
2.5	96.5±1.5	87.8±2.2	38.9±1.4	85.7±1.1
3.0	96.5±1.1	87.9±0.5	38.8±1.3	85.9±1.4

Table 6: ACC under different retained-rank ratios on four benchmarks.

Retained spectral rank. Table 6 varies $s_r = (r/k)_{\text{req}}$ in Eq. (5) from 1.0 to 3.0, setting the retained rank in Eq. (9). Macro ACC ranges from 77.2% to 77.4%, with the highest mean of 77.4% at the default $r/k = 2$. These consistent results demonstrate that BC-MKC's clustering performance is robust to variations in the retained rank. Other hyperparameters: Appendix F.5.

Lower-response regularization. We vary the regularization scale $s_\mu = \mu/\mu_0$ in Eq. (6) and damping α in Eq. (9). Fig. 5 evaluates their 5×5 grid on BBCSport and CUB. Averaged over the other parameter and both datasets, ACC is 89.0% at $\mu/\mu_0 = 10^{-2}$ and 88.6% at 10^{-1}, a relative difference of 0.5%. For damping, $\alpha = 0.4$ gives 89.1% versus 88.8% at 0.6, a relative difference of 0.3%. These minimal variations demonstrate that BC-MKC's performance is highly resilient to precise parameter calibration, alleviating exhaustive tuning efforts and ensuring reliable practical deployment. More datasets: Appendix F.4.

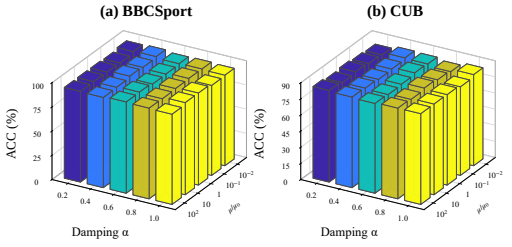


Figure 5: ACC sensitivity to lower-response regularization μ and damping α on BBCSport and CUB.

Certification failure. The budget ε bounds frozen-kernel perturbations in Eq. (13), while σ scales the solver-residual allowance in Eq. (14). Fig. 6(a) shows mean coverage across the four sensitivity benchmarks: at $\varepsilon = 10^{-4}$, coverage is 20.4% for $\sigma = 1.5$ and 17.4% for $\sigma = 3$, falling to 0% at $\varepsilon = 10^{-3}$. Fig. 6(b) shows abstention causes at $\sigma = 2$: the row-normalization limit is reached at $\varepsilon = 10^{-3}$, while insufficient corrected assignment margins also cause abstention at smaller budgets. Crucially, while severe perturbations may trigger theoretical abstention, BC-MKC safely defaults

to its nominal partition. This separation of formal verification from empirical execution ensures that strict certification limits do not compromise practical clustering utility.

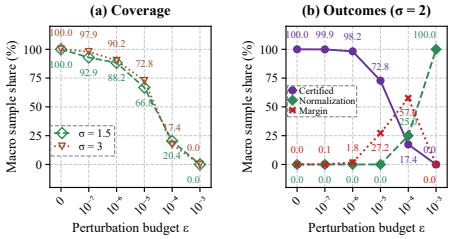


Figure 6: Certification coverage at $\sigma = 1.5, 3$ (a) and abstention causes at $\sigma = 2$ (b), averaged over four datasets and five runs each.

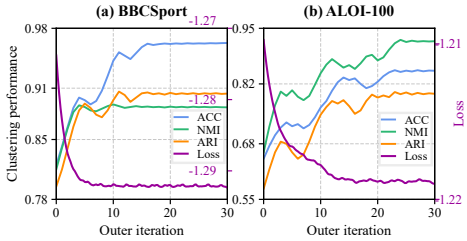


Figure 7: Evolution of loss and clustering performance during BC-MKC learning (five-run means).

Performance and convergence. We track the outer loss $\mathcal{J} = -\Phi$ in Eq. (8) and clustering performance at each iteration to examine the learning behavior of BC-MKC. Fig. 7 presents five-run means on BBCSport and ALOI-100 over 30 outer iterations. The loss drops sharply during the early iterations and then approaches a plateau with small fluctuations on both datasets. Meanwhile, ACC, NMI, and ARI improve during the early learning stage, despite transient fluctuations, and become more stable in later iterations. These trends validate the effectiveness of iterative adaptation in BC-MKC. Ultimately, this joint stabilization confirms that BC-MKC's guarded bilevel formulation effectively prevents optimization divergence, ensuring steady convergence toward a highly discriminative partition.

Running time. Fig. 8 compares main-pipeline time per complete run. Relative to the fastest baseline on each dataset, BC-MKC reduces runtime by 4.2%, 6.3%, 3.2%, and 3.1% on CUB, Reuters, YTF10, and NoisyMNIST, respectively. Together with the higher ACC in Table 2, these results indicate that BC-MKC achieves better clustering accuracy with less running time on the tested datasets. Six more datasets: Appendix F.1; memory and FLOPs: Appendix F.7.

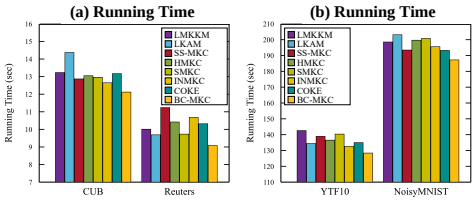


Figure 8: Main-pipeline running time per complete run (seconds).

4.5 Limitations and Future Work

i) Certification scope. The guarantee applies to frozen approximate kernels and fixed decoding under Assumptions 1–2. Extending it to raw-feature changes or landmark reselection requires additional perturbation bounds. *ii) Verification coverage.* Small eigengaps or narrow assignment margins can cause abstention (Fig. 6), while the nominal partition remains available. Tighter bounds may improve the coverage of certified assignments.

5 Conclusion

This paper presents Bilevel Certified Multiple Kernel Clustering (BC-MKC), combining adaptive kernel learning with assignment verification. Specifically, a trace-normalized Nyström bank stabilizes the approximation geometry, while a self-consistent lower response and guarded implicit updates guide kernel weighting. After training, a fixed-core verifier propagates kernel perturbations and numerical residuals to assignment margins, providing sufficient conditions for partition invariance or reporting abstention. The resulting certificates apply to frozen approximate kernels under the stated regularity and margin conditions. Experiments on ten benchmarks support the method's clustering performance, while ablation and sensitivity studies illustrate the contributions of its components and the limits of certification. Transferability and clustering visualizations are presented in Appendix F.6 and Appendix G.

REFERENCES

Francis Bach. Sharp analysis of low-rank kernel matrix approximations. In *Conference on learning theory*, pp. 185–209. PMLR, 2013. 1

Mathieu Blondel, Quentin Berthet, Marco Cuturi, Roy Frostig, Stephan Hoyer, Felipe Llinares-López, Fabian Pedregosa, and Jean-Philippe Vert. Efficient and modular implicit differentiation. *Advances in neural information processing systems*, 35:5230–5242, 2022. 3

Yan Chen, Bingbing Jiang, Peng Zhou, Lei Duan, Yuhua Qian, and Liang Du. Balanced multiple kernel clustering with discrete partition entropy auto regularization. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2197–2206, 2025. 1, 2, 6

Benoît Colson, Patrice Marcotte, and Gilles Savard. An overview of bilevel optimization. *Annals of operations research*, 153(1):235–256, 2007. 2

Corinna Cortes, Mehryar Mohri, and Ameet Talwalkar. On the impact of kernel approximation on learning accuracy. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 113–120. JMLR Workshop and Conference Proceedings, 2010. 1

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. 2, 3

Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 551–556, 2004. 1, 2

Liang Du, Peng Zhou, Lei Shi, Hanmo Wang, Mingyu Fan, Wenjian Wang, and Yi-Dong Shen. Robust multiple kernel k-means using l21-norm. In *IJCAI*, volume 15, pp. 3476–3482, 2015. 6

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, 2025. 1, 2, 6

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International conference on machine learning*, pp. 1568–1577. PMLR, 2018. 2, 3

Alex Gittens and Michael Mahoney. Revisiting the nystrom method for improved large-scale machine learning. In *International Conference on Machine Learning*, pp. 567–575. PMLR, 2013. 1

Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268, 2011. 1

Mehmet Gönen and Adam A Margolin. Localized data fusion for kernel k-means clustering with application to cancer biology. *Advances in neural information processing systems*, 27, 2014. 6

Ling Huang, Donghui Yan, Nina Taft, and Michael Jordan. Spectral clustering with perturbed data. *Advances in Neural Information Processing Systems*, 21, 2008. 1

Marius Kloft, Ulf Brefeld, Sören Sonnenburg, and Alexander Zien. Lp-norm multiple kernel learning. *The Journal of Machine Learning Research*, 12:953–997, 2011. 2

Abhishek Kumar, Piyush Rai, and Hal Daume. Co-regularized multi-view spectral clustering. *Advances in neural information processing systems*, 24, 2011. 2

Miaomiao Li, Xinwang Liu, Lei Wang, Yong Dou, Jianping Yin, and En Zhu. Multiple kernel clustering with local kernel alignment maximization. 2016. 6

Weixuan Liang, En Zhu, Shengju Yu, Huiying Xu, Xinzhong Zhu, and Xinwang Liu. Scalable multiple kernel clustering: Learning clustering structure from expectation. In *Forty-first International Conference on Machine Learning*, 2024. 1, 2, 6

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*, 2025. 1, 2, 6

Jiyuan Liu, Xinwang Liu, Siwei Wang, Sihang Zhou, and Yuexiang Yang. Hierarchical multiple kernel clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 8671–8679, 2021a. 1, 6

Xinwang Liu. Simplemkkm: Simple multiple kernel k-means. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):5174–5186, 2022. 1, 2, 6

Xinwang Liu. Hyperparameter-free localized simple multiple kernel k-means with global optimum. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8566–8576, 2023. 6

Xinwang Liu, Yong Dou, Jianping Yin, Lei Wang, and En Zhu. Multiple kernel k-means clustering with matrix-induced regularization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016. 1, 2, 6

Xinwang Liu, Sihang Zhou, Yueqing Wang, Miaomiao Li, Yong Dou, En Zhu, and Jianping Yin. Optimal neighborhood kernel clustering with multiple kernels. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017. 1, 6

Xinwang Liu, Li Liu, Qing Liao, Siwei Wang, Yi Zhang, Wenxuan Tu, Chang Tang, Jiyuan Liu, and En Zhu. One pass late fusion multi-view clustering. In *International conference on machine learning*, pp. 6850–6859. PMLR, 2021b. 2, 6

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International conference on artificial intelligence and statistics*, pp. 1540–1552. PMLR, 2020. 3

Jitao Lu, Yihang Lu, Rong Wang, Feiping Nie, and Xuelong Li. Multiple kernel k-means clustering with simultaneous spectral rotation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4143–4147. IEEE, 2022. 6

Cameron Musco and Christopher Musco. Recursive sampling for the nystrom method. *Advances in neural information processing systems*, 30, 2017. 1, 2

Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001. 1

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *International conference on machine learning*, pp. 737–746. PMLR, 2016. 3

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007. 1

Alain Rakotomamonjy, Francis Bach, Stéphane Canu, and Yves Grandvalet. Simplemkl. *Journal of Machine Learning Research*, 9:2491–2521, 2008. 2

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. *Advances in neural information processing systems*, 28, 2015. 1

Saverio Salzo. On the iteration complexity of hypergradient computation. *arXiv (Cornell University)*, 2020. 3

Peng Su, Yixi Liu, Shujian Li, Shudong Huang, and Jiancheng Lv. Robust contrastive multi-view kernel clustering. In *IJCAI*, pp. 4938–4945, 2024. 1

Ulrike von Luxburg. A tutorial on spectral clustering. statistics and computing. *Data Structures and Algorithms (cs. DS); Machine Learning*, pp. 395–416, 2007. 1

Siwei Wang, Xinwang Liu, En Zhu, Chang Tang, Jiyuan Liu, Jingtao Hu, Jingyuan Xia, and Jianping Yin. Multi-view clustering via late fusion alignment maximization. In *Ijcai*, pp. 3778–3784, 2019. 1, 2, 6

Yueqing Wang, Xinwang Liu, Yong Dou, and Rongchun Li. Approximate large-scale multiple kernel k-means using deep neural network. In *IJCAI*, pp. 3006–3012, 2017a. 6

Yueqing Wang, Xinwang Liu, Yong Dou, and Rongchun Li. Multiple kernel clustering framework with improved kernels. *Discover*, 1(2):3–4, 2017b. 1

Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. *Advances in neural information processing systems*, 13, 2000. 1

Xi Wu, Chuang Huang, Xinliu Liu, Fei Zhou, and Zhenwen Ren. Multiple kernel clustering with shifted laplacian on grassmann manifold. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 2448–2456, 2024. 6

Rongkai Xia, Yan Pan, Lei Du, and Jian Yin. Robust multi-view spectral clustering via low-rank and sparse decomposition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28, 2014. 2

Shuyin Xia, Yifan Wang, Lifeng Shen, and Guoyin Wang. Granular-ball-induced multiple kernel k-means. *arXiv preprint arXiv:2506.18637*, 2025. 1

Hao Yu, Weixuan Liang, Ke Liang, Suyuan Liu, Meng Liu, and Xinwang Liu. On the adversarial robustness of multi-kernel clustering. In *Forty-second International Conference on Machine Learning*, 2025. 2

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015. 2

Yi Zhang, Weixuan Liang, Xinwang Liu, Sisi Dai, Siwei Wang, Liyang Xu, and En Zhu. Sample weighted multiple kernel k-means via min-max optimization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1679–1687, 2022a. 6

Yi Zhang, Xinwang Liu, Jiyuan Liu, Sisi Dai, Changwang Zhang, Kai Xu, and En Zhu. Fusion multiple kernel k-means. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9109–9117, 2022b. 1, 2, 6

Yi Zhang, Siwei Wang, Jiyuan Liu, Shengju Yu, Zhibin Dong, Suyuan Liu, Xinwang Liu, and En Zhu. Dleft-mkc: Dynamic late fusion multiple kernel clustering with robust tensor learning via min-max optimization. In *International Conference on Learning Representations*, volume 2025, pp. 30267–30289, 2025. 1, 2, 6

Peng Zhou, Rongwen Li, and Liang Du. Fair kernel k-means: from single kernel to multiple kernel. *Advances in Neural Information Processing Systems*, 37:99686–99714, 2024. 1

Sihang Zhou, Xinwang Liu, Miaomiao Li, En Zhu, Li Liu, Changwang Zhang, and Jianping Yin. Multiple kernel clustering with neighbor-kernel subspace segmentation. *IEEE transactions on neural networks and learning systems*, 31(4):1351–1362, 2019. 6

Sihang Zhou, En Zhu, Xinwang Liu, Tianming Zheng, Qiang Liu, Jingyuan Xia, and Jianping Yin. Subspace segmentation-based robust multiple kernel clustering. *Information Fusion*, 53:145–154, 2020. 1, 6

SUMMARY OF THE APPENDIX

This appendix supplements *Bilevel Certified Multiple Kernel Clustering*.

- §A clarifies notation whose meaning is implicit in the main text.
- §B presents the lower-response solver and the overall BC-MKC algorithm.
- §C reviews representative MKC methods and their differences from BC-MKC.
- §D provides detailed proofs of the theoretical results.
- §E reports additional main-experiment results measured by ARI, F1, and NMI.
- §F presents supplementary running-time comparisons and convergence curves.
- §G visualizes the evolution of clustering structure on N-MNIST across training iterations.

A LIST OF SYMBOLS

A.1 LLM USAGE

We employ LLMs solely to refine the language and wording for improved fluency and precision, with no further application beyond this purpose.

A.2 LIST OF SYMBOLS

Table 7 clarifies notation used in the equations and figures. Symbols already explained in the main text are omitted.

Table 7: Additional notation used in BC-MKC.

Notation	Description	Index
ρ	Uniform/reliability weighting mixture coefficient	§3.1
R	Orthogonal change of subspace basis	§3.1
τ	Temperature for smooth training margins	Eq. (7)
τ_a	Temperature for soft cluster assignments	§3.1
χ_0	Target threshold in the contraction penalty	Eq. (7)
β_T, θ_T	Weight and prior variations in tangent coordinates	Eq. (11)
J_β	Tangent Jacobian of the response with respect to weights	Eq. (10)
J_θ	Tangent Jacobian of the response with respect to the prior	Eq. (10)
$g_{\theta,T}$	Implicit ascent direction in tangent coordinates	Eq. (11)
η	Outer backtracking step size	Eq. (12)
D_ε	Largest effective kernel-perturbation budget	Eq. (13)
A_ε	Direct kernel-perturbation contribution to weight error	Eq. (13)
B_K	Alignment-score sensitivity scaled by $1/\mu$	Assump. 1
κ	Weight-to-fused-kernel sensitivity coefficient	Assump. 1
$\bar{L}$	Projector sensitivity using the uniform gap bound	Assump. 1
L	Projector sensitivity using the nominal eigengap	§3.2
h_{ic}	Row/core geometry factor in the distance-error bound	Eq. (15)
δ_{ic}	Sample-to-core vector error allowance	Eq. (15)
e_{ic}	Squared sample-to-core distance error allowance	Eq. (15)
τ in Fig. 1(c)	Subspace-centrality threshold for view screening	Fig. 1(c)
m_i^τ	Smooth margin against competing training clusters	§3.1
$\mathcal{M}_\tau$	Smooth minimum of the samplewise training margins	Eq. (7)
p_{ic}	Soft probability of assigning sample i to cluster c	§3.1
$\mathcal{B}_\tau$	Penalty for nonuniform mean cluster probabilities	Eq. (7)
$\mathcal{H}$	Entropy of the kernel-weight prior	Eq. (7)
α	Damping factor of the lower-response update	Eq. (9)
χ	Feedback coefficient linking weight and projector errors	§3.2
R_K	Kernel-error bound before solver-residual correction	§3.2
σ	Safety multiplier for the reported solver residuals	Eq. (14)
$m_i(R_P)$	Worst residual-adjusted margin over competing clusters	Eq. (4)

B Optimization algorithms

B.1 Self-consistent lower response

Algorithm 1 Lower-response solver

Input: $\mathcal{K}, r, \boldsymbol{\beta}_0, \mu, \alpha$; solver settings.
Output: $\boldsymbol{\beta}, U, \gamma_r, \eta_{\mathrm{fp}}, \eta_{\mathrm{eig}}$.

$\boldsymbol{\beta} \leftarrow \boldsymbol{\beta}_0$.
while not converged and within the iteration limit **do**
  Update $U, \boldsymbol{\beta}$ by Eq. (9).
  Check weight and projector changes (§3.2).
Recompute U, γ_r at the final $\boldsymbol{\beta}$.
Evaluate $\eta_{\mathrm{fp}}, \eta_{\mathrm{eig}}$ (§3.2).
return $\boldsymbol{\beta}, U, \gamma_r, \eta_{\mathrm{fp}}, \eta_{\mathrm{eig}}$.

B.2 Overall procedure

Algorithm 2 BC-MKC

Input: $\{X^{(v)}\}_{v=1}^{V}, k, \boldsymbol{\theta}, \widetilde{\boldsymbol{\varepsilon}}$; model and solver settings.
Output: Labels $\boldsymbol{y}$, certified set S, and status.

// Initialization
Construct $\mathcal{K}$ by Eq. (5).
Set $\boldsymbol{\beta}_0 = \mathrm{softmax}(\boldsymbol{\theta})$; obtain the response by Algorithm 1.
Fix $A^{\mathrm{tr}}, \{q_i\}$ (§3.1).

// Bilevel updates
while not converged and within the iteration limit **do**
  Compute $\widehat{\boldsymbol{g}}_\theta$ by Eq. (11).
  if derivative checks fail **then**
    break.
  Backtrack $\boldsymbol{\theta}^+$ by Eq. (12); solve each trial by Algorithm 1.
  if no trial is accepted **then**
    break.
  Retain the accepted prior, response, and residuals.

// Fixed-core verification
Select A; freeze verification settings (§3.2).
$y_i \leftarrow \arg\min_c D_{ic}$; $S \leftarrow \varnothing$; status $\leftarrow$ abstain.
if spectral/residual checks fail **then**
  return $\boldsymbol{y}, S, \mathrm{status}$.
Compute R_P by Eq. (14) using Algorithm 1's residuals.
if $R_P < \min_i \nu_i$ **then**
  Compute e_{ic} by Eq. (15); collect S by Eq. (16).
  Set status to global, partial, or abstain according to S.
return $\boldsymbol{y}, S, \mathrm{status}$.

C Additional related work

C.1 Multiple kernel and multi-view clustering

Multiple kernel clustering (MKC) seeks a consensus partition by combining complementary similarity structures (Dhillon et al., 2004; Gönen & Alpaydın, 2011). A major line of work performs fusion directly at the kernel or similarity level. MKKM-MR regularizes inter-kernel relations to reduce redundancy, while LKAM emphasizes localized kernel alignment (Liu et al., 2016; Li et al., 2016). ONKC jointly optimizes neighborhood structure and cluster assignments (Liu et al., 2017), whereas SimpleMKKM formulates unsupervised kernel alignment through a minimization–maximization problem over kernel weights and the partition matrix (Liu, 2022). Despite different formulations, these methods primarily improve clustering by adapting similarities or their contributions to a shared partition.

An alternative line performs fusion in the partition space, avoiding direct aggregation of full kernel matrices. MVC-LFA aligns weighted base partitions to form a consensus, while OPLFMVC couples partition fusion with label generation (Wang et al., 2019; Liu et al., 2021b). HMKC introduces hierarchical partition fusion, DLEFT-MKC combines dynamic partition reconstruction with tensor low-rank modeling (Liu et al., 2021a; Zhang et al., 2025), and BMKC incorporates discrete partition entropy regularization for balanced clustering (Chen et al., 2025). These approaches differ in fusion granularity and structural regularization, but generally optimize the quality of the resulting consensus under a fixed computational configuration.

BC-MKC addresses a complementary issue: the clustering response itself may change when kernel approximations, weights, or retained spectral subspaces are perturbed. It aggregates configuration-level spectral projectors into a consensus kernel and jointly models kernel-weight re-optimization and eigenspace response under a learned prior. The outer objective evaluates this adaptive response using clustering margins and spectral criteria, while guarded updates stabilize prior learning. Beyond nominal clustering quality, a separate verifier propagates admissible kernel perturbations to assignment margins, identifying globally or partially invariant partitions under the specified perturbation budget.

C.2 Scalable kernel learning and assignment verification

Scalable MKC increasingly relies on approximate kernel constructions to reduce the cost of full similarity matrices. Nyström methods use sampled landmarks to obtain low-rank approximations, with established analyses of sampling and approximation error (Williams & Seeger, 2000; Bach, 2013; Musco & Musco, 2017). Representative clustering methods build on this principle in different ways: SMKC adopts low-rank kernel approximations for efficient clustering (Liang et al., 2024), INMKC combines leverage-score sampling with incremental consensus updates across heterogeneous landmark sets (Feng et al., 2025), and COKE uses compact core kernels to estimate fusion weights for large-scale extension (Liang et al., 2025). Meanwhile, bilevel optimization provides a natural framework for modeling how optimized lower-level variables respond to upper-level parameters (Colson et al., 2007; Franceschi et al., 2018), with implicit differentiation offering a general mechanism for differentiating through such solution maps (Blondel et al., 2022). BC-MKC brings these two lines together by learning a coupled response of kernel weights and spectral representations over approximate kernels, while using guarded updates to stabilize the resulting bilevel optimization.

Robustness of this response further requires understanding how perturbations propagate to the final clustering assignments. Classical Davis–Kahan theory and related results bound changes in eigenspaces through spectral separation (Davis & Kahan, 1970; Yu et al., 2015), but these guarantees concern continuous subspaces rather than discrete partitions. BC-MKC therefore treats spectral stability as an intermediate quantity rather than the final guarantee. Its post-training verifier propagates admissible kernel perturbations together with weight re-optimization and numerical residuals through the retained eigenspace, row normalization, and fixed-core decoding. The resulting sample-wise bounds are compared against assignment margins to certify global or partial partition invariance under the stated perturbation model, while abstaining when the available margin is insufficient.

D PROOF OF THEOREMS

Proof of Theorem 1

Proof. Fix the kernel bank and prior parameter $\boldsymbol{\theta}$. We use the main-text projector notation $P(\boldsymbol{\beta}) = U(\boldsymbol{\beta})U(\boldsymbol{\beta})^\top$, which removes the arbitrary choice of basis in the retained eigenspace. Vector norms are Euclidean; matrix norms carry their usual subscripts.

① **Let us first derive the weight response for a fixed subspace U.** Completing the square changes the objective only by a term independent of $\boldsymbol{\beta}$. The definition of Euclidean projection therefore gives

$$\begin{aligned}\boldsymbol{\beta} \in \operatorname*{argmin}_{\boldsymbol{b}\in\Delta_M} \left\{ \frac{\mu}{2}\|\boldsymbol{b}-\boldsymbol{\beta}_0\|_2^2 - \boldsymbol{s}(U)^\top \boldsymbol{b} \right\} &\iff \boldsymbol{\beta} \in \operatorname*{argmin}_{\boldsymbol{b}\in\Delta_M} \|\boldsymbol{b}-\boldsymbol{\beta}_0-\mu^{-1}\boldsymbol{s}(U)\|_2^2 \\ &\iff \boldsymbol{\beta} = \Pi_{\Delta_M}[\boldsymbol{\beta}_0 + \mu^{-1}\boldsymbol{s}(U)].\end{aligned} \tag{17}$$

Here $\boldsymbol{b}$ is only the minimization variable. The quadratic has Hessian μI_M, so its constrained minimizer is unique. This argument holds with U fixed. Once U depends on the weights, uniqueness of the coupled response requires the contraction argument below.

② **Next, we bound the change in the coupled response $\mathcal{R}_{\boldsymbol{\theta}}$.** For $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathcal{B}$, the triangle and Cauchy–Schwarz inequalities yield

$$\|K(\boldsymbol{\beta}) - K(\boldsymbol{\beta}')\|_2 \le \sum_j |\beta_j - \beta_j'| \|K_j\|_2 \le \kappa \|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_2. \tag{18}$$

The uniform gap $\underline{\gamma}$ and the population-gap form of the Davis–Kahan bound (Yu et al., 2015) imply

$$\|P(\boldsymbol{\beta}) - P(\boldsymbol{\beta}')\|_F \le \frac{2\sqrt{2r}}{\underline{\gamma}} \|K(\boldsymbol{\beta}) - K(\boldsymbol{\beta}')\|_2 \le \bar{L}\kappa \|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_2. \tag{19}$$

The factor $\sqrt{2}$ converts the principal-angle bound into a projector bound. Writing $s_j(P) = \operatorname{Tr}(K_j P)$, Frobenius duality gives $|\operatorname{Tr}(K_j(P-P'))| \le \|K_j\|_F \|P-P'\|_F$. Summing the squared component bounds yields

$$\mu^{-1}\|\boldsymbol{s}(P) - \boldsymbol{s}(P')\|_2 = \frac{\sqrt{\sum_j |\operatorname{Tr}(K_j(P-P'))|^2}}{\mu} \le B_K \|P-P'\|_F. \tag{20}$$

Euclidean projection onto the simplex is nonexpansive. Applying this property and Eqs. (19)–(20) gives

$$\|\mathcal{R}_{\boldsymbol{\theta}}(\boldsymbol{\beta}) - \mathcal{R}_{\boldsymbol{\theta}}(\boldsymbol{\beta}')\|_2 \le B_K \bar{L}\kappa \|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_2 = \chi \|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_2. \tag{21}$$

③ **We now establish the unique fixed point and the damped convergence rate.** The nonempty compact set $\mathcal{B}$ is complete, and $\mathcal{R}_{\boldsymbol{\theta}}$ maps it into itself. Since $\chi < 1$, the contraction mapping theorem gives a unique fixed point $\boldsymbol{\beta}^\star \in \mathcal{B}$. Convexity keeps each damped update in $\mathcal{B}$. Subtracting the fixed-point identity from Eq. (9) and applying Eq. (21) gives the one-step factor $(1-\alpha)+\alpha\chi$. Iteration yields

$$\|\boldsymbol{\beta}^{(s)} - \boldsymbol{\beta}^\star\|_2 \le [1-\alpha(1-\chi)]^s \|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^\star\|_2, \qquad 0 < \alpha \le 1. \tag{22}$$

For $\alpha > 0$, the damped and undamped maps have exactly the same fixed points:

$$\boldsymbol{\beta} = (1-\alpha)\boldsymbol{\beta} + \alpha\mathcal{R}_{\boldsymbol{\theta}}(\boldsymbol{\beta}) \quad\iff\quad \alpha[\boldsymbol{\beta} - \mathcal{R}_{\boldsymbol{\theta}}(\boldsymbol{\beta})] = \mathbf{0} \quad\iff\quad \boldsymbol{\beta} = \mathcal{R}_{\boldsymbol{\theta}}(\boldsymbol{\beta}). \tag{23}$$

Equation (19) also bounds the projector error along this sequence.

④ **Let us differentiate the fixed-point equation on the simplex tangent space.** Let $T \in \mathbb{R}^{M\times(M-1)}$ have orthonormal columns spanning $\mathbf{1}^\perp$, so $TT^\top = I_M - \mathbf{1}\mathbf{1}^\top/M$. Use β_T, θ_T for local tangent coordinates, as in Eq. (11). Both $\boldsymbol{\beta}$ and its response sum to one; hence their difference lies in the range of T, and

$$\boldsymbol{\beta} = \mathcal{R}_{\boldsymbol{\theta}}(\boldsymbol{\beta}) \quad\iff\quad T^\top[\boldsymbol{\beta} - \mathcal{R}_{\boldsymbol{\theta}}(\boldsymbol{\beta})] = \mathbf{0}. \tag{24}$$

The gap and unchanged face give a continuously differentiable tangent residual, whose partial derivatives are $I - J_\beta$ and $-J_\theta$. The same spectral and projection estimates imply $\|J_\beta\|_2 \le \chi < 1$.

Thus

$$(I - J_\beta)^{-1} = \sum_{t=0}^{\infty} J_\beta^t, \qquad \|(I - J_\beta)^{-1}\|_2 \le (1-\chi)^{-1}. \tag{25}$$

The implicit function theorem supplies a differentiable local response. Differentiating Eq. (24) and using invertibility gives

$$(I - J_\beta) D_{\theta_T} \beta_T^\star = J_\theta \quad \Longleftrightarrow \quad D_{\theta_T} \beta_T^\star = (I - J_\beta)^{-1} J_\theta. \tag{26}$$

Returning to the original coordinates yields $D_\theta \boldsymbol{\beta}^\star = T(I - J_\beta)^{-1} J_\theta T^\top$. The rightmost $T^\top$ follows from softmax shift invariance: changing $\boldsymbol{\theta}$ along $\mathbf{1}$ leaves the prior unchanged. Applying the chain rule to $\bar{\Phi}$ gives Eq. (11); the finite-difference and acceptance checks concern its numerical use.

⑤ **We next bound the response error using the fixed-point residual.** For $\widehat{\boldsymbol{\beta}} \in \mathcal{B}$, insert and subtract $\mathcal{R}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\beta}})$. The triangle inequality and Eq. (21) give the left-hand inequality below. Since $1 - \chi > 0$, rearranging it is equivalent to the stated error bound:

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 \le \eta_{\mathrm{fp}} + \chi \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 \quad \Longleftrightarrow \quad \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 \le \frac{\eta_{\mathrm{fp}}}{1-\chi}. \tag{27}$$

This completes the proof.

Proof of Theorem 2

Proof. Freeze the prior, bank construction, and decoding choices as in Assumption 2. Let $\widetilde{\boldsymbol{\beta}}, \widetilde{P}$ denote the exact re-optimized perturbed response. We retain one auxiliary projector, $P_0 = P(\widehat{\boldsymbol{\beta}})$, to distinguish the exact eigenspace at the computed nominal weights from its numerical approximation $\widehat{P}$. The nominal gap γ_r belongs to $K(\widehat{\boldsymbol{\beta}})$.

① **Let us first bound the direct change in the alignment scores.** For $E_j = \widetilde{K}_j - K_j$, the endpoint approximation allowances give $\|E_j\|_2 \le \varepsilon_j + 2\varrho_j = \widetilde{\varepsilon}_j$. Any rank-$r$ orthogonal projector P has nuclear norm r, so

$$|\operatorname{Tr}(E_j P)| \le r\|E_j\|_2, \qquad \mu^{-1}\|(\operatorname{Tr}(E_j P))_j\|_2 \le r\|\widetilde{\boldsymbol{\varepsilon}}\|_2/\mu = A_\varepsilon. \tag{28}$$

If no uniform approximation bound is supplied, $\varrho_j = 0$ and the argument concerns the frozen approximate operators.

② **Next, we bound the coupled changes in the weights and projector.** The nominal fixed-point defect is bounded by $\sigma\eta_{\mathrm{fp}}$ under Assumption 2. Compare the perturbed fixed-point equation with the nominal response at $\widehat{\boldsymbol{\beta}}$. Simplex nonexpansiveness, followed by splitting $\widetilde{K}_j = K_j + E_j$, gives

$$\begin{aligned} \|\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2 &\le \sigma\eta_{\mathrm{fp}} + \mu^{-1}\big\|(\operatorname{Tr}(K_j(\widetilde{P} - P_0)) + \operatorname{Tr}(E_j\widetilde{P}))_j\big\|_2 \\ &\le \sigma\eta_{\mathrm{fp}} + B_K\|\widetilde{P} - P_0\|_F + A_\varepsilon. \end{aligned} \tag{29}$$

The prior is the same at both endpoints. For the fused kernels, the difference decomposes into $\sum_j(\widetilde{\beta}_j - \widehat{\beta}_j)K_j + \sum_j \widetilde{\beta}_j E_j$. Since the perturbed weights belong to the simplex,

$$\|\widetilde{K}(\widetilde{\boldsymbol{\beta}}) - K(\widehat{\boldsymbol{\beta}})\|_2 \le \kappa\|\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2 + \sum_j \widetilde{\beta}_j\widetilde{\varepsilon}_j \le \kappa\|\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2 + D_\varepsilon. \tag{30}$$

The population-gap bound multiplies the right-hand side by L to bound $\|\widetilde{P} - P_0\|_F$. Substituting Eq. (29) and using $\chi = L\kappa B_K < 1$ gives the following equivalent inequalities:

$$\begin{aligned} & \|\widetilde{P} - P_0\|_F \le \chi\|\widetilde{P} - P_0\|_F + L(D_\varepsilon + \kappa A_\varepsilon) + \sigma L\kappa\eta_{\mathrm{fp}} \\ \Longleftrightarrow \quad & (1-\chi)\|\widetilde{P} - P_0\|_F \le L(D_\varepsilon + \kappa A_\varepsilon) + \sigma L\kappa\eta_{\mathrm{fp}} \\ \Longleftrightarrow \quad & \|\widetilde{P} - P_0\|_F \le \frac{L(D_\varepsilon + \kappa A_\varepsilon) + \sigma L\kappa\eta_{\mathrm{fp}}}{1-\chi}. \end{aligned} \tag{31}$$

This includes weight re-optimization. With exact solvers the same inequalities give $R_K = (D_\varepsilon + \kappa A_\varepsilon)/(1-\chi)$. Assumption 2 places this budget in the regular spectral regime. The projector estimate uses the nominal operator gap, without substituting an unchecked numerical gap.

③ **We now include the eigensolver error in the projector bound.** The valid eigensolver-error clause of Assumption 2 supplies

$$\|\widehat{P} - P_0\|_F \le 2\sqrt{2}\,\sigma\eta_{\text{eig}}/\gamma_r. \tag{32}$$

For clarity, let $U_\perp, \Lambda_\perp$ describe the complementary nominal eigenspace, and put $\widehat{\Lambda} = \widehat{U}^\top K\widehat{U}$, where $K = K(\widehat{\boldsymbol{\beta}})$. Multiplying the eigensolver residual by $U_\perp^\top$ yields

$$\Lambda_\perp(U_\perp^\top\widehat{U}) - (U_\perp^\top\widehat{U})\widehat{\Lambda} = U_\perp^\top(K\widehat{U} - \widehat{U}\widehat{\Lambda}). \tag{33}$$

When the retained Ritz values and complementary eigenvalues are separated by at least $\gamma_r/(2\sigma)$, the Sylvester equation bounds $\|U_\perp^\top\widehat{U}\|_F$ by $2\sigma\eta_{\text{eig}}/\gamma_r$. Multiplying by $\sqrt{2}$ proves Eq. (32) under this sufficient separation check. A small residual alone does not identify the intended spectral cluster. Combining the two projector errors and using $0 \le \chi < 1$ gives

$$\|\widetilde{P} - \widehat{P}\|_F \le \|\widetilde{P} - P_0\|_F + \|P_0 - \widehat{P}\|_F \le \frac{L(D_\varepsilon + \kappa A_\varepsilon) + \sigma L\kappa\eta_{\text{fp}} + 2\sqrt{2}\,\sigma\eta_{\text{eig}}/\gamma_r}{1-\chi} = R_P. \tag{34}$$

The last step conservatively divides the eigensolver contribution by $1-\chi$ as well. If the perturbed endpoint is also numerical, its errors must be included in valid total residual budgets.

④ **Let us convert projector error to normalized row error.** Use the main-text orthogonal rotation R to align $\widetilde{U}$ with $\widehat{U}$ by Procrustes minimization. For principal angles $\vartheta_1, \dots, \vartheta_r$,

$$\|\widetilde{U}R - \widehat{U}\|_F^2 = 2\sum_{a=1}^{r}(1 - \cos\vartheta_a) \le 2\sum_{a=1}^{r}\sin^2\vartheta_a = \|\widetilde{P} - \widehat{P}\|_F^2 \le R_P^2. \tag{35}$$

In particular, $\|\widetilde{u}_i R - \widehat{u}_i\|_2 \le R_P < \nu_i$ implies $\|\widetilde{u}_i\|_2 \ge \nu_i - R_P > 0$. Thus every perturbed row can be normalized. Adding and subtracting $\widetilde{u}_i R/\nu_i$ and using the reverse triangle inequality gives

$$\|\widetilde{z}_i R - \widehat{z}_i\|_2 \le \frac{\|\widetilde{u}_i R - \widehat{u}_i\|_2 + |\|\widetilde{u}_i\|_2 - \nu_i|}{\nu_i} \le \frac{2\|\widetilde{u}_i R - \widehat{u}_i\|_2}{\nu_i}. \tag{36}$$

A common right rotation preserves distances and rotates each core consistently because its coefficients $A_{:c}$ stay fixed.

⑤ **Next, we bound the change in each sample-to-core distance.** The coefficients of the difference between sample i and core c are $\mathbf{1}\{\ell = i\} - A_{\ell c}$. Their squared, row-scaled norm is $\sum_\ell(\mathbf{1}\{\ell = i\} - A_{\ell c})^2/\nu_\ell^2 = h_{ic}$. Applying Eq. (36) to the weighted row sum and then using Cauchy–Schwarz gives

$$\|(\widetilde{z}_i - \widetilde{g}_c)R - (\widehat{z}_i - \widehat{g}_c)\|_2 \le 2\sqrt{h_{ic}}\|\widetilde{U}R - \widehat{U}\|_F \le 2R_P\sqrt{h_{ic}} = \delta_{ic}. \tag{37}$$

Indeed, expanding the sum gives $\nu_i^{-2} + \sum_\ell A_{\ell c}^2\nu_\ell^{-2} - 2A_{ic}\nu_i^{-2} = h_{ic} \ge 0$. The cross term accounts for the sample's possible contribution to its own core. The reverse triangle inequality now gives $|\sqrt{\widetilde{D}_{ic}} - \sqrt{D_{ic}}| \le \delta_{ic}$. Factoring the difference of squares, and using $\sqrt{\widetilde{D}_{ic}} \le \sqrt{D_{ic}} + \delta_{ic}$, yields

$$|\widetilde{D}_{ic} - D_{ic}| \le \delta_{ic}(\sqrt{\widetilde{D}_{ic}} + \sqrt{D_{ic}}) \le 2\sqrt{D_{ic}}\,\delta_{ic} + \delta_{ic}^2 = e_{ic}. \tag{38}$$

The core representatives are recomputed from the perturbed embedding using the same A; their coordinates need not remain unchanged. This is why the allowance includes both the sample and its core.

⑥ **Finally, we use the distance bounds to establish assignment invariance.** The strict margin test is equivalently a separation between the nominal distance intervals:

$$D_{id} - D_{ic_i} > e_{ic_i} + e_{id} \quad\Longleftrightarrow\quad D_{ic_i} + e_{ic_i} < D_{id} - e_{id}, \qquad d \ne c_i. \tag{39}$$

For each competing core, the two-sided distance bounds therefore give

$$\widetilde{D}_{id} - \widetilde{D}_{ic_i} \ge (D_{id} - e_{id}) - (D_{ic_i} + e_{ic_i}) = D_{id} - D_{ic_i} - e_{id} - e_{ic_i} > 0. \tag{40}$$

Thus each passing sample retains its unique nearest-core label uniformly over admissible perturbations. For $i, i' \in S$, we obtain $\mathbf{1}\{\widetilde{y}_i = \widetilde{y}_{i'}\} = \mathbf{1}\{y_i = y_{i'}\}$, proving invariance of $C(\boldsymbol{y})_{S,S}$ and of the full matrix when $S = [n]$. A failed test gives no invariance claim for that sample.

E ADDITIONAL CLUSTERING RESULTS

Table 8: ARI comparison of BC-MKC with 22 reproduced MKC baselines on ten benchmarks (%). Δ denotes the gain over the best baseline in percentage points.

Algorithms	BBCSport	CUB	P-Fold	Reuters	Scene15	Fashion	ALOI	YTF10	N-MNIST	Winnipeg
LMKKM	57.2 ± 0.3	51.2 ± 0.4	26.5 ± 1.4	22.1 ± 2.2	29.4 ± 1.7	60.7 ± 1.2	61.6 ± 0.7	64.2 ± 1.3	52.1 ± 2.4	31.6 ± 0.7
RMKKM	61.6 ± 1.2	55.3 ± 0.1	25.6 ± 2.7	23.6 ± 1.2	26.5 ± 0.9	61.6 ± 0.7	59.3 ± 0.2	62.8 ± 1.2	50.3 ± 0.7	33.4 ± 1.2
MKKM-MR	48.4 ± 0.3	55.6 ± 2.8	29.2 ± 0.3	25.3 ± 0.8	27.0 ± 1.2	58.9 ± 2.4	58.3 ± 1.2	67.3 ± 0.5	48.4 ± 0.9	28.3 ± 2.4
LKAM	54.1 ± 2.7	47.1 ± 0.6	28.6 ± 0.6	23.7 ± 2.5	30.1 ± 0.6	54.3 ± 0.6	60.3 ± 2.4	61.4 ± 1.7	50.2 ± 1.3	34.8 ± 1.8
ONKC	47.2 ± 1.8	44.2 ± 1.7	27.8 ± 1.1	24.2 ± 0.7	26.2 ± 0.5	60.1 ± 1.4	56.2 ± 1.0	59.6 ± 2.2	49.2 ± 0.4	30.2 ± 0.9
NKSS	56.8 ± 0.9	50.2 ± 0.8	28.4 ± 1.7	27.8 ± 1.5	24.3 ± 0.8	57.4 ± 0.8	57.4 ± 1.5	62.4 ± 1.0	49.3 ± 1.2	36.5 ± 1.8
SS-MKC	61.6 ± 0.4	55.0 ± 2.6	23.1 ± 0.4	23.6 ± 1.4	29.7 ± 0.4	59.2 ± 1.1	59.7 ± 0.9	63.7 ± 1.1	51.0 ± 1.3	32.0 ± 1.4
SimpleMKKM	63.2 ± 0.8	57.2 ± 0.8	29.3 ± 1.2	25.4 ± 0.8	28.5 ± 0.7	60.2 ± 1.0	60.1 ± 1.6	62.8 ± 0.9	48.6 ± 0.6	30.6 ± 0.5
MKKM-SR	58.7 ± 0.2	54.2 ± 1.7	24.2 ± 0.9	27.2 ± 2.3	27.8 ± 1.5	56.3 ± 1.5	56.3 ± 0.2	63.2 ± 1.4	50.3 ± 1.4	28.4 ± 2.1
MVC-LFA	62.4 ± 1.3	61.4 ± 0.5	30.6 ± 2.1	23.5 ± 1.8	29.2 ± 0.6	59.9 ± 0.3	58.7 ± 2.6	65.3 ± 0.8	53.3 ± 0.7	34.2 ± 1.9
OPLFMVC	56.5 ± 2.3	49.3 ± 1.1	28.9 ± 0.5	24.1 ± 0.9	28.3 ± 1.1	60.0 ± 1.7	63.3 ± 0.8	58.3 ± 2.6	49.1 ± 2.5	31.6 ± 0.6
HMKC	70.1 ± 1.5	53.2 ± 1.4	27.4 ± 1.3	21.3 ± 1.0	26.7 ± 2.3	54.2 ± 2.5	61.5 ± 0.9	62.3 ± 1.5	50.2 ± 1.5	32.6 ± 0.8
FMKKM	66.3 ± 0.6	56.2 ± 0.3	28.7 ± 0.7	23.0 ± 0.8	29.1 ± 1.5	57.5 ± 0.9	59.2 ± 1.7	61.4 ± 0.7	51.6 ± 0.3	33.1 ± 1.6
DLEFT-MKC	73.6 ± 0.7	61.5 ± 0.8	30.4 ± 1.5	27.9 ± 1.7	30.3 ± 0.7	60.2 ± 1.3	62.4 ± 0.8	66.5 ± 0.8	47.8 ± 0.6	31.5 ± 0.8
Approx. MKKM	68.7 ± 0.4	58.3 ± 1.9	25.8 ± 1.2	25.4 ± 1.1	26.4 ± 1.8	57.3 ± 1.3	60.2 ± 0.5	63.1 ± 1.8	48.3 ± 1.2	31.7 ± 2.6
SMKC	59.3 ± 1.2	49.2 ± 0.3	27.6 ± 1.0	24.2 ± 0.4	27.1 ± 1.0	59.6 ± 0.7	61.3 ± 1.4	66.9 ± 0.4	50.2 ± 1.4	33.1 ± 1.5
INMKC	71.3 ± 0.2	60.2 ± 1.1	30.5 ± 0.6	26.3 ± 1.3	28.6 ± 0.9	61.1 ± 1.2	62.1 ± 1.6	66.2 ± 2.3	49.5 ± 1.0	32.3 ± 0.9
COKE	64.0 ± 1.6	65.0 ± 0.5	29.5 ± 2.4	26.6 ± 1.4	30.2 ± 2.5	57.8 ± 1.8	58.9 ± 1.3	61.0 ± 1.2	51.2 ± 0.6	33.4 ± 1.3
SWMKKM	70.5 ± 0.5	58.4 ± 1.7	28.2 ± 1.8	24.4 ± 0.5	27.4 ± 1.4	59.3 ± 0.4	60.2 ± 1.8	62.3 ± 1.3	52.3 ± 2.4	27.4 ± 0.9
HFLSMKKM	68.8 ± 0.6	61.2 ± 0.4	28.6 ± 0.9	25.1 ± 1.9	28.8 ± 0.6	59.1 ± 1.1	57.4 ± 0.8	64.6 ± 0.7	48.9 ± 1.3	33.5 ± 1.1
sLGm	67.4 ± 1.3	53.3 ± 2.0	26.5 ± 0.5	26.2 ± 2.3	26.3 ± 0.7	58.2 ± 0.8	61.4 ± 1.7	63.4 ± 0.9	49.0 ± 0.7	32.6 ± 0.8
BMKC	70.2 ± 1.1	58.3 ± 0.2	27.4 ± 1.3	24.7 ± 1.3	27.2 ± 1.3	57.1 ± 0.7	60.4 ± 0.4	64.7 ± 1.1	51.3 ± 1.2	32.8 ± 1.2
BC-MKC	$\mathbf{85.4 \pm 0.4}$	$\mathbf{77.4 \pm 0.6}$	$\mathbf{39.6 \pm 0.4}$	$\mathbf{36.5 \pm 0.9}$	$\mathbf{38.3 \pm 0.9}$	$\mathbf{77.5 \pm 0.7}$	$\mathbf{79.6 \pm 0.8}$	$\mathbf{81.8 \pm 0.5}$	$\mathbf{63.8 \pm 0.6}$	$\mathbf{48.7 \pm 0.7}$
Δ (%)	$\mathbf{\Delta + 11.8}$	$\mathbf{\Delta + 12.4}$	$\mathbf{\Delta + 9.0}$	$\mathbf{\Delta + 8.6}$	$\mathbf{\Delta + 8.0}$	$\mathbf{\Delta + 15.9}$	$\mathbf{\Delta + 16.3}$	$\mathbf{\Delta + 14.5}$	$\mathbf{\Delta + 10.5}$	$\mathbf{\Delta + 12.2}$

Table 9: F1 comparison of BC-MKC with 22 reproduced MKC baselines on ten benchmarks (%). Δ denotes the gain over the best baseline in percentage points.

Algorithms	BBCSport	CUB	P-Fold	Reuters	Scene15	Fashion	ALOI	YTF10	N-MNIST	Winnipeg
LMKKM	65.3 ± 0.9	64.1 ± 0.5	26.9 ± 2.9	37.9 ± 0.8	22.8 ± 3.2	59.7 ± 3.3	73.0 ± 2.1	64.7 ± 2.7	63.4 ± 3.4	49.0 ± 0.5
RMKKM	69.0 ± 1.9	73.3 ± 1.3	25.5 ± 1.2	39.3 ± 1.1	36.7 ± 1.4	73.7 ± 1.5	62.9 ± 3.2	68.0 ± 2.4	53.6 ± 2.1	62.6 ± 3.3
MKKM-MR	65.2 ± 2.8	59.8 ± 0.3	28.9 ± 0.2	37.2 ± 1.8	34.4 ± 2.4	65.0 ± 1.0	63.0 ± 1.5	74.1 ± 2.3	56.2 ± 2.2	59.5 ± 3.0
LKAM	79.6 ± 2.4	72.2 ± 1.2	24.3 ± 0.5	30.5 ± 2.0	29.0 ± 1.4	66.2 ± 1.2	73.8 ± 0.5	71.7 ± 0.4	65.2 ± 0.1	49.1 ± 2.4
ONKC	68.4 ± 0.8	72.9 ± 0.7	33.7 ± 1.9	39.6 ± 2.1	37.9 ± 2.3	64.4 ± 3.4	69.5 ± 2.7	69.4 ± 2.1	52.6 ± 0.5	61.5 ± 2.1
NKSS	74.9 ± 2.9	67.1 ± 2.4	16.6 ± 2.3	25.2 ± 1.7	23.2 ± 2.2	72.0 ± 1.8	65.3 ± 2.2	70.7 ± 0.4	54.1 ± 2.0	52.7 ± 1.4
SS-MKC	76.0 ± 2.5	69.4 ± 1.8	30.4 ± 3.2	28.1 ± 1.7	29.9 ± 2.2	69.5 ± 3.2	71.4 ± 2.4	72.1 ± 2.3	58.0 ± 3.2	51.7 ± 2.9
SimpleMKKM	69.9 ± 0.9	69.6 ± 0.6	26.8 ± 0.6	31.9 ± 1.4	27.8 ± 0.9	71.5 ± 3.3	66.2 ± 1.8	71.1 ± 1.3	55.5 ± 3.3	63.1 ± 3.3
MKKM-SR	75.3 ± 2.8	63.9 ± 0.8	17.2 ± 1.4	38.9 ± 2.9	27.5 ± 0.3	60.9 ± 0.9	74.2 ± 1.1	62.4 ± 3.2	59.5 ± 0.8	59.7 ± 1.0
MVC-LFA	76.6 ± 3.4	74.2 ± 2.0	29.0 ± 3.1	32.8 ± 1.4	35.3 ± 1.5	72.5 ± 1.3	69.6 ± 3.3	61.4 ± 2.5	64.3 ± 3.2	60.9 ± 2.0
OPLFMVC	66.2 ± 1.2	71.5 ± 1.5	27.5 ± 1.1	31.2 ± 2.6	29.4 ± 3.3	76.0 ± 1.9	65.4 ± 3.2	73.7 ± 1.6	64.1 ± 1.0	59.4 ± 2.2
HMKC	63.3 ± 1.6	73.3 ± 3.1	25.6 ± 2.8	27.7 ± 2.8	32.4 ± 1.2	70.3 ± 1.6	69.8 ± 3.3	68.9 ± 3.1	52.1 ± 0.2	50.2 ± 0.5
FMKKM	65.7 ± 2.7	68.5 ± 1.1	29.3 ± 2.7	33.6 ± 1.7	27.3 ± 2.6	64.4 ± 2.5	67.7 ± 2.2	72.0 ± 2.5	62.0 ± 3.1	52.3 ± 3.1
DLEFT-MKC	75.0 ± 2.9	61.7 ± 0.3	25.8 ± 0.9	24.8 ± 0.9	28.2 ± 3.3	61.8 ± 0.6	62.6 ± 2.1	68.7 ± 2.1	53.4 ± 3.2	55.0 ± 3.2
Approx. MKKM	73.2 ± 0.6	68.9 ± 2.4	25.7 ± 0.4	29.2 ± 1.1	31.7 ± 0.8	70.6 ± 2.4	60.7 ± 2.2	77.2 ± 2.3	56.1 ± 2.9	61.4 ± 0.9
SMKC	69.4 ± 2.1	73.0 ± 1.5	25.1 ± 1.6	26.9 ± 3.3	25.9 ± 1.3	75.7 ± 0.5	65.9 ± 2.4	71.8 ± 2.1	61.1 ± 0.6	57.7 ± 0.6
INMKC	67.2 ± 1.7	73.1 ± 0.7	24.5 ± 0.6	32.3 ± 0.4	36.7 ± 2.5	75.4 ± 0.4	72.9 ± 1.4	69.7 ± 0.8	56.6 ± 3.0	63.2 ± 1.5
COKE	65.8 ± 0.9	69.0 ± 0.7	28.4 ± 0.2	37.2 ± 1.2	41.1 ± 2.0	68.3 ± 2.3	74.0 ± 2.3	64.6 ± 1.1	51.7 ± 0.9	50.2 ± 0.7
SWMKKM	74.2 ± 0.5	75.5 ± 1.6	28.4 ± 1.4	28.7 ± 1.9	31.2 ± 1.6	64.4 ± 2.7	73.9 ± 1.3	74.1 ± 1.8	66.2 ± 1.9	53.3 ± 3.1
HFLSMKKM	72.0 ± 1.7	70.6 ± 2.2	23.1 ± 1.3	33.4 ± 2.2	33.0 ± 0.8	73.2 ± 1.7	61.0 ± 0.9	64.3 ± 2.9	59.6 ± 1.6	50.8 ± 2.8
sLGm	73.0 ± 2.8	74.4 ± 3.1	20.6 ± 2.6	39.1 ± 2.7	27.5 ± 2.2	57.8 ± 2.8	68.2 ± 3.2	69.7 ± 0.6	64.7 ± 0.8	54.0 ± 2.9
BMKC	69.8 ± 2.3	63.1 ± 0.3	21.5 ± 2.0	34.5 ± 2.4	32.1 ± 0.6	62.7 ± 1.3	64.4 ± 2.7	72.4 ± 2.0	62.5 ± 2.8	55.4 ± 1.6
BC-MKC	$\mathbf{90.6 \pm 0.3}$	$\mathbf{88.7 \pm 0.4}$	$\mathbf{43.8 \pm 0.5}$	$\mathbf{52.3 \pm 0.3}$	$\mathbf{50.5 \pm 0.5}$	$\mathbf{87.4 \pm 0.5}$	$\mathbf{86.9 \pm 0.9}$	$\mathbf{89.0 \pm 0.7}$	$\mathbf{79.5 \pm 0.6}$	$\mathbf{75.7 \pm 0.5}$
Δ (%)	$\mathbf{\Delta + 11.0}$	$\mathbf{\Delta + 13.2}$	$\mathbf{\Delta + 10.1}$	$\mathbf{\Delta + 12.7}$	$\mathbf{\Delta + 9.4}$	$\mathbf{\Delta + 11.0}$	$\mathbf{\Delta + 12.7}$	$\mathbf{\Delta + 11.8}$	$\mathbf{\Delta + 13.3}$	$\mathbf{\Delta + 12.5}$

Table 10: NMI comparison of BC-MKC with 22 reproduced MKC baselines on ten benchmarks (%). Δ denotes the gain over the best baseline in percentage points.

Algorithms	BBCSport	CUB	P-Fold	Reuters	Scene15	Fashion	ALOI	YTF10	N-MNIST	Winnipeg
LMKKM	70.2 ± 1.1	61.1 ± 1.4	29.9 ± 0.7	22.0 ± 1.4	28.9 ± 3.2	56.3 ± 1.9	78.5 ± 0.8	73.9 ± 1.8	53.4 ± 1.1	29.7 ± 3.2
RMKKM	62.2 ± 0.4	54.1 ± 1.1	25.7 ± 3.2	20.3 ± 3.1	25.4 ± 2.5	63.1 ± 0.4	66.3 ± 2.5	77.0 ± 0.5	44.5 ± 2.4	33.0 ± 0.5
MKKM-MR	65.0 ± 0.3	58.3 ± 1.6	22.7 ± 1.2	27.8 ± 0.7	24.1 ± 1.9	57.6 ± 1.9	58.8 ± 2.4	66.7 ± 0.2	46.7 ± 0.8	33.3 ± 1.1
LKAM	57.4 ± 1.5	61.9 ± 2.2	34.1 ± 2.8	21.9 ± 2.7	20.1 ± 3.0	63.3 ± 1.4	69.3 ± 3.3	77.1 ± 0.5	55.0 ± 0.9	29.0 ± 1.3
ONKC	65.1 ± 1.4	66.8 ± 2.6	27.7 ± 3.2	26.2 ± 1.2	24.8 ± 1.1	58.1 ± 2.1	77.2 ± 1.8	69.1 ± 2.1	46.2 ± 2.8	33.6 ± 1.1
NKSS	66.5 ± 0.4	55.6 ± 0.3	29.1 ± 2.7	22.5 ± 0.6	29.2 ± 1.4	61.4 ± 1.0	71.6 ± 3.2	68.5 ± 1.7	55.9 ± 3.4	26.9 ± 2.5
SS-MKC	59.9 ± 1.7	72.1 ± 3.0	30.2 ± 3.0	21.4 ± 0.5	34.7 ± 0.4	57.9 ± 0.2	73.7 ± 0.9	69.8 ± 1.0	51.4 ± 0.5	37.3 ± 0.5
SimpleMKKM	68.6 ± 2.4	65.1 ± 0.6	23.3 ± 1.3	20.7 ± 3.2	25.4 ± 2.3	57.6 ± 1.6	72.8 ± 1.7	68.4 ± 2.1	48.8 ± 2.7	33.7 ± 2.4
MKKM-SR	62.6 ± 2.1	69.0 ± 2.3	30.4 ± 3.1	19.7 ± 2.2	24.6 ± 1.6	59.2 ± 0.2	76.1 ± 2.2	67.0 ± 2.6	57.8 ± 1.6	34.6 ± 1.2
MVC-LFA	55.9 ± 2.2	64.6 ± 2.8	33.7 ± 1.9	21.8 ± 2.9	30.8 ± 0.2	73.4 ± 0.7	59.8 ± 0.6	73.9 ± 2.0	57.7 ± 1.6	30.9 ± 0.8
OPLFMVC	62.0 ± 0.5	58.3 ± 3.1	30.5 ± 1.5	18.5 ± 2.7	27.6 ± 1.0	68.6 ± 0.9	79.1 ± 1.1	68.8 ± 2.7	54.5 ± 0.3	23.5 ± 2.3
HMKC	69.1 ± 2.9	59.7 ± 1.3	31.8 ± 1.4	20.5 ± 0.6	30.3 ± 1.1	59.0 ± 2.2	77.5 ± 3.2	76.2 ± 3.2	55.4 ± 1.3	32.2 ± 1.5
FMKKM	71.2 ± 3.0	64.8 ± 0.6	27.0 ± 1.0	20.8 ± 1.4	29.3 ± 0.6	63.8 ± 2.9	67.0 ± 0.3	69.4 ± 1.5	48.5 ± 3.0	26.4 ± 1.5
DLEFT-MKC	62.8 ± 0.3	58.6 ± 0.4	22.9 ± 3.1	21.2 ± 0.5	31.3 ± 2.5	71.5 ± 2.0	72.5 ± 0.5	62.6 ± 2.1	57.3 ± 2.1	35.0 ± 3.1
Approx. MKKM	68.0 ± 1.9	59.2 ± 3.3	31.3 ± 0.4	20.1 ± 1.9	33.0 ± 2.8	68.2 ± 1.2	67.5 ± 2.7	60.8 ± 1.4	51.7 ± 2.8	37.1 ± 2.4
SMKC	63.4 ± 0.8	68.8 ± 2.4	32.4 ± 0.4	19.7 ± 2.4	28.9 ± 2.4	56.1 ± 2.5	72.0 ± 0.4	73.0 ± 1.0	53.2 ± 1.2	26.0 ± 2.6
INMKC	65.4 ± 2.1	70.9 ± 0.5	28.6 ± 0.8	17.1 ± 1.6	25.0 ± 3.3	57.6 ± 2.0	74.1 ± 0.2	74.0 ± 2.2	59.1 ± 0.5	31.5 ± 2.6
COKE	59.3 ± 2.4	66.1 ± 1.7	25.1 ± 1.5	22.7 ± 2.1	27.7 ± 1.2	69.4 ± 1.1	78.1 ± 1.8	62.4 ± 0.4	48.3 ± 2.2	32.0 ± 0.8
SWMKKM	63.6 ± 2.8	68.2 ± 2.0	30.2 ± 0.5	22.0 ± 1.9	25.5 ± 0.8	59.5 ± 3.0	66.4 ± 2.0	71.3 ± 2.1	44.3 ± 3.0	30.1 ± 1.1
HFLSMKKM	56.5 ± 3.2	70.3 ± 3.2	31.5 ± 1.6	21.5 ± 2.2	29.7 ± 1.8	67.6 ± 2.0	64.6 ± 3.2	68.4 ± 0.7	60.4 ± 0.7	33.5 ± 1.1
sLGm	63.3 ± 2.4	64.6 ± 1.6	33.3 ± 1.4	19.4 ± 2.3	26.0 ± 0.3	70.4 ± 0.9	69.9 ± 1.7	69.8 ± 2.3	58.0 ± 1.8	35.6 ± 3.2
BMKC	62.4 ± 1.5	57.0 ± 1.0	31.1 ± 1.1	22.4 ± 2.5	26.3 ± 0.7	52.8 ± 2.4	70.2 ± 1.2	67.6 ± 2.9	51.0 ± 1.9	29.3 ± 3.4
BC-MKC	$\mathbf{82.6 \pm 0.6}$	$\mathbf{82.7 \pm 0.8}$	$\mathbf{46.1 \pm 0.7}$	$\mathbf{34.6 \pm 0.6}$	$\mathbf{43.9 \pm 0.4}$	$\mathbf{82.1 \pm 0.6}$	$\mathbf{91.7 \pm 0.4}$	$\mathbf{88.1 \pm 0.3}$	$\mathbf{70.4 \pm 0.6}$	$\mathbf{51.9 \pm 0.9}$
Δ (%)	$\mathbf{\Delta + 11.4}$	$\mathbf{\Delta + 10.6}$	$\mathbf{\Delta + 12.0}$	$\mathbf{\Delta + 6.8}$	$\mathbf{\Delta + 9.2}$	$\mathbf{\Delta + 8.7}$	$\mathbf{\Delta + 12.6}$	$\mathbf{\Delta + 11.0}$	$\mathbf{\Delta + 10.0}$	$\mathbf{\Delta + 14.6}$

F ADDITIONAL EXPERIMENTS

F.1 RUNNING TIME ON ADDITIONAL DATASETS

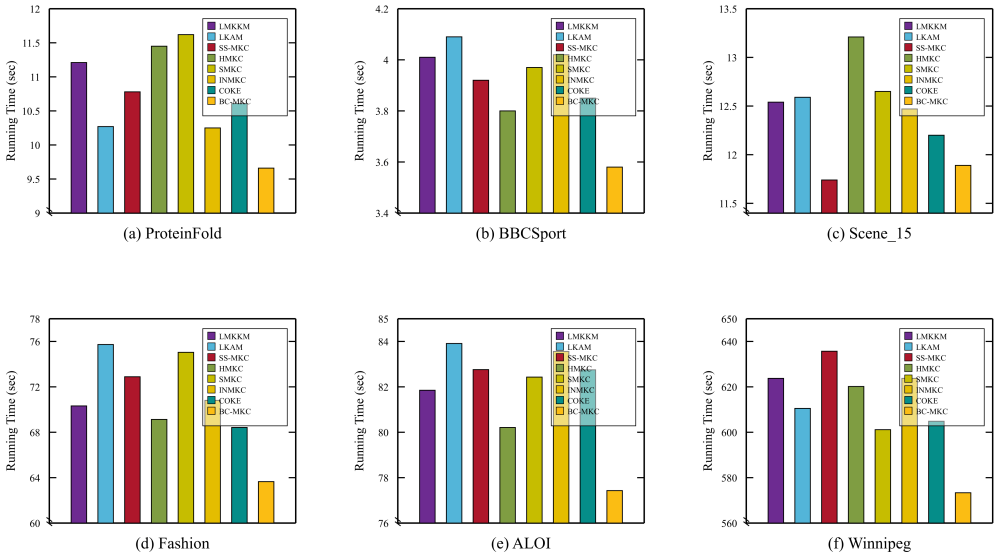


Figure 9: Running-time comparison (seconds) of BC-MKC and seven baselines on six additional datasets.

F.2 CONVERGENCE ON ADDITIONAL DATASETS

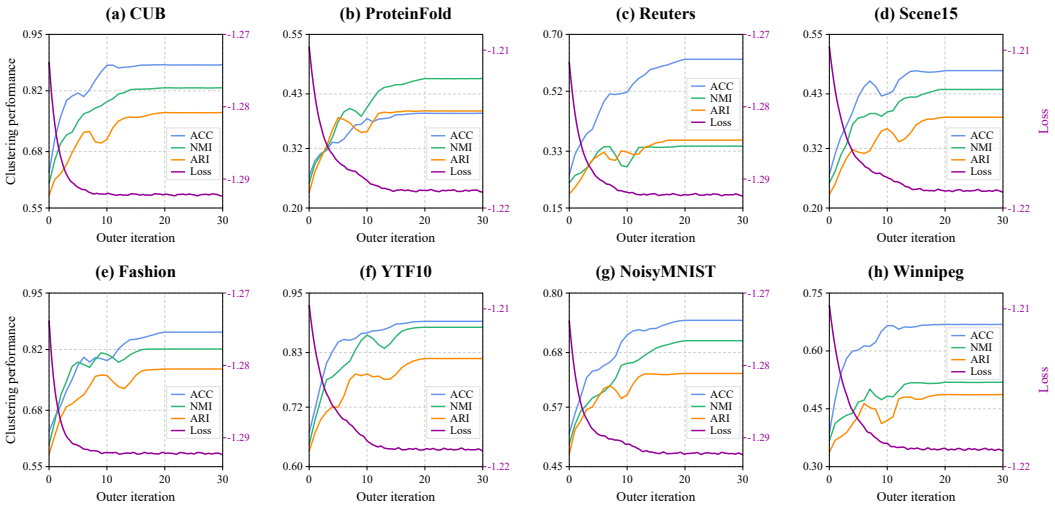


Figure 10: Evolution of loss and clustering performance (ACC, NMI, and ARI) during BC-MKC learning on eight additional datasets.

F.3 Additional ablation studies

Contribution of the consensus kernel. In Table 11, NoConsensus removes K_s from Eq. (5) and renormalizes the remaining initial prior weights. Full yields relative gains of 22.6% in ACC and 34.7% in ARI on ProteinFold. Higher scores across all four datasets support the consensus kernel's contribution to clustering quality.

Dataset	Variant	Method	ACC	NMI	F1	ARI
BBCSport	#1	**Full (Ours)**	$\mathbf{96.6 \pm 0.4}$	$\mathbf{82.6 \pm 0.6}$	$\mathbf{90.6 \pm 0.3}$	$\mathbf{85.4 \pm 0.4}$
BBCSport	#2	NoConsensus	87.4 ± 1.8	67.3 ± 2.3	77.0 ± 2.7	71.4 ± 3.9
CUB	#1	**Full (Ours)**	$\mathbf{88.0 \pm 0.5}$	$\mathbf{82.7 \pm 0.8}$	$\mathbf{88.7 \pm 0.4}$	$\mathbf{77.4 \pm 0.6}$
CUB	#2	NoConsensus	81.1 ± 0.4	72.3 ± 0.5	82.1 ± 0.22	72.7 ± 0.5
ProteinFold	#1	**Full (Ours)**	$\mathbf{39.1 \pm 2.0}$	$\mathbf{46.1 \pm 0.7}$	$\mathbf{43.8 \pm 0.5}$	$\mathbf{39.6 \pm 0.4}$
ProteinFold	#2	NoConsensus	31.9 ± 1.1	42.1 ± 0.2	35.8 ± 1.4	29.4 ± 0.8
NoisyMNIST	#1	**Full (Ours)**	$\mathbf{74.5 \pm 0.0}$	$\mathbf{70.4 \pm 0.6}$	$\mathbf{79.5 \pm 0.6}$	$\mathbf{63.8 \pm 0.6}$
NoisyMNIST	#2	NoConsensus	62.86 ± 0.8	62.5 ± 0.3	68.0 ± 0.3	57.5 ± 0.3

Table 11: Consensus-kernel ablation on four datasets.

Fixed training references. Table 12 compares fixed references with Refresh, which reselects them between outer steps. On Fashion, Full achieves 86.0% ACC and 82.1% NMI, versus 75.3% and 70.4% for Refresh. Higher scores on all three datasets support retaining fixed references during training.

Dataset	Variant	Method	ACC	NMI	F1	ARI
CUB	#1	**Full (Ours)**	$\mathbf{88.0 \pm 0.5}$	$\mathbf{82.7 \pm 0.8}$	$\mathbf{88.7 \pm 0.4}$	$\mathbf{77.4 \pm 0.6}$
CUB	#2	Refresh	82.4 ± 2.7	70.3 ± 1.2	75.3 ± 2.4	61.2 ± 3.1
Fashion	#1	**Full (Ours)**	$\mathbf{86.0 \pm 0.8}$	$\mathbf{82.1 \pm 0.6}$	$\mathbf{87.4 \pm 0.5}$	$\mathbf{77.5 \pm 0.7}$
Fashion	#2	Refresh	75.3 ± 1.4	70.4 ± 0.9	74.3 ± 1.8	64.1 ± 1.5
ALOI-100	#1	**Full (Ours)**	$\mathbf{84.6 \pm 0.1}$	$\mathbf{91.7 \pm 0.4}$	$\mathbf{86.9 \pm 0.9}$	$\mathbf{79.6 \pm 0.8}$
ALOI-100	#2	Refresh	73.0 ± 1.7	82.6 ± 0.3	77.5 ± 1.5	68.0 ± 1.1

Table 12: Training-reference ablation on three datasets.

Consensus-kernel structure visualization. We consider a controlled four-group comparison with 1,000 samples, 250 per group, using INMKC and BC-MKC. The comparison contrasts fusion without and with K_s, followed by the assignment indicator and the row-normalized class correspondence matrix. Samples share the same group-wise order in every panel; cluster IDs are matched to reference groups only for evaluation. Cosine-normalized affinities use a common scale, while dashed boxes mark within-group blocks.

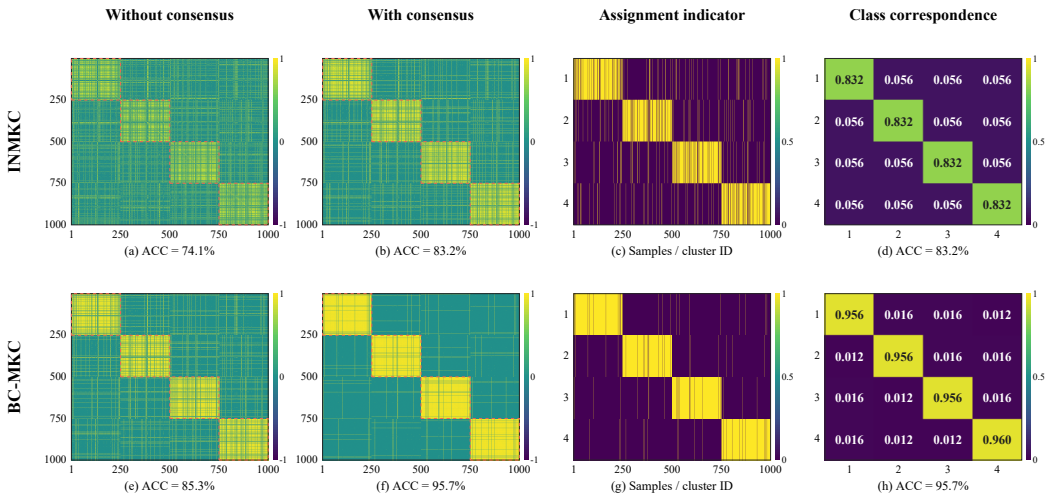


Figure 11: Consensus-kernel structure comparison for INMKC (top) and BC-MKC (bottom). Columns show affinities without and with K_s, assignment indicators after fusion with K_s, and class correspondence fractions (rows: reference groups; columns: assigned groups).

Visualization and analysis. Figure 11 highlights two structural differences. First, adding K_s reduces cross-group mixing in both rows: the displayed ACC gains are 9.1 percentage points for INMKC and 10.4 points for BC-MKC. Second, BC-MKC exhibits cleaner within-group blocks and fewer misplaced assignments, with 95.7% ACC versus 83.2% for INMKC. Its class correspondence matrix has a mean diagonal value of 0.957, reflecting more consistent grouping across the four classes. These patterns illustrate the intended benefit of consensus-guided fusion: strengthening shared cluster structure while limiting cross-group interference. Ultimately, this distinct structural clarity visually corroborates the enhanced clustering accuracy achieved by the proposed method.

F.4 LOWER-RESPONSE SENSITIVITY ON ADDITIONAL DATASETS

Figure 12 extends the 5×5 grid in Fig. 5 to the remaining eight datasets, varying $s_\mu = \mu/\mu_0$ in Eq. (6) and damping α in Eq. (9). Averaged over the other parameter and all eight datasets, ACC is 67.5% at $\mu/\mu_0 = 10^{-2}$ and 67.2% at 1, a relative difference of 0.5%. For damping, $\alpha = 0.8$ gives 67.6% versus 67.1% at 0.6, a relative difference of 0.7%. These minimal variations confirm that BC-MKC remains highly robust and practically insensitive to the choice of both parameters across the additional datasets.

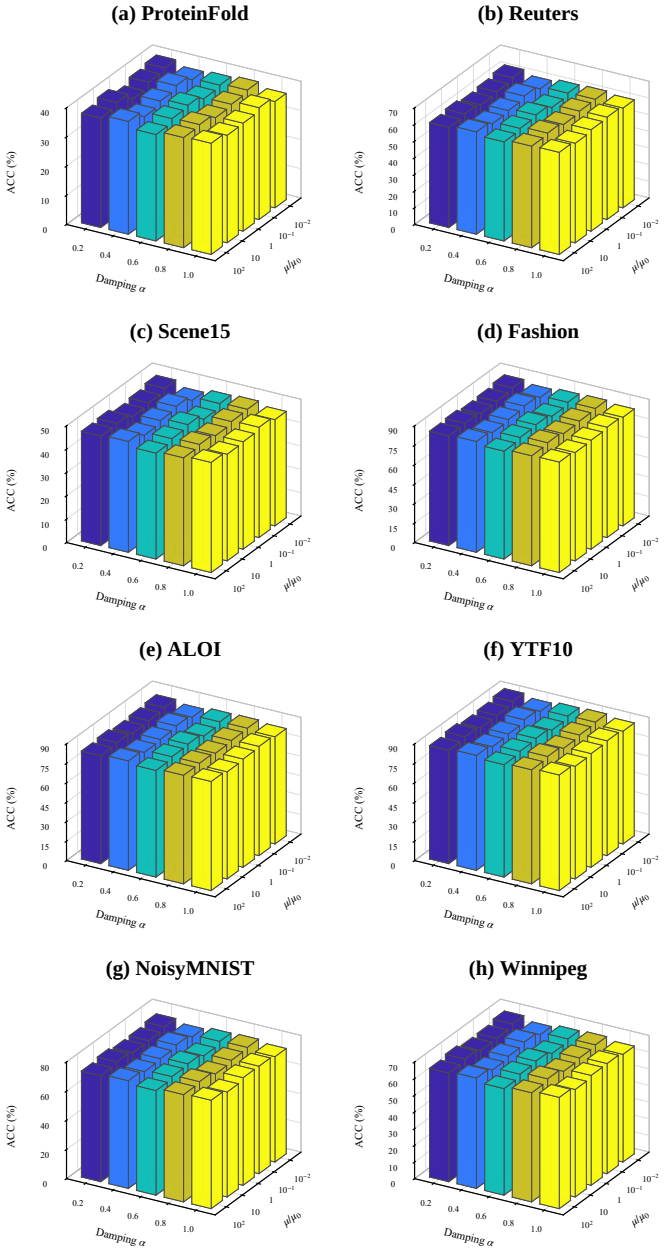


Figure 12: ACC sensitivity to lower-response regularization μ and damping α on eight additional datasets.

F.5 HYPERPARAMETER EXPERIMENTS AND STATISTICAL TESTS

Table 13: ACC sensitivity to w_m, w_g, and τ on three benchmarks.

(a) Sensitivity to w_m

w_m	CUB	Fashion	P-Fold
0.5	87.5±0.3	85.6±0.2	38.0±2.3
1.0	87.4±0.2	85.3±0.4	38.3±1.2
2.0	**88.0±0.5**	**86.0±0.8**	**39.1±2.0**
4.0	87.7±0.1	85.8±0.1	38.5±1.6
8.0	87.6±0.7	85.7±0.3	38.8±1.1

(b) Sensitivity to w_g

w_g	CUB	Fashion	P-Fold
0.125	87.6±0.6	85.2±0.9	38.4±2.0
0.25	87.3±0.2	85.5±0.7	38.6±2.3
0.5	**88.0±0.5**	**86.0±0.8**	**39.1±2.0**
1.0	87.7±0.7	85.3±0.4	38.4±1.7
2.0	87.2±0.4	85.6±0.1	38.7±2.2

(c) Sensitivity to τ

τ	CUB	Fashion	P-Fold
0.005	87.6±0.7	85.5±0.6	38.4±1.7
0.01	87.2±0.4	85.4±0.3	38.1±2.4
0.02	**88.0±0.5**	**86.0±0.8**	**39.1±2.0**
0.04	86.9±0.9	85.1±0.3	38.3±1.9
0.08	87.8±0.1	85.2±1.0	38.5±2.2

Margin and eigengap hyperparameters. Table 13 varies the margin weight w_m, eigengap weight w_g, and margin temperature τ in Eq. (7). The weights scale $\mathcal{M}_\tau$ and γ_r, respectively, while τ controls the smooth margin aggregation. We test $w_m \in [0.5, 8.0]$, $w_g \in [0.125, 2.0]$, and $\tau \in [0.005, 0.08]$, with defaults 2.0, 0.5, and 0.02. Averaged over CUB, Fashion, and ProteinFold, macro ACC ranges from 70.3% to 71.0%, 70.4% to 71.0%, and 70.1% to 71.0%, respectively. The default settings attain the highest mean ACC on each dataset in all three sweeps. These minimal variations confirm that BC-MKC's clustering performance remains highly robust and practically insensitive to the choice of these three hyperparameters.

Statistical significance tests.

Table 14: Statistical comparison of ACC and NMI on Fashion.

Method	ACC (%)↑			NMI (%)↑		
	mean±std	95% CI	p_{Holm}	mean±std	95% CI	p_{Holm}
SMKC	75.70 ± 1.10	[75.19, 76.21]	6.18×10^{-19}	56.10 ± 2.50	[54.93, 57.27]	9.78×10^{-21}
COKE	74.20 ± 1.30	[73.59, 74.81]	3.78×10^{-20}	69.40 ± 1.10	[68.89, 69.91]	1.08×10^{-20}
sLGm	73.40 ± 0.60	[73.12, 73.68]	1.84×10^{-25}	70.40 ± 0.90	[69.98, 70.82]	9.78×10^{-21}
BC-MKC	$\mathbf{86.00 \pm 0.80}$	[**85.63**, **86.37**]	—	$\mathbf{82.10 \pm 0.60}$	[**81.82**, **82.38**]	—

Table 14 reports means, sample standard deviations, and 95% confidence intervals for mean ACC and NMI over 20 paired observations per method. Each baseline is compared with BC-MKC using a two-sided paired t-test, with Holm correction across the six comparisons. BC-MKC exceeds SMKC, COKE, and sLGm by 10.3%, 11.8%, and 12.6% in ACC, and by 26.0%, 12.7%, and 11.7% in NMI, respectively. All six adjusted p-values are below 0.05. Within this comparison, BC-MKC has the highest mean scores on both metrics, with statistically significant differences from the three baselines.

F.6 GENERALITY OF THE CONSENSUS KERNEL

We use the consensus kernel K_s in Eq. (5) as a portable module and incorporate it into three existing kernel-fusion methods: MKKM-MR, FMKKM, and INMKC. Figure 13 compares their original and augmented variants on four datasets; "Method+BC-MKC" denotes the variant incorporating K_s. On CUB, incorporating K_s improves ACC by 8.9%, 7.1%, and 8.5% for MKKM-MR, FMKKM, and INMKC, respectively. All three methods also achieve higher ACC on the other tested datasets. These consistent gains support the portability of K_s across the three tested fusion methods and indicate that the consensus kernel can benefit clustering beyond the full BC-MKC framework.

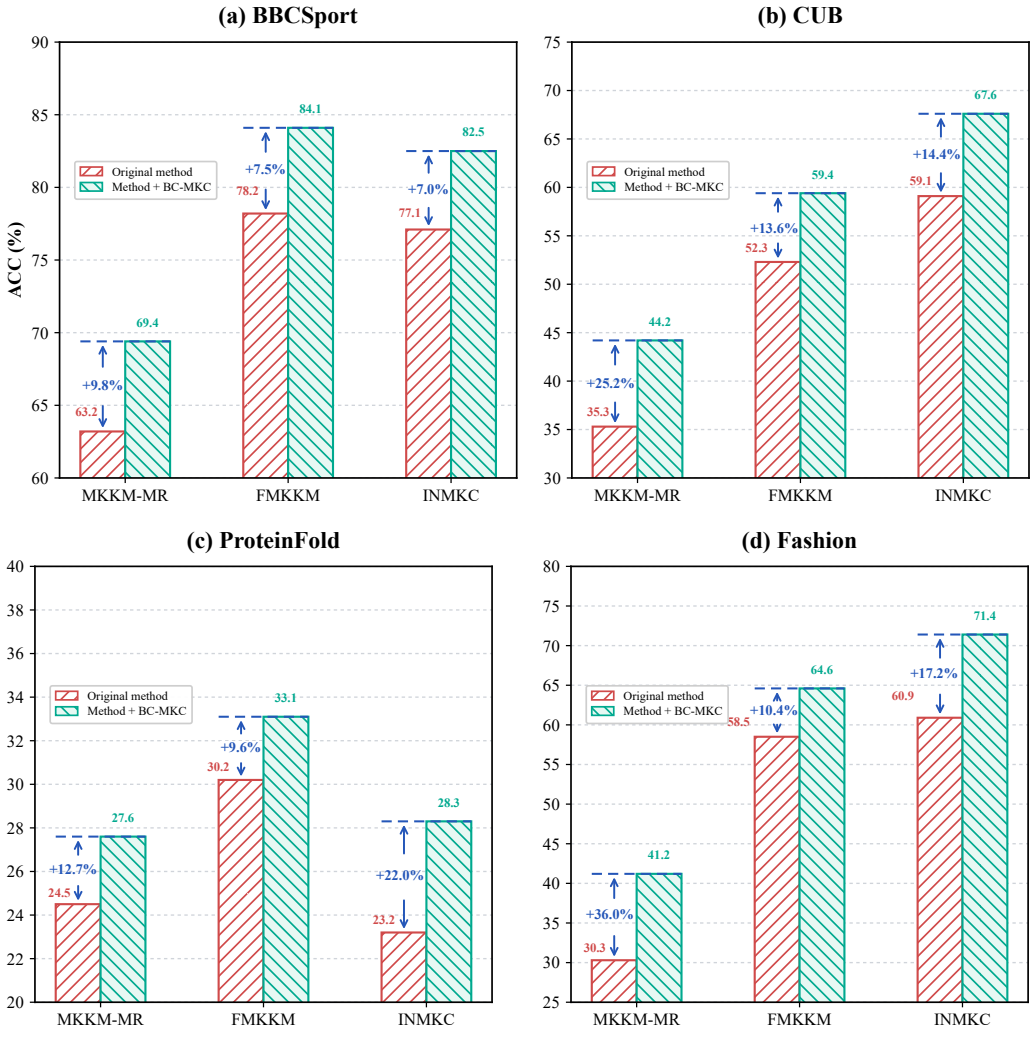


Figure 13: ACC comparison of three kernel-fusion methods with and without K_s.

F.7 COMPUTATIONAL EFFICIENCY

Table 15 compares computational cost and clustering performance on Fashion. BC-MKC uses 2.01 GB of memory and 17.22 G FLOPs, with a clustering time of 32.3 s, the lowest reported costs among the six methods. It also achieves the highest ACC, NMI, and ARI in this comparison. These results indicate a favorable balance between computational cost and clustering quality on the tested dataset.

Table 15: Computational cost and clustering performance on Fashion.

Method	Time (s)↓	Mem. (GB)↓	FLOPs (G)↓	ACC (%)↑	NMI (%)↑	ARI (%)↑
SimpleMKKM	40.5	4.24	32.62	61.0 ± 1.7	57.6 ± 1.6	60.2 ± 1.0
HMKC	39.1	3.82	25.51	67.1 ± 0.5	59.0 ± 2.2	54.2 ± 2.5
SMKC	37.4	3.17	28.73	75.7 ± 1.1	56.1 ± 2.5	59.6 ± 0.7
INMKC	38.3	4.06	31.45	60.9 ± 1.7	57.6 ± 2.0	61.1 ± 1.2
COKE	41.2	3.32	30.32	74.2 ± 1.3	69.4 ± 1.1	57.8 ± 1.8
BC-MKC	**32.3**	**2.01**	**17.22**	$\mathbf{86.0 \pm 0.8}$	$\mathbf{82.1 \pm 0.6}$	$\mathbf{77.5 \pm 0.7}$

G CLUSTERING VISUALIZATION ON N-MNIST

Figure 14 presents the two-dimensional visualizations of NoisyMNIST (N-MNIST) at iterations 8, 16, 24, and 30. The later snapshots show more compact groups with less visible mixing, alongside higher clustering scores. ACC increases by 6.3 percentage points and NMI by 3.7 points between iterations 8 and 30. The gains between iterations 24 and 30 are 0.3 points for both metrics, indicating that performance becomes more stable toward the end of the recorded interval. These observations illustrate the progression of clustering structure during iterative optimization.

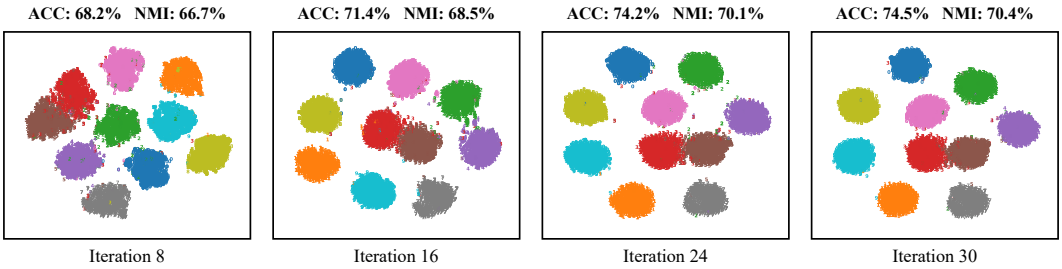


Figure 14: Clustering visualizations on N-MNIST at iterations 8, 16, 24, and 30. ACC and NMI are shown above each panel, with the iteration number below.