FAIR SUBSET SELECTION ACROSS CATEGORIES AND VIEWS VIA ADAPTIVE NEPV

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern machine learning increasingly relies on large datasets, but limited computational budgets can make training on all available examples impractical. In many cases, selecting a small training subset requires balancing category representation against feature diversity. This is especially difficult when categories overlap and one item contributes to several categories under a shared budget. We construct positive-definite category kernels by softening membership weights on a common feature geometry, then maximize a soft minimum of normalized log-determinants. Endpoint identities separate membership-count effects from diversity and distinguish zero-ridge and ridge-dominant regimes. A Stiefel relaxation yields an adaptive nonlinear eigenvalue formulation and a self-consistent-field solver. Its symmetric lifting has rank at most $2k$, an explicit spectrum, a uniform eigengap, and exact small-matrix compression; local convergence remains conditional on operator sensitivity. At a budget of 500 examples, trademark experiments report 81.91% mAP on OpenLogo retrieval and 53.67% category recall on USPTO description generation, compared with 76.15% and 51.36% under uniform sampling. The method does not lead on every metric. We distinguish these empirical comparisons from hard category quotas, global optimization guarantees, and downstream coreset guarantees.

1 INTRODUCTION

Limited training budgets make it important to choose which examples to use from large datasets. Selecting a small training subset requires considering both how examples differ and which categories they represent. Diversity alone can neglect rare categories, while allocating a fixed number of examples to each category can retain redundant observations. The allocation becomes harder when categories overlap: a trademark depicting both a bird and a tree consumes one budget slot but represents two categories. Selecting independently within each category duplicates decisions and leaves the shared budget unresolved.

We study the selection of k distinct items from a dataset with multiple, possibly overlapping categories. Our starting point is determinant-based selection. A principal determinant of a feature Gram matrix measures the squared volume spanned by selected feature vectors, favoring nonredundant examples (Kulesza & Taskar, 2012). A pooled determinant, however, does not explicitly account for category representation. Prescribed group quotas solve a different problem: they impose counts rather than decide the tradeoff between representation and diversity (Celis et al., 2018).

We assign every category a kernel on the same item index. Each kernel retains the shared feature geometry but attenuates items outside its category. A ridge makes all selected principal matrices positive definite. Every category can therefore evaluate the same subset, including subsets containing nonmembers. We maximize a soft minimum of their normalized log-determinants. Categories with weaker scores receive larger derivative weights, allowing selection to respond to the current subset rather than to a fixed allocation. Here *fairness* means emphasis on the weakest category-kernel score; it is not demographic parity, a hard quota, or a guarantee of equal counts.

The distinction between membership and geometry is explicit. At zero ridge, the category log-determinant separates into a membership-count term and a shared feature-volume term. With positive ridge, a different block identity and a perturbation bound describe the category-specific volume.

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.

These regimes must not be conflated. In particular, a very small membership floor does not by itself justify treating a regularized determinant as a pure count objective.

To optimize the shared criterion, we relax coordinate selectors to orthonormal frames. The resulting Stiefel objective gives an adaptive nonlinear eigenvalue problem (NEPv) and a self-consistent-field (SCF) solver. The contribution is its determinant-specific spectral structure, rather than the classical maximin principle itself (Dette et al., 2007). We evaluate category selection in two trademark tasks: brand-disjoint retrieval with partitioned categories, and description generation with overlapping design-code categories.

Contributions.

1. **A shared multicategory selection criterion.** Softened membership kernels combine category representation and feature diversity under one budget. Endpoint identities distinguish overlapping memberships, partition count bounds, and ridge-dependent interpretations.
2. **A structured adaptive solver.** The Stiefel stationarity equations admit a symmetric lifting of rank at most $2k$, an explicit spectrum, a uniform eigengap, and exact small-matrix compression. Local convergence of the idealized map is conditional; no global or rounding guarantee is claimed.
3. **Evidence in two category settings.** At a budget of 500 items, FAIRNEPV attains 81.91% mAP on OpenLogo and 53.67% category recall on USPTO descriptions. Matched selection comparisons distinguish coverage from within-category diversity, with measurement and reproducibility limitations stated explicitly.

2 RELATED WORK

Determinantal selection. DPPs model quality and diversity through principal determinants (Kulesza & Taskar, 2012). Their fixed-cardinality MAP problem is closely related to maximum-entropy sampling and maximum-volume subset selection, and is NP-hard (Ko et al., 1995; Çivril & Magdon-Ismail, 2009). Greedy and accelerated log-determinant methods (Han et al., 2017; Chen et al., 2018) address discrete selection; sampling methods address a different objective (Anari et al., 2016). Fractional-coordinate relaxations (Gillenwater et al., 2012; Nikolov, 2015; Singh & Xie, 2020) optimize item weights, whereas our relaxation optimizes a subspace containing all coordinate-selector frames. Orthogonality-constrained determinant objectives also appear in PCA (Hauser et al., 2018). Our rounding adapts established leverage-score and exchange ideas (Gu & Eisenstat, 1996; Belhadji et al., 2020); it has no new approximation guarantee.

Representation constraints and robust aggregation. Fair DPPs and partition-constrained determinant maximization enforce prescribed representation (Celis et al., 2018; Nikolov & Singh, 2016). We instead encode possibly overlapping memberships in category kernels and optimize their weakest score without specifying per-category quotas. A fixed-weight product of determinants is equivalent to a sum of log-determinants (Ohsaka & Matsuoka, 2020); high category scores can then compensate for weak ones. Averaging kernels before computing a determinant also removes explicit worst-category control. Multiple-kernel DPPs can learn or mix kernels (Chao et al., 2015; Matsuoka et al., 2021), which differs from evaluating one subset against every category.

Maximin design and robust subset selection are established principles (Dette et al., 2007; Krause et al., 2008; Udwani, 2018), and soft-min smoothing is standard (Nesterov, 2005). The present criterion applies them to normalized category determinants. Robust-submodular guarantees do not automatically apply: principal log-determinants need not be monotone, and the soft minimum does not automatically inherit a useful greedy guarantee. Fair PCA similarly balances group utilities on a common subspace (Samadi et al., 2018; Tantipongpipat et al., 2019), but uses reconstruction or variance criteria rather than category principal determinants.

Nonlinear eigenvalue optimization. General NEPv formulations and SCF safeguards for orthogonality-constrained optimization precede this work (Li, 2026; Yang et al., 2009; Bai et al., 2022). Our lifting exploits a determinant-specific gradient identity to obtain its exact spectrum and

small compression. Neither adaptive weights nor nonlinear stationarity alone is novel: even fixed-weight sums retain inverse compressed Gram factors. We use *coreset* only in the informal sense of a training subset, not as a downstream loss-approximation guarantee (Tremblay et al., 2019; Gollapudi et al., 2023).

3 Background and Notation

For a positive-semidefinite matrix L on $[n] = \{1, \dots, n\}$, an L-ensemble DPP assigns probability $\det(L_S)/\det(I + L)$ to subset S. If $\operatorname{rank}(L) \geq k$, conditioning on size k gives the MAP problem

$$\max_{|S|=k} \log\det(L_S), \tag{1}$$

where L_S is the principal submatrix and $\log 0 = -\infty$. We use natural logarithms and $\operatorname{St}(n, k) = \{V \in \mathbb{R}^{n\times k} : V^\top V = I_k\}$. Category indices are g or $q \in [M]$; all kernels use the same n item indices. A NEPv is a system $H(V)V = V\Lambda$ with $H(V)$ symmetric and dependent on V. Our SCF iteration recomputes its leading k-dimensional eigenspace.

Assumption 1 (Regularized category kernels)**.** Let $1 \leq k < n$ and $\tau > 0$. Every category kernel $L^{(q)} \in \mathbb{R}^{n\times n}$ is symmetric positive definite. The construction below uses a positive ridge, ensuring $V^\top L^{(q)} V \succ 0$ for every $V \in \operatorname{St}(n, k)$.

The zero-ridge identities below describe endpoints, not convergence guarantees at a singular kernel limit. All proofs are in section A.

4 Multicategory Determinant Selection

4.1 One budget with overlapping memberships

Let $G_1, \dots, G_M \subseteq [n]$ be nonempty categories. For $S \subseteq [n]$, write $n_g(S) = |S \cap G_g|$. Memberships can overlap, so $\sum_g n_g(S)$ may exceed $|S|$, while each selected item still consumes only one budget slot. A bird-and-tree trademark represents both categories without being duplicated.

Let $u_i \in \mathbb{R}^p$ be a shared unit feature vector and $K_{ij} = \langle u_i, u_j\rangle$. Categories identify memberships, not separate feature modalities. Independent category-wise selection chooses $S_g \subseteq G_g$ of prescribed size k_g to maximize $\log\det K_{S_g}$, then takes $\bigcup_g S_g$. Overlap can make this union smaller than $\sum_g k_g$; deduplication leaves an allocation problem. Even without overlap, fixing k_g in advance does not jointly value an item's representation and redundancy across the final subset.

4.2 Softened category kernels

Zero-padding a category kernel gives $L_0^{(g)} = D_g K D_g$, where $D_g = \operatorname{diag}(\mathbf{1}\{i \in G_g\})$. It preserves independent category selection:

Proposition 1. *For $1 \leq k_g \leq |G_g|$, the maximum of $\log\det((L_0^{(g)})_S)$ over all size-k_g subsets equals the maximum of $\log\det K_S$ over size-k_g subsets of G_g. If this maximum is finite, their maximizing sets coincide.*

Any nonmember creates a zero row and column. Thus a finite worst-category score under zero padding would require every selected item to belong to every category. We remove this restriction with a membership floor and ridge:

$$L^{(g)} = Q_g K Q_g + \rho I_n, \qquad Q_g = \operatorname{diag}(q_g(1), \dots, q_g(n)), \quad q_g(i) = \begin{cases} 1 & i \in G_g, \\ \mathrm{LO} & i \notin G_g, \end{cases} \tag{2}$$

where $0 < \mathrm{LO} \leq 1$ and $\rho > 0$. Nonmembers receive smaller positive feature weights when $\mathrm{LO} < 1$; a multi-category item receives weight one in each of its categories. Every candidate subset now has a finite score in every category.

Representation and diversity. At $\rho = 0$, assuming $K_S \succ 0$, factorization gives

$$\log\det L_S^{(g)} = 2n_g(S)\log(1/\mathrm{LO}) + \log\det K_S + 2k\log\mathrm{LO}. \tag{3}$$

The worst-category score trades the count floor $\min_g n_g(S)$ against shared feature volume. Smaller LO increases the count coefficient; at $\mathrm{LO} = 1$ every kernel is the same pooled kernel. For a partition, $\min_g n_g(S) \le \lfloor k/M \rfloor$; overlap invalidates this ceiling because one item can raise several counts. Neither this tradeoff nor its smooth version imposes a hard count guarantee.

With positive ridge, eq. (3) is not an exact identity. At the separate endpoint $\mathrm{LO} = 0$, writing $S_g = S \cap G_g$ gives

$$\log\det L_S^{(g)} = \log\det(K_{S_g} + \rho I) + (k - n_g(S))\log\rho. \tag{4}$$

The feature-volume term is now category specific. Lemma 4 bounds deviations from this block regime for small positive LO. In particular, if the floor is small relative to the ridge, one cannot infer pure count maximization merely from the floor's magnitude.

Calibration. All category kernels share K, LO, and ρ, fixed before selection. Rescaling one kernel by $c > 0$ adds $\log c$ to its normalized score and can change which category is weakest. Common feature normalization and common parameters make this comparison explicit.

4.3 A SOFT MINIMUM OVER CATEGORY SCORES

For one shared size-k subset, let $d_q(S) = k^{-1}\log\det L_S^{(q)}$. The maximin goal is $\max_{|S|=k}\min_q d_q(S)$. We optimize its smooth form

$$\max_{|S|=k} F_\tau(S), \qquad F_\tau(S) = -\tau\log\left[\frac{1}{M}\sum_{q=1}^{M} e^{-d_q(S)/\tau}\right]. \tag{5}$$

It satisfies $\min_q d_q(S) \le F_\tau(S) \le \min_q d_q(S) + \tau\log M$. Smaller τ tracks the weakest score more closely; finite temperature still allows limited compensation. Unlike a determinant of the averaged kernel, it retains each category's score. Unlike a fixed-weight sum of log-determinants, it adapts emphasis toward weaker categories.

Aggregation and interpretation. Writing $h_q(S) = \det(L_S^{(q)})^{1/k}$, the criterion is the logarithm of a negative-order power mean:

$$F_\tau(S) = \log\left[\left(\frac{1}{M}\sum_q h_q(S)^p\right)^{1/p}\right], \qquad p = -1/\tau < 0. \tag{6}$$

It emphasizes weak scores rather than equality for its own sake: a subset with equal but uniformly poor scores need not be preferred. In the zero-ridge identity, $2\log(1/\mathrm{LO})$ is an exchange rate between one extra member of the least-represented category and pooled log-volume. Allocations with the same minimum count and pooled volume can still differ substantially in their other counts. For example, partition counts $(2, 2, 4)$ at $k = 8$, $M = 3$ attain the maximum floor of two without being equal. Overlap makes this distinction stronger, since one item may raise several counts simultaneously. The exact minimum reads the worst value; its finite-temperature approximation also responds to the other category scores. These observations explain why the selection statistics in the experiments include more than coverage alone.

4.4 STIEFEL RELAXATION

A subset $S = \{i_1, \ldots, i_k\}$ has coordinate frame $V_S = [e_{i_1}, \ldots, e_{i_k}]$, with $V_S^\top L^{(q)} V_S = L_S^{(q)}$. Allowing every orthonormal frame gives

$$\max_{V \in \mathrm{St}(n,k)} f_\tau(V), \qquad f_\tau(V) = -\tau\log\left[\frac{1}{M}\sum_q e^{-g_q(V)/\tau}\right], \tag{7}$$

where $G_q(V) = V^\top L^{(q)} V$ and $g_q(V) = k^{-1}\log\det G_q(V)$.

Lemma 1. *Every size-k subset is feasible through V_S and $f_\tau(V_S) = F_\tau(S)$. The relaxed optimum upper-bounds the discrete optimum, and $f_\tau(VQ) = f_\tau(V)$ for any orthogonal Q.*

Thus the optimization is over subspaces, not over a particular basis. A dense frame must still be rounded to item indices. For one kernel the relaxed optimum is $k^{-1}\sum_{i=1}^{k}\log\lambda_i(L)$, attained by a leading eigenspace (proposition 4). For several categories, their preferred subspaces can differ, and the compressed inverses and soft-min weights both change with the current frame.

5 Adaptive NEPv Solver

5.1 Stationarity and spectral structure

Proposition 2 (Gradient and stationarity)**.** *Define*

$$w_q(V) = \frac{e^{-g_q(V)/\tau}}{\sum_r e^{-g_r(V)/\tau}}, \qquad A_\tau(V) = \frac{1}{k}\sum_q w_q(V)L^{(q)}VG_q(V)^{-1}. \tag{8}$$

Then $\nabla f_\tau(V) = 2A_\tau(V)$ and $V^\top A_\tau(V) = I_k/k$. Stationarity is equivalent to

$$A_\tau(V) = V/k. \tag{9}$$

A lower category score receives a larger weight. For a dense frame these are compressed-kernel scores, not integer category counts. The Riemannian gradient is $2A_\perp(V)$, where $A_\perp = A_\tau - V/k$ and $V^\top A_\perp = 0$.

Theorem 1 (Symmetric adaptive lifting)**.** *Let*

$$H_\tau(V) = A_\tau(V)V^\top + VA_\tau(V)^\top - k^{-1}VV^\top. \tag{10}$$

Then $H_\tau(V)V = A_\tau(V)$, so stationarity is exactly $H_\tau(V)V = V(I_k/k)$. Moreover, $H_\tau(VQ) = H_\tau(V)$ for orthogonal Q.

The particular subtractive projector makes the following spectrum explicit. General symmetric NEPv liftings are not unique (Li, 2026).

Proposition 3 (Uniform gap and exact compression)**.** *For $\sigma \geq 0$, write $d = 1/k + \sigma$ and $H_{\tau,\sigma} = H_\tau + \sigma VV^\top$. Then*

$$H_{\tau,\sigma} = dVV^\top + A_\perp V^\top + VA_\perp^\top. \tag{11}$$

If $s_1, \ldots, s_r > 0$ are the nonzero singular values of $A_\perp$, the spectrum of $H_{\tau,\sigma}$ consists of $(d \pm \sqrt{d^2 + 4s_j^2})/2$, $k - r$ copies of d, and $n - k - r$ zeros. Its rank is at most $2k$, it has exactly k positive eigenvalues, and its leading eigengap is at least d.

The top-k eigenspace is therefore uniquely defined. Its principal angles from the current subspace are

$$\theta_j = \tfrac{1}{2}\arctan(2s_j/d) < \pi/4. \tag{12}$$

The remaining angles are zero. Stationary subspaces are exactly fixed points of this idealized map, including stationary points that are not maxima. For $n \geq 2k$, an orthogonal residual basis $A_\perp = Q_2R$, $V^\top Q_2 = 0$, gives the exact compressed matrix

$$H_{\tau,\sigma} = [V, Q_2]\begin{pmatrix} dI_k & R^\top \\ R & 0 \end{pmatrix}[V, Q_2]^\top. \tag{13}$$

Rank deficiency requires orthogonal completion; for $n < 2k$ a compact basis has dimension at most n (lemma 7).

5.2 Iteration, rounding, and limitations

At each iteration we evaluate the category scores, compute stable soft-min weights, and form A_τ. If $\|A_\tau - V/k\|_F < \varepsilon$, we stop. Otherwise we obtain a leading frame $\widetilde{V}$ of $H_{\tau,\sigma}$, align its basis with $Q = \mathrm{Polar}(\widetilde{V}^\top V)$, and update

$$V^+ = \mathrm{Polar}\big(\alpha\widetilde{V}Q + (1-\alpha)V\big), \quad 0 < \alpha \leq 1. \tag{14}$$

Algorithm 1 FAIRNEPV: adaptive SCF for (9)

Require: Regularized feature-map kernels $L^{(q)} = \Phi_q \Phi_q^\top + \rho I$, $q \in [M]$; target size k; temperature $\tau > 0$; tolerance ε; maximum iterations $T_{\max}$; damping $\alpha \in (0, 1]$; shift $\sigma \geq 0$.

Initialize $V_0 \leftarrow$ a random orthonormal frame or a problem-specific warm start.
for $t = 0, \ldots, T_{\max} - 1$ **do**
  for $q = 1, \ldots, M$ **do**
    $G_{q,t} \leftarrow V_t^\top L^{(q)} V_t, \quad g_{q,t} \leftarrow k^{-1} \log\det(G_{q,t})$.
  end for
  $g_{\min,t} \leftarrow \min_q g_{q,t}, \quad w_{q,t} \leftarrow \exp(-(g_{q,t} - g_{\min,t})/\tau) / \sum_r \exp(-(g_{r,t} - g_{\min,t})/\tau)$.
  $A_t \leftarrow k^{-1} \sum_q w_{q,t} L^{(q)} V_t G_{q,t}^{-1}$.
  if $\|A_t - V_t/k\|_F < \varepsilon$ **then**
    return V_t
  end if
  Apply $H_\tau(V_t)X \leftarrow A_t(V_t^\top X) + V_t(A_t^\top X) - k^{-1} V_t(V_t^\top X)$ implicitly.
  $H_{\tau,\sigma}(V_t)(\cdot) \leftarrow H_\tau(V_t)(\cdot) + \sigma V_t(V_t^\top \cdot)$ ▷ level shift
  $\widetilde{V}_{t+1} \leftarrow$ top-k eigenvectors of $H_{\tau,\sigma}(V_t)$.
  $Q_t \leftarrow \mathrm{Polar}\big(\widetilde{V}_{t+1}^\top V_t\big)$ ▷ gauge alignment
  $V_{t+1} \leftarrow \mathrm{Polar}\big(\alpha \widetilde{V}_{t+1} Q_t + (1 - \alpha) V_t\big)$ ▷ damping
end for
return $V_{T_{\max}}$ ▷ iteration-cap endpoint

Damping and level shifting are safeguards, not convergence certificates. The full procedure is in algorithm 1. Iteration-cap termination must be distinguished from satisfying the stationarity tolerance.

Theorem 2 (Conditional local contraction)**.** *Let $V^\star$ be stationary and let $\mathcal{T}_{\tau,\sigma}(V) = \mathrm{TopEigvecs}_k(H_{\tau,\sigma}(V))$. Suppose locally $\|H_{\tau,\sigma}(V) - H_{\tau,\sigma}(V^\star)\|_2 \leq C_{\tau,\sigma} \|\sin\Theta(V, V^\star)\|_F$. If the Davis–Kahan constant C_{DK} satisfies*

$$C_{\mathrm{DK}} C_{\tau,\sigma} < 1/k + \sigma, \tag{15}$$

the idealized map shrinks chordal distance to $V^\star$ by factor $\gamma = C_{\mathrm{DK}} C_{\tau,\sigma} / (1/k + \sigma) < 1$.

Here Θ contains principal angles and the chordal distance is $\|\sin\Theta\|_F$. The sensitivity constant is not computed here. This theorem does not prove convergence of the damped/polar update, global convergence, or optimality. A larger shift reduces step angles but also changes operator sensitivity; it is not a general guarantee of faster convergence.

Why the spectral gap is not sufficient. The explicit spectrum describes $H_{\tau,\sigma}$ at a fixed frame, whereas convergence depends on how that operator changes with the frame. Each update changes both the inverse compressed Gram matrices and the category weights. Positive ridge controls invertibility, and positive temperature makes the weights smooth, but neither supplies the strict sensitivity inequality in eq. (15) for a particular run. The inverse-Gram bound in lemma 5 and the local proof in section A.6 make these dependencies explicit. Consequently, a well-separated leading eigenspace and a small per-step rotation should not be substituted for the stationarity-residual stopping test.

Returning a subset. Screen the $m = \min(n, \lceil ck \rceil)$ largest leverage scores $\ell_i = \|e_i^\top V\|_2^2$ for $c > 1$. Greedily select k items from this pool using the category soft-min, then perform improving one-item swaps against the full pool. For partial nonempty sets T, normalize by $|T|$; completed sets use k. This heuristic has no approximation guarantee: improving the relaxed objective does not establish an improvement after rounding or in a downstream model. Section B describes objective diagnostics.

Cost. For $L^{(q)} = \Phi_q \Phi_q^\top + \rho I$ with feature width p_q, a sweep costs $O(\sum_q (n p_q k + n k^2 + k^3))$ for kernel terms. Cholesky solves avoid explicit inverses; the exact compression costs $O(nk^2 + k^3)$ and avoids a dense $n \times n$ eigendecomposition. At most B full swap passes require $O(Bnk)$ candidate evaluations in addition to continuous optimization. These are operation counts, not measured speed or memory claims.

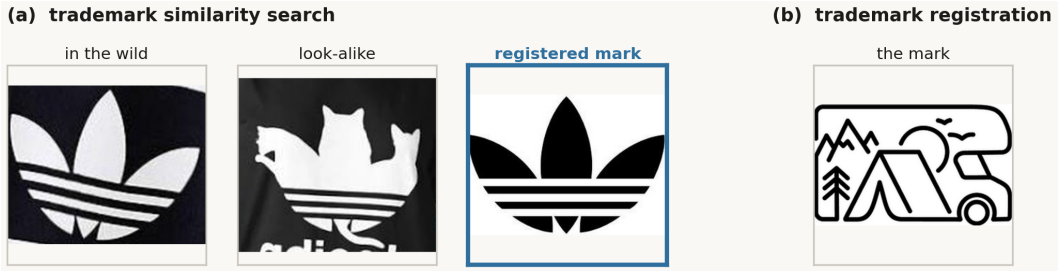

Figure 1: **Category-based trademark tasks.** Left: photographs and a registered mark illustrate retrieval. Right: one USPTO drawing combines several design-code categories, illustrating overlapping memberships for description generation. Images reproduce the supplied experimental examples.

Table 1: **OpenLogo,** $k = 500$**.** Selection counts refer to 158 training brands; retrieval uses 80 unseen test brands. Larger NN distance indicates less selected-set redundancy. Best downstream values are bold.

	Categories	Crops per brand			NN	mAP	R-prec.
Selector	covered	min	med.	max	distance	(%)	(%)
FAIRNEPV	158	3	3	8	0.948	**81.91**	**78.11**
Averaged-kernel	158	1	3	8	0.786	79.02	74.53
Stratified	158	3	3	4	0.826	78.30	75.10
Fixed-weight	141	0	2	15	0.960	78.86	75.11
Uniform	115	0	1	21	0.797	76.15	71.97

6 EXPERIMENTS: SELECTION ACROSS TRADEMARK CATEGORIES

We evaluate a budget of $k = 500$ in two category settings: OpenLogo brands form a partition, while USPTO design-code categories overlap. The downstream tasks are retrieval and description generation, respectively. These experiments compare category selection, not separate image and text kernels.

Comparisons. The shared controls are uniform sampling, a determinant of the averaged category kernel, and a fixed-weight average of category log-determinants. OpenLogo additionally uses stratified uniform allocation matched on category coverage and minimum count. Within each task, selectors use the same candidate pool and feature construction. Each selected subset trains the same downstream model with matched initial weights, learning rate, schedule, batch size, epoch budget, stopping rule, and seed. Selection statistics characterize representation and redundancy; downstream metrics test whether those properties transfer to prediction. These are distinct assessments.

6.1 OPENLOGO: PARTITIONED BRANDS

The supplied OpenLogo-derived corpus (Su et al., 2018) reports 43,195 annotated crops over 284 brands. The reported usable brand-disjoint split has 158 training brands and 18,353 crops, 40 validation brands and 5,245 crops, and 80 test brands and 11,233 crops. Each crop belongs to one brand. Selection features are unit residuals: subtract the registered-mark exemplar embedding from the crop embedding and renormalize. This construction removes a shared brand reference before measuring within-brand variation. The reported feature ablation lowers mAP by 4.50 percentage points when crop embeddings replace residuals with other factors held fixed.

Each subset fine-tunes a DINOv2 matcher (Oquab et al., 2023). Evaluation uses 80 registered-mark queries against the 11,233 held-out crops; training brands are absent from this test. Checkpoints are chosen by validation mAP. We report mAP and R-precision (Manning et al., 2008), alongside category counts and mean nearest-neighbor distance in the unit-residual selection space.

Table 2: **USPTO, $k = 500$.** All selectors cover all 29 categories. Downstream summaries are averaged over categories and three seeds. Recall is reference-conditioned judge recall, not precision or factuality.

	Marks per category			Category	Qwen3	BGE
Selector	min	med.	max	recall (%)	cos (%)	R@1 (%)
FAIRNEPV	85	86	278	**53.67**	86.57	**86.06**
Averaged-kernel	26	89	345	53.01	86.21	84.16
Fixed-weight	23	88	388	53.35	86.09	84.99
Uniform	20	30	248	51.36	**86.60**	85.21

Representation and diversity. FAIRNEPV covers all 158 brands with at least three crops each. Three is the largest possible minimum count, since $\lfloor 500/158 \rfloor = 3$. Stratification also attains this floor, while averaged-kernel selection covers all brands but leaves one with a single crop. Fixed-weight selection omits 17 brands and uniform sampling omits 43. These observations demonstrate the allocation differences; they do not make floor attainment a general theorem about the finite-temperature, regularized objective.

Among methods covering every brand, FAIRNEPV has the largest mean NN distance: 0.948 versus 0.826 for stratification and 0.786 for averaged-kernel selection. Fixed-weight selection is slightly more dispersed at 0.960 but has weaker coverage. FAIRNEPV improves mAP by 2.89 points over the next-best reported selector, averaged-kernel, by 3.61 over stratification, and by 5.76 over uniform. The stratified comparison shows that matching coverage and the count floor alone does not reproduce the observed retrieval performance. It does not isolate diversity as the sole cause, because the complete allocation profiles and selected images also differ. The experiment measures transfer to unseen brands, not memorization of the training brands.

6.2 USPTO: OVERLAPPING DESIGN CATEGORIES

USPTO trademark records provide a drawing, a recorded description, and design codes (USPTO, 2026b;c). A category is a design-code root from 01–29; a mark carries between 1 and 17 roots. Membership is presence or absence of that root, with the reported floor $\mathrm{LO} = 10^{-15}$. Thus one selected mark can represent several categories. Interpreting this small floor requires the ridge regime in section 4.2; it does not by itself establish that the objective is only the minimum category count.

Each selected set of 500 marks fine-tunes Qwen3-VL-4B (Bai et al., 2025) with LoRA (Hu et al., 2022) to generate descriptions from drawings. A vision–language judge receives the recorded reference, a design code, and its definition, and determines whether the generation describes that element. This supplies recall over 9,220 mark–code pairs per run, without a precision counterpart. Additional metrics are Qwen3 text-embedding cosine similarity to the reference (Zhang et al., 2025) and BGE-M3 retrieval of the generation's own reference, measured by R@1 (Chen et al., 2024).

Results and interpretation. Because all methods cover all categories, the minimum number of selected marks per category is more informative than coverage here. FAIRNEPV raises this minimum to 85, compared with 26, 23, and 20 for averaged-kernel, fixed-weight, and uniform selection. Overlapping counts need not sum to 500. The improvement in representation accompanies the highest category recall and BGE R@1 in table 2. Recall exceeds uniform by 2.31 points and fixed-weight by 0.32; BGE R@1 exceeds uniform by 0.85 points. The smaller differences among determinant selectors warrant a narrower interpretation than the larger comparison with uniform sampling.

There is no uniform win across metrics: uniform sampling has the highest Qwen3 cosine, 86.60 versus 86.57. All four cosine scores fall within 0.51 points; the table alone does not establish whether this reflects metric insensitivity or similar semantic performance. The judge's reference-conditioned recall can penalize correct elements absent from the recorded description, and longer generations can receive higher recall. The reported description reference score under the same procedure is about 78%, so it is an empirical reference level rather than a perfect ceiling. These issues need not affect all selectors equally. No precision, length-matched comparison, or statistical significance claim is inferred from the displayed means.



Figure 2: **USPTO category profiles.** (a) Selected marks per root category, with the pool rescaled to the subset budget. (b) Distinct roots per selected mark. Counts overlap: a single mark can contribute to several bars. The figure records selection composition, not downstream causal effects.

Scope of the evidence. Both tasks support the usefulness of category-aware selection at the reported budget and representations. They do not establish gains for arbitrary categories, datasets, budgets, or downstream losses. Complete configuration and per-run records are needed to reproduce the tables and to distinguish solver convergence from iteration-cap termination. Section C records the available protocol details and the missing information without supplying unverified settings.

7 Discussion and Conclusion

Multicategory selection must coordinate representation and diversity within a single budget, particularly when one example belongs to several categories. Softened membership kernels put these assessments on a common item index, and a soft minimum of normalized log-determinants emphasizes weaker category scores. The Stiefel formulation yields an adaptive NEPv with explicit spectral structure and exact small compression. Its local convergence result is conditional, and practical rounding remains heuristic.

The trademark results show improved retrieval and category recall under the reported comparisons, while the USPTO cosine metric favors uniform sampling slightly. These findings support a specific empirical claim, not a universal prediction or representation guarantee. Further work should test robustness to category imbalance, feature construction, ridge scale, and rounding, with complete configurations and uncertainty estimates.

Future work. The shared-index formulation could also compare aligned representations, such as an image and its associated description, by assigning one kernel to each. Establishing that extension empirically is left to future work; it is not an evaluated setting of the main paper. The full formulation and proposed protocol are retained in sections H and K.

Ethics Statement

Fairness here refers to category-kernel scores, not demographic parity or protected-group representation. Generated trademark descriptions and similarity rankings require human review; these experiments do not evaluate legal correctness or suitability for registration or infringement decisions. Reference-conditioned automated judging may omit valid content or reward verbosity. Public availability of trademark records does not remove dataset licensing, privacy, or downstream-use obligations; those obligations must be checked for any release.

Reproducibility Statement

Sections 4 and 5 specify the objective, assumptions, solver, and rounding rule. Section A provides the proofs, and section C separates recorded empirical settings from details not supplied in the current record. The numerical tables are retained from the supplied experiments; a complete anonymous artifact with split identifiers, feature versions, selected subsets, solver settings, and per-run outputs is not established by this manuscript alone.

References

N. Anari, S. Oveis Gharan, and A. Rezaei. Monte Carlo Markov chain algorithms for sampling strongly Rayleigh distributions and determinantal point processes. In *Conference on Learning Theory (COLT)*, 2016.

N. Anari, N. Haghtalab, J. Naor, S. Pokutta, M. Singh, and A. Torrico. Structured robust submodular maximization: Offline and online algorithms. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, 2019. URL https://proceedings.mlr.press/v89/anari19a.html.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Z. Bai, D. Lu, and B. Vandereycken. Robust Rayleigh quotient minimization and nonlinear eigenvalue problems. *SIAM Journal on Scientific Computing*, 40(5):A3495–A3522, 2018. doi: 10.1137/18M1167681.

Z. Bai, R.-C. Li, and D. Lu. Sharp estimation of convergence rate for self-consistent field iteration to solve eigenvector-dependent nonlinear eigenvalue problems. *SIAM Journal on Matrix Analysis and Applications*, 43(1):301–327, 2022. doi: 10.1137/20M136606X.

A. Belhadji, R. Bardenet, and P. Chainais. A determinantal point process for column subset selection. *Journal of Machine Learning Research*, 21(197):1–62, 2020. URL https://www.jmlr.org/papers/v21/19-080.html.

D. Braess and H. Dette. On the number of support points of maximin and Bayesian optimal designs. *The Annals of Statistics*, 35(2):772–792, 2007. doi: 10.1214/009053606000001307.

Y. Cai, L.-H. Zhang, Z. Bai, and R.-C. Li. On an eigenvector-dependent nonlinear eigenvalue problem. *SIAM Journal on Matrix Analysis and Applications*, 39(3), 2018.

L. E. Celis, V. Keswani, D. Straszak, A. Deshpande, T. Kathuria, and N. K. Vishnoi. Fair and diverse DPP-based data summarization. In *International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, 2018. URL https://proceedings.mlr.press/v80/celis18a.html.

W.-L. Chao, B. Gong, K. Grauman, and F. Sha. Large-margin determinantal point processes. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2015. URL https://auai.org/uai2015/proceedings/papers/209.pdf.

J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024.

L. Chen, G. Zhang, and E. Zhou. Fast greedy MAP inference for determinantal point process to improve recommendation diversity. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

A. Çivril and M. Magdon-Ismail. On selecting a maximum volume sub-matrix of a matrix and related problems. *Theoretical Computer Science*, 410(47–49), 2009.

C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. doi: 10.1137/0707001.

H. Dette, L. M. Haines, and L. A. Imhof. Maximin and Bayesian optimal designs for regression models. *Statistica Sinica*, 17(2):463–480, 2007. URL https://www3.stat.sinica.edu.tw/statistica/j17n2/j17n23/j17n23.html.

M. Elfeki, L. Wang, and A. Borji. Multi-stream dynamic video summarization. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 339–349, 2022. doi: 10.1109/WACV51458.2022.00026.

J. Gillenwater, A. Kulesza, and B. Taskar. Near-optimal MAP inference for determinantal point processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

S. Gollapudi, S. Mahabadi, and V. Sivashankar. Composable coresets for determinant maximization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

M. Gu and S. C. Eisenstat. Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM Journal on Scientific Computing*, 17(4):848–869, 1996. doi: 10.1137/0917055.

I. Han, K. Lee, and J. Shin. Faster greedy MAP inference for determinantal point processes. In *International Conference on Machine Learning (ICML)*, 2017.

R. A. Hauser, A. Eftekhari, and H. Matzinger. PCA by determinant optimization has no spurious local optima. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018. doi: 10.1145/3219819.3220069.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

C.-W. Ko, J. Lee, and M. Queyranne. An exact algorithm for maximum entropy sampling. *Operations Research*, 43(4):684–691, 1995. doi: 10.1287/opre.43.4.684.

A. Krause, H. B. McMahan, C. Guestrin, and A. Gupta. Robust submodular observation selection. *Journal of Machine Learning Research*, 9:2761–2801, 2008. URL https://www.jmlr.org/papers/v9/krause08b.html.

A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3), 2012.

R.-C. Li. A theory of the NEPv approach for optimization on the Stiefel manifold. *Foundations of Computational Mathematics*, 26:179–244, 2026. doi: 10.1007/s10208-024-09687-2.

C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. ISBN 978-0-521-86571-5.

T. Matsuoka, N. Ohsaka, and A. Yabe. On the convex combination of determinantal point processes. In *Asian Conference on Machine Learning (ACML)*, volume 157 of *Proceedings of Machine Learning Research*, 2021. URL https://proceedings.mlr.press/v157/matsuoka21a.html.

Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1): 127–152, 2005. doi: 10.1007/s10107-004-0552-5.

A. Nikolov. Randomized rounding for the largest simplex problem. In *Symposium on Theory of Computing (STOC)*, 2015.

A. Nikolov and M. Singh. Maximizing determinants under partition constraints. In *ACM Symposium on Theory of Computing (STOC)*, 2016. doi: 10.1145/2897518.2897649.

N. Ohsaka and T. Matsuoka. On the (in)tractability of computing normalizing constants for the product of determinantal point processes. In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, 2020. URL https://proceedings.mlr.press/v119/ohsaka20a.html.

M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: A method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318, 2002.

S. Samadi, U. Tantipongpipat, J. Morgenstern, M. Singh, and S. Vempala. The price of fair PCA: One extra dimension. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

M. Singh and W. Xie. Approximation algorithms for D-optimal design. *Mathematics of Operations Research*, 45(4), 2020.

N. Sogi, T. Shibata, M. Terao, M. Suganuma, and T. Okatani. MS-DPPs: Multi-source determinantal point processes for contextual diversity refinement of composite attributes in text-to-image retrieval. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2025. doi: 10.24963/ijcai.2025/207.

F. Spaeh and A. Miyauchi. An asymptotically optimal approximation algorithm for multiobjective submodular maximization at scale. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, 2025. URL https://proceedings.mlr.press/v267/spaeh25a.html.

H. Su, X. Zhu, and S. Gong. Open logo detection challenge. In *British Machine Vision Conference (BMVC)*, 2018.

U. Tantipongpipat, S. Samadi, M. Singh, J. Morgenstern, and S. Vempala. Multi-criteria dimensionality reduction with applications to fairness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

N. Tremblay, S. Barthelmé, and P.-O. Amblard. Determinantal point processes for coresets. *Journal of Machine Learning Research*, 20(168):1–70, 2019. URL https://www.jmlr.org/papers/v20/18-167.html.

R. Udwani. Multi-objective maximization of monotone submodular functions with cardinality constraint. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

USPTO. Trademark manual of examining procedure (TMEP). https://tmep.uspto.gov/, 2026a.

USPTO. Trademark bulk data and trademark status and document retrieval (TSDR) products. https://bulkdata.uspto.gov/, 2026b.

USPTO. Trademark design search code manual. https://tmdesigncodes.uspto.gov/, 2026c.

M. Xu, B. Jiang, W. Pu, Y.-F. Liu, and A. M.-C. So. An efficient alternating riemannian/projected gradient descent ascent algorithm for fair principal component analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7195–7199, 2024. doi: 10.1109/ICASSP48485.2024.10447172.

C. Yang, W. Gao, and J. C. Meza. On the convergence of the self-consistent field iteration for a class of nonlinear eigenvalue problems. *SIAM Journal on Matrix Analysis and Applications*, 30 (4):1773–1788, 2009. doi: 10.1137/080716293.

G. Zalcberg and A. Wiesel. Fair principal component analysis and filter design. *IEEE Transactions on Signal Processing*, 69:4835–4842, 2021. doi: 10.1109/TSP.2021.3099983.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*, 2020.

Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, F. Huang, and J. Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

A TECHNICAL APPENDIX

This appendix proves the embedding, stationarity, group-kernel identities, and local contraction claims for the multicategory relaxation. Throughout, $G_q(V) = V^\top L^{(q)} V$ and $g_q(V) = k^{-1} \log\det G_q(V)$.

A.1 DEFERRED PROOFS

Proof of lemma 1. The columns of a coordinate selector V_S are distinct standard basis vectors, so $V_S^\top V_S = I_k$. For each kernel,

$$(V_S^\top L^{(q)} V_S)_{j\ell} = e_{i_j}^\top L^{(q)} e_{i_\ell} = L^{(q)}_{i_j, i_\ell}.$$

Thus $G_q(V_S) = L_S^{(q)}$, so $g_q(V_S) = d_q(S)$ and $f_\tau(V_S) = F_\tau(S)$. Maximizing over all Stiefel frames can only improve on the coordinate selectors. Finally, replacing any frame V by VQ changes each Gram matrix by orthogonal congruence and leaves every determinant unchanged. □

Proposition 4 (Single-kernel spectral special case)**.** *Let $L \succeq 0$ have eigenvalues $\lambda_1(L) \geq \cdots \geq \lambda_n(L) \geq 0$. Interpret the objective as $-\infty$ when $V^\top L V$ is singular. If $\lambda_k(L) > 0$, then*

$$\max_{V \in \mathrm{St}(n,k)} \frac{1}{k} \log\det(V^\top L V) = \frac{1}{k} \sum_{i=1}^{k} \log \lambda_i(L),$$

attained by every dominant k-dimensional invariant subspace, with the usual freedom inside tied eigenspaces.

Proof of proposition 4. The multiplicative Rayleigh-Ritz principle gives

$$\det(V^\top L V) \leq \prod_{i=1}^{k} \lambda_i(L)$$

for every $V \in \mathrm{St}(n,k)$, with equality for any dominant k-dimensional invariant subspace. Taking logarithms in the extended-real sense and dividing by k proves the result. □

Proof of proposition 2. Jacobi's formula and the symmetry of $L^{(q)}$ give

$$\nabla g_q(V) = \frac{2}{k} L^{(q)} V G_q(V)^{-1}.$$

Differentiating the soft-min yields $\nabla f_\tau(V) = \sum_q w_q(V) \nabla g_q(V) = 2A_\tau(V)$. Moreover,

$$V^\top A_\tau(V) = \frac{1}{k} \sum_q w_q(V) G_q(V) G_q(V)^{-1} = \frac{1}{k} I_k.$$

The Stiefel tangent projection is

$$\Pi_V(\xi) = \xi - V \,\mathrm{sym}(V^\top \xi).$$

Substituting the Euclidean gradient $\xi = \nabla f_\tau(V) = 2A_\tau(V)$ and using the identity above gives

$$\begin{aligned}
\Pi_V(2A_\tau(V)) &= 2A_\tau(V) - V \,\mathrm{sym}\big(2V^\top A_\tau(V)\big) \\
&= 2A_\tau(V) - V \,\mathrm{sym}\left(\frac{2}{k} I_k\right) \\
&= 2A_\tau(V) - \frac{2}{k} V \\
&= 2\left(A_\tau(V) - \frac{1}{k} V\right),
\end{aligned}$$

where $(2/k)I_k$ is symmetric. Hence the Riemannian gradient vanishes exactly when $A_\tau(V) = V/k$, which is (9). □

Proof of theorem 1. Since $V^\top A_\tau = A_\tau^\top V = I_k/k$ and $V^\top V = I_k$,

$$H_\tau(V)V = A_\tau(V) + V/k - V/k = A_\tau(V).$$

Thus $H_\tau(V)V = V/k$ is equivalent to (9). For $Q \in \mathrm{O}(k)$, orthogonal congruence gives $G_q(VQ) = Q^\top G_q(V)Q$, so $g_q(VQ) = g_q(V)$ and $w_q(VQ) = w_q(V)$. Consequently $A_\tau(VQ) = A_\tau(V)Q$. Substituting this into (10) and cancelling $QQ^\top = I_k$ proves $H_\tau(VQ) = H_\tau(V)$. □

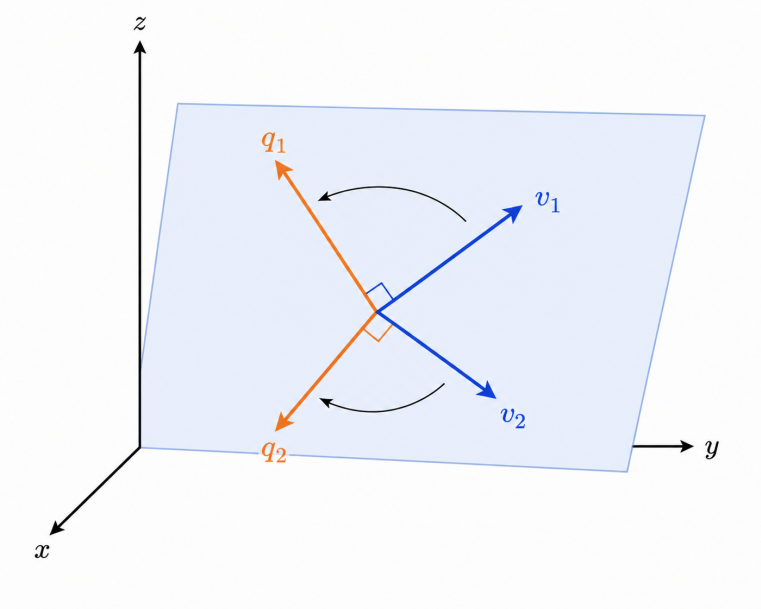

Figure 3: **Gauge invariance.** The blue and orange bases are rotations within the same plane, so they span the same subspace. Only the basis changes; the plane, objective value, and operator remain unchanged.

A.2 Group-kernel identities

The groups $G_g \subseteq [n]$ may overlap. Throughout this subsection, $n_g(S) = |S \cap G_g|$ and K is the Gram matrix of shared unit feature vectors. The identities hold category by category and do not require a partition; the count ceiling below explicitly treats the disjoint special case. Endpoint assumptions on LO and ρ are stated separately.

Remark 1 (Exact augmentation restricts selection to the intersection). For any nonempty subset S, if $S \not\subseteq \bigcap_{g=1}^{M} G_g$, then

$$\min_g \log\det\big((L_0^{(g)})_S\big) = -\infty.$$

Some selected item lies outside at least one category, creating a zero row and column in that category's principal submatrix. If instead $S \subseteq \bigcap_g G_g$, all selected submatrices equal K_S and the worst-category score is finite exactly when $K_S \succ 0$. Thus a finite size-k solution requires at least k common members with a nonsingular selected Gram matrix. For disjoint categories with $M \geq 2$, no nonempty subset meets this condition. With overlapping categories, the restriction can still be too severe: the bird–tree example would allow only trademarks containing both objects. The soft-min has the same restriction because one $-\infty$ category score makes its value $-\infty$.

Proof of proposition 1. If $S \not\subseteq G_g$, some $i \in S \setminus G_g$ indexes a zero row and column of $(L_0^g)_S$, so $\det((L_0^g)_S) = 0$ and the objective is $-\infty$. If $S \subseteq G_g$, then $(L_0^g)_S = K_S$ by construction. Hence every feasible set with finite value is contained in G_g, and on such sets the two objectives coincide. □

Lemma 2 (Quality factorization, $\rho = 0$)**.** *Assume* $0 < \mathrm{LO} \le 1$, $\rho = 0$, *and* $K_S \succ 0$. *Then*

$$\log\det\big(L_S^{(g)}\big) \;=\; 2\,n_g(S)\log\frac{1}{\mathrm{LO}} + \log\det(K_S) + 2k\log\mathrm{LO}. \tag{16}$$

Consequently

$$\min_g \log\det\big(L_S^{(g)}\big) \;=\; \Big(2\log\frac{1}{\mathrm{LO}}\Big)\min_g n_g(S) + \log\det(K_S) + 2k\log\mathrm{LO}. \tag{17}$$

The volume term $\log\det(K_S)$ *is identical across groups, so the maximin of the group scores combines the count floor* $\min_g n_g(S)$ *with a shared pooled-diversity score. For* $0 < \mathrm{LO} < 1$, *the coefficient on the count floor is strictly positive; at* $\mathrm{LO} = 1$, *this membership term vanishes.*

Proof of lemma 2. With $\rho = 0$, the principal submatrix is $L_S^{(g)} = Q_S K_S Q_S$, where $Q_S = \mathrm{diag}(q_g(i) : i \in S)$. Then

$$\det(Q_S K_S Q_S) = \det(Q_S)^2\det(K_S) = \left(\prod_{i\in S} q_g(i)\right)^2 \det(K_S).$$

The product equals $\mathrm{LO}^{k-n_g(S)}$, so

$$\det(L_S^{(g)}) = \mathrm{LO}^{2(k-n_g(S))}\det(K_S).$$

Taking logarithms and rewriting $2(k - n_g(S))\log\mathrm{LO} = 2n_g(S)\log(1/\mathrm{LO}) + 2k\log\mathrm{LO}$ gives (16). The identity (17) follows because $\log\det(K_S)$ and $2k\log\mathrm{LO}$ do not depend on g, while $2\log(1/\mathrm{LO}) \ge 0$ for $\mathrm{LO} \in (0, 1]$. At $\mathrm{LO} = 1$, the coefficient is zero, and the identity remains valid. □

Lemma 3 (Block form, $\mathrm{LO} = 0$)**.** *Assume* $\mathrm{LO} = 0$ *and* $\rho > 0$. *Permute* $S = S_g \cup S_{-g}$ *with* $S_g = S \cap G_g$ *and* $S_{-g} = S \setminus G_g$. *Then*

$$L_S^{(g)} \;=\; \begin{pmatrix} K_{S_g} + \rho I & 0 \\ 0 & \rho I_{k-n_g(S)} \end{pmatrix}, \tag{18}$$

and

$$\log\det\big(L_S^{(g)}\big) \;=\; \log\det\big(K_{S_g} + \rho I\big) + \big(k - n_g(S)\big)\log\rho. \tag{19}$$

Use $\det(0 \times 0) = 1$ *for an empty in-group block. If* $S \setminus G_g \neq \varnothing$ *and* $i \in G_g \setminus S$, *define* $s = K_{ii} - K_{iS_g}(K_{S_g} + \rho I)^{-1}K_{S_g i} > 0$. *Replacing one out-of-group selected item by* i *changes* (19) *by* $\log(1 + s/\rho) > 0$. *This is a statement about category* g *alone; the removed item may also represent other categories. The first term in* (19) *is the in-group volume.*

Proof of lemma 3. With $\mathrm{LO} = 0$, q_g is the indicator of G_g. After permuting $S = S_g \cup S_{-g}$, the membership matrix is $Q_S = \mathrm{blkdiag}(I_{n_g(S)}, 0)$, so $Q_S K_S Q_S = \mathrm{blkdiag}(K_{S_g}, 0)$. Adding the ridge produces (18). The determinant of a block-diagonal matrix is the product of the block determinants, which is (19).

For the residual claim, let U have rows $u_j^\top$ for $j \in S_g$, and let u_i be the unit feature vector of the added item. Define

$$s = K_{ii} - K_{iS_g}(K_{S_g} + \rho I)^{-1}K_{S_g i} = u_i^\top\big[I - U^\top(UU^\top + \rho I)^{-1}U\big]\,u_i.$$

The matrix in brackets equals $\rho(U^\top U + \rho I)^{-1} \succ 0$, so $s > 0$. The Schur complement for adding item i to the regularized in-group Gram matrix is $\rho + s$. Removing one out-of-group selected item removes a diagonal factor ρ, so the log-determinant changes by $\log(\rho + s) - \log\rho = \log(1 + s/\rho) > 0$. Other category scores can change as well when memberships overlap; the calculation concerns only category g. □

Lemma 4 (Ridge-dominant perturbation)**.** *Assume* $0 \le \mathrm{LO} \le 1$ *and* $\rho > 0$. *Let* A_0 *be the block matrix in* (18) *and write* $L_S^{(g)} = A_0 + \Delta$, *where* Δ *collects the* LO*-scaled off-blocks and the* LO^2 *out-group Gram. If* $\|A_0^{-1}\Delta\|_2 \le \eta < 1$, *then*

$$\big|\log\det(L_S^{(g)}) - \log\det(A_0)\big| \;\le\; -k\log(1 - \eta). \tag{20}$$

A sufficient condition is $\mathrm{LO}\sqrt{n_g(S)\,(k - n_g(S))} + \mathrm{LO}^2(k - n_g(S)) \le \eta\rho$. *In particular, whenever* $\mathrm{LO} \ll \rho/k$ *the operative expansion is* (19) *rather than* (16).

Proof of lemma 4. Write $n_g = n_g(S)$ and $m = k - n_g$. Expanding (2) on S gives

$$L_S^{(g)} = \begin{pmatrix} K_{S_g} + \rho I & \mathrm{LO}\, C \\ \mathrm{LO}\, C^\top & \mathrm{LO}^2 K_{S_{-g}} + \rho I \end{pmatrix} = A_0 + \Delta,$$

where $C_{ab} = \langle u_a, u_b \rangle$ for $a \in S_g$, $b \in S_{-g}$, and

$$\Delta = \begin{pmatrix} 0 & \mathrm{LO}\, C \\ \mathrm{LO}\, C^\top & \mathrm{LO}^2 K_{S_{-g}} \end{pmatrix}.$$

The off-diagonal block matrix with blocks C and $C^\top$ has operator norm $\|C\|_2$, and $|C_{ab}| \le 1$, so $\|C\|_2 \le \|C\|_F \le \sqrt{n_g m}$. Also $\|K_{S_{-g}}\|_2 \le m$ because $K_{S_{-g}}$ is a Gram matrix of m unit vectors. Hence $\|\Delta\|_2 \le \mathrm{LO}\sqrt{n_g m} + \mathrm{LO}^2 m$. Since $A_0 \succeq \rho I$, $\|A_0^{-1}\|_2 \le 1/\rho$, and the stated bound on LO implies $\|A_0^{-1}\Delta\|_2 \le \eta < 1$. Then

$$\log\det(A_0 + \Delta) - \log\det(A_0) = \log\det(I + A_0^{-1}\Delta),$$

The matrix $A_0^{-1}\Delta$ is similar to the symmetric matrix $A_0^{-1/2}\Delta A_0^{-1/2}$, so its eigenvalues are real; the norm bound places them in $[-\eta, \eta]$. Hence every eigenvalue of $I + A_0^{-1}\Delta$ lies in $[1 - \eta, 1 + \eta]$. Therefore

$$\left|\log\det(I + A_0^{-1}\Delta)\right| \le k \max\left\{-\log(1-\eta), \log(1+\eta)\right\} = -k\log(1-\eta),$$

which is (20). □

Corollary 1 (Integer floor ceiling for a partition)**.** *If $G_1, \ldots, G_M$ partition $[n]$, then for every size-k set S*

$$\min_g n_g(S) \;\le\; \lfloor k/M \rfloor.$$

This follows from $\sum_g n_g(S) = k$. Counts differing by at most one attain the bound, but that condition is not necessary in general. The partition assumption matters: with overlapping groups, $\sum_g n_g(S) = \sum_{i \in S} |\{g : i \in G_g\}|$ may exceed k, so $\lfloor k/M \rfloor$ is no longer a general upper bound.

Proof of corollary 1. If the groups partition $[n]$, then $\sum_g n_g(S) = k$, so their minimum cannot exceed $\lfloor k/M \rfloor$. If the counts differ by at most one, they equal $\lfloor k/M \rfloor$ or $\lceil k/M \rceil$, and the bound is attained. The converse is false: for $k = 8$ and $M = 3$, counts $(2, 2, 4)$ attain the bound despite differing by two. With overlapping groups, double-counting memberships gives

$$\sum_g n_g(S) = \sum_{i \in S} |\{g : i \in G_g\}|,$$

which can exceed k. In particular, if all k selected items belong to every category, then $\min_g n_g(S) = k$; the partition ceiling does not extend to arbitrary overlapping categories. □

A.3 Interpretation of the selection criterion

The main text compares category scores; Appendix H extends the criterion to aligned views. The power-mean interpretation and the two-kernel illustration in fig. 4 provide additional context for that aggregation.

Remark 2 (Generalized-mean interpretation). Let $h_q(S) := \det(L_S^{(q)})^{1/k} = \exp d_q(S)$ and $p := -1/\tau < 0$. Then the score is exactly

$$F_\tau(S) = \log\left[\left(\frac{1}{M}\sum_{q=1}^{M} h_q(S)^p\right)^{1/p}\right].$$

Thus (5) is the logarithm of a negative-order power mean, an established smooth maximin criterion in robust design and multiobjective optimization (Dette et al., 2007; Anari et al., 2019). Our use of it for separate principal minors and the resulting Stiefel/NEPv structure are the problem-specific parts.

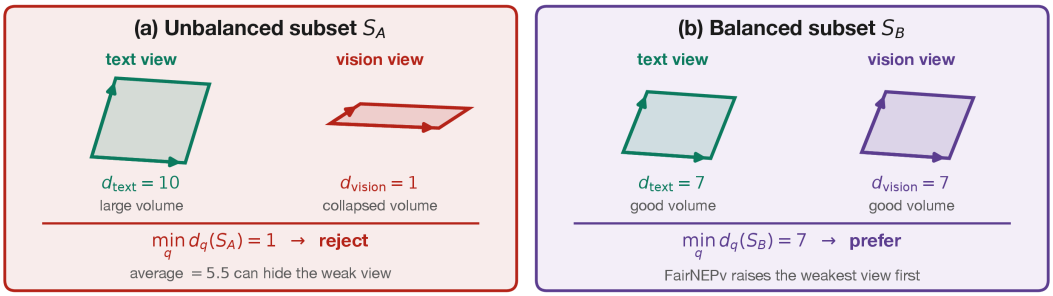

Figure 4: **Maximin compares the weakest kernel score.** Illustrative per-kernel log-determinants $d_q(S)$ for two size-k subsets, drawn with two representation kernels; the same comparison applies to two group kernels. (a) S_A is strong in one kernel but collapsed in the other, so an arithmetic average of per-kernel log-scores can hide the failure. (b) S_B is preferred by $\min_q d_q(S)$. For sufficiently small temperature, the soft minimum in (5) makes the same choice; the displayed score bound quantifies its deviation from the exact minimum. With membership kernels, a low score indicates weak representation under that category kernel, not necessarily zero selected category members.

Remark 3 (Exchange rate, not a balancer). In (17), the coefficient $2\log(1/\mathrm{LO})$ is the price in pooled log-volume of one additional item for the least represented category. Smaller positive floors increase this coefficient; at $\mathrm{LO} = 1$, all category kernels reduce to the same pooled kernel $K + \rho I$. The ridge-dominant endpoint instead has the count and in-group-volume terms in (19). The exact minimum-count interpretation in (17) does not distinguish allocations with the same minimum count and pooled volume, even if their other counts differ. Overlap also allows one item to raise several category counts, so the partition ceiling in corollary 1 applies only to the disjoint special case. The finite-temperature objective uses all category scores through its soft-min weights.

A.4 Per-kernel inverse-Gram control

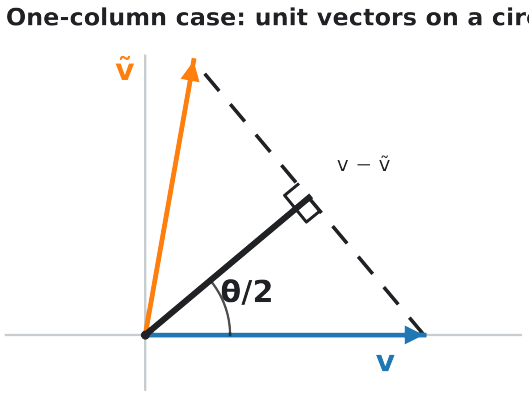

Figure 5: **One-column geometry for the Procrustes bound.** For aligned unit vectors v and $\widetilde{v}$ separated by angle θ, the chord is perpendicular to the central angle bisector at its midpoint, and $\|v - \widetilde{v}\|_2 = 2\sin(\theta/2)$.

Lemma 5 (Inverse-Gram perturbation). *Let $L \succ 0$ be one kernel and let $V, \widetilde{V} \in \mathrm{St}(n,k)$ be orthogonally Procrustes-aligned, meaning $V^\top \widetilde{V} \succeq 0$. If $\eta = \|\sin\Theta(V,\widetilde{V})\|_F < 1/2$, then*

$$\left\|(V^\top L V)^{-1} - (\widetilde{V}^\top L \widetilde{V})^{-1}\right\|_2 \leq \frac{\frac{7}{2}\|L\|_2}{\lambda_{\min}(V^\top L V)\lambda_{\min}(\widetilde{V}^\top L \widetilde{V})}\eta.$$

Proof. Write $A = V^\top L V$ and $\widetilde{A} = \widetilde{V}^\top L \widetilde{V}$. The resolvent identity gives

$$\|A^{-1} - \widetilde{A}^{-1}\|_2 \le \frac{\|A - \widetilde{A}\|_2}{\lambda_{\min}(A)\lambda_{\min}(\widetilde{A})}.$$

For aligned representatives, the largest principal angle satisfies

$$\|V - \widetilde{V}\|_2 = 2\sin(\theta_{\max}/2).$$

Every angle obeys $\sin\theta_j \le \eta < 1/2$, so $\theta_{\max} < \pi/6$. The double-angle identity gives the one-line derivation

$$\sin\theta = 2\sin(\theta/2)\cos(\theta/2) \quad \Longrightarrow \quad 2\sin(\theta/2) = \frac{\sin\theta}{\cos(\theta/2)}.$$

Therefore,

$$\|V - \widetilde{V}\|_2 = \frac{\sin\theta_{\max}}{\cos(\theta_{\max}/2)} \le \frac{\eta}{\cos(\pi/12)} \le 1.04\,\eta.$$

Expanding $A - \widetilde{A}$ by adding and subtracting $V^\top L\widetilde{V}$ gives $A - \widetilde{A} = V^\top L(V - \widetilde{V}) + (V - \widetilde{V})^\top L\widetilde{V}$, and since $\|V\|_2 = \|\widetilde{V}\|_2 = 1$,

$$\|A - \widetilde{A}\|_2 \le 2\|L\|_2\|V - \widetilde{V}\|_2 \le 2.08\,\|L\|_2\eta \le \tfrac{7}{2}\|L\|_2\eta.$$

Substitution proves the claim. Applying the lemma to each $L^{(q)}$ controls all inverse-Gram factors in A_τ. □

A.5 Spectrum of the shifted adaptive operator

This subsection computes the spectrum of $H_{\tau,\sigma}(V)$ in closed form at an arbitrary V. The consequence used elsewhere is that the top-k eigenspace is always well defined, with an eigengap bounded below by $1/k + \sigma$ *uniformly in* V, so no additional leading-eigenspace isolation hypothesis is needed.

The identity $V^\top A_\tau(V) = I_k/k$ gives the orthogonal split

$$A_\tau(V) = V/k + A_\perp(V), \qquad A_\perp(V) = (I - VV^\top)A_\tau(V), \qquad V^\top A_\perp(V) = 0. \tag{21}$$

The Grassmann gradient is $2A_\perp(V)$, so this residual vanishes exactly at stationary subspaces. Figure 6 illustrates the distinction between basis rotation and subspace motion. Substitution into (10) gives

$$H_{\tau,\sigma}(V) = d\,VV^\top + A_\perp V^\top + VA_\perp^\top, \qquad d = k^{-1} + \sigma. \tag{22}$$

The first term acts within the current subspace; the other two terms couple it to the orthogonal residual directions. The following proposition diagonalizes those two-dimensional couplings.

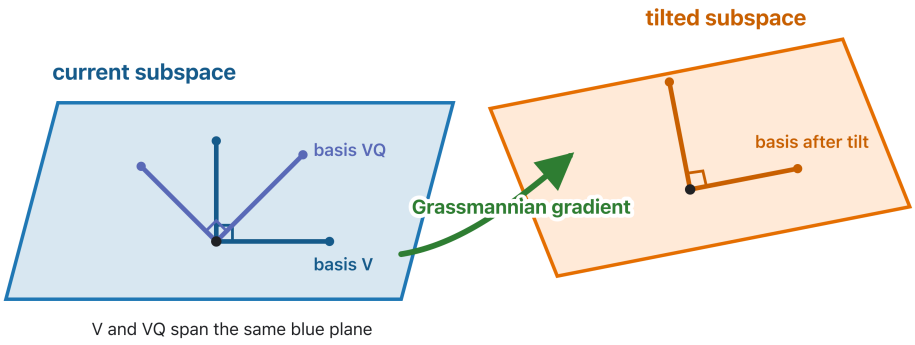

Figure 6: **Grassmannian gradient as subspace motion.** The orthonormal frames V and VQ span the same blue plane, so an in-plane basis rotation does not change the represented subspace. The Grassmannian gradient discards that rotation and retains only the normal component $2A_\perp(V)$, which tilts the plane toward a different subspace.

Proposition 5 (Closed-form spectrum). *Let $A_\perp = U\Sigma W^\top$ be a thin SVD with $r = \operatorname{rank} A_\perp \le \min(k, n-k)$ and singular values $s_1 \ge \cdots \ge s_r > 0$. Then the eigenvalues of $H_{\tau,\sigma}(V)$ are*

$$\lambda_j^{\pm} = \frac{d \pm \sqrt{d^2 + 4s_j^2}}{2} \quad (j \in [r]),$$

together with d of multiplicity $k-r$ and 0 of multiplicity $n-k-r$.

Proof. Put $p_j := Vw_j$. The p_j are orthonormal, span a subspace of $\operatorname{range}(V)$, and are orthogonal to the u_j because $V^\top A_\perp = 0$. From (22),

$$H_{\tau,\sigma} p_j = d\, p_j + s_j u_j, \qquad H_{\tau,\sigma} u_j = s_j p_j,$$

so each plane $\operatorname{span}\{p_j, u_j\}$ is invariant and carries the block $\left(\begin{smallmatrix} d & s_j \\ s_j & 0 \end{smallmatrix}\right)$, with eigenvalues $\lambda_j^{\pm}$. Complete $\{p_j\}_{j \le r}$ to an orthonormal basis of $\operatorname{range}(V)$; the added vectors satisfy $A_\perp^\top p = 0$, hence $H_{\tau,\sigma} p = d\, p$. Any vector orthogonal to $\operatorname{range}(V) + \operatorname{range}(A_\perp)$ is annihilated. Counting $2r + (k-r) + (n-k-r) = n$ exhausts the spectrum. □

Corollary 2 (Inertia and unconditional eigengap). *For every $V \in \operatorname{St}(n,k)$ and every $\sigma \ge 0$, the inertia of $H_{\tau,\sigma}(V)$ is $(k,\, n-k-r,\, r)$, written as (positive, zero, negative). In particular it has exactly k positive eigenvalues, so its top-k eigenspace is well defined and unique, and*

$$\lambda_k\big(H_{\tau,\sigma}(V)\big) \;\ge\; d = \frac{1}{k} + \sigma \;>\; 0 \;\ge\; \lambda_{k+1}\big(H_{\tau,\sigma}(V)\big),$$

where $\lambda_{k+1} = 0$ in the generic case $n > k+r$ and $\lambda_{k+1} = \lambda_r^- < 0$ in the boundary case $n = k+r$. The eigengap is therefore at least d in both cases. At a stationary point $V^\star$ we have $A_\perp = 0$, so $H_{\tau,\sigma}(V^\star) = d\, V^\star V^{\star\top}$ and the gap is attained, $\delta_\star = 1/k + \sigma$.

Proof. By proposition 5, $\lambda_j^+ > d > 0 > \lambda_j^-$, so the positive eigenvalues are the r values λ_j^+ together with the $k-r$ copies of d, giving exactly k; the negative ones are the r values λ_j^-. The smallest positive eigenvalue is at least d, and it equals d whenever $r < k$, so the stated bound is tight. The zero eigenvalue has multiplicity $n-k-r$, which is the inertia claim and holds for every $n \ge k+r$; when $n = k+r$ that multiplicity is zero and $\lambda_{k+1} = \lambda_r^- < 0$, so the gap only widens. □

Corollary 3 (Fixed points). $\operatorname{range}(\mathcal{T}_{\tau,\sigma}(V)) = \operatorname{range}(V)$ *if and only if $A_\perp(V) = 0$, that is, if and only if V is stationary for (7). Stationary subspaces and fixed points of the idealized map therefore coincide exactly.*

Proof. If $A_\perp = 0$ then $H_{\tau,\sigma} = d\, VV^\top$ by (22) and the top-k eigenspace is $\operatorname{range}(V)$. Conversely, if $A_\perp \ne 0$ then $s_1 > 0$ and, by the proof of proposition 5, the eigenvector for λ_1^+ has a nonzero component along $u_1 \perp \operatorname{range}(V)$, so the top-$k$ eigenspace differs from $\operatorname{range}(V)$. □

Lemma 6 (Per-step rotation bound). *The principal angles between $\operatorname{range}(V)$ and $\mathcal{T}_{\tau,\sigma}(V)$ are*

$$\theta_j = \tfrac{1}{2} \arctan\big(2s_j/d\big) \quad (j \in [r]),$$

and 0 otherwise. Hence $\theta_j < \pi/4$ for every j, and each nonzero angle ($j \le r$) is strictly decreasing in σ; the remaining angles stay zero.

Proof. Within the plane $\operatorname{span}\{p_j, u_j\}$ the eigenvector for λ_j^+ is $(\cos\theta_j, \sin\theta_j)$ with $\tan\theta_j = (\lambda_j^+ - d)/s_j$. The double-angle identity turns this into $\tan 2\theta_j = 2s_j/d$. The remaining $k-r$ basis vectors lie in $\operatorname{range}(V)$ and contribute zero angles. Since $\arctan < \pi/2$ we get $\theta_j < \pi/4$, and d is increasing in σ. □

Remark 4 (The $k = 1$ case). For $k = 1$ and $\sigma = 0$, writing $c = \|A_\perp\|_2$ in the basis (v, u) gives $H = \left(\begin{smallmatrix} 1 & c \\ c & 0 \end{smallmatrix}\right)$. At $c = 1$ this is the Fibonacci matrix, with eigenvalues the golden ratio φ and $-1/\varphi$, and a tilt of $\frac{1}{2}\arctan 2 \approx 31.7°$. Letting $c \to 0$ collapses the rank from $2k$ to k, which is the rank drop at stationarity.

Lemma 7 (Exact $2k$ compression and QR recipe)**.** *Assume $n \geq 2k$. Choose $Q_2 \in \mathbb{R}^{n\times k}$ with orthonormal columns such that $V^\top Q_2 = 0$ and $A_\perp = Q_2 R_{22}$, completing an economy QR basis arbitrarily if $A_\perp$ is rank deficient. Set $Q = [\,V,\ Q_2\,]$. Then*

$$H_{\tau,\sigma}(V) = Q\, M_\sigma\, Q^\top, \quad M_\sigma = \begin{pmatrix} d\, I_k & R_{22}^\top \\ R_{22} & 0 \end{pmatrix} \in \mathbb{R}^{2k\times 2k}.$$

The compression preserves every nonzero eigenvalue and the top-k eigenspace exactly. When $n > 2k$, the omitted orthogonal complement contributes only $n - 2k$ additional zero eigenvalues.

Proof. By proposition 2, the residual after projecting $A_\tau(V)$ onto $\mathrm{range}(V)$ is $A_\tau(V) - V/k = A_\perp$. The stated QR construction therefore exists, including at rank-deficient points. Evaluating (22) in the basis Q gives $V^\top H_{\tau,\sigma} V = d\, I_k$, $V^\top H_{\tau,\sigma} Q_2 = A_\perp^\top Q_2 = R_{22}^\top$, and $Q_2^\top H_{\tau,\sigma} Q_2 = 0$, since $V^\top Q_2 = 0$. Exactness follows because $\mathrm{range}(H_{\tau,\sigma}) \subseteq \mathrm{range}(Q)$ by (22). □

Proof of proposition 3. The decomposition is (22); the inertia, the count of positive eigenvalues, the uniqueness of the top-k eigenspace and the bound $\lambda_k - \lambda_{k+1} \geq 1/k + \sigma$ with equality at stationarity are corollary 2; the rotation angles (12) are lemma 6. □

Consequently one SCF step needs a single thin QR, at $O(nk^2)$, and one dense eigendecomposition of M_σ, at $O(k^3)$, after the feature-map products of section 5.2; the top-k eigenvectors of $H_{\tau,\sigma}$ are Q times those of M_σ. In a distributed setting the rows of each Φ_q are sharded, so each kernel requires an all-reduce of $\Phi_q^\top V \in \mathbb{R}^{p_q\times k}$. Compressed Gram matrices require $k \times k$ reductions, and residual tests require scalar reductions. Distributed QR reduces $O(k^2)$ triangular factors per reduction stage; total communication depends on the collective scheme. M_σ can then be diagonalized redundantly.

A.6 Local adaptive-SCF contraction

Proof of theorem 2. At a stationary point, theorem 1 gives $H_\tau(V^\star) = k^{-1} V^\star V^{\star\top}$, so after level shifting corollary 2 supplies the eigengap outright, $\delta_\star = 1/k + \sigma > 0$, and its leading invariant subspace is $V^\star$. No additional leading-eigenspace isolation hypothesis is required. For a sufficiently small neighborhood, Davis-Kahan gives

$$d_{\mathrm{ch}}\big(\mathcal{T}_{\tau,\sigma}(V), V^\star\big) \leq C_{\mathrm{DK}} \frac{\|H_{\tau,\sigma}(V) - H_{\tau,\sigma}(V^\star)\|_2}{\delta_\star}.$$

Lemma 5 controls each inverse Gram matrix. Since $\tau > 0$, the soft-min weights are smooth functions of the finite vector $(g_1(V), \ldots, g_M(V))$. Product-rule bounds in (10), together with the projector shift, give the local Lipschitz condition in theorem 2. Substitution produces the factor $C_{\mathrm{DK}} C_{\tau,\sigma}/\delta_\star = \gamma < 1$. Choose a closed ball around $V^\star$ on which both estimates hold. The one-step bound maps this ball into itself because $\gamma < 1$, so induction yields the geometric bound. □

Remark 5 (Scope). The contraction result is local. The isolation of the leading eigenspace is no longer a hypothesis, since corollary 2 supplies it unconditionally, but the Lipschitz constant $C_{\tau,\sigma}$ is not computed here: proposition 5 diagonalizes $H_{\tau,\sigma}$ at a *fixed* V and says nothing about how $A_\perp$ varies with V. The result does not establish global convergence, global optimality, or convergence of the damped update.

B Per-Kernel Diagnostics and Exhaustive Search

For a selected set S, report the vector $(d_1(S), \ldots, d_M(S))$, its minimum, and the common fair score $F_\tau(S)$. The single-kernel value

$$U_{k,q} := \frac{1}{k} \sum_{i=1}^{k} \log \lambda_i(L^{(q)})$$

upper-bounds the per-item log-determinant for kernel q. Since the soft-min is coordinatewise increasing, these values give the global certificate

$$\max_{V\in \mathrm{St}(n,k)} f_\tau(V) \leq -\tau \log\left[\frac{1}{M}\sum_{q=1}^{M} e^{-U_{k,q}/\tau}\right].$$

The bound is generally unattained because the kernel-wise maximizing eigenspaces can differ. For small instances, exhaustive enumeration of (5) gives the exact discrete fair optimum and measures the discrete suboptimality $\max_{|S|=k} F_\tau(S) - F_\tau(\widehat{S})$ of a returned subset directly. It requires evaluating all $\binom{n}{k}$ subsets and is not used at scale.

C Implementation and Empirical Record

C.1 Adaptive iteration

Algorithm 1 specifies the mathematical iteration. Parameters must be reported for an empirical run, including initialization, temperature, ridge, shift, damping, residual tolerance, iteration cap, screening factor, and swap budget. Reaching the cap returns an endpoint, not a convergence certificate. The projection and alignment operations preserve the frame constraint. The main-paper algorithm is supplemented by the expanded derivations below.

C.2 Available experimental details

The main-paper tables preserve the numerical results in the supplied trademark experiments. This section documents their stated setup and limitations; it is not a reconstruction of execution logs.

OpenLogo. The derived corpus is reported as 284 brands and 43,195 crops. The original OpenLogo benchmark reports 352 classes and 27,083 images (Su et al., 2018); images and extracted crops are different units. The extraction/version mapping behind the supplied crop totals is not specified. The usable split totals 278 brands and 34,831 crops across training, validation, and test. The record does not enumerate the filtering that accounts for this difference. The training subset contains 500 crops. The kernel features are normalized crop-minus-exemplar residuals. The downstream matcher uses DINOv2 and is selected by validation mAP; test queries are registered marks from 80 unseen brands and the gallery contains 11,233 crops. The reported comparison of residual versus crop embeddings is a 4.50-point mAP difference, but a complete ablation table and run-level uncertainty are not supplied. The stratified control matches category coverage and minimum count, not the full per-brand allocation profile.

USPTO. Descriptions are called recorded references here: their inclusion in a USPTO record does not establish examiner authorship. Applicants supply mark descriptions, which can be amended during examination (USPTO, 2026a). Categories are the 29 design-code roots, with 1–17 memberships per mark. All methods select 500 marks and cover all roots. The supplied floor is $\mathrm{LO} = 10^{-15}$. Qwen3-VL-4B is fine-tuned with LoRA, and the reported downstream summaries average three seeds. Judge recall evaluates 9,220 mark–code pairs per run against recorded descriptions. Qwen3 embedding cosine and BGE-M3 reference retrieval provide complementary readings. The record reports that the recall advantage over uniform holds across the three seeds, but does not supply their individual values or uncertainty intervals. Mean values alone therefore do not support a significance test.

Configuration and evidence still needed. The supplied text does not specify the complete feature-extractor versions, category-kernel ridge values, temperatures, selector implementations and stopping records for the trademark tables, or all downstream optimizer settings. It also does not supply the full split and selected-item identifiers, the USPTO pool and test sizes, LoRA configuration, judge model and prompt, or individual seed outcomes. In particular, the stated membership floor is insufficient to identify whether the zero-ridge count identity or the ridge-dominant block interpretation applies to those runs. These details must be recovered from experimental artifacts for exact reproduction; no values are inferred here.

Interpretive limits. Higher nearest-neighbor distance is evidence of dispersion in the chosen selection features, not proof of a causal explanation for retrieval gains. Likewise, higher category recall under a reference-conditioned judge is not a measure of hallucination rate, precision, or legal accuracy. Length-matched judging, per-category outcomes, per-seed uncertainty, and alternative judge

checks would be needed to resolve these questions. No such additional experiments are claimed in this revision.

C.3 Additional reported observations and application context

The supplied OpenLogo account describes photographic logo queries in the wild, including registered marks and look-alikes; its numerical retrieval protocol, however, uses registered marks as queries and held-out crops as the gallery. Those directions should not be interchanged. In the illustrated USPTO record, serial 98037493, the drawing combines a recreational vehicle, tree, tent, mountains, birds, and sun, associated with roots 01, 03, 05, 06, 07, and 18. The recorded description enumerates these pictorial elements (USPTO, 2026b;a). The task evaluates generated descriptions, rather than a complete registration decision.

The OpenLogo source also compares category-allocation profiles with a uniform allocation over 158 brands: it describes stratification as the flattest, FairNEPv as more even than averaged-kernel and fixed-weight selection, and uniform item sampling as reflecting the pool imbalance. It does not give numerical divergence values, so none are added here. The minimum, median, maximum, and coverage do not determine the entire count profile. The reported NN distances imply about 15% greater dispersion than stratification and 21% greater than averaged-kernel selection; fixed-weight is about 1% higher than FairNEPv while omitting categories. These percentages describe this feature-space statistic, not a guaranteed diversity effect in a trained model.

In the supplied USPTO account, one pool category contains more than half of the marks and the smallest category has fewer than one tenth as many. It reports a preference for multi-category marks under all three determinant selectors, with fixed-weight accumulating the most total category memberships. It also reports that the recall benefit over uniform holds in each of three seeds, exceeds the within-method seed spread by more than an order of magnitude, and occurs in most categories, with no category decreasing by more than one point. Those claims require per-seed and per-category outputs to substantiate; the aggregate table does not independently establish them. They are retained here as reported observations, not significance or causal conclusions. The account notes that the methods describing more codes also produce longer text, motivating the length-confounding limitation in the main discussion.

The original empirical interpretation sometimes treated the membership floor as making the selector a pure count rule and assigned downstream margins solely to balance or diversity. The category theory only supplies such exact count factorizations under stated endpoint assumptions; finite ridge and finite temperature require their full determinants. This revision retains the numerical evidence and profile observations while replacing those unqualified mechanistic assertions with the conditional interpretation in section 4.2. Likewise, similar cosine scores do not prove a metric is insensitive, and reference-conditioned judging need not bias all selectors equally.

D Extended Related-Work Discussion

Determinantal point processes, applications, and relaxations. DPPs were popularized in ML by Kulesza & Taskar (2012) as models for selecting subsets whose elements are both high-quality and mutually different, and are used for diverse recommendation, summarization, and coreset construction (Tremblay et al., 2019; Gollapudi et al., 2023). Sampling and normalization are tractable, but MAP inference (26) is NP-hard (Çivril & Magdon-Ismail, 2009). The same principal-submatrix problem is classical maximum-entropy sampling (MESP), with spectral and continuous upper bounds, branch-and-bound methods, and exact algorithms (Ko et al., 1995). Greedy and lazy-greedy algorithms exploit submodularity of a suitably regularized log-determinant, while faster variants approximate marginal gains (Chen et al., 2018; Han et al., 2017). MCMC methods target sampling rather than MAP (Anari et al., 2016). Leverage scores, volume sampling, projection DPPs, rank-revealing QR, and exchange methods turn a subspace or fractional design into coordinates (Gu & Eisenstat, 1996; Belhadji et al., 2020; Nikolov, 2015; Singh & Xie, 2020). Accordingly, our screen-and-swap rounding is a heuristic adaptation whose candidate evaluations use the shared-subset soft-min objective; we do not claim a new general rounding principle or approximation factor.

On the continuous side, the closest DPP-specific *indicator-vector* relaxation is the softmax extension of Gillenwater et al. (2012). For a fractional membership vector $x \in [0,1]^n$, their extension is

$$\widetilde{F}(x) \;=\; \log\det\big(I + \mathrm{diag}(x)(L - I)\big), \tag{23}$$

which is differentiable wherever its determinant is positive and can be optimized over a polytope, followed by rounding. Related log-determinant relaxations appear in maximum-entropy sampling, largest principal subdeterminant, and D-optimal design, where one solves

$$\max_{x\in[0,1]^n,\ \mathbf{1}^\top x = k} \log\det\left(\sum_{i=1}^{n} x_i\, a_i a_i^\top + \lambda I_d\right), \tag{24}$$

where $a_i \in \mathbb{R}^d$ and $\lambda > 0$. Under $p_i = x_i/k$, the equivalent constraints are $p_i \geq 0$, $\sum_i p_i = 1$, and $p_i \leq 1/k$; dropping the upper bounds gives the related approximate-design relaxation (Nikolov, 2015; Singh & Xie, 2020). Their geometry is simplex geometry, with Frank–Wolfe, projected gradient, exchange, or randomized rounding in the coordinates x_i. Orthogonality-constrained determinant objectives also occur in determinant formulations of PCA (Hauser et al., 2018). Our relaxation instead enlarges the set of coordinate-selector frames to all k-dimensional subspaces.

These single-objective formulations motivate different relaxations of selection. Fractional-coordinate methods and subspace methods retain different structures; neither alone specifies how to balance several kernel scores on one subset.

Category constraints and multiple kernels. For category-based selection, partition constraints or prescribed group counts restrict how many selected items belong to each group (Celis et al., 2018; Nikolov & Singh, 2016). Such methods can jointly enforce representation and favor diverse items; they are not equivalent to running independent selectors in every category. Quotas are appropriate when counts are requirements, whereas our objective compares category scores without requiring a quota vector. The independent baseline in section F.1 instead motivates the coordination problem when memberships overlap.

Multiple-kernel models address a related but distinct choice. Large-margin DPPs learn a kernel from several bases (Chao et al., 2015), and mixtures combine DPP distributions (Matsuoka et al., 2021). Multi-DPP summarizes events across video streams (Elfeki et al., 2022); that setting need not retain every view of each selected item, as required by section H. Product DPPs multiply principal determinants of several matrices (Ohsaka & Matsuoka, 2020); at fixed subset size, maximizing that product is equivalent to maximizing their sum of log-determinants. MS-DPP also combines attribute-specific determinant scores for multi-source retrieval (Sogi et al., 2025). Our distinction is the soft-min aggregation of separate scores, used with either category kernels or aligned-view kernels.

Aggregating kernel scores. Write $d_q(S) = k^{-1}\log\det(L_S^{(q)})$ for the per-item log-determinant of kernel q. A useful way to read the methods above is that each chooses an *aggregation* $\mathcal{A}$ mapping the per-kernel scores (or kernels) to one scalar objective. Averaging the kernels first, $\bar{L} = M^{-1}\sum_q L^{(q)}$, aggregates *before* the determinant and has an ordinary single-kernel spectral relaxation; its discrete subset problem remains combinatorial. A fixed-weight sum $\frac{1}{M}\sum_q d_q(S)$ aggregates *after* the determinant with constant weights; this is the arithmetic-mean (utilitarian) choice and permits compensation, since surplus volume in a well-served group offsets a weak one term for term. The maximin (egalitarian) choice aggregates by the minimum, $\mathcal{A} = \min_q d_q(S)$, approximated by the soft minimum F_τ of (D-Fair). The exact minimum scores only the weakest value, whereas finite-temperature smoothing depends on all scores, within the additive bound in section F.3. Maximizing the worst of several set functions under a cardinality constraint is classical robust or multiobjective subset selection (Krause et al., 2008; Udwani, 2018; Spaeh & Miyauchi, 2025), and soft-min/log-sum-exp smoothing is standard (Nesterov, 2005; Anari et al., 2019); many robust-submodular approximation guarantees require monotonicity, which principal log-determinants need not satisfy. Such guarantees therefore cannot be imported without checking their assumptions. Worst-model determinant criteria also have a long history in compound and maximin D-optimal design (Dette et al., 2007; Braess & Dette, 2007), where negative-order generalized means smoothly approach the minimum efficiency. The same worst-case idea appears in subspace optimization, where fair PCA and multi-criteria dimensionality reduction optimize one orthogonal

subspace against several group utilities (Samadi et al., 2018; Tantipongpipat et al., 2019; Zalcberg & Wiesel, 2021; Xu et al., 2024); their utilities are reconstruction losses, projected traces, or variances, whereas ours are compressed per-kernel log-determinants.

The exact minimum targets the weakest score, not equality of all category counts or view scores. Its soft approximation retains that emphasis without a hard representation guarantee. Both fixed-weight sums and the soft minimum generally have subspace-dependent stationarity equations; the adaptive weights are an additional feature of the latter. General NEPvframeworks for Stiefel optimization already exist (Li, 2026). Our contribution is the determinant-specific lifting and spectral structure in section G, rather than a new general maximin principle.

E Additional Background

E.1 Determinantal Point Processes

Given a positive semidefinite (PSD) kernel matrix $L \in \mathbb{R}^{n\times n}$ over a ground set $[n] = \{1, \dots, n\}$, an *L-ensemble* DPP has probability

$$\Pr(\mathbf{S} = S) = \frac{\det(L_S)}{\det(I + L)}, \qquad S \subseteq [n], \tag{25}$$

where L_S is the principal submatrix indexed by S. If $\operatorname{rank}(L) \geq k$, conditioning on $|S| = k$ defines a k-DPP. With the convention $\log 0 = -\infty$, DPP-*MAP* inference is

$$S^\star \;\in\; \arg\max_{|S|=k} \log\det(L_S), \tag{26}$$

which is NP-hard (Çivril & Magdon-Ismail, 2009). For regularized PSD kernels the set function $S \mapsto \log\det(L_S + \epsilon I)$ is submodular, and it is also monotone under the sufficient condition that every eigenvalue of $L+\epsilon I$ is at least one. With $f(\varnothing) = 0$ and an exact cardinality constraint, greedy selection then attains $1 - 1/e$ relative to the optimum of this regularized objective.

E.2 Nonlinear Eigenvalue Problems with Eigenvector Dependency

A NEPv problem (Cai et al., 2018; Bai et al., 2018; Li, 2026) takes the form

$$\begin{aligned} H(V)\,V \;&=\; V\,\Lambda, \\ V \in \mathbb{R}^{n\times k}, \quad V^\top V = I_k, \quad \Lambda = \Lambda^\top &\in \mathbb{R}^{k\times k}, \end{aligned} \tag{27}$$

where $H(V) \in \mathbb{R}^{n\times n}$ is symmetric and depends on V. A common *self-consistent field* (SCF) iteration recomputes the problem-specific target invariant subspace of $H(V_t)$. In our construction below, the target is the top-k subspace:

$$V_{t+1} \;=\; \text{TopEigvecs}_k\big(H(V_t)\big), \tag{28}$$

optionally with level-shifting or damping. Such safeguards and local operator-sensitivity/eigengap analyses predate the present problem (Yang et al., 2009; Bai et al., 2022).

E.3 Common notation and standing assumption

We use $q \in [M]$ to index a kernel: it represents a category in section F and a view in section H. The kernels share one ground set of n items, and k is the number of selected items. L_S denotes a principal submatrix, $\log\det$ uses natural logarithms, and $\mathrm{St}(n, k)$ is the set of $n \times k$ matrices with orthonormal columns.

The objective and differential identities use the following kernel assumption; theorem 4 adds a local sensitivity condition.

Assumption 2 (Regularized kernels)**.** Let $M \geq 1$, $1 \leq k < n$, and $\tau > 0$, and let $L^{(1)}, \dots, L^{(M)} \in \mathbb{R}^{n\times n}$ be symmetric positive-definite kernels. In practice we use $L^{(q)} = \Phi_q\Phi_q^\top + \epsilon_q I$ with $\epsilon_q > 0$. The ridge makes every compressed matrix $V^\top L^{(q)} V$ positive definite. The solver analysis assumes positive definiteness throughout; the zero-ridge identities in section A.2 are separate endpoint calculations, not a convergence analysis of that limit.

F Full Development of the Category Formulation

F.1 Overlapping categories and independent selection

Consider a dataset indexed by $[n]$ with M nonempty categories $G_1, \ldots, G_M \subseteq [n]$. We use *category* and *group* interchangeably. Memberships may overlap: the same item can belong to several categories. Given a budget k, our goal is to retain k distinct items that provide adequate representation across the categories while preserving feature diversity. The categories may differ substantially in size, redundancy, and overlap, making the allocation of this limited budget a central part of the selection problem.

As a running example, consider trademarks associated with design words that describe the objects they depict. A trademark containing a bird and a tree belongs to both the bird and tree categories. Selecting it consumes one budget slot but contributes an example to each category. Trademarks serve only as an illustration: the setting applies to any dataset with possibly overlapping category memberships. For a selected set S, write $n_g(S) := |S \cap G_g|$. These counts measure category representation; their sum can exceed $|S|$ because a single item may contribute to several groups.

A natural starting point is to run an independent DPP-based selector on the items belonging to each category. Let u_i denote a shared unit feature vector for item i, let $K_{ij} := \langle u_i, u_j \rangle$, and write $K^g := K_{G_g}$ for the kernel restricted to category g. The same feature representation is used for all categories; category membership does not require a separate modality or feature space. Prescribing a budget k_g for each category gives the following baseline.

Definition 1 (Independent category-wise selection)**.** For each category g, choose $1 \leq k_g \leq |G_g|$ and solve its fixed-size DPP MAP problem independently (assuming $\mathrm{rank}(K_{G_g}) \geq k_g$):

$$S_g^\star \in \operatorname*{arg\,max}_{S_g \subseteq G_g,\ |S_g| = k_g} \log\det(K_{S_g}).$$

The resulting selection is $S_{\mathrm{ind}} = \bigcup_{g=1}^{M} S_g^\star$.

In the running example, one selector chooses diverse bird trademarks and another chooses diverse tree trademarks. This approach addresses diversity within each category, but leaves three difficulties.

First, *overlapping memberships can produce repeated selections*. Both selectors may choose the same bird-and-tree trademark. That item represents both categories but remains a single selected example, so even when $\sum_g k_g = k$, the union can contain fewer than k distinct items. Removing duplicates therefore leaves a further allocation decision about how to use the remaining budget.

Second, *the category budgets must be fixed in advance*. The choice of k_g depends on category sizes, internal diversity, and overlap with other categories. Equal budgets need not reflect these differences, while unequal budgets require an allocation rule before the selected examples are known. In particular, deciding the numbers of bird and tree trademarks separately does not account for which examples can represent both.

Third, *the selectors do not coordinate their decisions*. A bird selector evaluates its candidates without accounting for what the tree selector retains or which other categories remain poorly represented. Consequently, an item's contribution to several categories and its redundancy with examples selected elsewhere are not assessed together.

F.2 Augmented and softened category kernels

A useful representational step is to place every category kernel on the same ground set $[n]$. This retains one index for each item, including items with several memberships, while preserving each category's original within-group similarities.

Definition 2 (Augmented category kernel)**.** Let $D_g = \mathrm{diag}(\mathbf{1}\{i \in G_g\})$. The augmented kernel is $L_0^{(g)} := D_g K D_g \in \mathbb{R}^{n \times n}$, or entrywise,

$$(L_0^{(g)})_{ij} = \begin{cases} K_{ij}, & i, j \in G_g, \\ 0, & \text{otherwise.} \end{cases}$$

Thus every category is represented by a matrix of the same dimension and with the same item ordering. The bird-and-tree trademark retains its index in both kernels and its similarities to other members of each category. An item outside a category has a zero row and column in that category's kernel. This construction makes overlapping memberships explicit without altering the feature geometry or duplicating items in the dataset.

Augmentation alone, however, does not change independent selection.

Proposition 6 (Augmentation preserves category-wise selection)**.** *With* $\log 0 = -\infty$*, for every* $1 \le k_g \le |G_g|$*,*

$$\max_{S\subseteq[n],\ |S|=k_g} \log\det\big((L_0^{(g)})_S\big) = \max_{S\subseteq G_g,\ |S|=k_g} \log\det(K_S).$$

Whenever the optimum is finite, the two problems have exactly the same maximizing subsets.

Indeed, any S containing an item outside G_g produces a zero row and column and hence determinant zero. For $S \subseteq G_g$, the selected principal submatrix is exactly K_S. The equivalence holds whether or not categories overlap. Separately optimizing the augmented kernels therefore reproduces the original category-wise selections: a common representation by itself resolves neither repeated selections, the choice of k_g, nor the lack of coordination between selectors.

Softening category membership. The augmented kernels retain the original category-wise problems, but their zero entries prevent a category from assigning a finite log-determinant to a subset containing any nonmember. For a shared subset to be evaluated finitely by all of these kernels, every selected item would have to belong to every category, and its selected Gram matrix would have to be nonsingular. In the bird–tree example, this would exclude bird-only and tree-only trademarks. We therefore soften the membership weights while keeping the common item index and feature geometry:

$$L^{(g)} = Q_g K Q_g + \rho I_n, \qquad Q_g = \operatorname{diag}(q_g(1),\dots,q_g(n)), \qquad q_g(i) = \begin{cases} 1, & i \in G_g, \\ \mathrm{LO}, & i \notin G_g. \end{cases} \tag{29}$$

Here $0 < \mathrm{LO} \le 1$ is the *membership floor* and $\rho > 0$ is a ridge parameter. The floor gives nonmembers a smaller positive feature weight when $\mathrm{LO} < 1$, and the ridge makes every category kernel positive definite. Thus every candidate subset has a finite score under every category. A bird-and-tree trademark receives weight one in both kernels; a tree-only trademark receives weight LO in the bird kernel and one in the tree kernel. Multiple memberships are expressed by the weights, without treating the categories as separate feature views.

The score combines representation and diversity. At the endpoint $\rho = 0$ with $K_S \succ 0$, diagonal factorization gives

$$\log\det(L_S^{(g)}) = 2n_g(S)\log(1/\mathrm{LO}) + \log\det(K_S) + 2k\log\mathrm{LO}.$$

For $0 < \mathrm{LO} < 1$, selecting more members of category g increases its count term, while the shared determinant rewards feature diversity. Smaller floors increase the relative weight of representation. At $\mathrm{LO} = 1$, all categories instead use the same pooled kernel $K + \rho I_n$. The displayed identity explains the zero-ridge endpoint; with positive ridge, we evaluate the full regularized determinant rather than assume this exact decomposition. Overlapping memberships allow one selected trademark to contribute to several category counts. The endpoint identities and ridge perturbation bound are established in section A.2. These determinant scores are surrogates for representation and diversity; optimizing them does not impose a hard lower bound on every category count.

Remark 6 (Kernel calibration). A common scale is also necessary for comparing categories. All kernels here use the same K, LO, and ρ, fixed before selection. Multiplying a single category kernel by a positive constant c adds $\log c$ to its normalized log-determinant and can change which category scores lowest. Sharing the construction fixes that calibration; the method does not treat independent category rescalings as immaterial.

F.3 From category scores to a soft minimum

Having placed the categories on a common ground set, we now ask how to evaluate one selected subset across all categories. We choose a single set $S \subseteq [n]$ with $|S| = k$, so each item consumes

one budget slot even when it belongs to several categories. Each category evaluates this same set through its own kernel $L^{(q)}$, giving the per-item score

$$d_q(S) := \frac{1}{k} \log\det(L_S^{(q)}).$$

The softened membership kernels above make every $d_q(S)$ finite.

Our aim is to improve the weakest category score. In the running example, a high score for bird trademarks should not compensate for a poor score for tree trademarks. This leads to the maximin objective

$$\max_{S\subseteq[n],\ |S|=k}\ \min_{q\in[M]} d_q(S).$$

The inner minimum identifies the weakest category for a given subset; the outer maximization chooses the subset with the best such score. Maximin criteria are classical in robust design (Dette et al., 2007) and also appear in fair subspace learning (Samadi et al., 2018; Zalcberg & Wiesel, 2021; Xu et al., 2024).

The minimum is generally nonsmooth where the identity of the weakest category changes. To obtain a smooth aggregation of the category scores for the continuous optimization below, we replace it with a *soft minimum*. Our selection objective is to maximize this soft minimum:

$$\max_{|S|=k} \quad F_\tau(S) := -\tau \log\left[\frac{1}{M}\sum_{q=1}^{M} \exp\left(-\frac{d_q(S)}{\tau}\right)\right], \tag{D-Fair}$$

where $\tau > 0$ is the smoothing temperature. Smaller temperatures make the objective track the weakest category more closely; lower-scoring categories receive greater weight in its derivatives. The soft minimum aggregates the category scores, while maximization selects S. For every subset S, the approximation error is controlled by

$$\min_q d_q(S) \le F_\tau(S) \le \min_q d_q(S) + \tau \log M.$$

This differs from two natural alternatives. Averaging the category kernels first and applying one DPP combines them before evaluating determinants and does not retain separate category scores; its continuous relaxation is an ordinary spectral problem, although discrete subset selection remains combinatorial. Averaging the category log-determinants with fixed weights retains separate scores but permits compensation between categories. The exact minimum prevents such compensation, while the bound above quantifies how closely the soft minimum approximates it. Section D places these aggregation choices in the literature.

F.4 From subset selection to a Stiefel relaxation

The soft minimum gives a smooth way to combine category scores, but the selection variable is still discrete: a direct search ranges over $\binom{n}{k}$ subsets. To obtain a continuous optimization problem, we first express a subset as a matrix. For $S = \{i_1, \dots, i_k\}$, let $V_S = [e_{i_1}, \dots, e_{i_k}]$, where e_i is the ith standard basis vector in $\mathbb{R}^n$. Its columns are orthonormal, and $V_S^\top L^{(q)} V_S$ extracts the principal submatrix $L_S^{(q)}$. In the trademark example, each column identifies one selected mark, regardless of how many design categories that mark belongs to.

We relax the requirement that the columns be coordinate vectors and allow any matrix in the Stiefel manifold

$$\mathrm{St}(n,k) := \{V \in \mathbb{R}^{n\times k} : V^\top V = I_k\}.$$

This enlarges the feasible set while preserving the score of every original discrete selection. Each category is now evaluated on the same continuous matrix V, through its compressed kernel. Write

$$G_q(V) := V^\top L^{(q)} V, \qquad g_q(V) := \frac{1}{k}\log\det G_q(V).$$

Our continuous relaxation is

$$\max_{V\in\mathrm{St}(n,k)} \quad f_\tau(V) := -\tau \log\left[\frac{1}{M}\sum_{q=1}^{M} \exp\left(-\frac{g_q(V)}{\tau}\right)\right]. \tag{P-Fair}$$

Lemma 8 ((P-Fair) relaxes (D-Fair))**.** *For every k-subset $S = \{i_1, \dots, i_k\} \subseteq [n]$, let V_S contain the coordinate columns $e_{i_1}, \dots, e_{i_k}$. Then $V_S \in \mathrm{St}(n,k)$ and, for every kernel q,*

$$V_S^\top L^{(q)} V_S = L_S^{(q)}.$$

Consequently $f_\tau(V_S) = F_\tau(S)$, and

$$\max_{V \in \mathrm{St}(n,k)} f_\tau(V) \;\geq\; \max_{|S|=k} F_\tau(S).$$

The objective is unchanged by $V \mapsto VQ$ for $Q \in \mathrm{O}(k)$. (Proof in the appendix.)

Thus the relaxation contains every size-k subset and its global optimum upper-bounds the discrete optimum. The basis invariance in the lemma means that the objective depends on the subspace spanned by V, rather than on a particular orthonormal basis. A general relaxed solution is not itself a list of selected items, so a rounding step is required to recover a discrete subset, as described in section G.4.

The remaining task is to solve this relaxed problem. For a single positive-definite kernel L with ordered eigenvalues $\lambda_1(L) \geq \cdots \geq \lambda_n(L) > 0$, the spectral benchmark is

$$\max_{V \in \mathrm{St}(n,k)} \frac{1}{k} \log\det(V^\top L V) = \frac{1}{k} \sum_{i=1}^{k} \log \lambda_i(L),$$

attained by a leading k-dimensional eigenspace (proposition 4). With several category kernels, their preferred subspaces may differ, and the soft-min weights change with the current category scores. The derivatives also involve each compressed kernel's inverse, which depends on V. Consequently, the first-order conditions do not generally reduce to an eigendecomposition of one fixed matrix. Section G derives these conditions as an adaptive nonlinear eigenvalue problem and develops an iterative solver, followed by rounding to a size-k subset. Separately, section H gives a second kernel construction to which the same objective and relaxation apply.

G Full Solver Derivation and Implementation Discussion

The solver takes the positive-definite kernels from either section F or section H. All statements below depend on these kernels and the shared relaxed problem, not on whether the kernel index denotes a category or a view.

G.1 The adaptive NEPv reformulation

Proposition 7 (Gradient and stationarity)**.** *Define the adaptive kernel weights*

$$w_q(V) := \frac{\exp(-g_q(V)/\tau)}{\sum_{r=1}^{M} \exp(-g_r(V)/\tau)}, \qquad A_\tau(V) := \frac{1}{k} \sum_{q=1}^{M} w_q(V) L^{(q)} V \, G_q(V)^{-1}. \tag{30}$$

Then $\nabla f_\tau(V) = 2A_\tau(V)$ and $V^\top A_\tau(V) = k^{-1} I_k$. Thus V is stationary for (P-Fair) exactly when

$$A_\tau(V) = \frac{1}{k} V. \tag{31}$$

(Proof in the appendix.)

Remark 7 (Weights on the weakest kernel scores). The weight $w_q(V)$ is the softmax of $-g_q(V)/\tau$. A kernel that currently has smaller compressed volume receives a strictly larger weight, and as $\tau \to 0$ the mass concentrates on $\arg\min_q g_q(V)$. For the membership kernels of section F.2, the endpoint identities in lemmas 2 and 3 explain how *discrete* category scores reflect membership counts and diversity. For a dense relaxed subspace, $g_q(V)$ is instead a compressed-kernel score; it need not have an integer-count interpretation. The adaptive weights emphasize whichever category or view currently has the weakest such score.

Theorem 3 (Adaptive NEPv form)**.** *Define the symmetric, gauge-invariant operator*

$$H_\tau(V) := A_\tau(V) V^\top + V A_\tau(V)^\top - \frac{1}{k} V V^\top. \tag{32}$$

Then V is stationary for (P-Fair) *if and only if*

$$\boxed{H_\tau(V)V = V\Lambda, \qquad \Lambda = \frac{1}{k} I_k.} \tag{Fair-NEPv}$$

Moreover, $H_\tau(VQ) = H_\tau(V)$ for every $Q \in \mathrm{O}(k)$. (Proof in the appendix.)

An orthogonal change of frame rotates the basis within the same subspace; fig. 3 in the appendix illustrates this distinction.

Remark 8 (What is specific to this lifting). General symmetric gradient-based constructions can represent orthogonality-constrained KKT systems as NEPvs, and such operators are not unique (Li, 2026). The claim here is narrower: the subtractive projector in (32), together with the identity $V^\top A_\tau(V) = I_k/k$, yields the exact low-rank decomposition, inertia, eigengap, and principal-angle formula in proposition 8. Those spectral properties select a unique leading invariant subspace and enable the exact small compression used by our implementation. More explicitly, the generic lifting in Li (2026, Eq. (6.5)) is $H_{\mathrm{gen}} = (\nabla f_\tau)V^\top + V(\nabla f_\tau)^\top$. Our choice is $H_\tau = \frac{1}{2}H_{\mathrm{gen}} - VV^\top/k$. The rank-at-most-$2k$ bound already follows from that generic construction; it is not by itself a new property. The constant identity $V^\top A_\tau = I/k$ and the chosen projector coefficient give the displayed spectral formulas. These algebraic facts alone do not establish an optimization advantage.

Remark 9 (Not orthogonal iteration). At a stationary point, $H_\tau(V) = k^{-1}VV^\top$. Away from one, both $A_\tau(V)$ and the weights $w_q(V)$ move with the current subspace, so the plain map is not orthogonal iteration on any fixed matrix. The inverse Gram matrix of every kernel and the soft-min weighting both enter the operator.

Remark 10 (Fixed aggregation and adaptive weights). A determinant of the averaged kernel has the single-kernel spectral relaxation in proposition 4. A fixed-weight sum of log-determinants generally retains subspace-dependent inverse-Gram factors. The soft minimum adds score-dependent weights to this coupling. Thus nonlinearity is not unique to maximin, although emphasis on weaker scores is central to the present objective.

Proposition 8 (Unconditional gap and bounded step)**.** *Write $A_\perp = A_\perp(V) := (I - VV^\top)A_\tau(V)$, half the Grassmann gradient of f_τ, let $s_1 \geq \cdots \geq s_r > 0$ be its singular values with $r = \mathrm{rank} A_\perp \leq \min(k, n-k)$, and set $d := k^{-1} + \sigma$. For every $V \in \mathrm{St}(n,k)$ and every $\sigma \geq 0$, the level-shifted operator $H_{\tau,\sigma}(V) := H_\tau(V) + \sigma VV^\top$ satisfies*

$$H_{\tau,\sigma}(V) = d\,VV^\top + A_\perp V^\top + VA_\perp^\top, \tag{33}$$

and has inertia $(k,\ n-k-r,\ r)$, written as (positive, zero, negative). It therefore has exactly k positive eigenvalues, its top-k eigenspace is always well defined and unique, and

$$\lambda_k - \lambda_{k+1} \ \geq\ \frac{1}{k} + \sigma,$$

with equality at a stationary point. Moreover one top-k step rotates the iterate by the principal angles

$$\theta_j = \tfrac{1}{2}\arctan\big(2s_j/d\big) < \tfrac{\pi}{4} \quad (j = 1, \ldots, r), \tag{34}$$

while $\theta_{r+1} = \cdots = \theta_k = 0$. (Proof in the appendix.)

Throughout, $\Theta(V, \widetilde{V})$ denotes the diagonal matrix of principal angles. The two subspaces are $\mathrm{range}(V)$ and $\mathrm{range}(\widetilde{V})$.

Remark 11 (Stationary points and SCF). By proposition 8 the leading eigenspace is always isolated, so the stationary points of (P-Fair) are *exactly* the fixed points of the idealized top-k map. The fair objective is nonconvex, so neither the theorem below nor the implementation claims global convergence or a global optimum.

For a full-column-rank matrix X, $\mathrm{Polar}(X) = X(X^\top X)^{-1/2}$. The angle bound in proposition 8 makes $\widetilde{V}^\top V$ invertible. After alignment, $V^\top \widetilde{V} Q$ is positive definite; hence $V^\top[\alpha\widetilde{V}Q + (1-\alpha)V] \succ 0$, so the damped matrix has full column rank and its polar factor is defined. In the category construction, $\Phi_q = Q_q U$, where row i of $U \in \mathbb{R}^{n\times p}$ is $u_i^\top$, and all ϵ_q equal the common ridge ρ.

Algorithm 2 FAIRNEPV: adaptive SCF for (Fair-NEPv)

Require: Regularized feature-map kernels $L^{(q)} = \Phi_q \Phi_q^\top + \epsilon_q I$, $q \in [M]$; target size k; temperature $\tau > 0$; tolerance ε; maximum iterations $T_{\max}$; damping $\alpha \in (0, 1]$; shift $\sigma \geq 0$.

Initialize $V_0 \leftarrow$ a random orthonormal frame or a problem-specific warm start.
for $t = 0, \ldots, T_{\max} - 1$ **do**
  for $q = 1, \ldots, M$ **do**
    $G_{q,t} \leftarrow V_t^\top L^{(q)} V_t, \quad g_{q,t} \leftarrow k^{-1} \log\det(G_{q,t})$.
  end for
  $g_{\min,t} \leftarrow \min_q g_{q,t}, \quad w_{q,t} \leftarrow \exp(-(g_{q,t} - g_{\min,t})/\tau) / \sum_r \exp(-(g_{r,t} - g_{\min,t})/\tau)$.
  $A_t \leftarrow k^{-1} \sum_q w_{q,t} L^{(q)} V_t G_{q,t}^{-1}$.
  if $\|A_t - V_t/k\|_F < \varepsilon$ **then**
    return V_t
  end if
  Apply $H_\tau(V_t)X \leftarrow A_t(V_t^\top X) + V_t(A_t^\top X) - k^{-1} V_t(V_t^\top X)$ implicitly.
  $H_{\tau,\sigma}(V_t)(\cdot) \leftarrow H_\tau(V_t)(\cdot) + \sigma V_t(V_t^\top \cdot)$ ▷ level shift
  $\widetilde{V}_{t+1} \leftarrow$ top-k eigenvectors of $H_{\tau,\sigma}(V_t)$.
  $Q_t \leftarrow \mathrm{Polar}\big(\widetilde{V}_{t+1}^\top V_t\big)$ ▷ gauge alignment
  $V_{t+1} \leftarrow \mathrm{Polar}\big(\alpha \widetilde{V}_{t+1} Q_t + (1 - \alpha) V_t\big)$ ▷ damping
end for
return $V_{T_{\max}}$ ▷ iteration-cap endpoint

G.2 SCF ITERATION AND VARIANTS

We consider three variants of the iteration in algorithm 2: **(i)** *Plain* SCF ($\alpha = 1, \sigma = 0$); **(ii)** *Damped-only* SCF ($\alpha < 1, \sigma = 0$), which can reduce oscillation when the weakest kernel changes; **(iii)** *Shifted-only* SCF ($\alpha = 1, \sigma > 0$), which raises the uniform eigengap lower bound to $k^{-1} + \sigma$ and damps the step (34). Damping and shifting may also be combined ($\alpha < 1, \sigma > 0$; see remark 13). Because H_τ is gauge-invariant, the eigensolver returns an arbitrary orthonormal basis of the top-k eigenspace, so $\widetilde{V}_{t+1}$ is aligned to V_t before damping; without that step the convex combination would depend on the basis the solver happened to return.

G.3 CONVERGENCE THEOREM FOR THE IDEALIZED SUBSPACE MAP

This subsection states a local convergence result for the *idealized* subspace map

$$\begin{aligned} \mathcal{T}_{\tau,\sigma}(V) &:= \mathrm{TopEigvecs}_k\big(H_{\tau,\sigma}(V)\big), \\ H_{\tau,\sigma}(V) &:= H_\tau(V) + \sigma V V^\top. \end{aligned} \tag{35}$$

It does *not* prove convergence of the damped/polar implementation in algorithm 2; damping and the polar retraction change the map being analyzed. The result is local and conditional: if the current subspace is already close to a stationary subspace $V^\star$ and the local sensitivity condition (36) holds, one application of $\mathcal{T}_{\tau,\sigma}$ shrinks the distance to $V^\star$ by a fixed factor. The proof follows the standard local SCF pattern of combining an operator Lipschitz bound with eigenspace perturbation (Davis & Kahan, 1970; Yang et al., 2009; Bai et al., 2022). Here the bound for H_τ must combine the inverse-Gram perturbation of each kernel with the sensitivity of the soft-min weights. Throughout, distances between subspaces are measured by the chordal metric $d_{\mathrm{ch}}(V, \widetilde{V}) := \|\sin\Theta(V, \widetilde{V})\|_F$.

Theorem 4 (Local contraction for adaptive Fair-SCF)**.** *Assume Assumption 2, and let $V^\star$ be a stationary point of (P-Fair). By proposition 8 the top-k eigenspace of $H_{\tau,\sigma}(V^\star)$ is $V^\star$, with eigengap $\delta_\star = k^{-1} + \sigma > 0$; this is automatic rather than assumed. Let $C_{\tau,\sigma}$ be a local Lipschitz constant such that, after shrinking a neighborhood of $V^\star$ if necessary,*

$$\|H_{\tau,\sigma}(V) - H_{\tau,\sigma}(V^\star)\|_2 \leq C_{\tau,\sigma}\, d_{\mathrm{ch}}(V, V^\star).$$

Let C_{DK} be the Davis–Kahan constant for the chordal distance (for example, $C_{\mathrm{DK}} = 2\sqrt{k}$ after shrinking the neighborhood so that the operator perturbation has norm less than $\delta_\star/2$). If

$$C_{\mathrm{DK}} C_{\tau,\sigma} < \delta_\star = \frac{1}{k} + \sigma, \tag{36}$$

then there is a neighborhood $\mathcal{N} \subset \mathrm{Gr}(n,k)$ of $V^\star$ such that the idealized iteration

$$V_{t+1} = \mathcal{T}_{\tau,\sigma}(V_t) = \mathrm{TopEigvecs}_k\big(H_{\tau,\sigma}(V_t)\big)$$

satisfies

$$d_{\mathrm{ch}}(V_{t+1}, V^\star) \le \gamma\, d_{\mathrm{ch}}(V_t, V^\star), \tag{37}$$

with $\gamma := C_{\mathrm{DK}} C_{\tau,\sigma}/\delta_\star < 1$. Thus $d_{\mathrm{ch}}(V_t, V^\star) \le \gamma^t d_{\mathrm{ch}}(V_0, V^\star)$ for all $V_0 \in \mathcal{N}$.

The full proof is deferred to the Technical Appendix.

Remark 12 (What the theorem does and does not say). The theorem is a local statement for the idealized subspace map $\mathcal{T}_{\tau,\sigma}$. It does not prove convergence of the damped/polar update in algorithm 2, nor global convergence or global optimality. The eigengap is no longer a hypothesis (proposition 8), but the Lipschitz constant $C_{\tau,\sigma}$ is not computed here. The theorem says only that a stationary subspace is attracting from a sufficiently small neighborhood when (36) holds.

Remark 13 (Role of the level shift). By proposition 8 the shift raises the uniform eigengap lower bound to $k^{-1} + \sigma$; at a stationary point the gap equals this value. It also shrinks every nonzero per-step rotation in (34) and changes the local Lipschitz constant $C_{\tau,\sigma}$, which we do not compute, so (36) still does not imply that a larger shift is faster. We use shifting as a numerical stabilizer and recommend choosing it on a validation split, consistent with established level-shifted SCF practice (Bai et al., 2022).

G.4 Rounding to a discrete subset

After algorithm 2 produces a dense k-dimensional subspace $\widehat{V}$, we recover a size-k set $S \subset [n]$ via *leverage-score screening + local greedy refinement*:

1. Compute leverage scores $\ell_i = \|e_i^\top \widehat{V}\|_2^2$; note $\sum_i \ell_i = k$.
2. Form a shortlist $\widetilde{S}$ from the $m = \min\{n, \lceil ck \rceil\}$ largest leverage scores, where $c > 1$.
3. Greedily construct a size-k subset from $\widetilde{S}$, then perform improving one-item swaps against the full candidate pool.

During greedy construction, a nonempty partial set T is scored using (D-Fair) with $d_q(T) = |T|^{-1} \log\det(L_T^{(q)})$; initialize with a singleton and compare candidates of the same size. Completed sets and swaps use the target normalization $1/k$. For aligned views, the returned indices are applied to every view together. The rule is a practical adaptation of established leverage-score, volume, rank-revealing QR, and exchange methods (Gu & Eisenstat, 1996; Belhadji et al., 2020; Singh & Xie, 2020); the problem-specific element is the use of F_τ for greedy additions and swaps. We prove nothing about this particular screen-then-refine procedure here. There is also no exact closed-form fair relaxation value analogous to the single-kernel quantity in proposition 4; section B gives a generally unattained spectral upper bound. We treat the oversampling factor c and the local refinement budget as rounding hyperparameters.

G.5 Scaling considerations

- **Low-rank-plus-ridge kernels.** Let $L^{(q)} = \Phi_q \Phi_q^\top + \epsilon_q I$ with $\Phi_q \in \mathbb{R}^{n \times p_q}$. A direct implementation forms $\Phi_q^\top V$ and $L^{(q)} V$ for each kernel, so one Fair-SCF sweep costs $O(\sum_q (np_qk + nk^2 + k^3))$ for the kernel terms. Cholesky solves can replace explicit Gram inverses, and this implementation need not form an $n \times n$ kernel.
- **Exact range-space compression.** When $n \ge 2k$, form $A_\perp = A_\tau(V) - V/k$ and a QR factorization $A_\perp = Q_2 R_{22}$ with $Q_2^\top V = 0$. Then $Q = [V, Q_2]$ gives $H_{\tau,\sigma}(V) = Q M_\sigma Q^\top$ *exactly*, with $M_\sigma \in \mathbb{R}^{2k \times 2k}$ defined in lemma 7. A rank-deficient residual requires an orthogonal completion within $\mathrm{range}(V)^\perp$. It preserves the nonzero spectrum and top-k eigenspace and replaces the $n \times n$ eigendecomposition by an $O(nk^2)$ QR and an $O(k^3)$ solve. For $n < 2k$, use a compact basis of dimension at most n.
- **Distributed kernels.** Shard rows of each Φ_q across GPUs. Each kernel requires an all-reduce of $\Phi_q^\top V \in \mathbb{R}^{p_q \times k}$ and reductions for its $k \times k$ compressed Gram matrix. Residual tests also

require scalar reductions. Distributed QR reduces $O(k^2)$ triangular factors per reduction stage; total communication depends on the worker count and collective scheme. The small eigenproblem can be solved redundantly.

- **Discrete refinement.** If at most B full one-swap passes are allowed, the refinement evaluates $O(Bnk)$ candidate swaps in addition to the continuous-solver cost; cached determinant updates can reduce the cost per evaluation.

H Aligned Views: Deferred Extension

This full formulation is retained as a deferred extension. The main paper's methodology and empirical claims concern categories. The material below preserves the aligned-view direction without treating it as an evaluated contribution of this revision.

The same selection criterion applies when each item has several aligned representations. Suppose item $i \in [n]$ is observed through M views, $z_i^{(1)}, \ldots, z_i^{(M)}$, where the common index identifies the same underlying item. The views may have different feature spaces and different similarity structures. We seek a single index set $S \subseteq [n]$ with $|S| = k$; view q then retains $\{z_i^{(q)} : i \in S\}$. Thus all views retain k observations with their correspondence preserved. Selecting a separate index set in each view would generally break this alignment.

For example, consider input–output pairs $\{(x_i, y_i)\}_{i=1}^n$. The inputs form one view and the outputs form a second. Selecting index i retains both x_i and its associated y_i, so the result is a subset of k intact pairs. An input kernel measures similarity between inputs, while an output kernel measures similarity between outputs. A subset that is diverse in its inputs may nevertheless contain highly redundant outputs, or vice versa. This two-view example illustrates the setting; the formulation allows any number of aligned views.

For an image-to-description task, the image is the input and its associated text is the output. Figure 7 illustrates the shared-index selection and the comparison of input and output scores. The example is schematic; its score values are illustrative rather than empirical results.

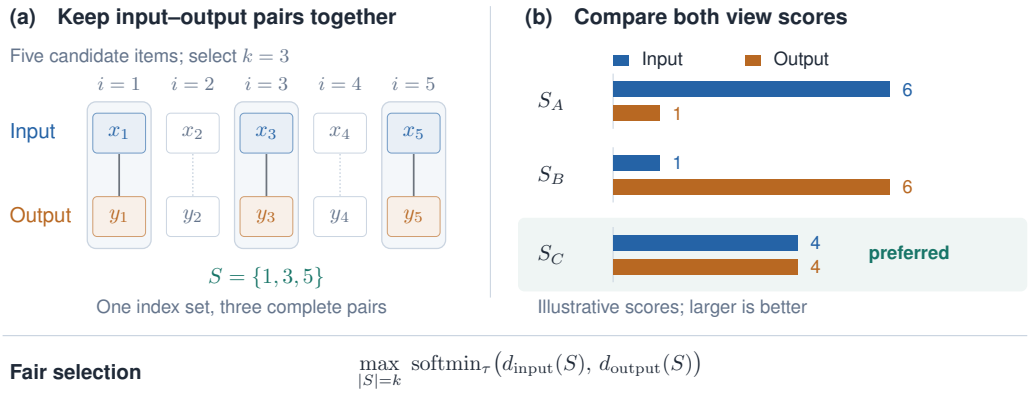

Figure 7: **Fair selection preserves correspondence and compares both views.** (a) Selecting three item indices retains three complete input–output pairs; the highlighted columns show $S = \{1, 3, 5\}$. (b) Illustrative normalized log-determinant scores for three candidate subsets of the same size. S_A and S_B each have a weak view, whereas S_C scores well in both. With $\tau = 0.15$, their soft-min scores are approximately 1.10, 1.10, and 4.00, respectively, so the criterion prefers S_C among these candidates. The construction extends to any number of aligned views.

View-specific kernels on a common index set. For each view q, construct a positive-definite kernel $L^{(q)} \in \mathbb{R}^{n \times n}$ over its observations, with rows and columns ordered by the common item index. For example, use $L^{(q)} = \Phi_q \Phi_q^\top + \rho I_n$ with $\rho > 0$, where row i of Φ_q represents $z_i^{(q)}$. Feature dimensions may differ across views. Here each kernel describes a different representation of all

items; the category-membership augmentation in section F.2 is not needed. As in remark 6, relative kernel scales must be fixed before comparing scores: rescaling one view changes its normalized log-determinant and hence its influence on the selection criterion. Unit-normalized features and a common ridge provide one concrete calibration convention.

The same objective and relaxation. Apply the normalized log-determinant score and soft-min objective of section F.3 to these view kernels. In (D-Fair), q now indexes views and every kernel evaluates the same selected index set S. For the input–output example, this maximizes the soft minimum of input diversity and output diversity, giving more emphasis to the weaker of the two scores. The selection variable remains one shared subset, rather than a collection of independently chosen subsets.

The coordinate selection matrix, Stiefel relaxation, and embedding result in section F.4 apply directly, yielding (P-Fair) with the view-specific kernels. One common relaxed matrix represents the selection across all views, and rounding returns one index set that is used in every view. No new relaxation or soft-min construction is required. Section G therefore provides a common solver for both the overlapping-category formulation and this aligned-view formulation.

I Retained Synthetic Aggregation Illustration

This original synthetic illustration and its numerical results are retained for completeness. It tests aggregation on two diagonal kernels, not the multicategory trademark hypothesis evaluated in the main paper.

We illustrate the aligned-view setting of section H with a deterministic quality-conflict stress test with $k = 5$ (fig. 8). Five text specialists have quality $(100, 0.1)$ in the (text, vision) views, five vision specialists have quality $(0.1, 100)$, and five balanced candidates have quality $(3, 3)$. A kernel average and a fixed geometric mean therefore reward specialist quality, whereas the worst-view criterion rewards a balanced mix. We compare FairNEPv with text-only and vision-only determinant selection, an averaged kernel $(L^{(\text{text})} + L^{(\text{vision})})/2$, and a fixed-weight sum-of-log-determinants ablation. The supplied record attributes the illustration to a `synthetic.py` prototype and uses the same soft-min objective as (D-Fair).

Synthetic findings. The example is built to isolate the fairness mechanism; it is not a general performance claim. With solver seed 17, FairNEPv reaches $F_\tau(S) = 1.120$, against 0.564 for the reported fixed-weight tie-breaking (the original text printed 0.565; the arithmetic correction is detailed in section J.1) and -2.199 for each single-view selector and for the reported averaged-kernel tie-breaking. Both aggregate baselines assign the same objective value to all ten specialists, so neither enforces a balanced tie-breaking. The averaged kernel run fills the budget from one side and leaves the other view at -2.303, while FairNEPv raises the weight on whichever view has low volume and settles on a mix of balanced and specialist items scoring 1.120 in both views. Exhaustive enumeration gives the discrete optimum 1.141, attained by two specialists from each side and one balanced item, so the reported screen-and-swap rounding gap is 0.021. These are soft-min values, not averages: a view that leads by Δ enters with relative weight $e^{-\Delta/\tau}$, so the 6.91-nat gap here suppresses the strong view by 10^{-20} and leaves $F_\tau \approx d_{\min} + \tau \log M = -2.30 + 0.10$. Section J gives the implementation settings and works out every row. This diagonal construction tests score aggregation, not the overlapping-category representation or performance on real input–output data.

J Synthetic Reproduction Details

The supplied account of the `synthetic.py` prototype describes five text specialists with qualities $(100, 0.1)$, five vision specialists with qualities $(0.1, 100)$, and five balanced candidates with qualities $(3, 3)$. It compares FairNEPv against text-only, vision-only, averaged-kernel, and fixed-weight multi-view selection. For every output subset it prints the selected candidate types, each normalized log determinant, and $F_\tau(S)$. The protocol is built to test emphasis on the weaker view; it is not a general ranking of single-kernel methods.

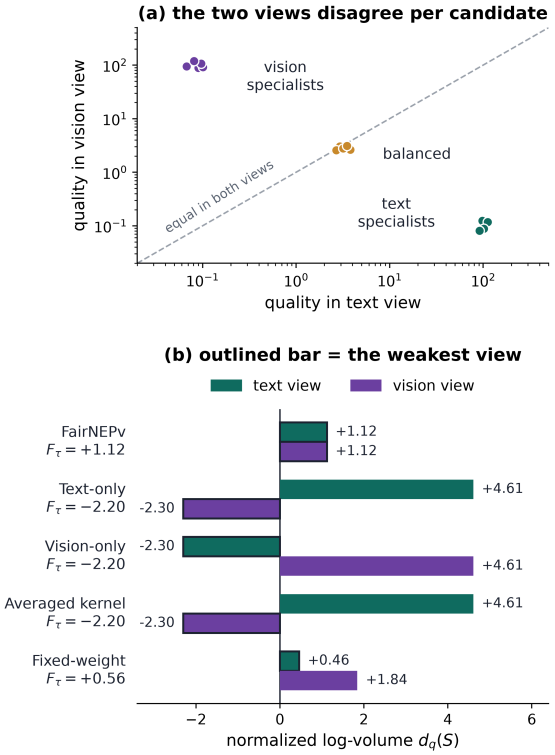

Figure 8: **Quality-conflict fairness stress test** ($k = 5$, $\tau = 0.15$). Both view kernels are diagonal here, so candidate coordinates carry no information and are not plotted. (a) Quality of each candidate in the two views (display jitter separates identical points and does not enter the kernels): specialists are strong in one view and negligible in the other, balanced candidates are moderate in both. (b) Per-view normalized log-determinant $d_q(S)$ of each selector's subset. The outlined bar is the weaker view, which the soft-min F_τ of (D-Fair) tracks, and both bars are outlined when the two views are tied. The single-view selectors collapse the other view. Both aggregate baselines are indifferent among the ten specialist candidates; the reported deterministic tie-breaking returns a one-sided averaged-kernel subset and a two-versus-three specialist split for the fixed-weight baseline.

Reproduction settings. The candidate pool has $n = 15$ and $k = 5$. The two kernels are diagonal with the stated positive qualities as their diagonal entries; these are kernel entries, not feature amplitudes to be squared again. The fair and fixed-weight prototypes use seed 17, at most 200 iterations, $\alpha = 0.7$, $\sigma = 0.05$, and $\tau = 0.15$. They stop when the chordal change between successive subspaces is below 10^{-9}, which is a stagnation criterion rather than the stationarity-residual test in algorithm 2. Rounding uses a leverage pool of $\min(6k, n) = 15$ items, up to ten singleton starts, and at most twenty improving one-swap passes. Small Cholesky jitter in the reference evaluator is not needed for the positive diagonal kernels. Direct calculations below use the stated diagonals; they correct the original fixed-weight rounding from 0.565 to 0.564. The prototype itself is not included in the current source package, so these calculations are not an execution replication. Baseline ties are resolved by the implementation's ordering and are not evidence of a unique aggregate optimum. The prototype uses dense matrices; it does not benchmark the scalable implementation outlined in section G.5.

J.1 Worked example: how the fair scores add up

The scores reported for the synthetic instance follow from (D-Fair), and it is worth seeing why they are not averages of the per-view values. Write $d_{\min} := \min_q d_q(S)$ and $\Delta_q := d_q(S) - d_{\min} \geq 0$. Factoring $\exp(-d_{\min}/\tau)$ out of the sum in (D-Fair) gives the exact identity

$$F_\tau(S) = d_{\min} + \tau \log M - \tau \log \Big[\textstyle\sum_{q=1}^{M} e^{-\Delta_q/\tau} \Big], \tag{38}$$

and hence the two-sided bound

$$d_{\min} \;\leq\; F_\tau(S) \;\leq\; d_{\min} + \tau \log M. \tag{39}$$

The lower bound is attained when all views are tied. For a unique minimum, the upper bound is approached when the sum of $e^{-\Delta_q/\tau}$ over all other views is negligible.

Here $M = 2$ and $\tau = 0.15$, so $\tau \log M = 0.104$. The view kernels are diagonal, so $d_q(S) = k^{-1} \sum_{i \in S} \log q_i$ with q_i the quality of candidate i in view q. Take the text-only subset, which is five text specialists: $d_{\text{text}} = \log 100 = 4.605$ and $d_{\text{vision}} = \log 0.1 = -2.303$. The gap is $\Delta = 6.908$, so the stronger view enters the soft-min with relative weight $e^{-\Delta/\tau} = e^{-46.05} \approx 10^{-20}$ and is negligible at the displayed precision:

$$\begin{aligned} F_\tau &= -0.15 \log \Big[\tfrac{1}{2} \big(e^{-30.70} + e^{15.35} \big) \Big] \\ &\approx -2.303 + 0.104 = -2.199. \end{aligned}$$

Surplus in the text view raises the soft-min by at most $\tau \log 2 = 0.104$ here, which does not erase the collapsed vision score.

Table 3 lists every selector against both ends of (39). Which end is attained is decided by the spread Δ_q alone, through the slack term of (38): $\tau \log[\sum_q e^{-\Delta_q/\tau}]$ equals its maximum $\tau \log M$ when the views are tied, and decays to zero when every nonminimum view leads the unique minimum by more than a few multiples of τ. FairNEPv is the only reported heuristic output with tied views, so its row is at the lower bound $d_{\min}$; the remaining selector rows have $\Delta/\tau \geq 9.2$ and sit on the upper bound to the displayed precision. Being at the lower bound is the signature of balance across views, not necessarily high absolute quality. Being near the upper bound means that there is a unique worst view separated at the scale τ; it does not, by itself, imply that the worst-view volume is small in absolute terms.

Arithmetic correction. For the two-text/three-vision specialist subset, the unrounded scores are 0.4605170186 and 1.8420680744. At $\tau = 0.15$, the soft-min is 0.5644740964 and its upper bound is 0.5644890957; both round to 0.564. The original reported 0.565 is retained here as a documented correction, not a new selection or training result. Rounding the per-view scores before aggregation can change the last displayed digit and should be avoided.

K Proposed Aligned-View Evaluation: Future Work

This proposed protocol illustrates the aligned-view setting of section H through trademark description generation (fig. 9). A *description of the mark* describes the visual and textual elements of a

	per view		$d_{\min} \le F_\tau \le d_{\min} + \tau \log M$		
Method/subset	d_{text}	d_{vis}	$d_{\min}$	F_τ	$d_{\min} + \tau \log M$
Exact enum.	1.141	1.141	1.141	**1.141**	1.245
FAIRNEPV	1.120	1.120	1.120	**1.120**	1.224
Text-only	4.605	−2.303	−2.303	−2.199	**−2.199**
Vision-only	−2.303	4.605	−2.303	−2.199	**−2.199**
Avg. kernel	4.605	−2.303	−2.303	−2.199	**−2.199**
Fixed-weight	0.461	1.842	0.461	0.564	**0.564**

Table 3: The fair score against both ends of (39), for the synthetic instance ($k = 5$, $\tau = 0.15$); an exact lower endpoint or an upper endpoint equal after displayed rounding is bold. The exact enumeration benchmark and FAIRNEPV have tied views; the latter is 0.021 below the former after the reported one-swap refinement. A tie drives F_τ to the lower bound because the slack $\tau \log[\sum_q e^{-\Delta_q/\tau}]$ is largest, $\tau \log M = 0.104$, when $\Delta_q \equiv 0$. The remaining selector rows have $\Delta/\tau \ge 9.2$, so their slack is at most 1.5×10^{-5} and they sit on the upper bound to the displayed precision.

logo. Such records in the public register supply two aligned modalities: the text that describes a mark and the logo image itself. A curation method that covers only description wording can still select the same visual designs repeatedly, and a vision-only method can neglect linguistic variation.

We specify the empirical study below but do not report it in this revision: the joined corpus, feature versions, implementation settings, and results must be supplied and verified by the experimenter before this becomes a completed study. The choices below are provisional rather than a record of an executed experiment.

K.1 DATASET

We will select from United States Patent and Trademark Office (USPTO) mark descriptions and their aligned logo images (USPTO, 2026b). The text view is a sparse sublinear-TF·IDF matrix B_{text} with normalized rows. The visual view is a normalized frozen vision-encoder feature matrix B_{vision} extracted from the corresponding logo image. After joining by serial number, the two views share exactly the same candidate index. We use

$$L^{(\text{text})} = B_{\text{text}} B_{\text{text}}^\top + \epsilon I, \quad L^{(\text{vision})} = B_{\text{vision}} B_{\text{vision}}^\top + \epsilon I.$$

Row normalization calibrates the two feature kernels to unit diagonal before the common ridge is added; as noted in remark 6, this convention is fixed before looking at any selected subset. The ridge ensures that both views satisfy Assumption 2. We will restrict descriptions to between 40 and 100 words, stratify the validation and test split by description length, and publish the joined identifiers, feature versions, split seed, and selected indices.

K.2 BASELINES AND METRICS

We will compare FAIRNEPV against text-only and vision-only determinant selection, a determinant of the averaged kernel, and a fixed-weight sum-of-log-determinants/product-DPP ablation (Ohsaka & Matsuoka, 2020; Sogi et al., 2025). Each method uses the same rounding budget. When scalable implementations permit, we will also report a discrete robust-selection baseline under the same oracle budget, while noting that off-the-shelf robust-submodular guarantees require monotonicity assumptions not automatic for these kernels. The primary intrinsic metric is the discrete fair score $F_\tau(S)$ in (D-Fair); we also report each per-view normalized log-determinant and their minimum. For normalized feature vectors $b_i^{(q)}$, we define the per-view pool-to-subset coverage distance

$$C_q(S) := \frac{1}{n} \sum_{i=1}^{n} \min_{j \in S} \bigl(1 - b_i^{(q)\top} b_j^{(q)}\bigr),$$

where lower is better. Together these metrics distinguish balance across views from high aggregate diversity with poor coverage in one view.

One possible downstream evaluation would train the same Qwen3-VL-8B-Instruct model with LoRA (Bai et al., 2025; Hu et al., 2022). We will report mean and standard deviation over matched

training seeds for BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and BERTScore-F1 (Zhang et al., 2020).



Figure 9: **An illustrative input for image-to-description selection.** The mark image and its corresponding description form two aligned observations of one item. This example illustrates the proposed protocol; it is not a selected-subset result.

L AI USE STATEMENT

The authors used generative AI as an assistive tool under their direction for manuscript redrafting, restructuring, and editing; literature discovery and suggesting comparisons with relevant prior work; better clarification of claims and mathematical proof checking; writing and running small numerical checks of mathematical identities and synthetic-score arithmetic; LaTeX formatting; assisting with the execution of experiments; and interpretation of reported results.