SIMPLEX-GATED TENSOR-PRODUCT REGULARIZATION FOR MULTI-VIEW MULTI-LABEL FEATURE SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-view multi-label feature selection requires balancing shared label structure with view-specific predictive information. We propose Simplex-Gated Tensor-Product Regularization (SG-TPR), which preserves the underlying linear feature-to-label function class while organizing shared task structure through tensor-product role–filler geometry. Each view-specific mapping combines a stabilized data-driven residual component with a simplex-gated low-rank structural component, where feature-side roles interact with shared label-side fillers. Signed training-label correlations initialize and calibrate the shared label geometry, while joint row-energy sparsity yields a unified global feature ranking. Across seven benchmarks and eight competitive baselines evaluated under a unified protocol, SG-TPR achieves the best cross-dataset average rank on Average Precision, Coverage, Hamming Loss, and Ranking Loss, with 21 first-place and three second-place results among 28 dataset–metric comparisons. Matched component controls and an independently retuned gate control support the structural branch and adaptive simplex allocation. These results are consistent with the proposed tensor-product-organized structured regularization while preserving the underlying linear function class.

1 INTRODUCTION

Multi-view multi-label data describe each instance through multiple heterogeneous feature sets while associating it with several, often correlated, labels. Such data arise naturally when an object is represented by complementary modalities, descriptors, sources, or feature families. The resulting representation is typically high-dimensional and redundant, whereas useful predictive evidence may be distributed unevenly across both views and labels. Feature selection in this setting must therefore identify a compact feature subset without discarding view-specific information or ignoring dependencies among prediction tasks (Zhang & Zhou, 2013; Qian et al., 2023; Yu et al., 2025).

A common strategy is to learn a feature-to-label prediction matrix and impose row sparsity so that entire feature rows are jointly suppressed across labels (Nie et al., 2010; Zhang et al., 2020).

Earlier multi-view formulations model common and private information, view-specific representations, view-specific labels, and explicit cross-view fusion (Tan et al., 2019; Wu et al., 2019; Zhao et al., 2022; Zhu et al., 2016; Lyu et al., 2022; Hao et al., 2024a). More recent multi-view multi-label feature-selection methods further exploit global-view reconstruction, feature redundancy, uncertainty, graph structure, and semantic dependencies (Hao et al., 2024c; 2025a;b; Han et al., 2024; Lyu et al., 2024; Liu et al., 2026; Chen et al., 2026a;b; Dai et al., 2026). Despite these advances, many existing methods couple parameters across view–label pairs primarily through a shared subspace, graph, reconstruction objective, or common regularizer.

Motivation for adaptive task-space structure. A multi-view multi-label predictor naturally defines a collection of related tasks indexed jointly by views and labels. Sharing information across these tasks can improve statistical efficiency, but the useful amount of shared structure need not be the same for every view. A fully independent treatment ignores reusable label structure, whereas forcing all views into a common representation can suppress view-specific predictive information.

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.

The central problem is therefore not only how to model shared view–label structure, but how to introduce it without constraining the view-specific mappings to the shared structure.

We address this issue through a subspace-structured formulation. For each view, the effective feature-to-label mapping is expressed as two additive components. A data-driven component is filtered by the stabilized residual operator $\mathbf{Q}_{\mathbf{B},\eta}$, while a low-rank structural component is parameterized through the shared label factor $\mathbf{B}$. The structural component therefore uses label-side directions induced by $\mathbf{B}$ and shares this geometry across views, while a learned probability-simplex gate induces adaptive view-specific structural regularization. This design allows shared label structure to regularize different views to different degrees while retaining the full linear mapping class for each view.

The structural branch admits an exact tensor-product role–filler interpretation (Smolensky, 1990). The tensor-product identity provides the representational form; the contribution of SG-TPR lies in coordinating the same shared factor $\mathbf{B}$ across structural prediction, residual filtering, signed label calibration, and row-wise regularization, with simplex gating adapting this bias across views. Unlike a conventional rank-r predictor, the resulting parameterization preserves the full linear mapping class and changes the induced regularization and optimization geometry rather than the hypothesis class.

We further inject signed label geometry through correlation-aware initialization and an explicit calibration objective, both derived only from the current training partition. Unlike nonnegative co-occurrence statistics, the calibration retains both positive and negative label relationships. Together with joint row-wise sparsity over the data-driven and structural components, this produces a single global ranking over features from all views.

Our contributions are threefold:

- **A linear-function-class-preserving residual–structural decomposition.** We introduce a two-branch parameterization that combines a view-specific residual component with a shared label-side structural component while retaining the full linear feature-to-label mapping class for each view.
- **Shared-factor task-space regularization beyond low-rank prediction.** A common label factor jointly shapes structural prediction, residual geometry, signed label calibration, and the joint row-wise sparsity penalty, while simplex weights adapt this structural bias across views.
- **A unified empirical evaluation with matched component controls.** Across seven datasets and eight competitive baselines evaluated under a common protocol, SG-TPR achieves the best average rank on all four evaluation metrics. Matched comparisons isolate the structural branch, residual coupling, SLCC calibration, and adaptive simplex allocation.

2 Related Work

Multi-label and multi-view feature selection. Embedded multi-label feature selection commonly relies on row-sparse regression, where an $\ell_{2,1}$ penalty suppresses complete rows of a feature-to-label prediction matrix and therefore provides a direct global feature ranking (Nie et al., 2010; Zhang et al., 2020). Beyond independent labels, label structure has been incorporated through local correlation models, manifold regularization, dependence maximization, latent representations, and label-specific features (Fan et al., 2021; Huang & Wu, 2021; Li et al., 2022; Zhang & Wu, 2015).

Multi-view learning introduces another source of structure: different views may contain both shared and view-specific predictive information. Representative approaches model commonality and individuality, view-specific information, view-specific labels, or explicit feature fusion (Tan et al., 2019; Wu et al., 2019; Zhao et al., 2022; Zhu et al., 2016; Lyu et al., 2022; 2024). Recent multi-view multi-label feature selectors further exploit global-view reconstruction, redundancy modeling, graph structure, uncertainty, and view-specific label relations (Hao et al., 2024a;b;c; 2025a;b; Han et al., 2024; Liu et al., 2026; Chen et al., 2026a; Dai et al., 2026). These methods provide increasingly rich mechanisms for information sharing across views. Our approach organizes a shared low-rank structural component through tensor-product role–filler structure and view-adaptive simplex regularization, providing a structured task-space bias for multi-view multi-label feature selection.

Structured multi-task learning. Multi-task learning provides several closely related perspectives on parameter sharing. Regularized multi-task formulations encourage related predictors to remain close (Evgeniou & Pontil, 2004); predictive-structure approaches learn shared low-dimensional representations across tasks (Ando et al., 2005); convex multi-task feature learning promotes shared feature structure (Argyriou et al., 2008); and task-relationship or clustered-task models explicitly learn dependencies among task parameters (Zhang & Yeung, 2012; Jacob et al., 2008). Matrix regularization provides a broader view of these structured sharing mechanisms (Kakade et al., 2012), while recent reviews emphasize that the useful amount and form of sharing can itself be task dependent (Upadhyay et al., 2024).

3 METHOD

3.1 PROBLEM SETUP

Let $\mathbf{X}^{(v)} \in \mathbb{R}^{n\times d_v}$, $v = 1,\dots,V$, denote the feature matrix of view v, and let $Y \in \{0,1\}^{n\times c}$ denote the label matrix.

For each view, SG-TPR learns a data-driven mapping

$$\mathbf{W}^{(v)} \in \mathbb{R}^{d_v\times c}$$

and a structural factor

$$\mathbf{U}^{(v)} \in \mathbb{R}^{d_v\times r}.$$

The label side is shared through

$$\mathbf{B} \in \mathbb{R}^{c\times r},$$

where r is the selected label-subspace dimension. We additionally learn $V-1$ free gate logits $\bar{\mathbf{g}} \in \mathbb{R}^{V-1}$.

3.2 LABEL SUBSPACE AND SIMPLEX GATE

Given $\mathbf{B}$, define the stabilized label-subspace operator

$$\mathbf{P}_{B,\eta} = \mathbf{B}\left(\mathbf{B}^\top\mathbf{B} + \eta\mathbf{I}_r\right)^{-1}\mathbf{B}^\top, \qquad \mathbf{Q}_{B,\eta} = \mathbf{I}_c - \mathbf{P}_{B,\eta}, \tag{1}$$

where $\eta = 10^{-8}$ in the implementation. Because $\eta > 0$, $\mathbf{Q}_{B,\eta}$ is a full-rank stabilized residual operator rather than an exact orthogonal-complement projector; hence no label-space direction is completely annihilated. An equivalent inverse form is given in Appendix B.

We obtain positive view weights that sum to one through

$$\boldsymbol{\pi} = \mathrm{softmax}\left(\begin{bmatrix}\bar{\mathbf{g}}\\ 0\end{bmatrix}\right). \qquad \sum_{v=1}^{V}\pi_v = 1, \qquad \pi_v > 0. \tag{2}$$

Fixing the last logit to zero removes the additive translation ambiguity of the softmax parameterization. This identifiability statement concerns only the gate coordinates and does not imply uniqueness of the complete factorization.

3.3 SUBSPACE-STRUCTURED EFFECTIVE MAPPING

For view v, define

$$\mathbf{F}^{(v)} = \mathbf{W}^{(v)}\mathbf{Q}_{B,\eta} + \pi_v\mathbf{U}^{(v)}\mathbf{B}^\top. \tag{3}$$

The first term is the view-specific mapping filtered by the stabilized residual operator $\mathbf{Q}_{\mathbf{B},\eta}$. The second term is a gated low-rank mapping whose label-side directions lie in the span induced by $\mathbf{B}$ and are shared across views.

Writing $\mathbf{u}_j^{(v)}$ for row j of $\mathbf{U}^{(v)}$ and $\mathbf{b}_\ell$ for row ℓ of $\mathbf{B}$,

$$[\mathbf{U}^{(v)}\mathbf{B}^\top]_{j\ell} = \langle\mathbf{u}_j^{(v)}, \mathbf{b}_\ell\rangle = \left\langle\mathbf{I}_r, \mathbf{u}_j^{(v)}\otimes\mathbf{b}_\ell\right\rangle_F. \tag{4}$$

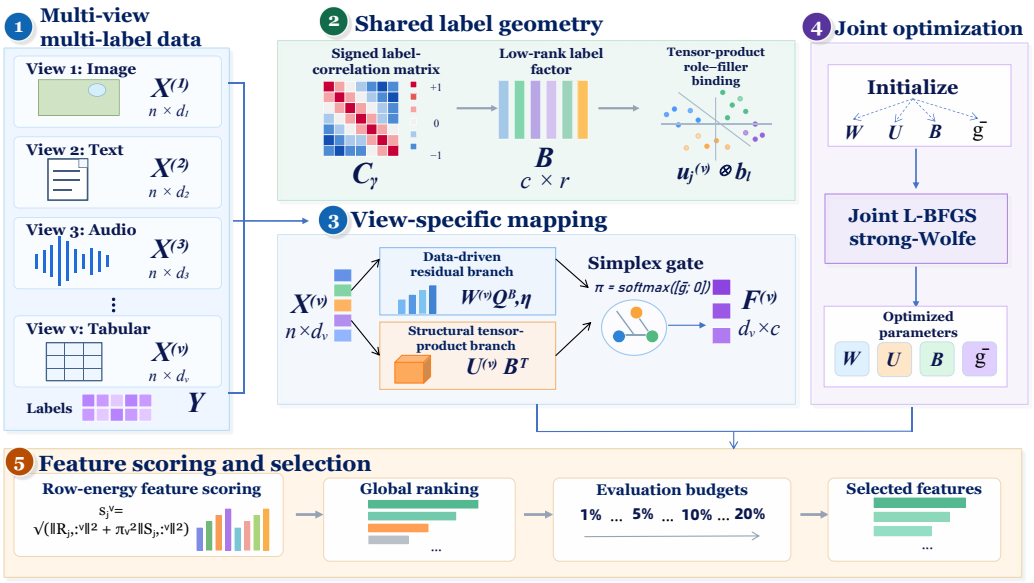

Figure 1: Overview of SG-TPR. Each view-specific mapping is decomposed into a data-driven residual branch and a simplex-gated shared structural branch parameterized by $\mathbf{B}$, whose bilinear form admits a tensor-product role–filler interpretation. Signed label geometry initializes and calibrates the shared factor, while all trainable variables are jointly optimized. The residual and structural branch energies are finally combined to produce a global feature ranking.

Equation (4) provides an exact role–filler tensor-product representation of the bilinear structural factorization. In SG-TPR, the structural effect arises from the coordinated use of the same shared factor $\mathbf{B}$ across structural prediction, residual filtering, signed label-geometry calibration, and row-wise feature regularization; these properties are formalized in Appendix B.

Thus, the structural branch corresponds to a contraction of feature-role and label-filler bindings. The tensor-product organization provides a structured representation of shared label-side geometry over the view–label task space. Combined with simplex gating and row-coupled sparsity, this role–filler component induces view-adaptive task-space regularization while preserving the underlying linear function class.

Preservation of the linear mapping class. Although the structural branch $\mathbf{U}^{(v)}\mathbf{B}^\top$ has rank at most r, the complete residual–structural mapping is not rank constrained. For any target linear mapping $\mathbf{F}_\star^{(v)} \in \mathbb{R}^{d_v \times c}$ and any $\pi_v > 0$, choose

$$\mathbf{W}^{(v)} = \mathbf{F}_\star^{(v)}$$

and

$$\mathbf{U}^{(v)} = \pi_v^{-1}\mathbf{F}_\star^{(v)}\mathbf{B}\left(\mathbf{B}^\top\mathbf{B} + \eta\mathbf{I}_r\right)^{-1}.$$

Then

$$\begin{aligned}\mathbf{W}^{(v)}\mathbf{Q}_{B,\eta} + \pi_v\mathbf{U}^{(v)}\mathbf{B}^\top &= \mathbf{F}_\star^{(v)}\left(\mathbf{Q}_{B,\eta} + \mathbf{P}_{B,\eta}\right)\\ &= \mathbf{F}_\star^{(v)}.\end{aligned} \tag{5}$$

Hence, SG-TPR preserves the original family of linear feature-to-label mappings even though its structural branch is low-rank. The observed gains are therefore not explained by an expansion of the underlying linear function class, and are instead consistent with the structured regularization and optimization geometry induced by the parameterization. Additional structural properties of the shared factor, including its induced row-energy geometry and latent-coordinate invariance, are established in Appendix B.

3.4 SIGNED LABEL-CORRELATION CALIBRATION

Let

$$\widetilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{1}\boldsymbol{\mu}_Y^\top, \qquad \boldsymbol{\mu}_Y = \frac{1}{n}\mathbf{Y}^\top\mathbf{1}. \tag{6}$$

The signed label-correlation matrix is

$$[\mathbf{C}_Y]_{ij} = \frac{\widetilde{\mathbf{Y}}_{:,i}^\top \widetilde{\mathbf{Y}}_{:,j}}{\|\widetilde{\mathbf{Y}}_{:,i}\|_2 \|\widetilde{\mathbf{Y}}_{:,j}\|_2 + \varepsilon_c}. \tag{7}$$

In the implementation, $\varepsilon_c = \mathrm{eps}(\mathrm{float64})$, the double-precision machine epsilon. The implementation symmetrizes $\mathbf{C}_Y$, clips its entries to $[-1, 1]$, and sets its diagonal to one. Only labels from the corresponding training partition are used.

We calibrate the shared label geometry by

$$\mathcal{L}_{\mathrm{SLCC}} = \left\|\mathbf{B}\mathbf{B}^\top - \mathbf{C}_Y\right\|_F^2. \tag{8}$$

3.5 LEARNING OBJECTIVE AND FEATURE SELECTION

We couple the data-driven and structural feature contributions through

$$\mathcal{R}_{\mathrm{FS}} = \sum_{v=1}^{V}\sum_{j=1}^{d_v} \sqrt{\left\|(\mathbf{W}^{(v)}\mathbf{Q}_{B,\eta})_{j,:}\right\|_2^2 + \pi_v^2 \left\|(\mathbf{U}^{(v)}\mathbf{B}^\top)_{j,:}\right\|_2^2 + \varepsilon_s}, \tag{9}$$

where $\varepsilon_s = 10^{-12}$.

$$\begin{aligned}\mathcal{L} = &\left\|\sum_{v=1}^{V}\mathbf{X}^{(v)}\mathbf{F}^{(v)} - \mathbf{Y}\right\|_F^2 \\ &+ \alpha\sum_{v=1}^{V}\left\|\mathbf{X}^{(v)}\mathbf{W}^{(v)} - \mathbf{Y}\right\|_F^2 \\ &+ \lambda\mathcal{R}_{\mathrm{FS}} + \rho\sum_{v=1}^{V}\sqrt{\|\mathbf{U}^{(v)}\|_F^2 + \varepsilon_s} \\ &+ \|\mathbf{B}\|_F^2 + \beta\left\|\mathbf{B}\mathbf{B}^\top - \mathbf{C}_Y\right\|_F^2.\end{aligned} \tag{10}$$

The first term fits the fused prediction from the effective mappings. The second encourages each unrestricted view-specific mapping to remain predictive. The joint row penalty couples the direct and structural contributions at the feature level. The remaining terms regularize the structural factors and calibrate the shared label geometry.

Because π_v scales both the effective structural mapping and its row-wise feature evidence, while $\mathbf{U}^{(v)}$ is regularized separately, the simplex gate should be interpreted as inducing adaptive view-specific structural regularization rather than as a pure mixture coefficient.

After optimization, feature j in view v is scored by

$$s_j^{(v)} = \sqrt{\left\|(\mathbf{W}^{(v)}\mathbf{Q}_{B,\eta})_{j,:}\right\|_2^2 + \pi_v^2\left\|(\mathbf{U}^{(v)}\mathbf{B}^\top)_{j,:}\right\|_2^2}. \tag{11}$$

This component-wise energy mirrors the row-sparsity regularizer and, unlike $\|\mathbf{F}_{j,:}^{(v)}\|_2$, does not permit cancellation between the data-driven and structural branches. The score vectors are concatenated in the registered original view order and sorted in descending order. Prefixes of the resulting global ranking are evaluated at all prescribed feature budgets.

Algorithm 1 Joint training and feature ranking with SG-TPR

Input: $\{\mathbf{X}^{(v)}\}_{v=1}^{V}$, $\mathbf{Y}$, and $\alpha, \lambda, \beta, \rho, r, T_{\max}, \varepsilon_{\text{stop}}$
Compute $\mathbf{C}_Y$ and initialize $\{\mathbf{W}^{(v)}, \mathbf{U}^{(v)}\}_{v=1}^{V}$, $\mathbf{B}$, and $\bar{\mathbf{g}}$.
$\Theta \leftarrow (\{\mathbf{W}^{(v)}, \mathbf{U}^{(v)}\}_{v=1}^{V}, \mathbf{B}, \bar{\mathbf{g}})$.
repeat
    Compute $\mathbf{P}_{B,\eta}$, $\mathbf{Q}_{B,\eta}$, $\boldsymbol{\pi}$, and $\{\mathbf{F}^{(v)}\}_{v=1}^{V}$.
    Evaluate Eq. (10) and its joint gradient with respect to Θ.
    Apply one strong-Wolfe L-BFGS chunk of at most 20 internal iterations.
until relative objective change is below $\varepsilon_{\text{stop}}$ or $T_{\max}$ is reached
Compute feature scores using Eq. (11), concatenate them, and sort in descending order.
return global feature ranking.

Table 1: Statistics of the registered benchmark matrices; feature dimensions are summed across views.

Dataset	Samples	Features	Labels	Views
yeast	2,417	103	14	2
SCENE	4,400	634	33	5
VOC07	3,817	712	20	3
MIRFlickr	4,053	712	38	3
3Sources	169	3,000	6	3
emotions	593	72	6	2
OBJECT	6,047	634	31	5

3.6 Optimization

SG-TPR uses deterministic structure-aware initialization: $\mathbf{W}^{(v)}$ and $\mathbf{U}^{(v)}$ are initialized by ridge regression, $\mathbf{B}$ from the leading eigenspace of $\mathbf{C}_Y$, and the free gate logits at zero. All trainable variables are then jointly optimized in double precision by strong-Wolfe L-BFGS, as summarized in Algorithm 1. Detailed initialization formulas are provided in Appendix A, while the optimization budgets and stopping rules are documented in Appendix F.

Complexity. A direct objective/gradient evaluation costs $\mathcal{O}(nDc + Dc(c+r) + c^2r + cr^2 + r^3)$; the cached implementation and full time–memory analysis are given in Appendix C.

4 Experiments

Datasets. We evaluate on seven multi-view multi-label benchmarks: yeast (Elisseeff & Weston, 2001), SCENE, VOC07 (Everingham et al., 2010), MIRFlickr (Huiskes & Lew, 2008), 3Sources (Greene & Cunningham, 2009), emotions (Trohidis et al., 2011), and OBJECT. Table 1 summarizes their dimensions. MIRFlickr denotes the registered 4K benchmark subset. SCENE and OBJECT are processed benchmark matrices derived from NUS-WIDE (Chua et al., 2009); detailed provenance is provided in Appendix F.

Baselines and evaluation protocol. We compare SG-TPR with Variance, RMAN-MMFS (Liu et al., 2026), GDRFS (Chen et al., 2026a), NR-CILL (Dai et al., 2026), MSFS (Zhang et al., 2020), SPLDG (Zhang et al., 2025), BHDG (Guo et al., 2026), and UGRFS (Hao et al., 2025b). All methods use the same five frozen outer folds and downstream evaluator. Hyperparameters are selected using outer-training data only: six datasets use one fixed 25% inner validation split, whereas 3Sources uses three frozen multilabel-stratified inner folds. The selected configuration is refitted on the complete outer-training partition before one evaluation on the held-out outer fold.

Preprocessing and model selection. Selectors use training-only MinMax scaling and one global ranking truncated at the registered 1%–20% feature budgets; selected features are evaluated with a common one-vs-rest logistic-regression pipeline using training-only standardization. For SG-TPR,

Table 2: Average Precision (AP, ↑) and Coverage (↓) on the seven benchmark datasets. Each value is averaged over the registered 1%–20% feature budgets within each of five frozen outer folds. Best and second-best results are determined using the unrounded values and are shown in bold and underlined, respectively. The proposed SG-TPR method is highlighted in gray. Lower average rank is better.

Dataset	Variance	RMAN-MMFS	GDRFS	NR-CILL	MSFS	SPLDG	BHDG	UGRFS	**SG-TPR**
Average Precision (AP, ↑)									
yeast	0.7261	0.7336	0.7381	0.7387	0.7367	0.7314	0.7231	0.7386	**0.7408**
SCENE	0.8364	0.8520	0.8595	0.8589	0.8553	0.8502	0.8483	0.8599	**0.8641**
VOC07	0.6736	0.6692	0.6830	0.6825	0.6862	0.6562	0.6773	0.6852	**0.6910**
MIRFlickr	0.7632	0.7476	0.7650	0.7601	0.7631	0.7216	0.7559	0.7656	**0.7668**
3Sources	0.4564	0.4336	0.4550	0.4138	0.4556	0.4353	**0.4616**	0.4363	0.4570
emotions	0.7073	0.6935	0.7552	0.7533	0.7376	0.7656	0.7304	0.7576	**0.7759**
OBJECT	0.6000	0.6420	0.6648	0.6743	0.6699	0.6396	0.6329	0.6687	**0.6763**
Avg. rank	6.8571	7.2857	4.0000	4.7143	4.2857	6.8571	6.5714	3.2857	**1.1429**
Coverage (↓)									
yeast	6.7502	6.6387	6.6018	**6.5990**	6.6001	6.6833	6.7689	6.6091	6.6067
SCENE	11.7355	11.0156	10.7291	10.7711	10.9511	11.1355	11.2735	10.7267	**10.5794**
VOC07	5.8508	5.8699	5.5981	5.6743	5.6333	6.2680	5.7596	5.5612	**5.4048**
MIRFlickr	18.3046	18.5233	18.1253	18.3691	18.1397	19.2601	18.6106	**18.0748**	18.1432
3Sources	2.5683	2.5641	2.5507	2.7425	**2.5427**	2.6240	2.5764	2.5853	2.5499
emotions	2.2075	2.3264	1.9938	1.9980	2.0848	1.9162	2.0866	1.9910	**1.9046**
OBJECT	6.5334	5.9341	5.5308	5.4286	5.4965	5.9294	6.0435	5.5102	**5.3633**
Avg. rank	7.2857	6.7143	3.2857	4.5714	3.4286	6.8571	7.4286	3.4286	**2.0000**

$\alpha = 1$ is fixed; (λ, ρ, r) is selected with $\beta = 0$, followed by selection of β with the base tuple fixed. Detailed budget construction, classifier settings, search spaces, stopping rules, and provenance are provided in Appendix F and the released configuration files.

4.1 Main Results

Tables 2 and 3 report the four evaluation metrics. Across the 28 dataset–metric comparisons, SG-TPR obtains 21 first-place and three second-place results and achieves the best cross-dataset average rank for every metric (AP 1.14, Coverage 2.00, HL 1.86, RL 1.14). Friedman tests reject the global equal-performance hypothesis for all four metrics (all $p < 4 \times 10^{-4}$); pairwise Wilcoxon–Holm tests are reported descriptively because seven dataset blocks limit corrected resolution (Appendix E).

Budget-wise behavior. Figure 2 shows that the aggregate results arise from different operating regimes. SG-TPR tracks the strongest competing curves across most budgets on VOC07 and increasingly leads on OBJECT as the retained subset grows, whereas MIRFlickr exhibits a crossover toward moderate-to-larger subsets. The remaining four datasets are reported in Appendix G.1, completing the seven-dataset budget-wise comparison.

Compression relative to all features. At the 20% feature budget, SG-TPR exceeds the all-feature AP on five of the seven datasets, is nearly unchanged on emotions, and differs only slightly on yeast. Complete dataset-level comparisons with the all-feature reference are reported in Appendix G.3 (Table 9). On these benchmarks, substantial feature reduction can preserve or improve held-out predictive performance.

4.2 Matched Ablations and Structural Control

Matched comparisons. Figure 3 summarizes the three component ablations and the Additive-LR structural control, using $\Delta\mathrm{AP} = \mathrm{AP}_{\mathrm{Full}} - \mathrm{AP}_{\mathrm{Control}}$. No-TPR removes the complete structural

Table 3: Hamming Loss (HL, ↓) and Ranking Loss (RL, ↓) on the seven benchmark datasets. Each value is averaged over the registered 1%–20% feature budgets within each of five frozen outer folds. Best and second-best results are determined using the unrounded values and are shown in bold and underlined, respectively. The proposed SG-TPR method is highlighted in gray. Lower average rank is better.

Dataset	Variance	RMAN-MMFS	GDRFS	NR-CILL	MSFS	SPLDG	BHDG	UGRFS	**SG-TPR**
Hamming Loss (HL, ↓)									
yeast	0.2199	0.2196	0.2168	0.2162	0.2177	0.2174	0.2219	0.2152	**0.2138**
SCENE	0.0977	0.0915	0.0876	0.0883	0.0894	0.0919	0.0924	0.0874	**0.0852**
VOC07	0.0821	0.0835	0.0818	0.0817	0.0812	0.0838	0.0820	0.0815	**0.0810**
MIRFlickr	0.1586	0.1648	0.1570	0.1599	0.1580	0.1768	0.1620	**0.1568**	0.1570
3Sources	0.2805	0.2814	0.2865	0.2973	0.2795	0.2978	**0.2729**	0.2914	0.2857
emotions	0.2633	0.2750	0.2326	0.2296	0.2440	0.2287	0.2483	0.2282	**0.2213**
OBJECT	0.0558	0.0531	0.0506	0.0506	0.0508	0.0537	0.0541	0.0505	**0.0503**
Avg. rank	7.0000	6.8571	4.0000	4.7143	4.2857	7.0000	6.5714	2.7143	**1.8571**
Ranking Loss (RL, ↓)									
yeast	0.1961	0.1901	0.1862	0.1861	0.1871	0.1921	0.1980	0.1861	**0.1849**
SCENE	0.0711	0.0609	0.0568	0.0573	0.0594	0.0622	0.0638	0.0567	**0.0547**
VOC07	0.1441	0.1453	0.1368	0.1386	0.1368	0.1572	0.1415	0.1349	**0.1306**
MIRFlickr	0.1122	0.1200	0.1109	0.1140	0.1114	0.1339	0.1162	**0.1104**	0.1104
3Sources	0.4783	0.4762	0.4741	0.5106	0.4716	0.4870	0.4772	0.4796	**0.4670**
emotions	0.2542	0.2784	0.2059	0.2099	0.2264	0.1950	0.2301	0.2039	**0.1889**
OBJECT	0.1252	0.1101	0.1000	0.0974	0.0987	0.1101	0.1130	0.0992	**0.0952**
Avg. rank	7.4286	6.7143	3.7143	4.8571	4.0000	7.0000	7.1429	3.0000	**1.1429**

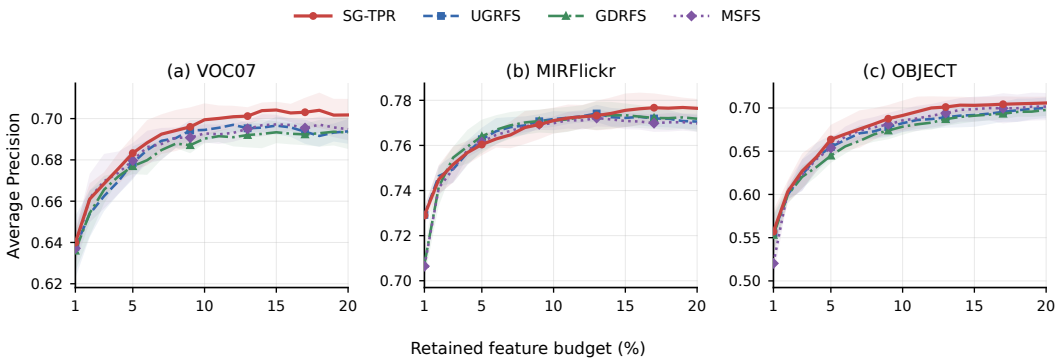

Figure 2: Budget-wise Average Precision on three representative datasets. Curves show the mean over the five frozen outer folds, with shaded regions denoting sample standard deviation. VOC07 shows broadly consistent performance across budgets, MIRFlickr exhibits a crossover at moderate-to-larger subsets, and OBJECT increasingly favors SG-TPR as more ranked features are retained. The four displayed methods use one global ranking per outer fold and are not retrained separately for individual budgets. UGRFS, GDRFS, and MSFS are shown as representative competitive structured baselines to preserve visual readability; complete numerical comparisons with all baselines are reported in Tables 2 and 3.

branch. Additive-LR replaces the B-coupled residual $\mathbf{W}^{(v)}\mathbf{Q}_{B,\eta}$ with the unrestricted mapping $\mathbf{W}^{(v)}$, while retaining the same gated low-rank structural branch, simplex gate, SLCC calibration, factor penalties, initialization, and parameter count. Full ($\beta = 0$) removes only the explicit SLCC loss while retaining correlation-aware initialization, and Uniform-Gate fixes $\pi_v = 1/V$ throughout

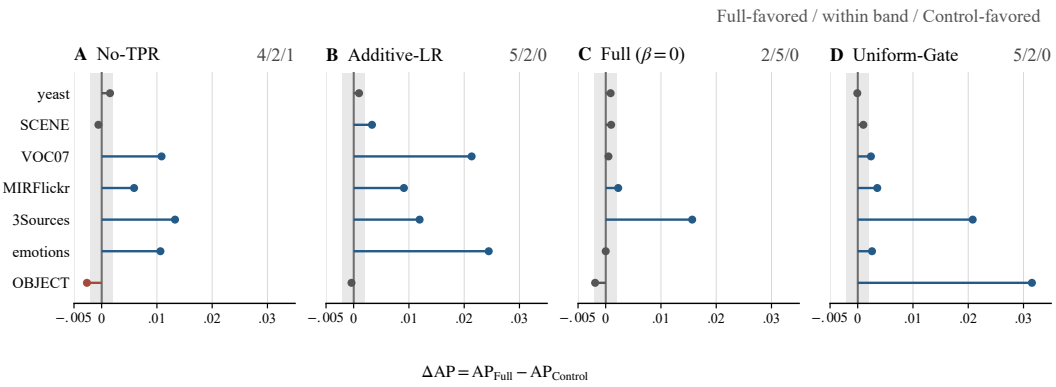

Figure 3: Matched comparisons with $\Delta\mathrm{AP} = \mathrm{AP}_{\mathrm{Full}} - \mathrm{AP}_{\mathrm{Control}}$. The gray band denotes the descriptive interval $[-0.002, +0.002]$. Full-favored/within-band/control-favored splits are $4/2/1$ for No-TPR, $5/2/0$ for the matched Additive-LR structural control, $2/5/0$ for Full ($\beta = 0$), and $5/2/0$ for Uniform-Gate. Absolute AP values are reported in Appendix D.

optimization. To isolate each intervention, all four matched comparisons inherit the Full-selected fold-specific configuration and differ only in the stated component. An independently retuned Uniform-Gate control is additionally reported under the same frozen inner-validation protocol.

Component effects. Relative to No-TPR, Full is favored on four datasets, remains within the reference band on two, and is control-favored on one $(4/2/1)$. Relative to Additive-LR, Full is favored on SCENE, VOC07, MIRFlickr, 3Sources, and emotions and remains within the band on yeast and OBJECT $(5/2/0)$. Since Additive-LR retains the same unrestricted mapping, gated low-rank branch, simplex gate, SLCC calibration, factor penalties, and parameter count, this matched comparison specifically tests the effect of coupling $\mathbf{B}$ to the residual geometry through $\mathbf{Q}_{B,\eta}$. Full versus Full ($\beta = 0$) yields $2/5/0$, while Full versus Uniform-Gate yields $5/2/0$. Together, these controls provide evidence for the contribution of the structural branch, residual-geometry coupling, and adaptive simplex allocation, with SLCC providing additional gains on selected datasets.

Independently retuned Uniform-Gate control. We further allow Uniform-Gate to perform its own training-side hyperparameter selection under the same frozen inner-validation protocol. After independent retuning, Full SG-TPR retains a 0.0049 average-AP advantage. Under the same reference band, the resulting Full-favored/within-band/control-favored split is $4/2/1$. Retuning narrows the matched-control gap by 0.0039 AP. Complete dataset-level results are reported in Appendix D.

5 Conclusion

We introduced SG-TPR, a residual–structural framework that preserves the underlying linear feature-to-label function class while organizing shared task structure through tensor-product role–filler geometry. The shared structural component couples feature-side roles with a common label-side factor, and a learned simplex gate adapts the strength of structural sharing across views.

Across seven benchmark datasets, SG-TPR achieves the best cross-dataset average rank on all four evaluation metrics. Matched comparisons support the shared structural branch, residual-geometry coupling, and adaptive simplex allocation, while the explicit SLCC loss provides additional gains on selected datasets. Because SG-TPR preserves the original linear mapping class, the observed gains cannot be attributed to increased linear expressivity and are instead consistent with the induced regularization and optimization geometry.

Future work will extend this adaptive structural-sharing formulation to incomplete-view and incomplete-label settings, broader downstream predictors, and heterogeneous multimodal representations.

Reproducibility Statement

We release code, frozen splits, fold-specific selections, matched controls, provenance manifests, and the available per-budget evaluation and audit artifacts. All selection and preprocessing are training-only. Full protocol and experimental provenance are documented in the appendix and release package. The released audit artifacts verify the matched-control results across all 35 dataset–outer-fold units against the corresponding Full-selected configurations, including hyperparameters, splits, optimization settings, rankings, and stored model identities.

AI Use Statement

Generative AI tools were used for language polishing, literature discovery, and limited assistance with code review, debugging, and presentation. All scientific decisions, mathematical derivations, experimental execution, result verification, and final manuscript content were independently reviewed and validated by the authors, who take full responsibility for the work.

References

Rie Kubota Ando, Tong Zhang, and Peter Bartlett. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of machine learning research*, 6(11), 2005.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine learning*, 73(3):243–272, 2008.

Jiaming Chen, Ping Zhang, Yonghao Li, and Wanfu Gao. Graph-fusion guided data reconstruction for multi-view multi-label feature selection. *Pattern Recognition*, pp. 113692, 2026a.

Zhiqi Chen, Yuzhou Liu, Jiarui Liu, and Wanfu Gao. Combining LLM semantic reasoning with GNN structural modeling for multi-view multi-label feature selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20454–20462, 2026b.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. NUS-WIDE: A real-world web image database from National University of Singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 1–9, 2009.

Jianhua Dai, Qiuyu Wu, and Witold Pedrycz. Noise-resistant commonality and individuality label learning for multiview multilabel feature selection using fuzzy mutual information. *IEEE Transactions on Cybernetics*, 56(3):1479–1491, 2026. doi: 10.1109/TCYB.2025.3627933.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

André Elisseeff and Jason Weston. A kernel method for multi-labelled classification. *Advances in neural information processing systems*, 14, 2001.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 109–117, 2004.

Yuling Fan, Jinghua Liu, Wei Weng, Baihua Chen, Yannan Chen, and Shunxiang Wu. Multi-label feature selection with local discriminant model and label correlations. *Neurocomputing*, 442: 98–115, 2021.

Derek Greene and Pádraig Cunningham. A matrix factorization approach for integrating multiple data views. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 423–438. Springer, 2009.

Cong Guo, Changqin Huang, Wenhua Zhou, and Xiaodi Huang. Multi-label feature selection based on binary hashing learning and dynamic graph constraints. *Pattern Recognition*, pp. 113603, 2026.

Qingqi Han, Liang Hu, and Wanfu Gao. Feature relevance and redundancy coefficients for multi-view multi-label feature selection. *Information Sciences*, 652:119747, 2024.

Pingting Hao, Weiping Ding, Wanfu Gao, and Jialong He. Exploring view-specific label relationships for multi-view multi-label feature selection. *Information Sciences*, 681:121215, 2024a.

Pingting Hao, Kunpeng Liu, and Wanfu Gao. Double-layer hybrid-label identification feature selection for multi-view multi-label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12295–12303, 2024b.

Pingting Hao, Kunpeng Liu, and Wanfu Gao. Anchor-guided global view reconstruction for multi-view multi-label feature selection. *Information Sciences*, 679:121124, 2024c.

Pingting Hao, Wanfu Gao, and Liang Hu. Embedded feature fusion for multi-view multi-label feature selection. *Pattern Recognition*, 157:110888, 2025a.

Pingting Hao, Kunpeng Liu, and Wanfu Gao. Uncertainty-aware global-view reconstruction for multi-view multi-label feature selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 17068–17076, 2025b.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979.

Rui Huang and Zhejun Wu. Multi-label feature selection via manifold regularization and dependence maximization. *Pattern Recognition*, 120:108149, 2021.

Mark J Huiskes and Michael S Lew. The MIR Flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pp. 39–43, 2008.

Laurent Jacob, Jean-philippe Vert, and Francis Bach. Clustered multi-task learning: A convex formulation. *Advances in neural information processing systems*, 21, 2008.

Sham M Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Regularization techniques for learning with matrices. *The Journal of Machine Learning Research*, 13:1865–1890, 2012.

Junlong Li, Peipei Li, Xuegang Hu, and Kui Yu. Learning common and label-specific features for multi-label classification with correlation information. *Pattern recognition*, 121:108259, 2022.

Yuzhou Liu, Jiarui Liu, and Wanfu Gao. Redundancy-optimized multi-head attention networks for multi-view multi-label feature selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23945–23953, 2026.

Gengyu Lyu, Xiang Deng, Yanan Wu, and Songhe Feng. Beyond shared subspace: A view-specific fusion for multi-view multi-label learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 7647–7654, 2022.

Gengyu Lyu, Weiqi Kang, Haobo Wang, Zheng Li, Zhen Yang, and Songhe Feng. Common-individual semantic fusion for multi-view multi-label learning. In *IJCAI*, pp. 4715–4723, 2024.

Feiping Nie, Heng Huang, Xiao Cai, and Chris Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. *Advances in neural information processing systems*, 23, 2010.

Wenbin Qian, Jintao Huang, Fankang Xu, Wenhao Shu, and Weiping Ding. A survey on multi-label feature selection from perspectives of label fusion. *Information Fusion*, 100:101948, 2023.

Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, 46(1-2):159–216, 1990.

Qiaoyu Tan, Guoxian Yu, Jun Wang, Carlotta Domeniconi, and Xiangliang Zhang. Individuality- and commonality-based multiview multilabel learning. *IEEE transactions on cybernetics*, 51(3): 1716–1727, 2019.

Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, and Ioannis Vlahavas. Multi-label classification of music by emotion. *EURASIP Journal on Audio, Speech, and Music Processing*, 2011(1):4, 2011.

Richa Upadhyay, Ronald Phlypo, Rajkumar Saini, and Marcus Liwicki. Sharing to learn and learning to share; fitting together meta, multi-task, and transfer learning: A meta review. *IEEE Access*, 12:148553–148576, 2024.

Xuan Wu, Qing-Guo Chen, Yao Hu, Dengbao Wang, Xiaodong Chang, Xiaobo Wang, and Min-Ling Zhang. Multi-view multi-label learning with view-specific information extraction. In *IJCAI*, pp. 3884–3890, 2019.

Zhiwen Yu, Ziyang Dong, Chenchen Yu, Kaixiang Yang, Ziwei Fan, and CL Philip Chen. A review on multi-view learning. *Frontiers of Computer Science*, 19(7):197334, 2025.

Min-Ling Zhang and Lei Wu. LIFT: Multi-label learning with label-specific features. *IEEE transactions on pattern analysis and machine intelligence*, 37(1):107–120, 2015.

Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2013.

Yao Zhang, Jun Tang, Ziqiang Cao, and Han Chen. Sparse multi-label feature selection via pseudo-label learning and dynamic graph constraints. *Information Fusion*, 118:102975, 2025.

Yongshan Zhang, Jia Wu, Zhihua Cai, and Philip S Yu. Multi-view multi-label learning with sparse feature selection for image annotation. *IEEE Transactions on Multimedia*, 22(11):2844–2857, 2020.

Yu Zhang and Dit-Yan Yeung. A convex formulation for learning task relationships in multi-task learning. *arXiv preprint arXiv:1203.3536*, 2012.

Dawei Zhao, Qingwei Gao, Yixiang Lu, and Dong Sun. Learning view-specific labels and label-feature dependence maximization for multi-view multi-label classification. *Applied Soft Computing*, 124:109071, 2022.

Xiaofeng Zhu, Xuelong Li, and Shichao Zhang. Block-row sparse multiview multilabel learning for image classification. *IEEE transactions on cybernetics*, 46(2):450–461, 2016. doi: 10.1109/TCYB.2015.2403356.

A Initialization Details

Initialization details. All ridge initializations use $\tau_{\text{init}} = 10^{-3}$. Let $\mathbf{C}_Y = \mathbf{E}\,\mathrm{diag}(\boldsymbol{\xi})\mathbf{E}^\top$. Using the leading r eigenpairs,

$$\mathbf{B}_0 = \mathbf{E}_r\,\mathrm{diag}\left(\sqrt{\max(\xi_1, 10^{-6})}, \ldots, \sqrt{\max(\xi_r, 10^{-6})}\right).$$

The latent initialization target is

$$\mathbf{H}_0 = \mathbf{Y}\mathbf{B}_0(\mathbf{B}_0^\top\mathbf{B}_0 + \eta\mathbf{I}_r)^{-1}.$$

Each $\mathbf{W}^{(v)}$ is initialized by ridge regression toward $\mathbf{Y}/V$, and each $\mathbf{U}^{(v)}$ by ridge regression from $\mathbf{X}^{(v)}$ to $\mathbf{H}_0$. The $\mathbf{Y}/V$ target is used only during initialization. Subsequent joint optimization follows Eq. (10), whose auxiliary view-specific prediction terms use the full target $\mathbf{Y}$. The free gate logits are initialized as $\bar{\mathbf{g}}_0 = \mathbf{0}$.

B Structural Properties

This appendix establishes the two structural properties most relevant to SG-TPR: the full linear mapping class is preserved despite the low-rank structural branch, and the shared factor $\mathbf{B}$ affects the regularization geometry beyond the additive term $\mathbf{U}^{(v)}\mathbf{B}^\top$. Additional algebraic and identifiability properties are summarized at the end of the section.

B.1 Preservation of the full linear mapping class

Proposition 1 (Low-rank structural branch with full linear mapping class). For any fixed $\mathbf{B} \in \mathbb{R}^{c\times r}$, any target mapping $\mathbf{F}_\star^{(v)} \in \mathbb{R}^{d_v\times c}$, $\eta > 0$, and $\pi_v > 0$, there exist $\mathbf{W}^{(v)}$ and $\mathbf{U}^{(v)}$ such that

$$\mathbf{F}^{(v)} = \mathbf{W}^{(v)}\mathbf{Q}_{B,\eta} + \pi_v\mathbf{U}^{(v)}\mathbf{B}^\top = \mathbf{F}_\star^{(v)}.$$

Consequently,

$$\mathcal{F}_{\text{SG-TPR}}^{(v)} = \mathbb{R}^{d_v\times c}.$$

Proof. Choose

$$\mathbf{W}^{(v)} = \mathbf{F}_\star^{(v)}, \qquad \mathbf{U}^{(v)} = \pi_v^{-1}\mathbf{F}_\star^{(v)}\mathbf{B}\left(\mathbf{B}^\top\mathbf{B} + \eta\mathbf{I}_r\right)^{-1}.$$

Since $\eta > 0$ and $\pi_v > 0$, the construction is well defined. Using $\mathbf{Q}_{B,\eta} + \mathbf{P}_{B,\eta} = \mathbf{I}_c$,

$$\begin{aligned}\mathbf{F}^{(v)} &= \mathbf{F}_\star^{(v)}\mathbf{Q}_{B,\eta} + \mathbf{F}_\star^{(v)}\mathbf{P}_{B,\eta}\\ &= \mathbf{F}_\star^{(v)}.\end{aligned}$$

For comparison, a conventional rank-r factorized predictor has

$$\mathcal{F}_{\text{LR}} = \left\{\mathbf{F} \in \mathbb{R}^{d_v\times c} : \mathrm{rank}(\mathbf{F}) \le r\right\}.$$

Hence, when $r < \min(d_v, c)$,

$$\mathcal{F}_{\text{LR}} \subsetneq \mathcal{F}_{\text{SG-TPR}}^{(v)} = \mathbb{R}^{d_v\times c}.$$

Thus, the structural branch is low-rank, but the complete SG-TPR parameterization does not restrict the underlying linear hypothesis class.

B.2 Shared-factor regularization geometry

The shared factor $\mathbf{B}$ affects both structural feature energy and the data-driven residual geometry. For feature j in view v,

$$\left\|(\mathbf{U}^{(v)}\mathbf{B}^\top)_{j,:}\right\|_2^2 = (\mathbf{u}_j^{(v)})^\top\mathbf{B}^\top\mathbf{B}\mathbf{u}_j^{(v)}.$$

Defining

$$\mathbf{G}_B = \mathbf{B}^\top\mathbf{B} \succeq 0,$$

the feature score in Eq. (11) becomes

$$\left(s_j^{(v)}\right)^2 = \left\|(\mathbf{W}^{(v)}\mathbf{Q}_{B,\eta})_{j,:}\right\|_2^2 + \pi_v^2(\mathbf{u}_j^{(v)})^\top \mathbf{G}_B \mathbf{u}_j^{(v)}.$$

Thus, $\mathbf{B}^\top\mathbf{B}$ induces a shared quadratic form on the latent structural coordinates.

The two Gram matrices associated with $\mathbf{B}$ play distinct roles: $\mathbf{B}\mathbf{B}^\top \in \mathbb{R}^{c\times c}$ is calibrated toward the signed label geometry $\mathbf{C}_Y$ through

$$\left\|\mathbf{B}\mathbf{B}^\top - \mathbf{C}_Y\right\|_F^2,$$

whereas $\mathbf{B}^\top\mathbf{B} \in \mathbb{R}^{r\times r}$ governs structural row energy.

The dependence on $\mathbf{B}$ is not confined to the low-rank structural branch. Because $\eta > 0$, the stabilized residual operator satisfies

$$\mathbf{Q}_{B,\eta} = \left(\mathbf{I}_c + \eta^{-1}\mathbf{B}\mathbf{B}^\top\right)^{-1}, \qquad \mathbf{Q}_{B,\eta}^{-1} = \mathbf{I}_c + \eta^{-1}\mathbf{B}\mathbf{B}^\top.$$

Let

$$\widetilde{\mathbf{W}}^{(v)} = \mathbf{W}^{(v)}\mathbf{Q}_{B,\eta}.$$

Then

$$\mathbf{F}^{(v)} = \widetilde{\mathbf{W}}^{(v)} + \pi_v \mathbf{U}^{(v)}\mathbf{B}^\top,$$

but the auxiliary view-specific prediction term becomes

$$\left\|\mathbf{X}^{(v)}\widetilde{\mathbf{W}}^{(v)}\left(\mathbf{I}_c + \eta^{-1}\mathbf{B}\mathbf{B}^\top\right) - \mathbf{Y}\right\|_F^2.$$

Therefore, even under the additive unconstrained-plus-low-rank reparameterization, $\mathbf{B}$ remains coupled to the data-driven branch. Its role is consequently broader than that of a conventional label-side factor in $\mathbf{U}^{(v)}\mathbf{B}^\top$.

This distinction is directly examined by the matched Additive-LR structural control, which removes the $\mathbf{B}$-dependent residual coupling while retaining the same low-rank branch, gate, label calibration, penalties, and parameter count.

B.3 Additional structural properties

Role–filler form. For each view, $\mathbf{S}^{(v)} = \mathbf{U}^{(v)}\mathbf{B}^\top$ has rank at most r, and all of its rows lie in the common label-side subspace induced by $\mathbf{B}$. For feature j and label ℓ,

$$S_{j\ell}^{(v)} = (\mathbf{u}_j^{(v)})^\top \mathbf{b}_\ell = \left\langle \mathbf{I}_r, \mathbf{u}_j^{(v)} \otimes \mathbf{b}_\ell \right\rangle_F,$$

which gives the exact tensor-product role–filler form used in Eq. (4).

Reference-logit gate identifiability. For

$$\boldsymbol{\pi} = \mathrm{softmax}([\bar{\mathbf{g}}; 0]),$$

any $\boldsymbol{\pi} \in \operatorname{int}\Delta^{V-1}$ uniquely determines the free logits through

$$\bar{g}_v = \log\frac{\pi_v}{\pi_V}, \qquad v = 1, \ldots, V-1.$$

Thus, fixing the final logit to zero removes the additive translation ambiguity of the softmax coordinates.

Orthogonal latent-coordinate invariance. Let $\mathbf{R} \in \mathbb{R}^{r\times r}$ be orthogonal and define

$$\mathbf{B}' = \mathbf{B}\mathbf{R}, \qquad \mathbf{U}'^{(v)} = \mathbf{U}^{(v)}\mathbf{R}.$$

Then

$$\mathbf{U}'^{(v)}\mathbf{B}'^\top = \mathbf{U}^{(v)}\mathbf{B}^\top, \qquad \mathbf{B}'\mathbf{B}'^\top = \mathbf{B}\mathbf{B}^\top,$$

and

$$\mathbf{P}_{B',\eta} = \mathbf{P}_{B,\eta}, \qquad \mathbf{Q}_{B',\eta} = \mathbf{Q}_{B,\eta}.$$

Moreover, because

$$\mathbf{B}'^\top\mathbf{B}' = \mathbf{R}^\top\mathbf{B}^\top\mathbf{B}\mathbf{R}, \qquad \mathbf{u}_j'^{(v)} = \mathbf{R}^\top\mathbf{u}_j^{(v)},$$

the structural row energy is unchanged. Hence, the effective mappings, residual operators, SLCC term, row-energy regularizer, feature scores, and Frobenius factor penalties are invariant to a common orthogonal rotation of the latent coordinates.

C Computational and Memory Complexity

Let

$$D = \sum_{v=1}^{V} d_v, \qquad P = Dc + Dr + cr + (V-1).$$

The quantity P is the number of trainable scalar parameters.

Objective evaluation. Constructing the stabilized label-subspace operator requires

$$\mathcal{O}(c^2 r + cr^2 + r^3)$$

operations. Across all views, computing $\mathbf{W}^{(v)}\mathbf{Q}_{B,\eta}$ and $\mathbf{U}^{(v)}\mathbf{B}^\top$ requires

$$\mathcal{O}(Dc^2 + Dcr).$$

Without cached sufficient statistics, the fused and per-view prediction terms require $\mathcal{O}(nDc)$ time. Hence one objective/gradient evaluation has complexity

$$C_{\mathrm{dir}} = \mathcal{O}\big(nDc + Dc^2 + Dcr + c^2 r + cr^2 + r^3\big). \tag{12}$$

Cached evaluation. When $D \le n$, the implementation precomputes the design Gram matrix and label cross-product at one-time cost

$$\mathcal{O}(nD^2 + nDc)$$

and memory cost $\mathcal{O}(D^2 + Dc)$. Thereafter, the prediction term is evaluated through quadratic forms, giving

$$C_{\mathrm{cache}} = \mathcal{O}\big(D^2 c + Dc^2 + Dcr + c^2 r + cr^2 + r^3\big). \tag{13}$$

Because this mode is activated only when $D \le n$, $D^2 c \le nDc$.

Optimization. Let E denote the total number of objective/gradient closure evaluations generated by strong-Wolfe L-BFGS. The optimization stage therefore costs

$$\mathcal{O}(EC_{\mathrm{dir}})$$

in direct mode, or

$$\mathcal{O}(nD^2 + nDc + EC_{\mathrm{cache}})$$

in cached mode, in addition to deterministic initialization.

Memory. Let h be the L-BFGS history size ($h = 50$ in all formal runs). The direct implementation requires

$$\mathcal{O}(nD + nc + c^2 + hP) \tag{14}$$

memory up to constant factors. Cached mode additionally stores the $D \times D$ design Gram matrix, giving

$$\mathcal{O}(nD + nc + c^2 + D^2 + hP). \tag{15}$$

Initialization. Let $m_v = \min(n, d_v)$. The primal/dual ridge implementation used for a q-dimensional target has cost

$$R_v(q) = \mathcal{O}(nd_v m_v + m_v^3 + nd_v q + m_v^2 q).$$

Including signed-correlation construction, eigendecomposition, and the two ridge initializations gives

$$C_{\mathrm{init}} = \mathcal{O}\left(nc^2 + c^3 + ncr + \sum_{v=1}^{V}[R_v(c) + R_v(r)]\right)$$

up to lower-order $r \times r$ operations.

D COMPLETE ABLATION AND GATE-RETUNING RESULTS

Table 4: Average Precision (AP) for Full SG-TPR, three matched component ablations, the matched Additive-LR structural control, and an independently retuned Uniform-Gate control. The matched comparisons inherit the Full-selected fold-specific configuration except for their stated interventions, whereas Retuned Uniform-Gate performs its own training-side model selection under the same frozen inner-validation protocol. Best and second-best results are determined using the unrounded values and are shown in bold and underlined, respectively. The final column reports the unweighted mean AP across the seven datasets.

Variant	yeast	SCENE	VOC07	MIRFlickr	3Sources	emotions	OBJECT	Mean
Full SG-TPR	<u>0.7408</u>	<u>0.8641</u>	**0.6910**	<u>0.7668</u>	**0.4570**	**0.7759**	0.6763	**0.7103**
No-TPR	0.7393	**0.8647**	0.6802	0.7609	0.4438	0.7652	**0.6790**	0.7047
Additive-LR	0.7398	0.8608	0.6697	0.7577	0.4451	0.7515	0.6767	0.7002
Full ($\beta = 0$)	0.7399	0.8631	0.6905	0.7645	0.4414	<u>0.7759</u>	<u>0.6782</u>	<u>0.7077</u>
Uniform-Gate	**0.7409**	0.8631	0.6886	0.7633	0.4362	0.7733	0.6448	0.7015
Retuned Uniform-Gate	0.7405	0.8529	<u>0.6908</u>	**0.7698**	<u>0.4540</u>	0.7695	0.6603	0.7054

Table 4 reports the absolute Average Precision values of Full SG-TPR, three matched component ablations, the matched Additive-LR structural control, and an independently retuned Uniform-Gate control. The matched comparisons inherit the Full-selected configuration and isolate specific structural interventions, whereas Retuned Uniform-Gate provides an independently tuned comparison under the same training-side model-selection protocol. Throughout this section, the reference band denotes the descriptive interval defined in Figure 3.

Structural branch. Relative to No-TPR, Full SG-TPR is favored on VOC07, MIRFlickr, 3Sources, and emotions and remains within the reference band on yeast and SCENE. The resulting Full-favored/within-band/control-favored split is $4/2/1$. Complete dataset-level values are reported in Table 4.

Explicit SLCC loss. Relative to Full ($\beta = 0$), Full SG-TPR exceeds the positive reference threshold on MIRFlickr and 3Sources and remains within the reference band on the other five datasets, yielding a $2/5/0$ Full-favored / within-band / control-favored split. Because Full ($\beta = 0$) retains the complete structural branch, learned simplex gate, and correlation-aware initialization, this comparison isolates the marginal effect of the explicit $\beta\|\mathbf{B}\mathbf{B}^\top - \mathbf{C}_Y\|_F^2$ calibration term.

Learned versus uniform simplex allocation. The Uniform-Gate control fixes $\pi_v = 1/V$ throughout optimization while retaining all other components of Full. No inner search or control-specific retuning is performed. Relative to this control, Full SG-TPR is favored on VOC07, MIRFlickr, 3Sources, emotions, and OBJECT and remains within the reference band on yeast and SCENE. The resulting Full-favored/within-band/control-favored split is $5/2/0$.

Matched Additive-LR control. Additive-LR replaces $\mathbf{W}^{(v)}\mathbf{Q}_{B,\eta}$ with $\mathbf{W}^{(v)}$ in both the effective mapping and the corresponding row-energy feature term. All other components are unchanged, including the gated structural branch $\pi_v\mathbf{U}^{(v)}\mathbf{B}^\top$, auxiliary prediction loss, SLCC calibration, factor penalties, initialization, optimizer, preprocessing, splits, and downstream evaluator. For every outer fold, Additive-LR inherits the exact Full-selected $(\lambda, \rho, r, \beta)$ configuration; no control-specific hyperparameter retuning is performed.

Effect of independent Uniform-Gate retuning. After independent inner-only retuning, Full SG-TPR retains a 0.0049 average-AP advantage over Uniform-Gate. Under the same reference band, Full is favored on SCENE, 3Sources, emotions, and OBJECT, remains within the band on yeast and VOC07, and the retuned control is favored on MIRFlickr, yielding a $4/2/1$ Full-favored/within-band/control-favored split. Retuning narrows the average gap relative to the matched Uniform-Gate control by 0.0039 AP. The largest Full-minus-Control difference is observed on OBJECT ($+0.0160$ AP).

E Cross-Dataset Statistical Details

For each metric, the seven datasets are treated as independent blocks following the standard multi-dataset comparison framework of Demšar (2006). We first apply the Friedman test across all nine methods. SG-TPR is then compared with each of the eight baselines using two-sided Wilcoxon signed-rank tests, followed by Holm step-down correction (Holm, 1979) separately within each metric. Wilcoxon tests use `zero_method="wilcox"` and `method="auto"`. With seven nonzero dataset-level pairs, the smallest attainable two-sided exact Wilcoxon p-value under complete directional agreement is $2/2^7 = 0.015625$. With eight within-metric SG-TPR–baseline comparisons, the first Holm step therefore has a best-case adjusted value of 0.125. Given seven datasets, the exact pairwise tests have limited Holm-corrected resolution and are therefore used as descriptive supplementary comparisons.

Table 5: Cross-dataset statistical summary. Friedman tests compare all nine methods; the minimum adjusted p-value is taken over the eight SG-TPR–baseline Wilcoxon comparisons after within-metric Holm correction.

Metric	Friedman p	Min. Holm-adj. p	SG-TPR rank
Average Precision	1.04×10^{-4}	0.125	**1.1429**
Coverage	8.34×10^{-5}	0.125	**2.0000**
Hamming Loss	3.77×10^{-4}	0.125	**1.8571**
Ranking Loss	1.50×10^{-5}	0.125	**1.1429**

F Experimental Protocol and Provenance

Dataset provenance. SCENE and OBJECT are processed derivatives of the corresponding NUS-WIDE scene and object subsets (Chua et al., 2009). The registered SCENE matrix contains 4,400 instances, 33 labels, and five feature views of dimensions 64, 225, 144, 73, and 128. The registered OBJECT matrix contains 6,047 instances, 31 labels, and the same five registered families. We evaluate directly on the registered released SCENE and OBJECT benchmark matrices. Their exact matrix files and cryptographic checksums are recorded in the reproducibility package to identify the evaluation data unambiguously. Because the registered evaluation units are percentage budgets rather than unique subset cardinalities, adjacent percentages may map to the same integer k_p on low-dimensional datasets. Such repeated cardinalities are intentionally retained for all methods to preserve the frozen percentage-grid protocol.

Frozen inner partitions. For yeast, SCENE, VOC07, MIRFlickr, emotions, and OBJECT, each outer-training partition uses the released fixed inner-training/validation split, with 25% of the outer-training samples assigned to validation. For 3Sources, each outer-training partition uses three released greedy multilabel-stratified inner folds. All split indices are frozen and distributed with the reproducibility package.

Inner model selection. Hyperparameters are selected independently within each outer fold using only the corresponding frozen training-side partitions. Each candidate produces one global feature ranking on every applicable inner-training partition and is evaluated over all 20 feature budgets. For all methods, hyperparameter selection uses mean inner-validation Average Precision as the sole selection criterion. Coverage, Hamming Loss, and Ranking Loss are not used for model selection and are reported only in the final outer-fold evaluation.

SG-TPR hyperparameter selection. We fix $\alpha = 1$. For each outer-training fold, model selection is performed using training-side validation data only.

In the first stage, $\beta = 0$ is fixed and one deterministic coordinate-search sweep selects the base tuple (λ, ρ, r) in the order

$$\lambda \to \rho \to r.$$

The continuous coordinates use

$$\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\},$$

and the admissible label-subspace dimensions are selected from

$$\{2, 4, 8, 16\},$$

with values exceeding the number of labels removed.

After the base tuple is frozen, β alone is selected from

$$\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}.$$

Candidate quality is mean inner-validation AP over the 20 feature budgets. For 3Sources the score is additionally averaged over its three stored inner folds.

Randomness and optimization. Each SG-TPR training unit uses the deterministic structure-aware initialization specified in the method section. All trainable variables are jointly optimized in double precision using strong-Wolfe L-BFGS. Optimization proceeds in chunks of at most 20 internal iterations and uses fixed dataset-specific computational budgets and relative-objective stopping tolerances distributed with the released configuration files.

Outer-fold refitting. After model selection, the chosen fold-specific configuration is refitted from scratch on the complete outer training partition, comprising both its inner-training and inner-validation samples. This refit uses the same deterministic initialization procedure and optimization settings as candidate training. It produces a single global feature ranking, which is evaluated on the held-out outer test fold at all 20 feature budgets. No parameter, initialization, stopping checkpoint, or feature budget is selected using outer-test performance. Accordingly, candidate objectives are defined over the registered percentage-budget grid rather than over unique subset cardinalities.

Preprocessing boundaries. For selector fitting, a `MinMaxScaler(clip=True)` is fitted exclusively on the relevant selector-training partition and then applied to the associated validation or test partition. Thus inner candidate evaluation uses an inner-training-only scaler, whereas the final refit uses a scaler fitted on the complete outer-training partition.

After feature selection, the retained dimensions are standardized with a training-only StandardScaler for the downstream classifier. No validation- or test-sample statistics enter either preprocessing stage.

Feature ranking and downstream evaluator. After optimization, feature j in view v receives score

$$s_j^{(v)} = \sqrt{\left\|(\mathbf{W}^{(v)}\mathbf{Q}_{B,\eta})_{j,:}\right\|_2^2 + \pi_v^2 \left\|(\mathbf{U}^{(v)}\mathbf{B}^\top)_{j,:}\right\|_2^2}. \tag{16}$$

The view-specific score vectors are concatenated in the registered original view order and sorted in descending order to obtain one global ranking.

Downstream classifier. For every selector and retained feature budget, we use the same one-vs-rest logistic-regression evaluator after standardizing the selected features with a `StandardScaler` fitted only on the corresponding training partition. Each binary classifier uses `solver=liblinear`, ℓ_2 regularization, `C=1.0`, `fit_intercept=True`, `class_weight=None`, `max_iter=1000`, `tol=1e-4`, and `random_state=0`. The effective configuration was verified under scikit-learn 1.7.2; `fit_intercept=True` and `tol=1e-4` correspond to the defaults of that environment, while the remaining listed parameters are explicitly specified by the evaluator. The identical downstream evaluator is used for SG-TPR, its controls, and all baselines.

Aggregation and provenance. Within each outer fold, every metric is first averaged over the 20 registered percentage budgets. The reported dataset-level value is then the mean over the five outer folds. The released artifacts retain fold-level results, frozen sample indices, fold-specific selected configurations, candidate-level validation evidence, source snapshots, and cryptographic hashes.

Baseline implementations. Baselines are implemented from available source code or published algorithmic descriptions, as appropriate. For methods without complete executable author releases, we use paper-guided implementations based on the published objective, main update structure, and

Table 6: Baseline implementation bases and model-selection spaces. All hyperparameter searches use frozen inner-validation data only.

Method	Implementation basis	Searched parameters
Variance	Exact population-variance definition evaluated on selector-training data.	None.
RMAN-MMFS	Paper-guided implementation based on the AAAI 2026 formulation.	Center-plus-one-factor search (13 tuples): λ_{static} and β_{dynamic} in $\{10^{-3}, \ldots, 10^{3}\}$ around center 0.1; `mi_neighbors`=3.
GDRFS	Implementation guided by the available author code (`GDRFS.py`, commit `99e40af...`).	Two-round coordinate search from $(1, 1, 1, 1, 16)$: $\alpha, \beta, \gamma, \delta \in \{10^{-3}, \ldots, 10^{3}\}$ and $V_{\text{dim}} \in \{4, 8, 16, 32, 64\}$.
NR-CILL	Paper-guided approximate implementation of the IEEE TCYB method.	Center-plus-one-factor search (16 tuples): $\alpha_{\text{common}}, \beta_{\text{individual}} \in \{10^{-2}, \ldots, 10^{2}\}$; $\gamma_{\text{noise}} \in \{10^{-2}, \ldots, 10^{1}\}$; $\lambda_{\text{sparse}} \in \{10^{-3}, \ldots, 10^{1}\}$.
MSFS	Paper-guided approximate implementation of the IEEE TMM method.	Center-plus-one-factor search (13 tuples): $\alpha_{\text{graph}} \in \{10^{-2}, \ldots, 10^{2}\}$; $\beta_{\text{sparse}}, \gamma_{\text{view}} \in \{10^{-3}, \ldots, 10^{1}\}$; $k = 5$.
SPLDG	Paper-guided approximate implementation, with implementation details cross-checked against an independent public implementation where applicable.	Center-plus-one-factor search (13 tuples): $\alpha_{\text{pseudo}} \in \{10^{-2}, \ldots, 10^{2}\}$; $\beta_{\text{feature_graph}}, \gamma_{\text{dynamic}} \in \{10^{-3}, \ldots, 10^{1}\}$; $k = 5$.
BHDG	Paper-guided approximate implementation of the published method.	Center-plus-one-factor search (26 tuples): $\lambda_{1:3} \in \{10^{-2}, \ldots, 10^{3}\}$; $\rho \in \{0.01, 0.05, \ldots, 0.25\}$; $\alpha \in \{0.1, 0.3, \ldots, 1\}$; hash bits $= L/2$, $k = 10$, $\sigma = 1$.
UGRFS	Extension of the available AAAI 2025 implementation under the common evaluation protocol.	$\alpha, \beta, \gamma, \lambda$ over pre-specified dataset-specific inner-validation candidate sets (37–87 tuples); the exact candidate sets are stored in the released evidence files.

Table 7: Baseline optimization settings and implementation notes for the methods evaluated under the common experimental protocol.

Method	Optimization setting	Implementation notes
Variance	One deterministic pass; no iterative stopping rule.	Training-fold computation only with deterministic tie handling under the common ranking/evaluation pipeline.
RMAN-MMFS	Deterministic sequential ranking; no numerical convergence tolerance. Generic runner iteration budgets are non-binding.	The executable implementation uses the label–feature attention construction together with a vectorized Gaussian-MI computation.
GDRFS	Relative-objective stopping at 10^{-3}; effective cap 500 for inner and outer runs; coordinate search terminates when a full round is unchanged.	Protocol-controlled seeds are used throughout; the executable implementation uses a 500-iteration cap, and V_{dim} is selected using inner-validation data.
NR-CILL	Relative-objective stopping at 10^{-4} after at least three iterations; caps 500 inner / 1000 outer.	A deterministic normalized kNN-MI construction is used for the membership component in the approximate implementation; the latent width and search grid are fixed implementation choices.
MSFS	Relative-objective stopping at 10^{-4} after at least three iterations; caps 500 inner / 1000 outer.	The implementation uses a half-label latent width and the reported log-scale model-selection grid under the common protocol.
SPLDG	Full-objective relative-change stopping at 10^{-4} after at least three iterations; caps 500 inner / 1000 outer.	The implementation uses kNN graph construction and a nonnegative projected-gradient realization with objective backtracking for the signed-centering optimization.
BHDG	Objective-only relative-change stopping at 10^{-4} after at least three iterations; caps 500 inner / 1000 outer.	Sparse kNN algebra is used to avoid dense $n \times n$ storage while preserving the published update order and graph-construction timing.
UGRFS	Final-change stopping at 10^{-3}; caps 500 inner / 1000 outer; within each outer fold, the configuration is selected independently using mean inner-validation AP over the 1%–20% budgets.	The available implementation is extended to the common 500/1000-iteration protocol using vector-diagonal storage, nonnegative selector inputs, and final-iterate row-ℓ_2 ranking.

feature-ranking mechanism. Tables 6 and 7 summarize the implementation bases, search spaces, optimization settings, and key implementation notes. All methods share the same frozen outer splits, training-only selection boundary, prespecified feature budgets, and downstream evaluator.

Search-space sizes follow the parameterizations and available implementations of the respective methods; all candidate selection is restricted to the same training-side validation boundary.

G Additional Empirical Diagnostics

G.1 Budget-Wise Performance

Figure 2 in the main paper and Figure 4 together report budget-wise Average Precision on all seven datasets. All curves use the same frozen outer partitions, registered feature budgets, and common downstream evaluator as the main comparison.

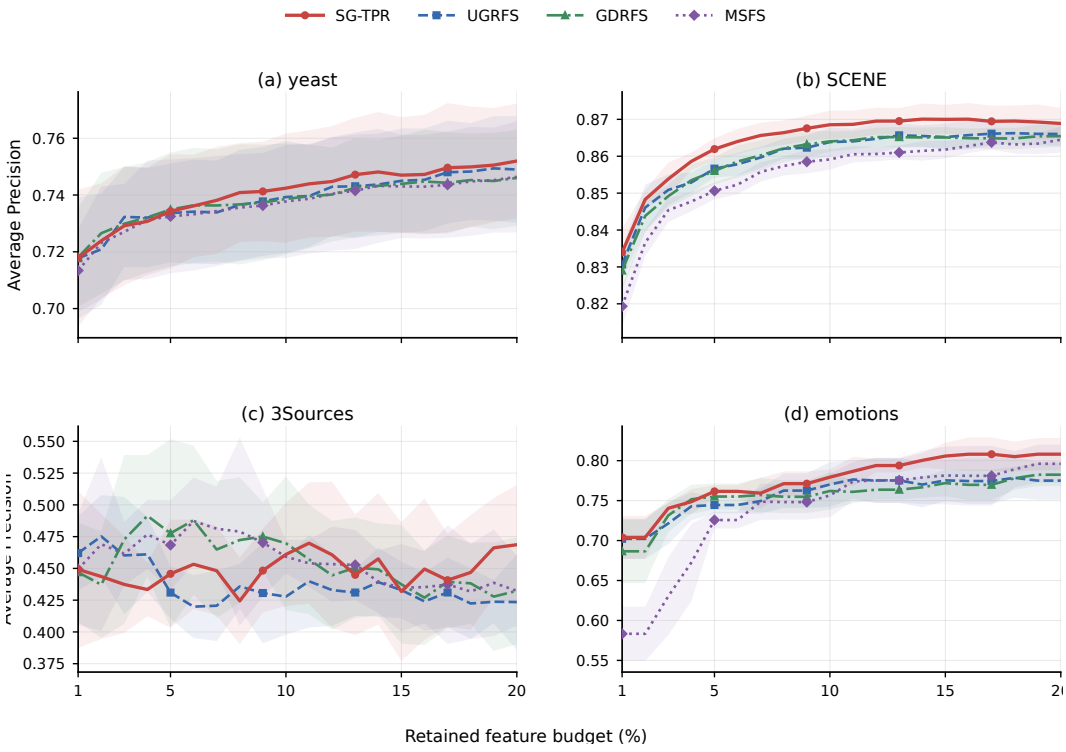

Figure 4: Additional budget-wise Average Precision curves on yeast, SCENE, 3Sources, and emotions. Curves show the mean over five frozen outer folds, with shaded regions denoting sample standard deviation. Together with Figure 2, the two figures cover all seven benchmark datasets.

G.2 Learned Simplex-Gate Diagnostics

For each outer fold, we summarize gate concentration using normalized entropy

$$H_{\mathrm{norm}}(\boldsymbol{\pi}) = -\frac{1}{\log V} \sum_{v=1}^{V} \pi_v \log \pi_v.$$

Thus $H_{\mathrm{norm}} = 1$ for uniform allocation, while values approaching zero indicate increasingly concentrated allocation. Table 8 reports the fold-wise mean and sample standard deviation of both π_v and H_{norm}.

Table 8: Learned simplex-gate diagnostics across the five outer folds. Entries for each view report mean $\pm$ sample standard deviation. $\bar{H}_{\mathrm{norm}}$ is the mean normalized gate entropy; $H_{\mathrm{norm}} = 1$ corresponds to uniform allocation and smaller values indicate more concentrated structural allocation. A dash denotes an unavailable view.

Dataset	View 1	View 2	View 3	View 4	View 5	$\bar{H}_{\mathrm{norm}}$
yeast	0.790 ± 0.189	0.210 ± 0.189	–	–	–	0.628 ± 0.260
SCENE	0.001 ± 0.001	0.969 ± 0.031	0.000 ± 0.000	0.030 ± 0.031	0.000 ± 0.000	0.082 ± 0.066
VOC07	0.463 ± 0.229	0.282 ± 0.198	0.255 ± 0.050	–	–	0.857 ± 0.219
MIRFlickr	0.765 ± 0.294	0.124 ± 0.169	0.111 ± 0.128	–	–	0.441 ± 0.482
3Sources	0.264 ± 0.147	0.468 ± 0.294	0.269 ± 0.146	–	–	0.807 ± 0.430
emotions	0.283 ± 0.291	0.717 ± 0.291	–	–	–	0.627 ± 0.206
OBJECT	0.096 ± 0.088	0.413 ± 0.286	0.052 ± 0.056	0.429 ± 0.275	0.010 ± 0.010	0.580 ± 0.298

Gate concentration. The learned gates exhibit substantially different allocation regimes across datasets. SCENE is the most concentrated case, assigning 0.969 ± 0.031 of the structural weight to View 2 and yielding a mean normalized entropy of only 0.082 ± 0.066. In contrast, VOC07 and 3Sources have considerably higher mean normalized entropies (0.857 and 0.807, respectively), indicating more distributed structural allocation.

Gate allocations vary across outer folds on MIRFlickr, emotions, and OBJECT. Accordingly, $\boldsymbol{\pi}$ is interpreted as a training-partition-dependent mechanism for adaptive structural regularization rather than as a fixed semantic ranking of view importance.

G.3 All-Feature Reference

To contextualize the registered feature-selection budgets, we evaluate the same downstream classifier using all available features under the same frozen outer folds and training-only preprocessing. No selector is fitted for this reference. Table 9 compares the all-feature AP with SG-TPR at the 10% and 20% budgets.

Table 9: Comparison with the all-feature downstream reference. The final column reports $\mathrm{AP}_{20\%} - \mathrm{AP}_{\mathrm{All}}$.

Dataset	All features	SG-TPR 10%	SG-TPR 20%	20% − All
yeast	0.7551	0.7424	0.7520	-0.0031
SCENE	0.8399	0.8685	0.8689	$+0.0290$
VOC07	0.6337	0.6994	0.7018	$+0.0680$
MIRFlickr	0.7217	0.7710	0.7764	$+0.0547$
3Sources	0.3863	0.4610	0.4687	$+0.0824$
emotions	0.8097	0.7793	0.8081	-0.0016
OBJECT	0.6726	0.6913	0.7058	$+0.0332$

At the 20% budget, SG-TPR exceeds the all-feature AP on five of seven datasets. On emotions, the difference from the all-feature reference is only -0.0016 AP, while yeast is lower by approximately 0.0031 AP. The largest gains occur on 3Sources, VOC07, and MIRFlickr, showing that substantial feature reduction can preserve or improve held-out predictive performance.