CONCEPT-AWARE VIEW ROUTING WITH SPARSE COMPENSATION FOR MULTI-VIEW MULTI-LABEL FEATURE SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years, multi-view multi-label learning (MVML) has attracted increasing attention due to its ability to characterize complex real-world objects by exploiting complementary information from multiple views and rich multi-label semantics. However, selecting discriminative features from high-dimensional multi-view data remains a critical challenge in MVML. Most existing methods adopt unified view fusion and shared feature selection strategies, making it difficult to capture the heterogeneous view dependencies associated with different label semantics and potentially overlooking specific features required for challenging semantic concepts. To address these issues, we propose View-Conditioned Route Selection for Multi-View Multi-Label Feature Selection (VCRS). Specifically, VCRS first constructs a semantic space that integrates atomic concepts and relational concepts, thereby jointly preserving label information and higher-order label semantics. Concept confidence, graph constraints, and adaptive view routing are further incorporated to achieve fine-grained concept-level multi-view fusion. In addition, VCRS learns a shared set of base features via nonconvex sparse learning, identifies under-represented concepts according to their concept reconstruction errors and importance, and selects concept-specific features through a sparse compensation mechanism. In this way, VCRS enhances the feature representation of challenging semantic concepts while retaining the advantages of shared feature selection. Experimental results on multiple multi-view datasets demonstrate the effectiveness of the proposed method.

1 INTRODUCTION

Modern intelligent systems increasingly rely on heterogeneous data sources to characterize complex real-world objects, where each object is typically described by multiple complementary views and associated with multiple correlated semantic labels. Multi-view multi-label learning (MVML) has attracted considerable attention as an important paradigm for overcoming the limitations of traditional single-view single-label learning and capturing rich semantic structures (Fu et al., 2024; Liu et al., 2023b; Zhang & Zhou, 2014; Zhang et al., 2018). However, the coexistence of multi-view heterogeneity, complex label dependencies, and high-dimensional feature spaces also introduces substantial redundant information, making it a critical challenge to effectively select discriminative features while preserving complementary semantic information (Luo et al., 2018; Chang & Yang, 2017; Li et al., 2017; Jian et al., 2016; Zhang et al., 2024a).

Existing Multi-View Multi-Label Feature Selection (MVMLFS) methods generally exploit the consistency among different views and learn shared representations through global fusion strategies to reduce redundancy across views (Xu et al., 2013; Sun, 2013; Liu et al., 2023c). Alternatively, they model feature-label dependencies based on label correlations to enhance semantic supervision during feature selection (Liu et al., 2023a; Zhang et al., 2020a). Although these approaches have achieved certain improvements in feature selection, most of them implicitly assume that different label semantics or latent objectives can share similar view contributions and feature selection patterns. However, in real-world multi-view scenarios, different semantic concepts often depend on distinct information sources (Lyu et al., 2022). For example, for the same image object, color-related views may be more suitable for describing appearance attributes, while texture or structural views may

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.

better characterize high-level semantics such as material and shape. In such cases, conventional MVMLFS methods may average the critical information required by different concepts, weakening the representation of certain concepts. Therefore, how to characterize the view dependency differences among different semantic objectives at the concept level remains a crucial issue in MVML.

Indeed, it is important to acknowledge that even if fine-grained view modeling can be performed according to different concepts, shared feature selection strategies still suffer from potential limitations. Since MVMLFS aims to select compact and discriminative feature representations from high-dimensional multi-view data, the optimization process generally tends to preserve shared features that can explain multiple semantic objectives. However, some low-frequency or difficult-to-represent semantic concepts may not be sufficiently captured by shared features, and their important discriminative information may be ignored during unified sparse selection, resulting in under-represented concepts (Tan et al., 2021; Zhao et al., 2022; Nie et al., 2010). Simply adding more features can alleviate this issue but will reintroduce substantial redundant information and weaken the sparsity of feature selection. Therefore, providing necessary incremental feature compensation only for concepts that remain insufficiently represented while maintaining the advantages of shared feature selection poses a new challenge for further improving MVMLFS.

To address the above issues, we revisit the semantic-view mismatch problem in MVMLFS and propose View-Conditioned Route Selection for Multi-View Multi-Label Feature Selection (VCRS). Unlike existing methods that fuse multi-view information in a unified space and learn shared feature representations, VCRS starts from concept-level semantic requirements, explores the compatibility between different semantic objectives and view information, and further considers the concept representation deficiency caused by shared selection. By jointly modeling concept semantic structures, concept-view dependencies, and concept-specific compensation requirements, VCRS obtains discriminative feature representations that better satisfy semantic requirements while maintaining feature selection sparsity. Overall, the main contributions of this paper can be summarized as follows:

- VCRS introduces a concept-aware semantic representation learning framework. It expands the original label space into a concept space containing both atomic concepts and relational concepts, and integrates concept confidence, concept graph structures, and concept-label prior constraints to achieve more comprehensive and stable semantic representation learning.
- We design a view-conditioned concept routing strategy. Different from existing unified view fusion approaches, VCRS estimates concept-view observability from the concept level and learns concept-conditioned view routing relationships, enabling different semantic concepts to utilize more suitable view information for representation.
- To address the problem of under-represented concepts, VCRS constructs a sparse compensation mechanism. Through a concept-adaptive sparse compensation strategy, VCRS selects limited complementary features, enhancing the representation capability of under-represented concepts while preserving the efficiency of shared feature selection.

2 Related Work

2.1 Multi-label Semantic Modeling and Feature Selection

Multi-label data typically contains multiple semantic labels with potential correlations. Exploiting label dependencies to obtain more informative supervision has become an important research direction in multi-label feature selection (Zhang et al., 2024a). To understand the feature differences corresponding to different labels, MB-MCF (Wu et al., 2020) analyzed information propagation mechanisms in multi-label data from the perspective of causal relationships between labels and features, and further distinguished shared features from label-specific features. In addition, some studies have explored latent label spaces and structural constraints (Huang & Zhou, 2012; Shan et al., 2026). LRDG (Zhang et al., 2024b) introduced hidden semantic structures into the feature selection process by learning latent representations and constructing dynamic graph constraints, demonstrating that latent spaces can provide supervision beyond the original label matrix. Meanwhile, methods based on label fusion and label structure modeling have further promoted the development of multi-label feature selection (Qian et al., 2023). Although existing methods based on global label spaces

or shared latent representations have achieved progress in enhancing overall semantic representation capability, how to characterize the heterogeneous information patterns relied on by different semantic concepts and identify and compensate for concepts that are insufficiently represented during feature selection remain challenging issues.

2.2 Multi-View Multi-Label Feature Selection

MVMLFS has attracted increasing research attention in recent years (Chen et al., 2026b; Liu et al., 2026). This direction needs to simultaneously handle information complementarity among multiple views and semantic correlations among multiple labels. Some methods mainly focus on learning shared components across all views. MSFS (Zhang et al., 2020b) jointly exploited multi-view information and sparse feature learning mechanisms to achieve discriminative feature selection for high-dimensional multi-view multi-label data. However, MVMLFS also faces the challenge of view-specific information. Utilizing only shared representations may ignore discriminative information unique to certain views (Wu et al., 2019; Lyu et al., 2024). Therefore, some studies jointly learn shared information and view-specific information. DHLI (Hao et al., 2024b) divided labels into common labels, view-specific labels, and noisy labels, and considered view consistency and complementarity through hybrid-label modeling. These studies indicate that different semantic concepts may depend on completely different views in complex real-world scenarios (Wei et al., 2025). Unified fusion strategies may average the view dependencies among different concepts, while shared sparse selection processes may cause low-frequency or difficult concepts to be insufficiently represented. Motivated by these observations, we attempt to propose a concept-level MVMLFS framework, enabling different concepts to exploit view information that better matches their semantic requirements while preserving both shared representation capability and concept completeness.

3 Method

3.1 Problem Formulation

Given a multi-view multi-label dataset $(\mathcal{X}, Y)$ with n instances, V views, and L labels, the v-th view is represented by $X^{(v)} \in \mathbb{R}^{n\times d_v}$, where d_v denotes its feature dimensionality. The multi-view feature set is $\mathcal{X} = \{X^{(1)}, X^{(2)}, \ldots, X^{(V)}\}$, and the label matrix is $Y \in \{0,1\}^{n\times L}$, with $Y_{:,j}$ denoting the j-th label. Our goal is to select discriminative features from high-dimensional multi-view data while exploiting complementary information across views and multi-label semantic structure. The main workflow of VCRS is shown in Figure 1.

3.2 Concept-Aware Semantic Space Construction

The original label space provides direct supervision but cannot explicitly capture higher-order label semantics. VCRS therefore extends it into a joint concept space consisting of *atomic concepts* and *relational concepts*. Atomic concepts directly preserve the original label information, i.e., $C_a = Y$. To construct relational concepts, VCRS applies K-means to group labels with similar response patterns into latent semantic clusters and the detailed process is provided in the Appendix E. For the i-th cluster G_i containing at least two labels, the corresponding relational concept is

$$c_i = \frac{1}{|G_i|} \sum_{j\in G_i} Y_{:,j}. \tag{1}$$

Let G_r denote the number of obtained relational concepts. Then, we have the relational concept matrix $C_r = [c_1, c_2, \ldots, c_{G_r}]$ and the final concept matrix $C = [C_a, C_r] \in \mathbb{R}^{n\times K}$, where $K = L + G_r$. This construction preserves direct label supervision while explicitly encoding latent label relationships, the specific proof process can be found in Appendix A.

To retain the structural correspondence between concepts and labels, we further construct a concept–label prior matrix $P \in \{0,1\}^{K\times L}$, where $P_{kj} = 1$ indicates that label j contributes to concept k, and $P_{kj} = 0$ otherwise. For atomic concepts, $P_a = I_L$; for each relational concept, the corresponding row is nonzero only at labels in its semantic cluster.

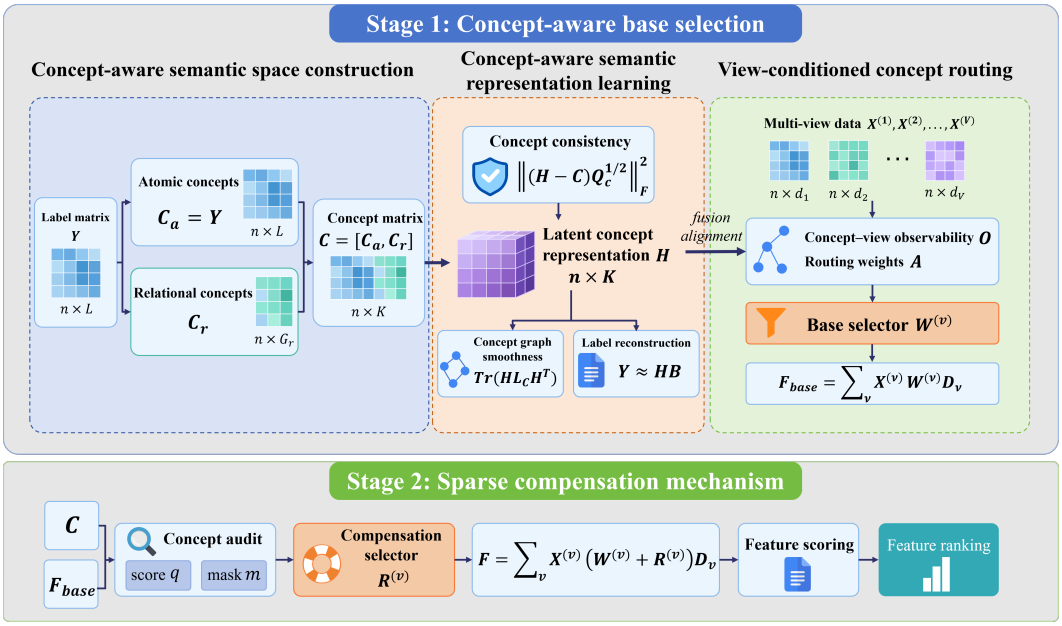

Figure 1: Overall workflow of the VCRS framework. Stage I performs concept-aware semantic modeling and view-conditioned routing for base feature selection, while Stage II identifies under-represented concepts and applies sparse compensation to obtain the final feature ranking.

3.3 Concept-Aware Semantic Representation Learning

Let $H \in \mathbb{R}^{n\times K}$ denote the latent concept representation matrix learned by the model, initialized as C. Considering that concepts may differ in reliability, VCRS introduces a response-pattern consistency matrix $Q_c = \mathrm{diag}(\rho_1, \rho_2, \ldots, \rho_K)$, where ρ_k denotes the confidence of concept k. Atomic concepts are directly derived from ground-truth labels and assigned a confidence value of 1, whereas relational concept confidence is estimated from within-cluster label agreement and statistical stability, the detailed calculation process is provided in the Appendix B. Accordingly, we formulate the concept consistency constraint as a weighted Frobenius-norm penalty:

$$\mathcal{L}_{\mathrm{con}} = \left\| (H - C)Q_c^{1/2} \right\|_F^2 . \tag{2}$$

VCRS also preserves relational structure among concepts. Let S_{ij} denote the correlation between concepts i and j across instances. Negative correlations are set to zero and self-connections are removed. With $D_{ii} = \sum_j S_{ij}$, we construct the normalized graph Laplacian $L_C = I - D^{-1/2}SD^{-1/2}$, where isolated concepts with $D_{ii} = 0$ are handled by setting their corresponding entries in $D^{-1/2}$ to zero and subsequently resetting their diagonal entries in L_C to zero, and I is the identity matrix. To encourage semantically related concepts to have similar latent representations, we impose the following concept graph smoothness regularization:

$$\mathcal{L}_{\mathrm{graph}} = \mathrm{Tr}(HL_CH^T). \tag{3}$$

To keep the expanded concept space anchored to the original labels, VCRS introduces a concept-to-label mapping matrix $B \in \mathbb{R}^{K\times L}$ and reconstructs Y with HB. The label reconstruction term is $\mathcal{L}_{\mathrm{lab}} = \|Y - HB\|_F^2 + \eta_0\|B\|_F^2$, where $\eta_0 > 0$ is a small ridge coefficient for numerical stability. Meanwhile, P penalizes mappings unsupported by the structural prior through $\mathcal{L}_{\mathrm{path}} = \|(1 - P) \odot B\|_F^2$, where $\odot$ denotes element-wise multiplication. Thus, H is jointly constrained by concept targets, concept structure, and ground-truth labels.

3.4 VIEW-CONDITIONED CONCEPT ROUTING

Different semantic concepts often depend on different views, making unified global fusion inadequate for concept-dependent view heterogeneity. VCRS therefore estimates the observability of each concept from each view and constructs a concept–view observability matrix:

$$\begin{aligned} S_{vk} &= \max_i \left| \mathrm{corr}\left(X_{:,i}^{(v)}, C_{:,k}\right) \right|, \\ O_{vk} &= \mathbb{I}(S_{vk} \geq \delta_k), \qquad \delta_k = 0.6 \max_u S_{uk}, \end{aligned} \tag{4}$$

where $\mathrm{corr}(\cdot,\cdot)$ denotes the correlation coefficient, and $O \in \{0,1\}^{V\times K}$ retains views with sufficiently strong responses to each target concept. The full details are provided in Appendix C.

Based on the concept–view observability matrix O, VCRS learns a concept-conditioned routing matrix $A \in \mathbb{R}^{V\times K}$, where A_{vk} denotes the routing weight assigned to view v for concept k. We enforce $A_{vk} = 0$ when $O_{vk} = 0$ and impose a nonnegative simplex constraint over observable views, such that $\sum_{v=1}^{V} A_{vk} = 1$ and $A_{vk} \geq 0$. We then define $D_v = \mathrm{diag}(A_{v,:} \odot O_{v,:})$, jointly incorporating learned view preferences and the concept observability prior into feature selection.

For view v, we introduce a base feature selection matrix $W^{(v)} \in \mathbb{R}^{d_v\times K}$. To jointly suppress irrelevant features across concept dimensions, we impose the nonconvex row-sparsity regularizer

$$\Omega_W = \sum_{v=1}^{V}\sum_{i=1}^{d_v} \left(\|W_{i,:}^{(v)}\|_2^2 + \varepsilon\right)^{p/2}, \qquad p = 0.5, \tag{5}$$

where p controls the degree of nonconvexity and $\varepsilon > 0$ is a small smoothing constant. In Stage I, the routed base concept representation is

$$F_{\mathrm{base}} = \sum_{v=1}^{V} X^{(v)} W^{(v)} D_v. \tag{6}$$

Hence, different concepts aggregate information along different view-specific routes, while $W^{(v)}$ identifies a shared set of base features within each view. The corresponding Stage-I routing constraint is $\mathcal{L}_{\mathrm{route}}^{\mathrm{base}} = \|H - F_{\mathrm{base}}\|_F^2 + \|A\|_F^2$.

3.5 SPARSE COMPENSATION MECHANISM

Although concept-conditioned routing selects more suitable views for individual concepts, shared feature selection may still fail to adequately represent some challenging semantic concepts. VCRS therefore identifies under-represented concepts after Stage I and applies concept-specific sparse compensation only to them.

For concept k, the concept reconstruction error is

$$e_k = \frac{\|C_{:,k} - F_{\mathrm{base},:,k}\|_2}{\|C_{:,k}\|_2 + \varepsilon}. \tag{7}$$

To account for semantic coverage and rarity in the data distribution, we define the empirical frequency $\pi_k = \frac{1}{n}\sum_{i=1}^{n} C_{ik}$ and the concept importance factor $r_k = \|P_{k,:}\|_2/\sqrt{\pi_k + \delta_f}$, where $\delta_f = 0.05$ is a frequency-smoothing constant. The concept compensation score is $q_k = e_k r_k$. After normalizing it to $\widehat{q}_k$, we set $\tau = \mathrm{mean}(\widehat{q})$ and define $m_k = \mathbb{I}(\widehat{q}_k > \tau)$. Thus, compensation accounts for reconstruction quality, semantic coverage, and rarity jointly.

For selected under-represented concepts, we introduce sparse compensation matrices $R^{(v)} \in \mathbb{R}^{d_v\times K}$ and update the final concept representation as

$$F = \sum_{v=1}^{V} X^{(v)} \left(W^{(v)} + R^{(v)}\right) D_v. \tag{8}$$

Here, $W^{(v)}$ models the shared base component, whereas $R^{(v)}$ provides concept-specific additive corrections for concepts identified as under-represented after Stage I. The routing consistency constraint becomes $\mathcal{L}_{\mathrm{route}} = \|H - F\|_F^2 + \|A\|_F^2$.

To restrict compensation to a small number of necessary features, we impose a concept-adaptive weighted ℓ_1 regularizer:

$$\Omega_R = \sum_{v=1}^{V} \sum_{k=1}^{K} m_k O_{vk} \xi_k \|R_{:,k}^{(v)}\|_1. \tag{9}$$

The weight ξ_k adaptively adjusts the sparsity penalty according to the degree of concept under-representation; its full computation is provided in the Appendix C. For any concept–view pair with $m_k O_{vk} = 0$, we set $R_{:,k}^{(v)} = \mathbf{0}$. Consequently, VCRS restricts sparse concept-specific corrections to audited concept–view pairs rather than applying compensation indiscriminately.

3.6 Optimization

VCRS integrates the above components into the unified objective

$$\begin{aligned} \min_{H,B,A,\{W^{(v)},R^{(v)}\}_{v=1}^{V}} \quad & \mathcal{L} = \|Y - HB\|_F^2 + \alpha \left[\|(H-C)Q_c^{1/2}\|_F^2 + \mathrm{Tr}(HL_CH^T) \right] \\ & \quad + \beta \left[\|H - F\|_F^2 + \|(1-P) \odot B\|_F^2 + \|A\|_F^2 \right] \\ & \quad + \eta_0 \|B\|_F^2 + \gamma \Omega_W + \lambda \Omega_R, \\ \text{s.t.} \quad & A_{vk} \geq 0, \quad \sum_{v=1}^{V} A_{vk} = 1, \quad A_{vk} = 0 \text{ if } O_{vk} = 0, \\ & R_{:,k}^{(v)} = 0 \text{ if } m_k O_{vk} = 0. \end{aligned} \tag{10}$$

where α, β, γ, and λ balance concept semantic modeling, multi-view concept fusion with path and routing regularization, nonconvex row-sparse base feature selection, and sparse compensation, respectively.

Owing to the coupled variables and nonconvex/nonsmooth regularizers in the objective, VCRS adopts a two-stage alternating optimization strategy. In Stage I, we set $R^{(v)} = 0$ and alternately update H, B, A, and W until the objective stabilizes. We then perform the concept under-representation audit once and fix m and ξ. In Stage II, H, B, A, W, and R are jointly updated. Specifically, H and B are updated by solving linear systems, A by projected gradient descent on the probability simplex, W by MM/IRLS for the nonconvex row-sparsity term, and R by ISTA for the weighted ℓ_1 subproblem. Complete update rules, initialization, and stopping criteria are provided in the Appendix D.

After optimization, VCRS computes feature scores from concept importance, routing weights, and the activation strength of base and compensation branches. Let $\omega_k = \|B_{k,:}\|_2 / \sqrt{\pi_k + \delta_f}$ and $g_{vk} = \widetilde{\omega}_k A_{vk} O_{vk}$, where $\widetilde{\omega}_k$ denotes the normalized concept weight. The score of feature i in view v is

$$s_i^{(v)} = \sqrt{\sum_{k=1}^{K} g_{vk} \left(W_{ik}^{(v)} \right)^2} + \sum_{k=1}^{K} g_{vk} \left| R_{ik}^{(v)} \right|. \tag{11}$$

Scores from all views are concatenated and sorted in descending order to obtain the global feature ranking.

Complexity. Let d denote the average feature dimensionality across views, $r = \min(n, d)$, T the number of outer iterations, T_W and T_R the numbers of inner MM and ISTA iterations, and K_r the average number of compensated concepts per view. The dominant training complexity is conservatively bounded by

$$\mathcal{O}\left(T \left[nKL + LK^3 + KVnd + T_W VK(ndr + r^3) \right] + T_2 T_R V K_r nd \right), \tag{12}$$

where T_2 denotes the number of Stage-II outer iterations. The additional R-update cost scales with K_r, which is typically smaller than K.

Table 1: Experimental results (mean ± std) in terms of AP and Coverage, where the 1st/2nd best results are shown in boldface/underline.

Dataset	VCRS	GRAFS	UGRFS	MSFS	NAIM3L	TM3L	GDRFS
				AP ↑			
emotions	$\mathbf{0.7871 \pm 0.015}$	0.7738 ± 0.033	0.7569 ± 0.015	0.7856 ± 0.015	0.7631 ± 0.039	0.7510 ± 0.047	$\underline{0.7867 \pm 0.015}$
yeast	$\mathbf{0.7604 \pm 0.008}$	0.7558 ± 0.005	0.7563 ± 0.005	0.7525 ± 0.007	0.7518 ± 0.005	0.7520 ± 0.005	$\underline{0.7581 \pm 0.005}$
Corel5K	$\mathbf{0.3974 \pm 0.045}$	$\underline{0.3730 \pm 0.042}$	0.3703 ± 0.038	0.3302 ± 0.032	0.3207 ± 0.035	0.3549 ± 0.033	0.3628 ± 0.041
SCENE	$\mathbf{0.8571 \pm 0.009}$	0.8466 ± 0.008	0.8532 ± 0.009	0.8399 ± 0.013	0.8423 ± 0.010	0.8500 ± 0.011	$\underline{0.8544 \pm 0.008}$
VOC07	$\mathbf{0.7011 \pm 0.019}$	0.6852 ± 0.015	$\underline{0.6874 \pm 0.018}$	0.6539 ± 0.014	0.6843 ± 0.017	0.6514 ± 0.005	0.6867 ± 0.019
MIRFlickr	$\mathbf{0.7760 \pm 0.011}$	$\underline{0.7736 \pm 0.013}$	0.7662 ± 0.012	0.7203 ± 0.011	0.7661 ± 0.013	0.7252 ± 0.007	0.7689 ± 0.010
				Coverage ↓			
emotions	2.9647 ± 0.041	3.0277 ± 0.167	3.1018 ± 0.064	$\mathbf{2.9333 \pm 0.058}$	3.0853 ± 0.195	3.1714 ± 0.247	$\underline{2.9418 \pm 0.084}$
yeast	$\mathbf{7.4377 \pm 0.082}$	$\underline{7.4589 \pm 0.036}$	7.4805 ± 0.038	7.5041 ± 0.052	7.5393 ± 0.032	7.5374 ± 0.030	7.4590 ± 0.040
Corel5K	$\mathbf{69.9665 \pm 6.026}$	$\underline{72.3159 \pm 5.787}$	72.3493 ± 6.352	83.7084 ± 1.069	87.1764 ± 2.030	78.5717 ± 2.886	73.7915 ± 5.768
SCENE	$\mathbf{12.0593 \pm 0.346}$	12.5526 ± 0.200	12.1569 ± 0.400	12.7496 ± 0.514	12.6730 ± 0.320	12.2491 ± 0.420	$\underline{12.1033 \pm 0.323}$
VOC07	$\mathbf{6.3563 \pm 0.442}$	6.6711 ± 0.316	$\underline{6.6564 \pm 0.404}$	7.3616 ± 0.403	6.7484 ± 0.315	7.5348 ± 0.159	6.6619 ± 0.447
MIRFlickr	$\underline{19.4200 \pm 0.237}$	19.5019 ± 0.290	19.5435 ± 0.244	20.7117 ± 0.090	19.6946 ± 0.233	20.7197 ± 0.148	$\mathbf{19.4022 \pm 0.248}$

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate VCRS on six representative multi-view multi-label datasets, namely emotions (Trohidis et al., 2011), yeast (Elisseeff & Weston, 2001), Corel5K (Duygulu et al., 2002), SCENE (Chua et al., 2009), VOC07 (Everingham et al., 2010), and MIRFlickr (Huiskes & Lew, 2008). These datasets cover diverse application domains, including image analysis, bioinformatics, and multi-source text analysis. They contain 2–5 views with sample sizes ranging from hundreds to thousands and different feature dimensionalities. The detailed information of the experimental dataset can be found in the Appendix E.

Comparing Methods. To comprehensively evaluate VCRS in multi-view multi-label scenarios, we compare it with six representative methods: GRAFS (Hao et al., 2024a), UGRFS (Hao et al., 2025), MSFS (Zhang et al., 2020b), GDRFS (Chen et al., 2026a), NAIM3L (Li & Chen, 2022), and TM3L (Zhao et al., 2021). All comparison methods learn multi-view representations or feature-label relationships, and can obtain feature weights for ranking, and select key features by weight ranking.

Evaluation Metrics. We adopt four widely used multi-label learning metrics to evaluate the effectiveness of VCRS, including Average Precision (AP), Hamming Loss (HL), Ranking Loss (RL), and Coverage. Higher AP indicates better performance, while lower values of the other three metrics indicate better performance. All experimental results are averaged over five independent runs. The parameters of all methods are tuned within $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$ to ensure fair comparison.

4.2 EXPERIMENTAL RESULTS

Feature Selection Performance. Table 1- 2 shows the performance comparison between VCRS and other methods. Each value represents the average result obtained from selected feature subsets with feature ratios ranging from 1% to 20%. VCRS achieves superior overall performance.

Specifically, VCRS obtains the highest average AP value of 0.7132 on six datasets, which is 1.03 percentage points higher than the best-performing baseline GDRFS. This demonstrates that VCRS can more accurately identify discriminative features related to label semantics. VCRS achieves the lowest average values of HL and RL, with scores of 0.1271 and 0.1269, respectively. These results demonstrate that VCRS can effectively reduce label prediction errors while preserving the semantic ranking structures induced by label dependencies. For Coverage, VCRS also achieves the lowest average value of 19.7007, which is 0.5140 lower than the best competing method UGRFS, showing that the selected features can more effectively characterize the label space structure and improve label ranking coverage.

To further evaluate the stability of VCRS under different feature budgets, Figure 2 illustrates the performance variation with feature ratios ranging from 1% to 20%. Feature ratio analysis shows that

Table 2: Experimental results (mean ± std) in terms of HL and RL, where the 1st/2nd best results are shown in boldface/underline.

Dataset	VCRS	GRAFS	UGRFS	MSFS	NAIM3L	TM3L	GDRFS
				HL ↓			
emotions	0.2279 ± 0.011	0.2320 ± 0.021	0.2305 ± 0.014	0.2322 ± 0.014	0.2369 ± 0.030	0.2533 ± 0.029	**0.2244 ± 0.014**
yeast	**0.2043 ± 0.010**	0.2085 ± 0.008	0.2046 ± 0.007	0.2113 ± 0.007	0.2100 ± 0.006	0.2152 ± 0.006	0.2067 ± 0.008
Corel5K	**0.0124 ± 0.000**	0.0128 ± 0.000	0.0127 ± 0.000	0.0129 ± 0.000	0.0128 ± 0.000	0.0127 ± 0.000	0.0127 ± 0.000
SCENE	**0.0877 ± 0.005**	0.0932 ± 0.004	0.0895 ± 0.005	0.0962 ± 0.006	0.0958 ± 0.005	0.0918 ± 0.005	0.0895 ± 0.004
VOC07	**0.0782 ± 0.003**	0.0802 ± 0.001	0.0799 ± 0.002	0.0832 ± 0.000	0.0798 ± 0.002	0.0839 ± 0.001	0.0804 ± 0.002
MIRFlickr	**0.1518 ± 0.005**	0.1537 ± 0.006	0.1547 ± 0.005	0.1756 ± 0.005	0.1568 ± 0.005	0.1729 ± 0.004	0.1554 ± 0.004
				RL ↓			
emotions	**0.1828 ± 0.014**	0.1985 ± 0.035	0.2148 ± 0.017	0.1838 ± 0.016	0.2109 ± 0.041	0.2276 ± 0.048	0.1893 ± 0.017
yeast	0.1706 ± 0.006	0.1729 ± 0.003	0.1727 ± 0.003	0.1758 ± 0.006	0.1762 ± 0.004	0.1769 ± 0.004	**0.1703 ± 0.004**
Corel5K	**0.1151 ± 0.013**	0.1199 ± 0.012	0.1212 ± 0.013	0.1428 ± 0.002	0.1510 ± 0.003	0.1336 ± 0.006	0.1227 ± 0.012
SCENE	**0.0591 ± 0.005**	0.0658 ± 0.004	0.0606 ± 0.005	0.0693 ± 0.009	0.0678 ± 0.006	0.0627 ± 0.006	0.0600 ± 0.005
VOC07	**0.1259 ± 0.014**	0.1348 ± 0.010	0.1353 ± 0.012	0.1582 ± 0.011	0.1384 ± 0.009	0.1644 ± 0.004	0.1365 ± 0.014
MIRFlickr	**0.1079 ± 0.005**	0.1093 ± 0.006	0.1118 ± 0.005	0.1367 ± 0.004	0.1129 ± 0.005	0.1351 ± 0.002	0.1100 ± 0.005

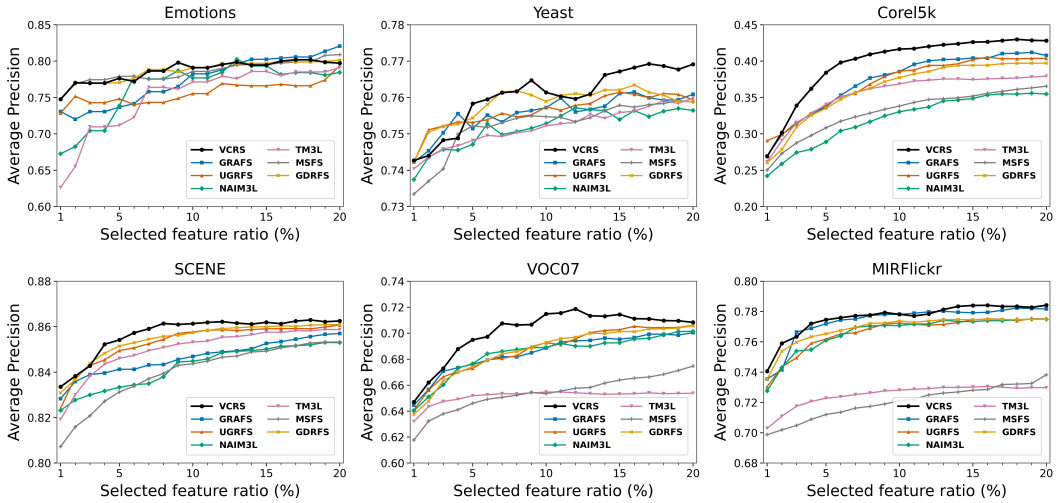

Figure 2: Performance variation with different feature ratios.

VCRS maintains leading performance under most feature ratios, demonstrating that its advantage does not rely on a specific feature budget but comes from a stable and effective feature selection mechanism. Even when only a small number of features are retained, VCRS can preserve sufficient discriminative information through concept-aware semantic modeling and sparse compensation mechanism.

Parameter Analysis. The four parameters α, β, γ, and λ in VCRS are associated with three key components: concept-aware semantic representation learning, view-conditioned concept routing, and sparse compensation mechanism. We adjust each parameter independently while keeping the remaining parameters fixed to evaluate the influence of different parameter configurations. As shown in Figure 3, AP increases steadily with more selected features. Across a wide parameter range, the performance varies smoothly without significant performance degradation, and only in a few extreme values does the performance decline occur. The results demonstrate that VCRS is relatively insensitive to parameter variations and can preserve stable feature selection capability over a broad parameter range.

Ablation Study. To verify the effectiveness of each key component, we construct three variants of the original VCRS model. VCRS$_{v1}$ degrades the concept matrix into the original label representation and removes relational concepts, concept graph constraints, and concept-label prior information.

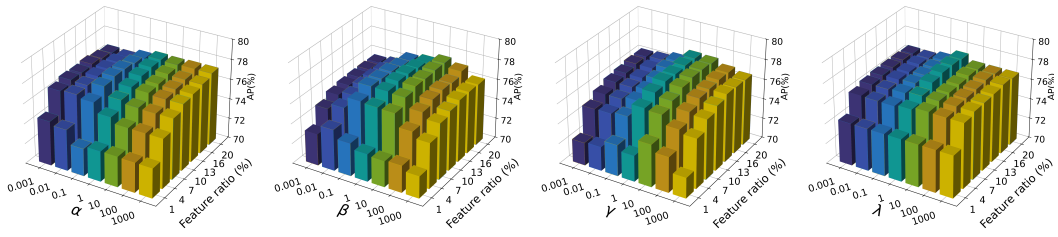

Figure 3: Parameter sensitivity studies on the MIRFlickr dataset.

Table 3: Ablation studies on the yeast, VOC07 and MIRFlickr datasets for Average Precision and Hamming loss.

	$AP \uparrow$				$HL \downarrow$			
Dataset	VCRS	VCRS_{v1}	VCRS_{v2}	VCRS_{v3}	VCRS	VCRS_{v1}	VCRS_{v2}	VCRS_{v3}
yeast	**0.7604**	0.7572	0.7599	0.7587	**0.2043**	0.2064	0.2044	0.2055
VOC07	**0.7011**	0.6948	0.6986	0.6985	**0.0782**	0.0790	0.0787	0.0787
MIRFlickr	**0.7760**	0.7744	0.7648	0.7742	**0.1518**	0.1526	0.1569	0.1525

VCRS_{v2} removes concept-view observability constraints and replaces adaptive routing with uniform view weights, making all concepts share the same view contributions. VCRS_{v3} removes the optimization of the sparse compensation term to evaluate the effectiveness of the sparse compensation mechanism.

Table 3 reports the AP and HL results of VCRS and its variants on yeast, VOC07, and MIRFlickr. VCRS achieves the best performance among all variants. VCRS_{v2} exhibits obvious degradation on datasets with more views and labels due to ignoring the importance of multi-view complementary information, especially on MIRFlickr, where the AP decreases by 1.12 percentage points. VCRS_{v1} and VCRS_{v3} also show performance degradation, while VCRS_{v1} generally performs worse than VCRS_{v3}, highlighting the importance of concept-aware semantic modeling and validating the effectiveness of sparse compensation mechanism. Overall, the ablation results demonstrate that each component of the proposed method is indispensable for achieving optimal performance.

5 Conclusion

This paper investigates the challenges of characterizing heterogeneous view dependencies and insufficient semantic concept representation in MVMLFS from a novel perspective. By constructing a concept semantic space integrating atomic concepts and relational concepts, it enhances the modeling capability of label information and higher-order label semantic relationships. Furthermore, concept-aware semantic representation learning and view-conditioned concept routing are employed to achieve fine-grained concept-level multi-view fusion. In addition, it incorporates a sparse compensation mechanism to provide targeted compensation for under-represented concepts, thereby improving the discriminative capability and semantic completeness of feature selection. Extensive experimental results demonstrate that the proposed method VCRS exhibits stable performance and superior feature selection capability. Although VCRS effectively alleviates the issue of insufficient representation of complex semantic concepts, its performance still depends on the quality of concept construction. Therefore, its adaptability in larger-scale and more diverse data scenarios remains to be further improved. Future work will explore more flexible concept discovery strategies and investigate efficient feature selection methods for dynamic multi-view multi-label environments.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing, related-work retrieval and discovery, and research ideation and execution. The authors were responsible for designing the proposed method, establishing the experimental protocol, conducting and auditing the experiments, and verifying all mathematical formulations and numerical results. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI tools.

REPRODUCIBILITY STATEMENT

We release the code, experimental configurations, preprocessing procedures, and evaluation protocols for reproducing VCRS. All feature normalization, model selection, and experimental decisions are performed using training partitions only. The released implementation follows the complete optimization pipeline described in the paper. Detailed parameter settings, convergence criteria, feature ranking procedures, and experimental protocols are documented in the appendix. The released artifacts enable reproduction of the main experimental comparisons, ablation studies, and parameter analyses under the same evaluation settings.

REFERENCES

Xiaojun Chang and Yi Yang. Semisupervised feature analysis by mining correlations among multiple tasks. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10):2294–2305, 2017. doi: 10.1109/TNNLS.2016.2582746.

Jiaming Chen, Ping Zhang, Yonghao Li, and Wanfu Gao. Graph-fusion guided data reconstruction for multi-view multi-label feature selection. *Pattern Recognition*, 179:113692, 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.

Pinar Duygulu, Kobus Barnard, Joao FG de Freitas, and David A Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European conference on computer vision*, pp. 97–112. Springer, 2002.

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.

Kaicheng Fu, Changde Du, Shengpei Wang, and Huiguang He. Multi-view multi-label fine-grained emotion decoding from human brain activity. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):9026–9040, 2024. doi: 10.1109/TNNLS.2022.3217767.

Pingting Hao, Kunpeng Liu, and Wanfu Gao. Anchor-guided global view reconstruction for multi-view multi-label feature selection. *Information Sciences*, 679:121124, 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. 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, 2025.

Sheng-Jun Huang and Zhi-Hua Zhou. Multi-label learning by exploiting label correlations locally. In *Proceedings of the AAAI conference on artificial intelligence*, volume 26, pp. 949–955, 2012.

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.

Ling Jian, Jundong Li, Kai Shu, and Huan Liu. Multi-label informed feature selection. In *Ijcai*, volume 16, pp. 1627–33, 2016.

Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6), 2017.

Xiang Li and Songcan Chen. A concise yet effective model for non-aligned incomplete multi-view and missing multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5918–5932, 2022. doi: 10.1109/TPAMI.2021.3086895.

Bo Liu, Weibin Li, Yanshan Xiao, Xiaodong Chen, Laiwang Liu, Changdong Liu, Kai Wang, and Peng Sun. Multi-view multi-label learning with high-order label correlation. *Information Sciences*, 624:165–184, 2023a.

Chengliang Liu, Jie Wen, Xiaoling Luo, Chao Huang, Zhihao Wu, and Yong Xu. Dicnet: Deep instance-level contrastive network for double incomplete multi-view multi-label classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 8807–8815, 2023b.

Wei Liu, Jiazheng Yuan, Gengyu Lyu, and Songhe Feng. Label driven latent subspace learning for multi-view multi-label classification. *Applied Intelligence*, 53(4):3850–3863, 2023c.

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.

Minnan Luo, Xiaojun Chang, Liqiang Nie, Yi Yang, Alexander G. Hauptmann, and Qinghua Zheng. An adaptive semisupervised feature analysis for video semantic recognition. *IEEE Transactions on Cybernetics*, 48(2):648–660, 2018. doi: 10.1109/TCYB.2017.2647904.

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.

Yi Shan, Liyang Gao, Yuena Lin, Zhen Yang, Gengyu Lyu, and Honggui Han. Hypergraph-based multi-view multi-label classification via adaptive high-order semantic fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 25269–25277, 2026.

Shiliang Sun. A survey of multi-view machine learning. *Neural computing and applications*, 23(7): 2031–2038, 2013.

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, 2021. doi: 10.1109/TCYB.2019.2950560.

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.

Hao Wei, Yongjian Deng, Qiuru Hai, Yuena Lin, Zhen Yang, and Gengyu Lyu. Multi-view multi-label classification via view-label matching selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21456–21464, 2025.

Xingyu Wu, Bingbing Jiang, Kui Yu, Huanhuan Chen, and Chunyan Miao. Multi-label causal feature selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 6430–6437, 2020.

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.

Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

Changqing Zhang, Ziwei Yu, Qinghua Hu, Pengfei Zhu, Xinwang Liu, and Xiaobo Wang. Latent semantic aware multi-view multi-label classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Fangwen Zhang, Xiuyi Jia, and Weiwei Li. Tensor-based multi-view label enhancement for multi-label learning. In *IJCAI*, pp. 2369–2375, 2020a.

Jia Zhang, Yidong Lin, Min Jiang, Shaozi Li, Yong Tang, Jinyi Long, Jian Weng, and Kay Chen Tan. Fast multilabel feature selection via global relevance and redundancy optimization. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4):5721–5734, 2024a. doi: 10.1109/TNNLS.2022.3208956.

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, 2014. doi: 10.1109/TKDE.2013.39.

Yao Zhang, Wei Huo, and Jun Tang. Multi-label feature selection via latent representation learning and dynamic graph constraints. *Pattern Recognition*, 151:110411, 2024b.

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, 2020b.

Dawei Zhao, Qingwei Gao, Yixiang Lu, and Dong Sun. Two-step multi-view and multi-label learning with missing label via subspace learning. *Applied Soft Computing*, 102:107120, 2021.

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.

A Group-Level Label Structure in the Concept Space

Let $Y = [y_1, \ldots, y_L] \in \mathbb{R}^{n \times L}$, where y_j is the response vector of label j across n instances. VCRS organizes labels with similar response patterns into groups and constructs a relational concept for each retained group. We formalize the label-group information represented by these concepts and examine how the resulting coordinates relate to the original label space.

Definition A.1 (Group-level label information). For a nonempty label group $G_g \subseteq \{1, \ldots, L\}$, define its relational concept response as

$$c_g = \frac{1}{|G_g|} \sum_{j \in G_g} y_j. \tag{13}$$

The group-level label information associated with this concept is the pair

$$\mathcal{I}_g = (G_g, c_g). \tag{14}$$

Here, G_g identifies the labels organized into one semantic unit, whereas c_g summarizes their responses across instances. We define the order of this information by the number of labels in the group:

$$\text{size}(I_g) = |G_g|. \tag{15}$$

An atomic concept has group size one, whereas a relational concept may organize two or more labels. Group size describes label organization, not the degree of a statistical or polynomial label interaction.

Proposition A.1 (Group-aware coordinates preserve the label span). Let $G_1, \ldots, G_{G_r}$ be the groups retained as relational concepts. Define their incidence matrix $E \in \{0,1\}^{L \times G_r}$ by

$$E_{jg} = \begin{cases} 1, & j \in G_g, \\ 0, & j \notin G_g. \end{cases}$$

Let

$$D_G = \text{diag}(|G_1|, \ldots, |G_{G_r}|), \qquad M = E D_G^{-1}. \tag{16}$$

Then the relational and complete concept matrices satisfy

$$C_r = YM, \qquad C = [Y, C_r] = Y[I, M]. \tag{17}$$

Consequently,

$$\text{span}(C) = \text{span}(Y), \qquad \text{rank}(C) = \text{rank}(Y). \tag{18}$$

Moreover, each relational coordinate retains an explicit label-set identity through $\text{supp}(M_{:,g}) = G_g$.

Proof. By the definition of E and D_G, the g-th column of YM is

$$(YM)_{:,g} = \frac{1}{|G_g|} \sum_{j \in G_g} y_j = c_g.$$

Hence $C_r = YM$, which gives Eq.(17). Every column of C belongs to the column span of Y, so $\text{span}(C) \subseteq \text{span}(Y)$. Conversely, C contains all columns of Y, implying $\text{span}(Y) \subseteq \text{span}(C)$. The equality of spans and ranks follows. Finally, the nonzero entries of $M_{:,g}$ occur exactly at the indices in G_g, establishing $\text{supp}(M_{:,g}) = G_g$. Thus, the relational coordinates record label-group membership without expanding the column span or rank of Y.

Proposition A.2 (The group organization is determined by label responses). To remove the influence of label frequency magnitude and obtain scale-invariant label response patterns, each label response vector is L2-normalized as

$$\tilde{y}_j = \frac{y_j}{\max(\|y_j\|_2, 10^{-12})}, \tag{19}$$

where the small constant 10^{-12} avoids numerical instability for zero-frequency labels. VCRS constructs label groups by applying K-means to $\tilde{y}_1, \ldots, \tilde{y}_L$. For a cluster G_g, its centroid is

$$\tilde{\mu}_g = \frac{1}{|G_g|} \sum_{j \in G_g} \tilde{y}_j. \tag{20}$$

Its within-cluster objective admits the pairwise form

$$\sum_{j \in G_g} \|\tilde{y}_j - \tilde{\mu}_g\|_2^2 = \frac{1}{2|G_g|} \sum_{a,b \in G_g} \|\tilde{y}_a - \tilde{y}_b\|_2^2. \tag{21}$$

For binary label responses, let $N_a = \sum_{i=1}^n Y_{ia}$ denote the frequency of label a, and let $N_{ab} = \sum_{i=1}^n Y_{ia} Y_{ib}$ denote the co-occurrence count of labels a and b. For labels with $N_a, N_b > 0$, we have

$$\|\tilde{y}_a - \tilde{y}_b\|_2^2 = 2 - \frac{2N_{ab}}{\sqrt{N_a N_b}}. \tag{22}$$

Therefore, the learned grouping reflects both individual label frequencies and pairwise co-occurrence patterns. If $\mathcal{G}(Y)$ denotes the groups obtained through clustering of the normalized label column, the relational concept construction can be expressed as

$$C_r(Y) = Y\, M\big(\mathcal{G}(Y)\big). \tag{23}$$

Proof. Expanding the definition of $\tilde{\mu}_g$ in Eq.(20) gives the standard pairwise-distance identity in Eq.(21). For binary label responses with $N_a, N_b > 0$, the normalized vectors satisfy

$$\tilde{y}_a^\top \tilde{y}_b = \frac{y_a^\top y_b}{\|y_a\|_2 \|y_b\|_2} = \frac{N_{ab}}{\sqrt{N_a N_b}}, \tag{24}$$

where $\|y_a\|_2 = \sqrt{N_a}$ and $\|y_b\|_2 = \sqrt{N_b}$. Since $\|\tilde{y}_a\|_2 = \|\tilde{y}_b\|_2 = 1$, we have

$$\|\tilde{y}_a - \tilde{y}_b\|_2^2 = \|\tilde{y}_a\|_2^2 + \|\tilde{y}_b\|_2^2 - 2\tilde{y}_a^\top \tilde{y}_b = 2 - \frac{2N_{ab}}{\sqrt{N_a N_b}},$$

which proves Eq.(22). The K-means objective therefore depends on label-response distances determined by these frequency and co-occurrence statistics. The resulting groups determine the incidence matrix and, in turn, the averaging matrix in Eq.(23). For a fixed grouping, multiplication by M is linear in Y; the grouping used to construct M depends on the observed label responses.

Proposition A.3 (Relational concepts aggregate label dependencies). Let $m_g = |G_g|$. The squared norm of a relational concept satisfies

$$\|c_g\|_2^2 = \frac{1}{m_g^2} \left[\sum_{a \in G_g} \|y_a\|_2^2 + 2 \sum_{\substack{a<b \\ a,b \in G_g}} y_a^\top y_b \right]. \tag{25}$$

For response vectors $u, v \in \mathbb{R}^n$, define empirical covariance and variance consistently by

$$\mathrm{Cov}_n(u, v) = \frac{1}{n}(u - \bar{u}\mathbf{1})^\top (v - \bar{v}\mathbf{1}), \qquad \mathrm{Var}_n(u) = \mathrm{Cov}_n(u, u),$$

where $\bar{u} = n^{-1}\mathbf{1}^\top u$. Then

$$\mathrm{Var}_n(c_g) = \frac{1}{m_g^2} \left[\sum_{a \in G_g} \mathrm{Var}_n(y_a) + 2 \sum_{\substack{a<b \\ a,b \in G_g}} \mathrm{Cov}_n(y_a, y_b) \right], \tag{26}$$

and, for two groups G_g and G_h,

$$\mathrm{Cov}_n(c_g, c_h) = \frac{1}{m_g m_h} \sum_{a \in G_g} \sum_{b \in G_h} \mathrm{Cov}_n(y_a, y_b). \tag{27}$$

Proof. Substituting Eq.(13) into $\|c_g\|_2^2$ and expanding the inner product gives

$$\|c_g\|_2^2 = \frac{1}{m_g^2} \sum_{a,b \in G_g} y_a^\top y_b.$$

Separating diagonal terms from off-diagonal terms yields Eq.(25). The empirical covariance is bilinear. Applying bilinearity to the average in Eq.(13) gives Eq.(26), and applying it to the averages of c_g and c_h gives Eq.(27). Thus, the response statistics of each relational concept aggregate individual-label and pairwise-label statistics within its group, while covariance between concepts aggregates pairwise dependencies across groups. The group-level information in Eq. (14) consists of explicit label-group membership and the group's mean response. Eq. (25)–(27) aggregate individual-label and pairwise statistics; they do not establish irreducible interactions among three or more labels.

The concept graph extends this organization across concepts. Its nodes include both atomic labels and relational concepts, while its affinities describe relationships between their response vectors. VCRS thus represents label structure at two connected levels: labels are organized within relational concepts, and the resulting concepts participate in a graph. The incidence matrix specifies which labels are organized together, whereas concept responses and graph affinities characterize how these groups behave across instances and relate to one another.

B Relational Concept Response-Pattern Consistency

To characterize the response-pattern consistency of different relational concepts, let $Y_g = Y_{:,G_g}$. We first compute the within-cluster label agreement:

$$a_g = 1 - \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}(Y_{g,i:}) \,. \tag{28}$$

We then define the label frequency stability:

$$s_g = 1 - \mathrm{Std}\left(\frac{1}{n}\sum_{i=1}^{n} Y_{g,i:}\right). \tag{29}$$

The relational concept confidence is

$$\rho_g = \mathrm{clip}\left(\frac{a_g + s_g}{2},\, 0.25,\, 0.90\right). \tag{30}$$

Here, agreement intentionally includes shared zeros, and frequency stability measures the balance of empirical label frequencies within a group. The method treats co-presence and co-absence as complementary evidence of a similar binary response pattern. Consequently, rare labels with disjoint positive instances may still form a high-consistency relational concept when their complete response vectors are close. Thus, ρ_g is a heuristic response-pattern consistency weight rather than an estimate of positive co-occurrence or causal association. This yields the response-pattern consistency matrix $Q_c = \mathrm{diag}(\rho_1, \rho_2, \ldots, \rho_K)$.

C Concept Observability and Sparse Compensation Details

C.1 Concept-view Observability

For feature i in view v and concept k, we define the centered vectors

$$x_i^{(v)} = X_{:,i}^{(v)} - \frac{1}{n}\sum_{r=1}^{n} X_{ri}^{(v)}, \qquad c_k = C_{:,k} - \frac{1}{n}\sum_{r=1}^{n} C_{rk}. \tag{31}$$

The correlation coefficient is implemented as

$$\mathrm{corr}\left(X_{:,i}^{(v)}, C_{:,k}\right) = \frac{\left(x_i^{(v)}\right)^{\top} c_k}{\left(\left\|x_i^{(v)}\right\|_2 + \varepsilon\right)\left(\|c_k\|_2 + \varepsilon\right)}. \tag{32}$$

The strongest response of view v to concept k is then

$$S_{vk} = \max_i \left|\mathrm{corr}\left(X_{:,i}^{(v)}, C_{:,k}\right)\right|. \tag{33}$$

Let

$$v_k^{\star} = \arg\max_v S_{vk}, \qquad S_k^{\star} = \max_v S_{vk}. \tag{34}$$

If $S_k^{\star} > \varepsilon$, then

$$O_{vk} = \mathbb{I}(S_{vk} \geq 0.6 S_k^{\star}) \,, \tag{35}$$

and the strongest-response view $v_k^{\star}$ is retained in all cases. Therefore, the 60% threshold in the main text is a relative threshold with respect to the response of the strongest view for the concept, rather than a fixed absolute correlation threshold. If $S_k^{\star} \leq \varepsilon$, only the view corresponding to $v_k^{\star}$ is retained, preventing all views from being activated simultaneously in degenerate cases.

C.2 Compensation Weights

After Stage I is completed, VCRS performs a one-time under-represented concept audit. For concept k, the concept compensation score is defined as $q_k = e_k r_k$. It is normalized by its maximum value:

$$\widehat{q}_k = \frac{q_k}{\max_j q_j + \varepsilon}. \tag{36}$$

We subsequently define

$$\tau = \frac{1}{K} \sum_{k=1}^{K} \widehat{q}_k, \tag{37}$$

and $m_k = \mathbb{I}(\widehat{q}_k > \tau)$. If no concept satisfies the strict inequality, the concept $k^\star = \arg\max_k \widehat{q}_k$ is forcibly included in the compensation set. The sparse compensation weight is first defined as

$$\xi_k^{(0)} = \frac{1}{\widehat{q}_k + 0.10}, \tag{38}$$

and is further normalized as

$$\xi_k = \frac{\xi_k^{(0)}}{\frac{1}{|M_{comp}|} \sum_{j \in M_{comp}} \xi_j^{(0)} + \varepsilon}, \tag{39}$$

where $M_{\text{comp}} = \{k | m_k = 1\}$ denote the set of concepts identified as under-represented and requiring sparse compensation. After Stage II begins, m and ξ are kept fixed. For any concept–view pair satisfying $m_k O_{vk} = 0$, we fix $R_{:,k}^{(v)} = \mathbf{0}$.

D Complete Update Rules for Two-Stage Optimization

D.1 Initialization

Within VCRS, each view is independently min–max scaled to $[0, 1]$. The model variables are initialized as

$$H^{(0)} = C, \qquad B^{(0)} = P, \qquad R^{(v,0)} = \mathbf{0}. \tag{40}$$

The base selection matrix is initialized using ridge regression:

$$W^{(v,0)} = \left(X^{(v)\top} X^{(v)} + 10^{-3} I \right)^{-1} X^{(v)\top} C. \tag{41}$$

When $d_v > n$, the equivalent form

$$W^{(v,0)} = X^{(v)\top} \left(X^{(v)} X^{(v)\top} + 10^{-3} I \right)^{-1} C \tag{42}$$

is used.

D.2 Parameter Update Rules

Stage I fixes $R^{(v)} = \mathbf{0}$, thereby deactivating the sparse compensation term, and sequentially updates $H \to B \to A \to W$. After each outer iteration, the full objective function is recomputed and convergence is assessed. After Stage I ends, the under-represented concept audit is performed only once, and m and ξ are fixed.

With the other variables fixed, the subproblem associated with H is

$$\begin{aligned} \min_H \quad & \|Y - HB\|_F^2 + \alpha \left\| (H - C) Q_c^{1/2} \right\|_F^2 \\ & + \alpha \operatorname{Tr}\left(H L_C H^\top \right) + \beta \|H - F\|_F^2 . \end{aligned} \tag{43}$$

Setting its first-order derivative to zero yields

$$H \leftarrow \left(Y B^\top + \alpha C Q_c + \beta F \right) \left(B B^\top + \alpha Q_c + \alpha L_C + \beta I \right)^{-1} . \tag{44}$$

If the matrix is singular, the pseudoinverse is used as a fallback. Similarly, B can be directly updated by solving the linear system

$$B_{:,l} \leftarrow \left(H^\top H + \beta \operatorname{diag}(1 - P_{:,l})^2 + \eta_0 I\right)^{-1} H^\top Y_{:,l}. \tag{45}$$

For the routing matrix A, each concept is independently optimized only over its observable views. Let a_k denote the routing vector of concept k over its observable views, and let Z_k consist of the current representations of that concept in these views. We use the projected gradient update

$$a_k \leftarrow \Pi_\Delta\left[a_k - \eta_A \left(Z_k^\top Z_k a_k - Z_k^\top H_{:,k} + a_k\right)\right], \tag{46}$$

where $\Pi_\Delta(\cdot)$ denotes projection onto the probability simplex. For the nonconvex row-sparsity term in $W^{(v)}$, MM/IRLS is used for iterative reweighting. Define

$$s_i^{(v)} = \frac{p}{2}\left(\left\|W_{i,:}^{(v)}\right\|_2^2 + \varepsilon\right)^{p/2-1}, \qquad S_v = \operatorname{diag}\left(s_1^{(v)}, \ldots, s_{d_v}^{(v)}\right). \tag{47}$$

The base selection coefficient for concept k can then be written as

$$W_{:,k}^{(v)} \leftarrow \left(\beta d_{vk}^2 X^{(v)\top} X^{(v)} + \gamma S_v\right)^{-1} \beta d_{vk} X^{(v)\top} \widetilde{H}_{vk}, \tag{48}$$

where $d_{vk} = A_{vk} O_{vk}$, and $\widetilde{H}_{vk}$ denotes the residual target after removing the contributions of the other views and the current compensation term.

When $\lambda > 0$ and the compensation set is nonempty, Stage II is entered and warm-started from the Stage-I solution. The sparse compensation matrix is updated only for concept–view pairs satisfying $m_k O_{vk} = 1$. The residual is obtained after removing the previous contribution of $R_{:,k}^{(v)}$. The corresponding subproblem is a weighted ℓ_1 sparse regression problem, which is solved using ISTA and the soft-thresholding operator:

$$R_{:,k}^{(v)} \leftarrow \mathcal{S}_{\lambda \xi_k / L_{vk}}\left[R_{:,k}^{(v)} - \frac{1}{L_{vk}} \nabla_{R_{:,k}^{(v)}}\left(\beta \left\|H - F\right\|_F^2\right)\right], \tag{49}$$

where L_{vk} is an upper bound on the gradient Lipschitz constant and $\mathcal{S}_\theta(\cdot)$ denotes the soft-thresholding operator.

D.3 Outer Convergence Criterion

Both Stage I and Stage II use the full objective function value to monitor outer convergence:

$$\Delta_t = \frac{|\mathcal{L}_t - \mathcal{L}_{t-1}|}{\max(|\mathcal{L}_{t-1}|, 1)}. \tag{50}$$

In the current final implementation, early stopping is applied after the initial iteration when $\Delta_t < \varepsilon_{\text{out}}$, where $\varepsilon_{\text{out}} = 10^{-3}$. The maximum number of outer iterations is 100 for both Stage I and Stage II.

D.4 Algorithm Overview

The complete algorithm is shown in Algorithm 1.

E Experimental Protocol

E.1 Experimental Run Settings

Following common experimental protocols for multi-view multi-label feature selection, each method first generates a feature ranking sequence according to the learned feature weights and selects key features with different feature ratios. The selected features are then evaluated using the same multi-label prediction protocol. To evaluate the feature selection capability under limited feature budgets,

Algorithm 1 Two-stage optimization procedure of VCRS

Require: $\{X^{(v)}\}_{v=1}^{V}$, Y, hyperparameters α, β, γ, λ, maximum iterations $T_{\max}$, and convergence threshold $\varepsilon_{\text{stop}}$

Stage I: Base Feature Selection
Construct concept matrix C, concept–label prior P, response-pattern consistency matrix Q_c, concept graph Laplacian L_C, and concept–view observability matrix O
Initialize H, B, A, and $\{W^{(v)}\}_{v=1}^{V}$, and set $R^{(v)} = \mathbf{0}$
repeat
    Alternately update $H \to B \to A \to W$
    Check convergence according to the relative objective change
until the objective converges or $T_{\max}$ is reached
Obtain the base representation F_{base}
Perform the concept under-representation audit and compute compensation indicators m and ξ
Fix m and ξ for the second optimization stage
Stage II: Sparse Compensation
Warm-start from the solution obtained in Stage I
repeat
    Alternately update $H \to B \to A \to W \to R$
    Activate $R_{:,k}^{(v)}$ only when $m_k O_{vk} = 1$
    Check convergence according to the relative objective change
until the objective converges or $T_{\max}$ is reached
Compute final feature scores and obtain the global feature ranking
return the global feature ranking

Table 4: Statistics of the registered benchmark datasets. Features are summed across views, and view dimensions list the feature counts of individual views.

Dataset	Samples	Features	Labels	Views	View dimensions
emotions	593	72	6	2	8 / 64
yeast	2,417	103	14	2	79 / 24
Corel5K	4,999	1,312	260	5	100 / 300 / 512 / 100 / 300
SCENE	4,400	634	33	5	64 / 225 / 144 / 73 / 128
VOC07	3,817	712	20	3	100 / 512 / 100
MIRFlickr	4,053	712	38	3	100 / 512 / 100

we further analyze the performance variation with different selected feature ratios, where the selected feature ratio is varied from 1% to 20%. The experiment first reserves a fixed 20% test set. The remaining 80% is further split such that 25% of it is used for validation, yielding an overall 60% Train + 20% Validation + 20% Test split. In every validation or test evaluation, the framework fits the `MinMaxScaler` only on the fitting partition and subsequently transforms the evaluation partition using the same scaler; therefore, evaluation features do not participate in estimating the scaling parameters.

VCRS internally splits the data again by view and performs per-view min–max scaling. Because the selector receives only the current fitting partition, this internal normalization likewise does not access validation or test data.

All methods select the corresponding number of features according to their feature rankings and use the same downstream classifier. A `LinearSVC` is trained separately for each label.

The detailed information of the datasets is shown in Table 4.

E.2 Baseline Feature Ranking Adaptation

Some comparison methods do not directly output feature rankings; therefore, we design a unified conversion strategy.

NAIM3L learns a prediction matrix W. We use $s_i = \|W_{i,:}\|_2$ as the feature importance and obtain the feature ranking by sorting s_i in descending order.

TM3L learns $Q^{(v)}$ and the latent-to-label mapping U. We define

$$s_i^{(v)} = \left\| \left(Q^{(v)} U \right)_{i,:} \right\|_2 . \tag{51}$$

This score jointly considers the original features, latent representations, and label semantics, thereby converting them into a unified original-feature ranking.

E.3 Dataset View Dimensions and Concept Construction Settings

The feature dimensions of different views and the total feature number are provided in Table 4. During preprocessing, each view is independently normalized and preserved according to its original dimensionality.

For relational concept construction, VCRS applies K-means clustering to L2-normalized label response vectors. The number of relational concept candidates is automatically determined as

$$G_r = \max(1, \min(\lceil \sqrt{L} \rceil, \lfloor L/2 \rfloor)), \tag{52}$$

where L denotes the number of labels. K-means uses k-means++ initialization with $n_{\text{init}} = 10$ and the same random seed as the corresponding experimental run to ensure reproducibility. After clustering, singleton clusters containing only one label are removed, since they cannot introduce additional relational semantic information beyond atomic concepts. Only clusters with at least two labels are retained to construct relational concepts.