Redundancy-optimized Multi-head Attention Networks for Multi-View Multi-Label Feature Selection

Yuzhou Liu[1, 2], **Jiarui Liu**[1, 2], **Wanfu Gao**[1, 2*]

[1] College of Computer Science and Technology, Jilin University, China
[2] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, China
liuyuzhou@jlu.edu.cn, liujr24@mails.jlu.edu.cn, gaowf@jlu.edu.cn

Abstract

Multi-view multi-label data offers richer perspectives for artificial intelligence, but simultaneously presents significant challenges for feature selection due to the inherent complexity of interrelations among features, views and labels. Attention mechanisms provide an effective way for analyzing these intricate relationships. They can compute importance weights for information by aggregating correlations between Query and Key matrices to focus on pertinent values. However, existing attention-based feature selection methods predominantly focus on intra-view relationships, neglecting the complementarity of inter-view features and the critical feature-label correlations. Moreover, they often fail to account for feature redundancy, potentially leading to suboptimal feature subsets. To overcome these limitations, we propose a novel method based on **R**edundancy-optimized **M**ulti-head **A**ttention **N**etworks for **M**ulti-view **M**ulti-label **F**eature **S**election (RMAN-MMFS). Specifically, we employ each individual attention head to model intra-view feature relationships and use the cross-attention mechanisms between different heads to capture inter-view feature complementarity. Furthermore, we design static and dynamic feature redundancy terms: the static term mitigates redundancy within each view, while the dynamic term explicitly models redundancy between unselected and selected features across the entire selection process, thereby promoting feature compactness. Comprehensive evaluations on six real-world datasets, compared against six multi-view multi-label feature selection methods, demonstrate the superior performance of the proposed method.

Introduction

The increasing prevalence of multi-modal data acquisition has led to the widespread use of multi-view multi-label data, where samples are described by multiple feature views and associated with multiple labels simultaneously (Han et al. 2025b,a; Wang et al. 2025). For example, an image can be represented by various visual descriptors such as HOG, color histograms and SIFT features, while being annotated with multiple tags like "sky", "river" and "desert" (Deng et al. 2025; Wang, Zhang, and Zhou 2025). Feature selection for such data aims to identify a discriminative and non-redundant feature subset by leveraging complementary information across views while balancing the relevance between features and multiple labels. This process is crucial to improving the performance of the model (Xie et al. 2024). Consequently, multi-view multi-label learning has gained significant traction in addressing complex real-world classification problems within domains such as machine learning and computer vision (Hao et al. 2024b; Xu et al. 2025).

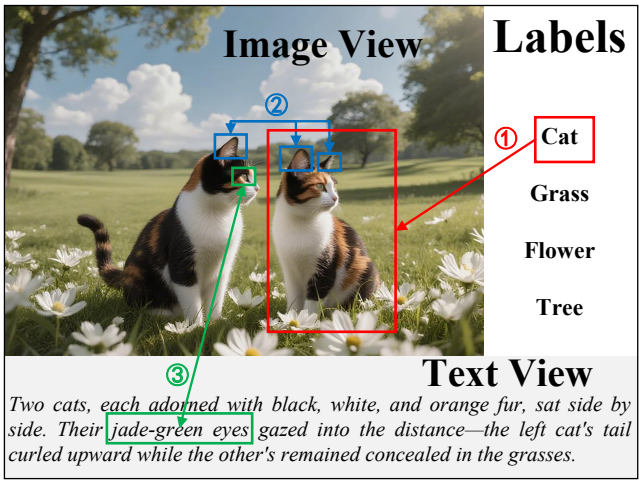

Figure 1: Example of relationships in MVML data. ① Feature-label correlations determine which features require attention. ② Inter-view relations may indicate redundant features (all of them are "*ear*"). ③ Cross-view relations could show complementary features (the *eyes* are "*jade-green*").

Multi-view multi-label data typically exhibits high dimensionality with redundant and noisy features (Han, Hu, and Gao 2024; Dong et al. 2025). More critically, the richness of its representation introduces intricate relationships. Figure 1 shows an example: (1) Correlations between features and labels influence selection results. For instance, the label "cat" would select features from the red box in the image view; (2) Features within the same view may exhibit higher redundancy due to shared origins, like features "ear" appearing many times in one image; (3) Features across different views can provide complementary information, for example, the feature "jade-green" in text view supplements the fea-

*Corresponding author
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ture "eye" in image view. Effectively capturing these multifaceted relationships is essential to identify the most representative and discriminative features, yet it significantly complicates data storage, analysis, and application (Hao, Liu, and Gao 2024a).

Deep learning-based feature selection effectively mitigates the "curse of dimensionality" (Zhao, Hu, and Wang 2015). Among these, attention mechanisms enable models to dynamically focus on the most informative components while suppressing irrelevant details, aligning naturally with feature selection goals by assigning importance weights to features, facilitating the identification of relevant features and the suppression of irrelevant ones (Gui, Ge, and Hu 2019). However, current attention-based approaches for multi-view feature selection focus mainly on feature-label relationships within individual views. They largely overlook potential interactions between views and fail to adequately address feature redundancy, both intra-view and inter-view. This oversight result in highly redundant selected features, thus reducing the precision and efficiency of the feature selection process (Li, Xue, and Du 2023a; Shao et al. 2022a).

To address these critical gaps, we propose RMAN-MMFS (Redundancy-optimized Multi-head Attention Networks for Multi-view Multi-label Feature Selection). For the features-label relationships within views, our framework leverages multi-head attention, where each head specializes in analyzing feature-label relationships within a single view. To capture the complementarity of features between views, we incorporate cross-attention mechanisms between different heads to explicitly model complementary relationships among features across views. Furthermore, recognizing the detrimental impact of redundancy, we introduce dedicated static and dynamic redundancy optimization terms. The static term minimizes redundancy among features within each view, while the dynamic term actively reduces redundancy between unselected and already selected features during the iterative selection process. This combined strategy significantly enhances the quality and compactness of the selected feature subset.

Extensive experiments conducted on six real-world datasets, comparing RMAN-MMFS against six representative multi-view multi-label feature selection baselines, validate its effectiveness and superiority.

In summary, the main contributions of this paper can be summarized as follows:

- A novel attention fusion framework: We integrate multi-head attention for intra-view analysis and cross-attention for inter-view complementarity modeling within a unified feature selection architecture.
- Comprehensive redundancy modeling: We design both static and dynamic redundancy optimization terms to effectively capture and minimize redundancy within views and between selected/unselected features throughout the selection process.
- Empirical validation: Rigorous experimentation on diverse real-world datasets demonstrates the superior performance of RMAN-MMFS over six existing state-of-the-art methods.

Related Work

Multi-View Multi-Label Research

Based on whether model training is involved, existing multi-view feature selection strategies can be broadly categorized into two types: (1) Non-training strategies (e.g. filter methods); (2) Training-integrated strategies (e.g. wrapper or embedded methods). The latter suffers from high time complexity, overfitting risks, and limited scalability to high-dimensional data, often converging to local optima. In contrast, non-training methods are the most widely used feature selection due to their low computational cost and robustness against overfitting, making them preferable for high-dimensional data. Method (Vinh et al. 2012) normalizes mutual information in the mRMR criterion to mitigate the dominance of correlations or redundancy. Information-theoretic method like (Qian et al. 2020) integrates random variable distributions with granular computing, developing a algorithm based on mutual information and label enhancement. Methods (Dai et al. 2024; Ma et al. 2023) leverage conditional mutual information to dynamically evaluate the relevance between selected features and candidates. They further propose novel label redundancy terms to better assess its effect on candidate feature relevance.

However, existing methods capture global view significance but neglect local importance, like inter-view complementarity or feature-specific relevance (Hao et al. 2024a; Lin et al. 2022). Additionally, high dimensionality of multi-view data introduces optimization challenges due to increased parameters. To address these, attention-based feature selection has emerged.

Attention-Based Multi-View Feature Selection

Attention mechanisms have been used in common feature selection (Wang et al. 2022). For example, (Pham et al. 2024) proposed a multi-head attention feature selector with KQV correspond feature, label and a transformed feature matrix respectively. (Wang et al. 2014) employed an attention module under cognitive bias constraints. However, these methods are limited to single-view data and cannot address the complexities of multi-view multi-label scenarios.

Existing attention-based multi-view feature selection methods primarily adopt non-training strategies. Based on Formula (1), methods (Li, Xue, and Du 2023b) utilize distinct attention networks to learn important features while suppressing noise. Method (Hennequin et al. 2022) employs multi-head attention networks to evaluate feature importance within domain-adaptive views. The MVFC method (Shao et al. 2022b) denoises multi-view features via subspace learning. Cross-view strategies capture richer correlations and complementarity. For example, method (Shen et al. 2020) (MML-DAN) uses dual-attention networks to model deep interactions between label-specific views while balancing view-label relevance.

While these algorithms may achieve reasonable performance, they inadequately model complex data relationships. Furthermore, cross-view attention remain relatively scarce, and no existing method simultaneously considers intra-view feature relevance, inter-view feature complementarity and

cross-feature redundancy. To address these limitations, we propose a cross-view attention-based approach that reduces information redundancy and ensures more comprehensive modeling of feature relationships.

The Proposed Method

Definitions and Overall Method

Mathematically, the attention mechanism can be defined as follows: for an input sequence $X = (x_1, x_2, ..., x_n)$, the attention mechanism calculates a set of attention weights $W = (\omega_1, \omega_2, ..., \omega_n)$, where each ω_i denotes the importance of the corresponding input element x_i. These weights are computed by a compatibility function F between the query matrix Q and the input X , such as the scaled dot product. The unnormalized scores are normalized via the softmax function, as formalized in:

$$W = \text{softmax}\left(\frac{Q^T K}{\sqrt{d_k}}\right). \tag{1}$$

Here Q is the query matrix, K is the key vector for X, and d_k denotes the dimensionality of K for scaling stability. The resulting weights W are then used to calculate the weighted sum of the input values, yielding the attention score $Y = WV$, where V is the value vector for X. In essence:

$$Attention(K, Q, V) = \text{softmax}\left(\frac{Q^T K}{\sqrt{d_k}}\right) V. \tag{2}$$

Applying the attention mechanism is to feature selection lacks explicit K, Q and V matrices, requiring designed representations to capture the feature-label relationship for effective use.

Consider a multi-view dataset with H views, denoted as $\{X^{(v)}\}_{v=1}^{H}$, where $X^{(v)} \in \mathcal{R}^{n \times d_v}$ is the feature matrix of the v-th view, n is the sample count, and d_v is its feature dimensionality. The multi-label matrix is $Y \in \mathcal{R}^{n \times c}$, with c being the total label count.

Then, we map this to the attention framework: $K^{(v)}$ is the standardized feature matrix $X_{norm}^{(v)}$ for the v-th view, where $X_{norm}^{(v)} = \frac{X^{(v)} - \mu^{(v)}}{\sigma^{(v)}}$, $\mu^{(v)}$ and $\sigma^{(v)}$ are the mean and standard deviation vector of $X^{(v)}$ respectively. Q is the label matrix Y, and V is X_{norm}. Based on such matching, we design our approach's framework as shown in Figure 2.

Overall, RMAN-MMFS comprises two parts: multi-head attention and redundancy optimization. In multi-head attention, each head computes: (1) View-self attention for feature-label relations within one view; (2) Cross attention for relations between the current view and others. Redundancy optimization also has two types: static redundancy considers relations among features within one view, while dynamic redundancy considers relations between candidate features across views. Details follow in this Section.

Multi-Head Attention for Multi-View Data

We apply multi-head attention to multi-view data by assigning one head per view. Each head independently learns view-specific feature-label and view-view interaction patterns.

View-Self Attention. For each view, we calculate its view-self attention, using the label matrix $Querry$, where Key and $Value$ are obtained from the features $X^{(v)} \in \mathcal{R}^{n \times d_v}$. The resulting attention values become the initial importance scores $scores_{intra}^{(v)}$ for features in this view:

$$scores_{intra}^{(v)} = \text{softmax}\left(\frac{Q^T K^{(v)}}{\sqrt{d_k}}\right) V^{(v)}. \tag{3}$$

Each attention head corresponds to one view to avoid overfitting from direct raw data computation, yielding attention weights per view. Learning distinct attention patterns in separate subspaces enables model to capture richer information through multi-perspective feature modeling and independent parameterization for heterogeneous data distributions.

However, the core value of multi-view data is inter-view complementarity: different views provide relatively independent perspectives that characterize samples comprehensively, improving feature selection thoroughness. Lacking cross-view interaction leads to incomplete feature importance assessment. Thus, we introduce cross-view attention.

Cross-Views Attention. Cross-attention models relationships between distinct sequences or modalities. Unlike self-attention, it enables one sequence to attend to another, proving valuable for integrating information from different sources, like machine translation and image captioning (Liu et al. 2024).

To enable mutual attention across views: For the view $X^{(v)}$, features from all other views are concatenated into the context key matrix $K_{context}^{(v)} = Concat\left(K^{(1)}, \ldots, K^{(v-1)}, K^{(v+1)}, \ldots, K^{(H)}\right)$, where $K_{context}^{(v)} \in \mathcal{R}^{n \times D_{context}}$, $D_{context} = \sum_{u \neq v} d_u$; The label-driven query matrix Q interacts with K to compute cross-view feature relevance:

$$\alpha_{cross}^{(v)} = \text{softmax}\left(\frac{Q^T K_{context}^{(v)}}{\sqrt{d_k}}\right) V^{(v)}. \tag{4}$$

Mapping cross-view attention to the current view feature space:

$$scores_{cross}^{(v)} = \alpha_{cross}^{(v)} \cdot K_{context}^{(v)} \cdot X_{norm}^{(v)}. \tag{5}$$

This dynamically quantifies inter-view complementarity through weight allocation. High weights indicate strong relevance between the current view and context features, facilitating synergistic information integration. This supplements view-self attention weights with cross-view feature complementarity, providing a robust global representation for redundancy penalization and feature selection.

Redundancy Optimization

As mentioned, solely the attention mechanism inadequately addresses feature redundancy. Thus, we introduce redundancy optimization terms. Typically, multi-view data redundancy manifests in two forms: (1) Intra-view redundancy: High correlation among features within the same view. (2) Global redundancy: Information overlap between

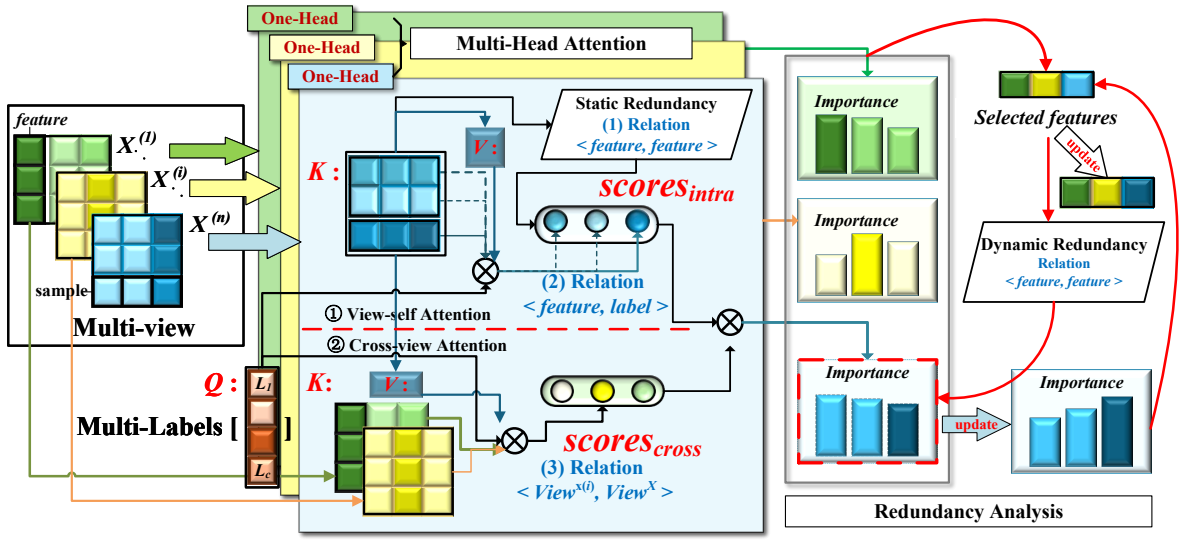

Figure 2: An illustration of the proposed RMAN-MMFS is presented, showcasing: (1) Multi-head attention for multi-view data analysis, which captures feature correlations and complementarity; (2) Static and dynamic redundancy penalty terms that quantify inter-feature redundancy. Feature weights are subsequently computed through this comprehensive integration process.

candidate features and selected features. Redundant features reduce model generalizability and obscure discriminative features. To address this, we introduce static and dynamic redundancy optimizations.

Static Redundancy Optimization. This measures intra-view linear redundancy using the mean absolute correlation coefficient (MACC) between features. For each feature i in current view, it is calculated as:

$$R^{(v)}_{static}(i) = \frac{1}{d_v - 1} \sum_{i \neq j} \left| Corr\left(X^{(v)}_i, X^{(v)}_j\right) \right|. \quad (6)$$

This optimizes highly correlated features within view by statically adjusting the attention weights, enabling rapid intra-view redundancy reduction, and we use it to adjust the initial $scores^{(v)}_{intra}$ to get a more reasonable results.

Dynamic Redundancy Optimization. Maintains a global set S tracking indices of selected features f (view and position). For candidate feature i, its mutual information (MI) with all features in S is computed:

$$R^{(v)}_{dynamic}(i) = \frac{1}{|S|} \Sigma_{i \in C} \sum_{f \in S} MI(i, f). \quad (7)$$

The generated attention weights are dynamically adjustable, capturing nonlinear dependence between cross-view and global features while accurately suppressing global redundancy in candidate and selected features, avoiding invisible repeated selection of redundant features. Based on the dynamic redundancy optimization, the feature selection results are updated dynamically in an iterative way as shown in Figure 2.

The two redundant optimizations work in synergy. The static term rapidly eliminates intra-view redundancy, while the dynamic term refines global feature selection. Combining linear (MACC) and nonlinear (MI) metrics balances efficiency and precision. This design effectively mitigates redundancy in multi-view multi-label tasks, yielding compact, discriminative feature subsets for downstream classifiers.

Objective Function

Integrating multi-head attention and redundancy analysis, the unified objective function is:

$$Importance^{(v)} = \left\| \left[\left(scores^{(v)}_{\text{intra}} - \lambda \cdot R^{(v)}_{\text{static}} \right) + scores^{(v)}_{\text{cross}} \right] - \beta \cdot R^{(v)}_{\text{dynamic}} \right\|_2. \quad (8)$$

Where λ and β are penalty coefficients. A detailed breakdown of this process is elucidated in Algorithm 1.

Experiments

Datasets

We evaluate our method on six public multi-view multi-label datasets: SCENE, VOC07, MIRFlickr, OBJECT, Yeast and Mfeat. The first four are real-world image datasets from (Zhang et al. 2020). Yeast contains gene expression data of yeast under different conditions, with each sample belonging to one of 14 categories. Mfeat includes six feature views of handwritten digits 0-9, with each sample belonging to one of ten categories. Table 1 details each dataset, including the number of views, features, samples and labels .

Baseline Methods

Information-theoretic and sparsity-based methods are mature and competitive feature selection baselines. Both information-theoretic methods and attention-based methods are lightweight and do not require training data. Therefore,

Algorithm 1: RMAN-MMFS

Input: Data matrices $\{X^{(v)}\}_{v=1}^{H}$, Label matrix Y
Parameter: Parameters λ and β
Output: Set of selected features F

Initialize $F \leftarrow \emptyset$ {Selected features set}
$X_{\text{norm}} \leftarrow \frac{X-\mu}{\sigma}$ {Normalize data}
for $v = 1$ **to** H **do**
  Define $Q \leftarrow Y$, $d_k \leftarrow \dim(X^{(v)})$
  Compute cross-view attention via Formula (4)
  Map to feature space via Formula (5)
  Define $K^{(v)} \leftarrow F(X^{(v)})$, $V^{(v)} \leftarrow X^{(v)}$
  Calculate view-self attention via Formula (3)
  for each feature i in $X^{(v)}$ **do**
    Compute $R^{(v)}_{\text{static},i}$ via Formula (6)
  end for
  $scores^{(v)}_{\text{intra}} \leftarrow scores^{(v)}_{\text{intra}} - \lambda R^{(v)}_{\text{static}}$
  $scores^{(v)} \leftarrow scores^{(v)}_{\text{intra}} + scores^{(v)}_{\text{cross}}$
  if $F \neq \emptyset$ **then**
    for each feature i in $X^{(v)}$ **do**
      Compute $R^{(v)}_{\text{dynamic},i}$ via Formula (7)
    end for
  end if
  $scores^{(v)} \leftarrow scores^{(v)} - \beta R^{(v)}_{\text{dynamic}}$
  $Importance^{(v)} \leftarrow \|scores^{(v)}\|_2$ {L2 norm}
  Sort $Importance^{(v)}$ descending $\rightarrow I_{\text{sorted}}$
  Add top features from I_{sorted} to F
end for
return F

we compare RMAN against three information-theoretic and three sparsity-based feature selection methods which are prominent: STFS (Gao et al. 2023), ENM (Gonzalez-Lopez, Ventura, and Cano 2020), MLSMFS (Zhang et al. 2021), MSFS (Zhang et al. 2020), DHLI (Hao, Liu, and Gao 2024b) and EF2FS (Hao, Gao, and Hu 2025).

Evaluation Metrics

We selected four common metrics in this field for evaluation:

- Average Precision (AP): Measures precision across recall levels via the area under the precision-recall curve for each label.
- Macro-Average Area Under Curve (AUC): Averages each label's AUC value, evaluating the model's positive-negative distinction across classification thresholds.
- Coverage Error (CE): Measures the average extra labels needed to cover all true labels.
- Ranking Loss (RL): Reflects the probability of irrelevant labels ranking above relevant ones in the predicted ranking across samples.

Higher values indicate better performance for AP and AUC, while lower values are better for CE and RL. Additionally, AP and AUC are label-based metrics, whereas CE and RL are sample-based, providing multidimensional evaluation (Hancer, Xue, and Zhang 2025; Zhang and Zhou 2013).

Views	SCENE	VOC07	MIRFlickr	OBJECT	Yeast	Mfeat
View1	64	100	100	64	79	76
View2	225	512	512	225	24	216
View3	144	100	100	144	-	64
View4	73	-	-	73	-	240
View5	128	-	-	128	-	47
View6	-	-	-	-	-	6
Features	634	712	712	634	103	649
Samples	4400	3817	4053	6047	2417	2000
Labels	31	20	38	31	14	10

Table 1: Description of datasets.

Experiment Settings

Python was used as the programming language. We selected MLKNN (k=10) as the classifier. Feature selection ranged from 2% to 20% of total features, increasing by 2% per iteration. In each experiment, 30% of samples were for testing and 70% for training. Experiments were repeated 10 times with averaged results. Hyperparameters for other methods were set according to their original papers. Multi-view datasets were concatenated for feature selection in other information-theoretic methods.

Experimental Results and Analysis

Tables 2-3 show method performance on six datasets using four metrics. The "Average" row displays mean results across all datasets. Bold values indicate the best classification performance on a specific dataset, while underlined values denote second-best. Up/down arrows signify higher/lower values indicate better performance.

Our method achieves the highest performance on five datasets (SCENE, VOC07, Yeast, OBJECT and MIRFlickr) due to RMAN's comprehensive relationship modeling. However, RMAN underperforms MLSMFS on Mfeat. Mfeat exhibits significantly higher feature-label correlations than others, causing greater information overlap between features. This overlap complicates redundancy avoidance, weakening optimization adjustment and causing excessive penalization of discriminative features. Nevertheless, considering overall performance across all six datasets, RMAN outperforms others on AP, AUC, CE, and RL. Figure 3 further presents one dataset performance across all evaluation metrics to clearly demonstrate our results.

Ablation Experiments

Ablation Study To further validate the effectiveness of each component, we conduct ablation experiments by removing cross-attention weight calculation and redundancy optimization terms from the objective function. Since view-self attention forms the architectural foundation, it is not ablated. This result in Formula (9) without cross-attention ($RMAN_1$):

$$\left\| \left(scores^{(v)}_{intra} - \lambda \cdot R^{(v)}_{static} \right) - \beta \cdot R^{(v)}_{dynamic} \right\|_2, \quad (9)$$

Formula (10) without static redundancy penalty ($RMAN_2$):

$$\left\| \left[scores^{(v)}_{intra} + scores^{(v)}_{cross} \right] - \beta \cdot R^{(v)}_{dynamic} \right\|_2, \quad (10)$$

Datasets	RMAN	MLSMFS	DHLI	STFS	MSFS	EF2FS	ENM
				AP ↑			
SCENE	**0.260 ± 0.010**	0.231 ± 0.013	0.244 ± 0.015	0.229 ± 0.014	0.233 ± 0.010	0.250 ± 0.006	0.227 ± 0.014
VOC07	**0.136 ± 0.003**	0.115 ± 0.002	0.128 ± 0.002	**0.136 ± 0.007**	0.111 ± 0.001	0.113 ± 0.006	0.131 ± 0.005
Yeast	**0.324 ± 0.006**	0.316 ± 0.011	0.321 ± 0.006	0.314 ± 0.012	0.311 ± 0.003	0.315 ± 0.006	0.323 ± 0.010
MIRFlickr	**0.296 ± 0.009**	0.260 ± 0.001	0.254 ± 0.002	0.288 ± 0.012	0.243 ± 0.001	0.269 ± 0.037	0.293 ± 0.011
Mfeat	0.844 ± 0.100	**0.944 ± 0.060**	0.729 ± 0.075	0.891 ± 0.082	0.658 ± 0.151	0.780 ± 0.032	0.651 ± 0.131
OBJECT	**0.128 ± 0.018**	0.098 ± 0.011	0.117 ± 0.019	0.104 ± 0.011	0.107 ± 0.009	0.125 ± 0.010	0.102 ± 0.013
Average	**0.331**	0.327	0.299	0.327	0.278	0.309	0.288
				AUC ↑			
SCENE	**0.658 ± 0.016**	0.596 ± 0.023	0.622 ± 0.032	0.613 ± 0.037	0.593 ± 0.023	0.607 ± 0.013	0.603 ± 0.033
VOC07	**0.612 ± 0.020**	0.536 ± 0.005	0.575 ± 0.008	0.609 ± 0.021	0.500 ± 0.001	0.505 ± 0.008	0.595 ± 0.020
Yeast	0.541 ± 0.015	**0.542 ± 0.018**	0.536 ± 0.016	0.536 ± 0.020	0.512 ± 0.007	0.526 ± 0.010	0.540 ± 0.021
MIRFlickr	**0.618 ± 0.017**	0.550 ± 0.001	0.546 ± 0.005	0.603 ± 0.015	0.509 ± 0.001	0.564 ± 0.009	0.613 ± 0.020
Mfeat	0.965 ± 0.020	**0.985 ± 0.012**	0.878 ± 0.026	0.978 ± 0.014	0.973 ± 0.057	0.623 ± 0.034	0.909 ± 0.048
OBJECT	**0.653 ± 0.030**	0.610 ± 0.023	0.634 ± 0.033	0.619 ± 0.028	0.595 ± 0.019	0.634 ± 0.011	0.619 ± 0.030
Average	**0.675**	0.637	0.632	0.660	0.612	0.577	0.647

Table 2: Experimental results of all methods in terms of AP and AUC (mean ± std).

Datasets	RMAN	MLSMFS	DHLI	STFS	MSFS	EF2FS	ENM
				CE ↓			
SCENE	**13.877 ± 0.253**	14.786 ± 0.350	14.361 ± 0.468	14.806 ± 0.589	14.967 ± 0.490	13.906 ± 0.230	14.792 ± 0.506
VOC07	**8.332 ± 0.270**	9.445 ± 0.044	8.744 ± 0.106	8.443 ± 0.284	10.489 ± 0.437	10.481 ± 0.246	8.546 ± 0.275
Yeast	**8.786 ± 0.142**	8.798 ± 0.353	8.997 ± 0.274	8.873 ± 0.347	9.298 ± 0.351	9.076 ± 0.172	8.878 ± 0.273
MIRFlickr	**21.575 ± 0.296**	22.834 ± 0.020	23.318 ± 0.279	21.884 ± 0.210	25.701 ± 1.283	22.940 ± 0.365	21.641 ± 0.383
Mfeat	1.418 ± 0.180	**1.221 ± 0.153**	2.210 ± 0.247	1.288 ±0.155	2.443 ± 0.488	4.779 ± 0.291	1.959 ± 0.412
OBJECT	**9.353 ± 0.500**	10.092± 0.354	9.728 ± 0.621	9.803 ± 0.441	10.292 ± 0.219	9.664 ± 0.201	9.564 ± 0.554
Average	**10.572**	11.196	11.226	12.172	12.918	11.808	10.897
				RL ↓			
SCENE	**0.090 ± 0.004**	0.104 ± 0.005	0.097 ± 0.006	0.103 ± 0.008	0.108 ± 0.007	0.091 ± 0.004	0.103 ± 0.007
VOC07	**0.192 ± 0.009**	0.231 ± 0.002	0.207 ± 0.003	0.197 ± 0.010	0.267 ± 0.017	0.266 ± 0.010	0.201 ± 0.009
Yeast	**0.250 ± 0.009**	0.258 ± 0.011	0.257 ± 0.013	0.254 ± 0.015	0.274 ± 0.028	0.261 ± 0.009	0.252 ± 0.011
MIRFlickr	**0.152 ± 0.006**	0.178 ± 0.001	0.178 ± 0.003	0.160 ± 0.006	0.230 ± 0.039	0.181 ± 0.008	0.153 ± 0.007
Mfeat	0.046 ± 0.022	**0.025 ± 0.017**	0.134 ± 0.027	0.032 ± 0.017	0.160 ± 0.054	0.420 ± 0.032	0.107 ± 0.046
OBJECT	**0.175 ± 0.014**	0.199 ± 0.008	0.187 ± 0.017	0.187 ± 0.012	0.204 ± 0.006	0.182 ± 0.005	0.187 ± 0.015
Average	**0.151**	0.166	0.177	0.156	0.207	0.234	0.167

Table 3: Experimental results of all methods in terms of CE and RL (mean ± std).

and Formula (11) without dynamic redundancy penalty ($RMAN_3$):

$$\left\| \left(scores_{intra}^{(v)} - \lambda \cdot R_{static}^{(v)} \right) + scores_{cross}^{(v)} \right\|_2 . \quad (11)$$

These formulations were evaluated across six datasets.

Table 4 shows components in in each variant and AP metrics for ablation methods. VSA, CVA, SRP, and DRP represent view-self attention, cross-view attention, static redundancy term and dynamic redundancy term, respectively. Results show removing any component harms performance, highlighting their importance. Our full method achieves optimal performance on six datasets, proving its effectiveness.

Redundancy Optimization Design Feature selection aims to identify the most relevant and least redundant feature subset. Redundancy penalties reduce inter-feature redundancy. Our algorithm uses : (1) correlation coefficients (2) mutual information. To validate this design, we test different combinations for static/dynamic penalties. Table 5 shows CE metrics under different combinations. Results confirm our current configuration achieves optimal performance.

Parameter Analysis

Figure 4 shows sensitivity analysis for two parameters on MIRFlickr. Optimize the grid by fixing other parameters. The X-axis represents grid search range (0.001 to 1000), Y-axis shows number of selected features, and Z-axis indicates AUC metric values. For parameter α and β, as the number of features increases, the indicator value rises slightly and tends to stabilize. Performance remains stable across feature subsets on MIRFlickr.

	VSA	CRA	SRP	DRP	SCENE	VOC07	Yeast	MIRFlickr	Mfeat	OBJECT
RMAN	√	√	√	√	**0.260**	**0.136**	**0.324**	**0.296**	**0.844**	**0.128**
RMAN_1	√		√	√	0.257	0.133	0.314	0.294	0.830	0.124
RMAN_2	√	√		√	0.259	0.134	0.314	0.295	0.842	0.119
RMAN_3	√	√	√		0.256	0.133	0.316	0.294	0.825	0.123

Table 4: Ablation experimental results of RMAN on six datasets.

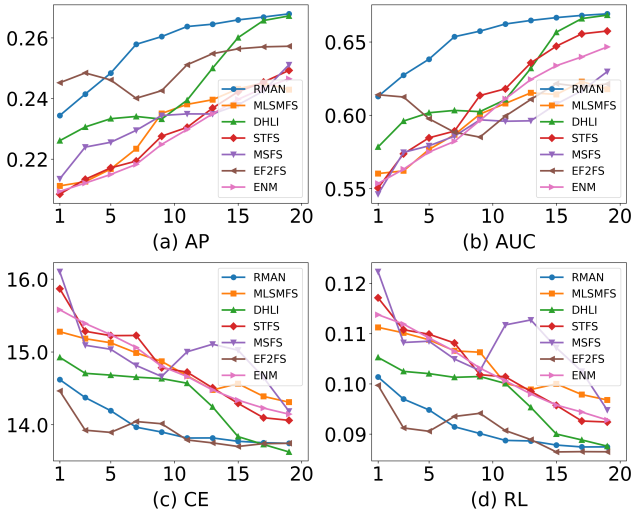

Figure 3: Seven methods on SCENE in terms of AP, AUC, CE and RL.

	SRP	DRP	SCENE	VOC07	Yeast
RMAN	Corr	MI	**13.877**	**8.332**	**8.786**
RMAN-α	MI	Corr	13.996	8.424	8.977
RMAN-β	MI	MI	13.934	8.428	8.980
RMAN-γ	Corr	Corr	13.956	8.441	9.018

Table 5: The impact of different redundant computations on the CE metrics on three datasets.

Time Complexity of Methods

We conducted complexity analysis of RMAN and performed comparative assessments against three training-data-free information-theoretic baselines. Complexity is determined by sample size (n), feature size (d), label count (q) and selected feature count (k). For RMAN, multi-head attention and cross-attention both have $O(nqd)$ complexity. Static and dynamic redundancy penalties have $O(d^2)$ and $O(ndk)$ complexity respectively. Thus RMAN's total complexity is $O(nqd + d^2 + ndk)$. Similarly, STFS, ENM and MLSMSF have complexities $O(ndq^2 + kndq + ndq)$, $O(ndq)$ and $O(ndq^2 + ndk + ndq)$ respectively. RMAN and ENM are cubic complexity, while the other two involve quartic terms, indicating our method's computational advantage. Actual runtimes on six datasets are shown in Table 6. Compared to ENM, RMAN exhibits marginally longer runtime, but RMAN delivers superior overall performance, demonstrating its ability to maintain optimal performance with lightweight computation.

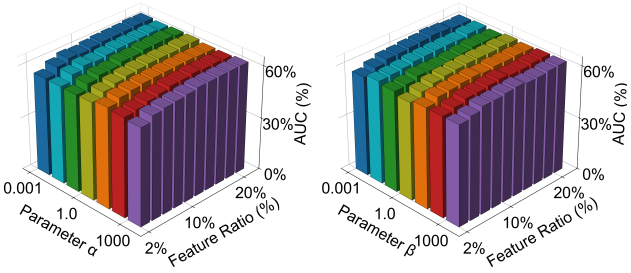

Figure 4: Parameter sensitivity studies on the SCENE datasets.

Methods	SCENE	VOC07	Yeast	MIRFlickr	Mfeat	OBJECT
ENM	558	345	123	616	137	659
MLSMFS	26911	13015	562	36103	1622	4532
STFS	215944	87146	742	207207	18470	301418
RMAN	2502	2150	200	2358	2047	2868

Table 6: The running time (in seconds).

Conclusion

This paper proposes a novel redundancy-optimized multi-head attention networks for multi-view multi-label feature selection, named RMAN-MMFS. Previous attention-based feature selection methods fail to adequately consider inter-view feature complementarity and overlook the guiding role of feature redundancy. To address these issues, RMAN-MMFS introduces multi-head attention to compute view-self feature-label relevance, employs cross-attention to handle inter-view feature complementarity, and designs both static and dynamic redundancy penalty terms. Experimental results demonstrate that our method outperforms existing state-of-the-art approaches across multiple aspects. Our research improves feature selection accuracy by capturing relationships among features, views, and labels as comprehensively as possible. Future work will focus on exploring relationships between labels to further enhance the method's performance.

Acknowledgments

This work is funded by: Jilin Provincial Science and Technology Development Plan Project No.20240302084GX, and Changchun Scienceand Technology Bureau Project 23YQ05.

References

Dai, J.; Huang, W.; Zhang, C.; and Liu, J. 2024. Multi-label feature selection by strongly relevant label gain and label mutual aid. *Pattern Recognition*, 145: 109945.

Deng, D.; Xu, J.; Deng, Z.; Wan, J.; Xia, D.; Cao, Z.; and Li, T. 2025. Feature selection based on fuzzy joint entropy and feature interaction for label distribution learning. *Information Processing & Management*, 62(6): 104234.

Dong, Z.; Liu, M.; Wang, S.; Liang, K.; Zhang, Y.; Liu, S.; Jin, J.; Liu, X.; and Zhu, E. 2025. Enhanced then Progressive Fusion with View Graph for Multi-View Clustering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 15518–15527.

Gao, W.; Hao, P.; Wu, Y.; and Zhang, P. 2023. A unified low-order information-theoretic feature selection framework for multi-label learning. *Pattern Recognition*, 134: 109111.

Gonzalez-Lopez, J.; Ventura, S.; and Cano, A. 2020. Distributed multi-label feature selection using individual mutual information measures. *Knowledge-Based Systems*, 188: 105052.

Gui, N.; Ge, D.; and Hu, Z. 2019. AFS: An attention-based mechanism for supervised feature selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 3705–3713.

Han, Q.; Hu, L.; and Gao, W. 2024. Feature relevance and redundancy coefficients for multi-view multi-label feature selection. *Information Sciences*, 652: 119747.

Han, Q.; Shi, R.; Hu, L.; and Gao, W. 2025a. Multi-label feature selection based on positive sample information weighting. *Knowledge-Based Systems*, 114373.

Han, Q.; Zhao, Z.; Hu, L.; and Gao, W. 2025b. Enhanced multi-label feature selection considering label-specific relevant information. *Expert Systems with Applications*, 264: 125819.

Hancer, E.; Xue, B.; and Zhang, M. 2025. A Survey on Evolutionary Feature Selection in Multi-Label Classification. *IEEE Transactions on Evolutionary Computation*.

Hao, P.; Ding, W.; Gao, W.; and He, J. 2024a. Exploring view-specific label relationships for multi-view multi-label feature selection. *Information Sciences*, 681: 121215.

Hao, P.; Gao, W.; and Hu, L. 2025. Embedded feature fusion for multi-view multi-label feature selection. *Pattern Recognition*, 157: 110888.

Hao, P.; Liu, K.; and Gao, W. 2024a. Anchor-guided global view reconstruction for multi-view multi-label feature selection. *Information Sciences*, 679: 121124.

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

Hao, P.; Zhang, P.; Feng, Q.; and Gao, W. 2024b. Label generation with consistency on the graph for multi-label feature selection. *Information Sciences*, 677: 120890.

Hennequin, M.; Benabdeslem, K.; Elghazel, H.; Ranvier, T.; and Michoux, E. 2022. Multi-view self-attention for regression domain adaptation with feature selection. In *International Conference on Neural Information Processing*, 177–188. Springer.

Li, L.; Xue, Z.; and Du, X. 2023a. ASCRB: Multi-view based attentional feature selection for CircRNA-binding site prediction. *Computers in Biology and Medicine*, 162: 107077.

Li, L.; Xue, Z.; and Du, X. 2023b. ASCRB: Multi-view based attentional feature selection for CircRNA-binding site prediction. *Computers in Biology and Medicine*, 162: 107077.

Lin, Y.; Liu, H.; Zhao, H.; Hu, Q.; Zhu, X.; and Wu, X. 2022. Hierarchical feature selection based on label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(6): 5964–5976.

Liu, S.; Zhang, Y.; Li, W.; Lin, Z.; and Jia, J. 2024. Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8599–8608.

Ma, X.-A.; Jiang, W.; Ling, Y.; and Yang, B. 2023. Multi-label feature selection via maximum dynamic correlation change and minimum label redundancy. *Artificial Intelligence Review*, 56(Suppl 3): 3099–3142.

Pham, H.; Tan, Y.; Singh, T.; Pavlopoulos, V.; and Patnayakuni, R. 2024. A multi-head attention-like feature selection approach for tabular data. *Knowledge-Based Systems*, 301: 112250.

Qian, W.; Huang, J.; Wang, Y.; and Shu, W. 2020. Mutual information-based label distribution feature selection for multi-label learning. *Knowledge-Based Systems*, 195: 105684.

Shao, S.; Xing, L.; Wang, Y.; Liu, B.; Liu, W.; and Zhou, Y. 2022a. Attention-based multi-view feature collaboration for decoupled few-shot learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(5): 2357–2369.

Shao, S.; Xing, L.; Wang, Y.; Liu, B.; Liu, W.; and Zhou, Y. 2022b. Attention-based multi-view feature collaboration for decoupled few-shot learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(5): 2357–2369.

Shen, J.; Zhang, Y.; Yu, C.; and Wang, C. 2020. Multi-view multi-label learning with dual-attention networks for stroke screen. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1124–1128. IEEE.

Vinh, L. T.; Lee, S.; Park, Y.-T.; and d'Auriol, B. J. 2012. A novel feature selection method based on normalized mutual information. *Applied Intelligence*, 37: 100–120.

Wang, Q.; Tao, Z.; Gao, Q.; and Jiao, L. 2022. Multi-view subspace clustering via structured multi-pathway network. *IEEE Transactions on Neural Networks and Learning Systems*, 35(5): 7244–7250.

Wang, Q.; Zhang, J.; Song, S.; and Zhang, Z. 2014. Attentional neural network: Feature selection using cognitive feedback. *Advances in neural information processing systems*, 27.

Wang, X.; Zhang, Y.; Zhang, J.; and Zhou, Y. 2025. Incomplete Multiview Clustering using Discriminative Feature Recovery and Tensorized Matrix Factorization. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wang, X.; Zhang, Y.; and Zhou, Y. 2025. Bidirectional Probabilistic Multi-graph Learning and Decomposition for Multi-view Clustering. *IEEE Transactions on Image Processing*.

Xie, J.; Wang, M.; Grant, P. W.; and Pedrycz, W. 2024. Feature selection with discernibility and independence criteria. *IEEE Transactions on Knowledge and Data Engineering*.

Xu, X.; Wei, F.; Yu, T.; Lu, J.; Liu, A.; Zhuo, L.; Nie, F.; and Wu, X. 2025. Embedded multi-label feature selection via orthogonal regression. *Pattern Recognition*, 163: 111477.

Zhang, M.-L.; and Zhou, Z.-H. 2013. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8): 1819–1837.

Zhang, P.; Liu, G.; Gao, W.; and Song, J. 2021. Multi-label feature selection considering label supplementation. *Pattern recognition*, 120: 108137.

Zhang, Y.; Wu, J.; Cai, Z.; and Yu, P. S. 2020. Multi-view multi-label learning with sparse feature selection for image annotation. *IEEE Transactions on Multimedia*, 22(11): 2844–2857.

Zhao, L.; Hu, Q.; and Wang, W. 2015. Heterogeneous feature selection with multi-modal deep neural networks and sparse group lasso. *IEEE Transactions on Multimedia*, 17(11): 1936–1948.