Neighbor-aware Contrastive Disambiguation for Cross-Modal Hashing with Redundant Annotations

Chao Su [1], Likang Peng [1], Yuan Sun [2], Dezhong Peng [1,3], Xi Peng [1,2], Xu Wang [1,4]*
[1]College of Computer Science, Sichuan University, Chengdu, China
[2]National Key Laboratory of Fundamental Algorithms and Models
for Engineering Numerical Simulation, Sichuan University, Chengdu, China
[3]Tianfu Jincheng Laboratory, Chengdu, China
[4]Centre for Frontier AI Research (CFAR), A*STAR, Singapore
suchao.ml@gmail.com, likangpeng@stu.scu.edu.cn, sunyuan_work@163.com,
pengdz@scu.edu.cn, pengx.gm@gmail.com, wangxu.scu@gmail.com

Abstract

Cross-modal hashing aims to efficiently retrieve information across different modalities by mapping data into compact hash codes. However, most existing methods assume access to fully accurate supervision, which rarely holds in real-world scenarios. In fact, annotations are often redundant, i.e., each sample is associated with a set of candidate labels that includes both ground-truth labels and redundant noisy labels. Treating all annotated labels as equally valid introduces two critical issues: (1) the sparse presence of true labels within the label set is not explicitly addressed, leading to overfitting on redundant noisy annotations; (2) redundant noisy labels induce spurious similarities that distort semantic alignment across modalities and degrade the quality of the hash space. To address these challenges, we propose that effective cross-modal hashing requires explicitly identifying and leveraging the true label subset within all candidate annotations. Based on this insight, we present Neighbor-aware Contrastive Disambiguation (NACD), a novel framework designed for robust learning under redundant supervision. NACD consists of two key components. The first, Neighbor-aware Confidence Reconstruction (NACR), refines label confidence by aggregating information from cross-modal neighbors to distinguish true labels from redundant noisy ones. The second, Class-aware Robust Contrastive Hashing (CRCH), constructs reliable positive and negative pairs based on label confidence scores, thereby significantly enhancing robustness against noisy supervision. Moreover, to effectively reduce the quantization error, we incorporate a quantization loss that enforces binary constraints on the learned hash representations. Extensive experiments conducted on three large-scale multimodal benchmarks demonstrate that our method consistently outperforms state-of-the-art approaches, thereby establishing a new standard for cross-modal hashing with redundant annotations. Code is available at https://github.com/Rose-bud/NACD.

1 Introduction

With the explosion of large-scale and diverse data on the Internet [1–7], efficiently retrieving semantically relevant data across modalities has become increasingly important [8–16]. For large-scale datasets, cross-modal hashing (CMH) offers an effective solution by encoding heterogeneous data into compact binary hash codes, enabling high retrieval efficiency and low storage cost. The core

[*] Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

challenge of CMH lies in effectively leveraging available supervision while minimizing semantic discrepancies between different modalities.

Existing CMH methods can be broadly categorized into unsupervised and supervised approaches based on whether the label information is available. The unsupervised CMH methods [17–22] learn hash functions by exploring the intrinsic structure and similarity of the data without access to labels. For instance, CIRH [21] jointly preserves the multimodal correlation and identity semantics into binary hash codes based on a heterogeneous graph network. Moreover, UCCH [22] proposes a contrastive learning-based unsupervised CMH method with a momentum optimizer and cross-modal ranking learning loss to improve performance. However, the lack of supervision limits their ability to learn semantically discriminative representations. In contrast, supervised methods [23–31] leverage label information to learn more discriminative hash codes to improve retrieval performance. Within a probabilistic modality alignment framework, MIAN [27] investigates the preservation of asymmetric similarities both within and between modalities, thereby fully utilizing multi-level semantic information throughout the entire database. RSHNL [31] designs a Robust Self-paced Hashing mechanism that mitigates the misleading effects of noisy labels on the model by simulating the human cognitive process. While these supervised methods have achieved satisfactory retrieval performance by leveraging label information, they implicitly rely on two assumptions: (1) all labels in the training data are accurate; (2) noisy annotations are simulated by replacing correct labels with incorrect ones. However, in real-world applications, data annotations are often redundant, i.e., each instance is labeled with a candidate label set that includes both true and spurious labels. We refer to this setting as redundant annotations. As shown in Fig. 1, among the annotations of the anchor sample pair, only "Sea, Plant, Beach, Cloud" are correct labels, while the others are additional noisy labels. Such redundancy can severely distort semantic similarity estimation and hinder effective hash learning by introducing spurious correlations across modalities. Crucially, existing methods fail to explicitly distinguish true labels within the candidate label set, leading to degraded performance under redundant noisy supervision. Although partial multi-label learning (PML) [32–35] provides a potential solution for redundant annotations, most PML methods assume a shared feature space and overlook cross-modal semantic divergence. In contrast, CMH must address both accurate label disambiguation and modality-robust contrastive pair construction, which remains a rarely explored challenge.

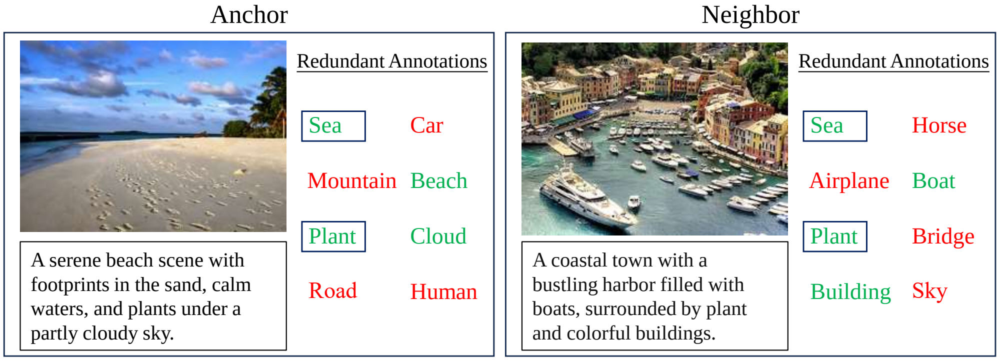

Anchor
Redundant Annotations
Sea
Car
Mountain
Beach
Plant
Cloud
Road
Human
A serene beach scene with footprints in the sand, calm waters, and plants under a partly cloudy sky.
Neighbor
Redundant Annotations
Sea
Horse
Airplane
Boat
Plant
Bridge
Building
Sky
A coastal town with a bustling harbor filled with boats, surrounded by plant and colorful buildings.

Figure 1: The redundant annotations in cross-modal hashing. Among the annotations, the correct labels and redundant noisy labels are represented by green and red, respectively. Shared labels between the anchor sample pair and its neighbor sample pair are enclosed in dashed boxes.

In this paper, we focus on the practical scenario of redundant annotations and propose a novel framework, Neighbor-aware Contrastive Disambiguation (NACD), for robust cross-modal hashing. NACD benefits from an efficient Neighbor-aware Confidence Reconstruction (NACR) module and a novel Class-aware Robust Contrastive Hashing (CRCH) module. Specifically, NACR integrates sample label confidence with its cross-modal neighborhood label confidence to identify the ground-truth labels within the entire annotations. CRCH dynamically constructs positive and negative sample pairs based on class confidence thresholds, reducing erroneous associations caused by misleading labels and significantly enhancing the model's robustness to redundant supervision. Moreover, to effectively reduce the quantization error, we employ an effective quantization loss that enforces binary constraints on the learned hash representations. The main contributions of the proposed NACD are as follows:

- We first focus on the practical scenario of redundant annotations in cross-modal hashing and propose a novel Neighbor-aware Contrastive Disambiguation (NACD) framework that addresses the challenges caused by redundant noisy supervision.
- To achieve accurate label disambiguation, an efficient Neighbor-aware Confidence Reconstruction (NACR) module is presented, which integrates the label confidence of the anchor sample pair with that aggregated from its cross-modal neighbors to identify the ground-truth labels within the entire annotations.
- We design an innovative Class-aware Robust Contrastive Hashing (CRCH) module which dynamically constructs positive and negative sample pairs based on class-wise confidence thresholds, reducing erroneous associations caused by redundant noisy labels and significantly enhancing the model's robustness to incorrect supervision.
- Comprehensive experiments on three multimodal datasets, i.e., MIRFlickr-25k, NUS-WIDE, and MS-COCO, demonstrate that NACD consistently outperforms state-of-the-art CMH methods across various redundancy levels.

2 Related Work

2.1 Cross-Modal Hashing

Cross-modal hashing aims to retrieve semantically relevant data across different modalities within a shared Hamming space. The key challenge lies in bridging the modality gap. To address this challenge, numerous approaches have been proposed, which can be broadly divided into two categories: unsupervised CMH methods and supervised CMH methods. More specifically, unsupervised CMH methods [19–22] learn modality-specific transformations by maximizing cross-modal correlations without label supervision. For example, UCCH [22] integrates contrastive learning into unsupervised CMH to enhance retrieval performance and robustness. However, these unsupervised methods suffer from limited performance due to the absence of explicit supervision. Supervised CMH methods [29–31, 36] are typically based on two assumptions: (1) all labels in the training data are accurate; (2) noisy annotations are simulated by replacing correct labels with incorrect ones. For example, HCCH [36] introduces a coarse-to-fine hierarchical hashing strategy to effectively utilize hierarchical features and accurate labels across modalities. RSHNL [31] proposes a robust self-paced hashing mechanism that emulates human cognition, thereby reducing the negative impact of noisy labels and improving model performance.

However, in real-world applications, multimodal annotations often contain "redundant annotations". Therefore, this paper focuses on a largely unexplored yet challenging problem: cross-modal hashing with redundant annotations.

2.2 Learning with Redundant Annotations

Partial multi-label learning (PML) trains models using redundantly annotated data, where each instance is associated with a candidate label set containing both true and redundant noisy labels. The key challenge in PML lies in filtering out incorrect labels and identifying reliable ones, thereby recovering the true label distribution for supervision. To achieve this, a number of methods have been developed. These methods can be broadly categorized into smoothness assumption-based, low-rank constraint-based, and sparsity regularization-based approaches. Smoothness assumption-based approaches [35, 37, 38] are based on the assumption that neighboring samples in the feature space are more likely to have similar labels. The low-rank constraint-based approaches [39–41] leverage the low-rank property to achieve disambiguation. The sparsity regularization-based approaches [32, 33, 40] impose sparsity on the candidate label set, effectively suppressing noisy labels and facilitating disambiguation.

In contrast to the aforementioned PML methods, our approach integrates the label confidence of an anchor sample pair with that aggregated from its cross-modal neighbors to identify true labels within all annotations. Furthermore, a dynamically updated class-wise threshold enables more accurate and adaptive label disambiguation.

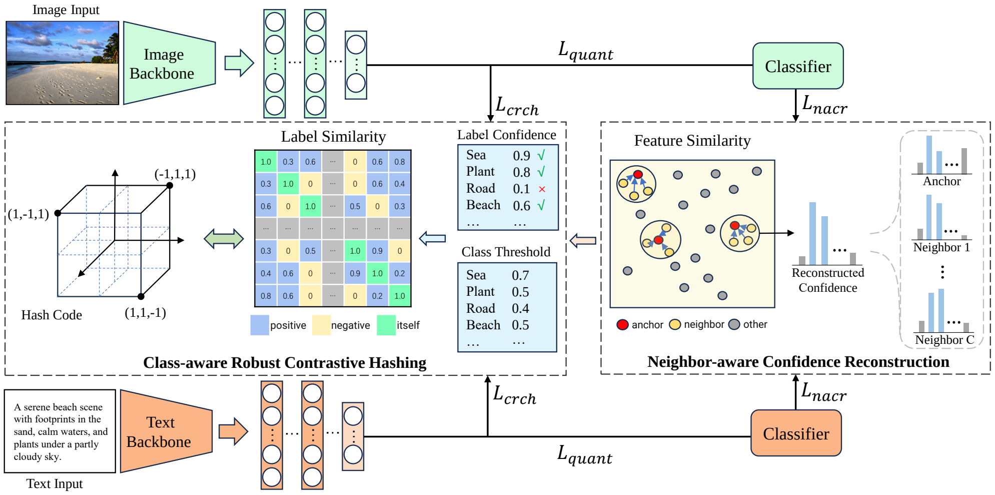


Figure 2: The pipeline of the proposed framework NACD for cross-modal hashing with redundant annotations. NACR refines label confidence by aggregating information from cross-modal neighbors to distinguish true labels from redundant noisy ones. Meanwhile, CRCH constructs reliable positive and negative pairs based on the learned label confidence, which significantly improves robustness against noisy supervision.

3 Proposed Approach

3.1 Problem Definition

For ease of presentation, we first give some definitions of cross-modal hashing of redundant annotations. Suppose the input space is denoted as $\mathcal{X}$, and the label space as $\mathcal{Y} = \{1, 2, ..., K\}$, where K indicates the total number of classes. Denote $\left\{\{x_j^m\}_{m=1}^2, y_j\right\}_{j=1}^N$ as the training set with N sample pairs, where x_j^m represents the j-th instance from the m-th modality ($m = 1$ for image and $m = 2$ for text). $y_j \in \{0, 1\}^K$ denotes the candidate label vector of the j-th sample pair, which encodes both ground-truth and redundant noisy labels. The k-th element of y_j equals 1 if the corresponding sample pair is annotated as class k. Here, p_j is the label confidence vector of the candidate label vector y_j. The hash codes are denoted as $\{\boldsymbol{b}_j^m\}_{j=1}^N \in \{-1, 1\}^L$, where L is the hash code length. CMH leverages hash functions to map data from different modalities into a common Hamming space, enabling efficient similarity search across modalities through compact binary codes. Let the hash functions be f^m, where $m \in \{1, 2\}$. Due to the NP-hard problem in binary optimization, we calculate the hash representations by $h_j^m = tanh(f_m(x_j^m)), m \in \{1, 2\}$ in the training process. Thus, the final binary hash codes are obtained by applying the $sign$ function: $b_j^m = sign(h_j^m), m \in \{1, 2\}$. Additionally, a linear classifier with a sigmoid activation function $g(\cdot)$ is employed to obtain the probability distribution $z_j^m = g(h_j^m)$, where $m \in \{1, 2\}$.

3.2 Neighbor-aware Confidence Reconstruction

Under redundant supervision, confidence estimates obtained from individual samples may be unreliable. Meanwhile, semantically similar samples across modalities often share the same true labels. Motivated by this observation, we present a confidence-mixture (CM) strategy to reconstruct label confidence by balancing neighborhood consensus and self-prediction.

First, assume that the prediction probability of the j-th target sample $\{x_j^m\}_{m=1}^2$ is z_j, under the supervision of redundant annotations y_j. We define the confidence of each sample's prediction by the model as:

$$p_j \leftarrow \gamma p_j + (1 - \gamma)\frac{1}{2}\sum_{m=1}^{2} z_j^m \circ y_j, \tag{1}$$

where $\circ$ denotes Hadamard product, and γ is a momentum parameter that decays from 0.95 to 0.8 during training.

Since model predictions may be unreliable under redundant supervision, and similar samples tend to share the same true labels, we exploit the predictions of cross-modal neighbors to refine the confidence estimation. Given the C nearest neighbors $\mathcal{N}_j$ of the anchor sample pair x_j, the neighbor-aggregated confidence q_j is calculated as:

$$q_j = \frac{\sum_{c \in N_j} s_{jc} p_c}{\sum_{c' \in N_j} s_{jc'}}, \qquad \text{where } s_{jc} = \tfrac{1}{2}\left(s_{jc}^1 + s_{jc}^2\right), \tag{2}$$

where s_{jc} denotes the fused similarity between the anchor j and its neighbor c, obtained by averaging the similarities from the image and text modalities. To mitigate the issue that inaccurate estimates during training may hinder model optimization, we reconstruct the anchor confidence p_j by mixing it with the neighbor-aggregated confidence q_j. The reconstructed confidence is computed as:

$$p_j \leftarrow \lambda q_j + (1 - \lambda) p_j, \tag{3}$$

where λ is a mixture coefficient that balances neighborhood consensus and self-prediction. After obtaining reconstructed confidence by Eq. (3), the disambiguation loss $\mathcal{L}_{nacr}$ can be formulated as:

$$\mathcal{L}_{nacr} = -\frac{1}{2N} \sum_{m=1}^{2} \sum_{j=1}^{N} \sum_{k=1}^{K} \left[p_{jk} \log\left(z_{jk}^m\right) + (1 - p_{jk}) \log\left(1 - z_{jk}^m\right) \right], \tag{4}$$

where z_{jk}^m denotes the predicted probability that the j-th sample belongs to the k-th class under modality m, and p_{jk} represents the reconstructed confidence of the j-th sample pair for class k.

Empirically, incorporating neighborhood consensus into label confidence estimation improves disambiguation accuracy, especially with abundant redundant noisy annotations. It enables the model to construct semantically faithful contrastive pairs, thus significantly enhancing the model's robustness.

3.3 Class-aware Robust Contrastive Hashing

Although NACR reconstructs label confidence effectively, it does not explicitly incorporate class-level information when constructing robust positive and negative pairs. Therefore, we propose a Class-aware Robust Contrastive Hashing (CRCH) module to adaptively build reliable pseudo-labels thereby enhancing the stability of contrastive optimization.

First, we establish a class-wise threshold for each class based on the reconstructed label confidence. The threshold t_k for the k-th class is calculated as follows:

$$t_k = \frac{1}{N_k} \sum_{j=1}^{N} p_{jk}, \tag{5}$$

where p_{jk} denotes the reconstructed label confidence of the j-th sample pair on class k, and N_k indicates the number of samples for which $p_{jk} > 0$ among all N samples. This class-specific averaging adaptively captures the distributional characteristics of each class rather than relying on a fixed threshold. To build class-wise supervision, we derive the pseudo-label $\hat{y}_j = [\hat{y}_{j1}, \ldots, \hat{y}_{jK}]$ by comparing each reconstructed confidence p_{jk} with its corresponding class-wise threshold t_k:

$$\hat{y}_{jk} = \begin{cases} 1, & \text{if } p_{jk} \geq t_k \\ 0, & \text{otherwise} \end{cases}, \tag{6}$$

where $\hat{y}_{jk} \in \{0, 1\}$ indicates whether class k is considered positive for the j-th sample pair. For a mini-batch containing n sample pairs, we compute a label similarity matrix $T \in [0, 1]^{n \times n}$ based on the intersection-over-union (IoU) between pseudo-label vectors:

$$T_{ij} = \frac{\hat{y}_i \cap \hat{y}_j}{\hat{y}_i \cup \hat{y}_j}, \tag{7}$$

where T_{ij} measures the semantic similarity between the i-th and j-th sample pair based on their pseudo-labels. A higher T_{ij} indicates stronger semantic consistency, while $T_{ij} = 0$ means that the sample pairs are semantically disjoint. Accordingly, positive and negative pairs are determined by the indicator functions $\mathbb{I}[T_{ij} > 0]$ and $\mathbb{I}[T_{ij} = 0]$, respectively. However, directly relying on T_{ij} for pair construction can be unreliable due to redundant noisy annotations. To mitigate this issue, we follow [22] and introduce a margin-based thresholding mechanism that adaptively adjusts cross-modal similarities matrix S to reduce the impact of negative pairs with overly high similarities. Specifically, the adjusted similarity matrix N_{ij}^{*} is defined as:

$$N_{ij}^{*} = \begin{cases} S_{ij}^{*}, & \text{if } S_{ij}^{*} \geq S_{ii} - \delta \\ S_{ij}^{*} - \xi, & \text{otherwise} \end{cases}, \tag{8}$$

where $* \in \{12, 21\}$ denotes image-to-text and text-to-image retrieval directions. $S_{ij}^{12} = h_i^1 \cdot h_j^2$, $S_{ij}^{21} = h_i^2 \cdot h_j^1$. The margin parameter δ distinguishes hard and easy negative pairs, while the shift parameter ξ suppresses the influence of overly easy negatives. The loss of CRCH is defined as:

$$\mathcal{L}_{crch} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \left[\mathbb{I}[T_{ij} = 0] \cdot \sum^{*} \exp\left(N_{ij}^{*}\right) + \mathbb{I}[T_{ij} > 0] \cdot \exp(-S_{ij} - T_{ij}) \right] - \frac{1}{n} \sum_{i=1}^{n} S_{ii}, \tag{9}$$

where $\mathcal{L}_{crch}$ consists of three components: (1) $\mathbb{I}[T_{ij} = 0] \cdot \sum^{*} \exp(N_{ij}^{*})$ penalizes all negative cross-modal sample pairs, especially those that are easily confusable. (2) $\mathbb{I}[T_{ij} > 0] \cdot \exp(-S_{ij} - T_{ij})$ treats cross-modal sample pairs with non-zero label similarity T_{ij} as positive pairs and encourages greater semantic alignment. The higher the label similarity, the stronger the penalty imposed for insufficient feature similarity S_{ij}. (3) $-\frac{1}{n} \sum_{i=1}^{n} S_{ii}$ encourages consistency within each sample pair by increasing the similarity between the image and text representations.

The CRCH module dynamically constructs positive and negative sample pairs based on class-wise confidence thresholds, enabling the model to effectively capture nuanced relationships between samples. This design significantly reduces erroneous associations caused by redundant noisy labels and enhances the model's robustness to incorrect supervision.

3.4 Optimization

Benefiting from the NACR and CRCH modules, our model effectively learns discriminative hash codes within the Hamming space. However, the discrepancy between continuous representations and discrete binary codes inevitably leads to quantization errors, which can substantially degrade retrieval performance in CMH. To address this issue, we present an effective quantization loss as follows:

$$\mathcal{L}_{quant} = \frac{1}{n \cdot L} \sum_{m=1}^{2} \sum_{j=1}^{n} \sum_{l=1}^{L} \left| h_{jl}^{m} - sign(h_{jl}^{m}) \right|, \tag{10}$$

where h_{jl}^{m} denotes the l-th element of the hash representation h_j^m. This loss penalizes the deviation between continuous hash values and their corresponding binary codes $\mathrm{sign}(h_{jl}^{m})$, thereby encouraging each element to approach ± 1 and effectively reducing the quantization error, as demonstrated in our ablation studies. Thus, the final loss function of NACD can be defined as:

$$\mathcal{L} = \mathcal{L}_{nacr} + \alpha \mathcal{L}_{crch} + \beta \mathcal{L}_{quant}, \tag{11}$$

where α and β are hyperparameters that balance the contributions of $\mathcal{L}_{nacr}$, $\mathcal{L}_{crch}$, and $\mathcal{L}_{quant}$. Additional optimization details of NACD are provided in Appendix Sec. A.

4 Experiments

4.1 Experimental Settings

To evaluate the effectiveness of NACD, we conducted experiments on three benchmark datasets: MIRFlickr-25k (Flickr) [42], NUS-WIDE (NUS) [43], and MS-COCO (COCO) [44]. For all methods, we evaluate the Mean Average Precision (MAP) on both Image-to-Text (I2T) and Text-to-Image (T2I) retrieval tasks. Note that all MAP scores are computed over the entire retrieval set (i.e., MAP@ALL). Additionally, precision–recall curves under the hash lookup protocol are employed to visually assess the performance of CMH. To distinguish between different levels of redundant annotations, we define a redundant rate, which represents the ratio between the number of redundant noisy labels and the number of ground-truth labels in the entire annotations. Experiments were conducted with hash code lengths of 32, 64, and 128 bits, under redundant rates of 1.0, 1.5, 2.0, and 2.5. Additional details about the datasets are provided in Appendix Sec. B.

4.2 Implementation Details

In the proposed NACD, the image modality adopts the VGG19 model [45], pre-trained on ImageNet, as its CNN backbone. For text processing, the pre-trained Doc2Vec model [46] is employed as the backbone. For cross-modal shared representation learning, the image and text modalities employ three and two hidden layers, respectively. Each fully connected (FC) layer is succeeded by a ReLU activation layer, except for the final layer, which uses a tanh function. Each hidden layer contains 8,192 units, followed by an output layer of dimension L representing the shared embedding space. The model is trained using the RMSprop optimizer [47], with an initial learning rate of $1e-5$ and a maximum of 100 epochs. The parameters δ and ξ in Eq. (8) are set to 0.2 and 1.0, respectively. Additionally, we employ a batch size n of 128. The model is evaluated every 20 epochs, with the first 10 epochs serving as a warm-up phase during which the CM strategy and class-wise threshold update are disabled. The number of neighbors C is set to 20 to ensure accurate neighborhood information. Our NACD is implemented using the PyTorch framework [48] and all experiments are carried out with 4 NVIDIA V100 GPUs.

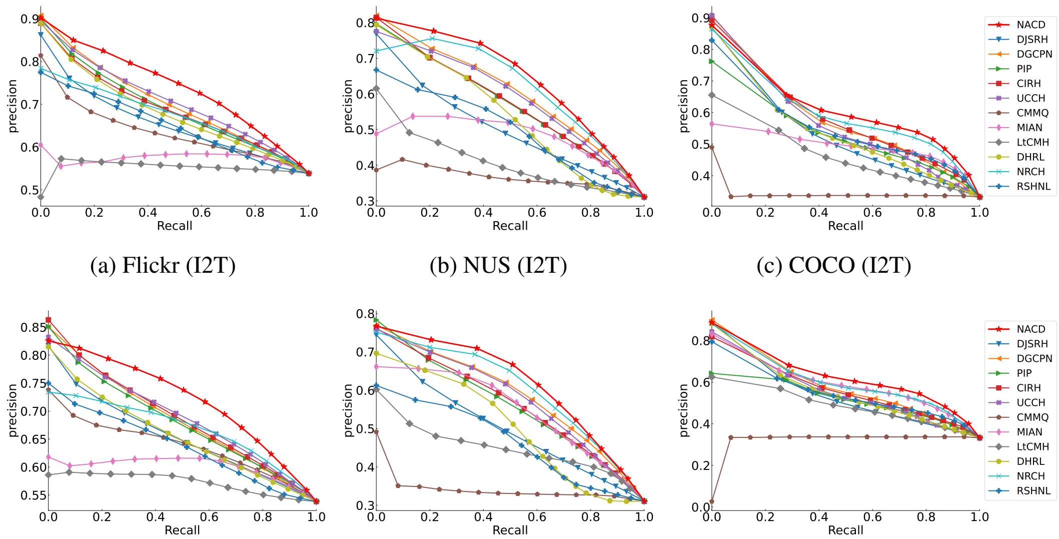
4.3 Comparison with State-of-the-Arts

In this work, we compare our NACD against 11 state-of-the-art CMH methods, including five unsupervised methods: DJSRH [18], DGCPN [20], PIP [49], CIRH [21], and UCCH [22]; and six supervised methods: CMMQ [50], MIAN [27], LtCMH [28], DHRL [29], NRCH [30], and RSHNL [31]. The average MAP scores for the I2T and T2I tasks are reported in Table 1. Additionally, the experiment results with 8 and 16 bits can be found in Appendix Sec. C.1. Fig. 3 presents precision-recall curves on three datasets for a hash code length of 128 bits and a redundant rate of 2.5. Based on these results, we make the following observations:

- As the hash code length increases, the performance of almost all methods improves, since longer codes contain more discriminative information in the Hamming space.
- As the redundant rate increases, the performance of all supervised CMH methods deteriorates, since the progressively redundant noisy supervision misleads model training. In contrast, NACD maintains stable and superior performance by effectively extracting correct supervision from redundant annotations. Meanwhile, unsupervised CMH methods remain unaffected as they do not rely on label information.
- From Table 1, we can see that the proposed NACD consistently outperforms all competing methods across all settings. For instance, when the hash code length is 128 and the redundant rate is 2.5, NACD exceeds the second-best methods by 3.3%, 1.7%, and 2.6% on the Flickr, NUS, and COCO datasets, respectively.
- As illustrated in Fig. 3, the area under the precision–recall curves indicates that NACD consistently outperforms all other state-of-the-art methods in both I2T and T2I tasks, demonstrating its stable and superior performance.

4.4 Ablation Study

To verify the effectiveness of each component in NACD, we conducted extensive ablation studies on the NUS and COCO datasets with a hash code length of 128 bits across various redundant rates. We compared the full NACD with six ablated variants: (1) only $\mathcal{L}_{nacr}$; (2) only $\mathcal{L}_{crch}$; (3) NACD without the CM strategy; (4) NACD without the class-wise threshold t_k fixed at 0.5; (5) NACD without $\mathcal{L}_{quant}$; (6) NACD without the warm-up phase; and (7) the full NACD. As shown in Table 2, NACR, CRCH, and $\mathcal{L}_{quant}$ all effectively enhance the performance of NACD. Additionally, the warm-up strategy, the CM strategy, and dynamic class-wise threshold updating are all crucial for achieving optimal model performance.



(a) Flickr (I2T)

(b) NUS (I2T)

(c) COCO (I2T)

(d) Flickr (T2I)

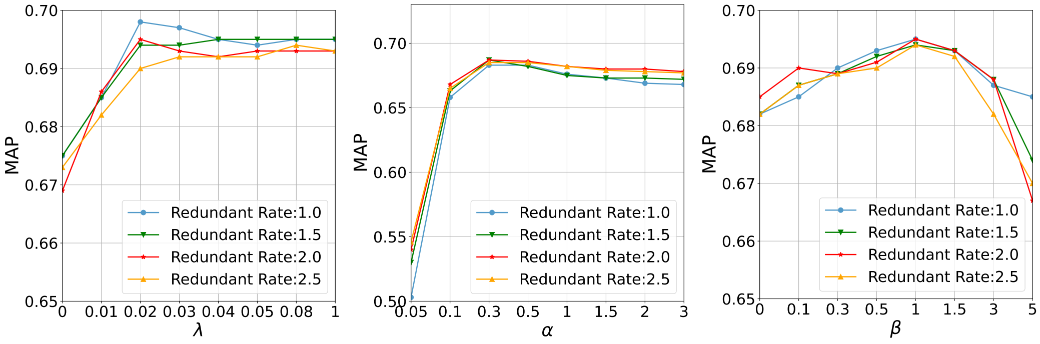
(e) NUS (T2I)

(f) COCO (T2I)

Figure 3: The precision-recall curves on three datasets. Note that the hash code length is 128bits and the redundant rate is 2.5.

4.5 Parameter Analysis

To evaluate the impact of the coefficient λ in Eq. (3), as well as α and β in Eq. (11), we conducted extensive experiments on the COCO dataset with a hash code length of 128 bits under various redundant rates. As shown in Fig. 4, λ yields optimal results within the range of $[0.01, 0.04]$, demonstrating the effectiveness of NACR. Moreover, the model achieves superior performance when α lies within $[0.1, 0.5]$ and β within $[0.5, 1.5]$, further validating the effectiveness of both CRCH and the quantization loss $\mathcal{L}_{quant}$. Additional parameter analyses on the Flick and NUS datasets are provided in Appendix Sec. C.2.



(a) λ on COCO.

(b) α on COCO.

(c) β on COCO.

Figure 4: The performance of NACD in terms of average MAP scores versus different values of λ, α, and β on COCO dataset with 128 bits.

4.6 Model Analysis

Robustness Analysis. To intuitively demonstrate the robustness of NACD, we compared it with two of its variants and the NRCH method. Their average MAP scores on the COCO dataset were plotted under the settings of a hash code length of 128 bits and redundant rates of 2.0 and 2.5. Specifically, NACD-1 denotes the variant without the CM strategy, while NACD-2 represents the variant in which the class-wise threshold t_k is fixed at 0.5, meaning that no dynamic threshold update is performed.

Table 1: The performance comparison in terms of average MAP scores (%) of I2T and T2I tasks under different redundant rates and various bit lengths on the MIRFlickr-25K(Flickr), NUS-WIDE(NUS), and MS-COCO(COCO) datasets. The highest and second highest MAP scores among all methods are shown in **bold** and underline respectively.

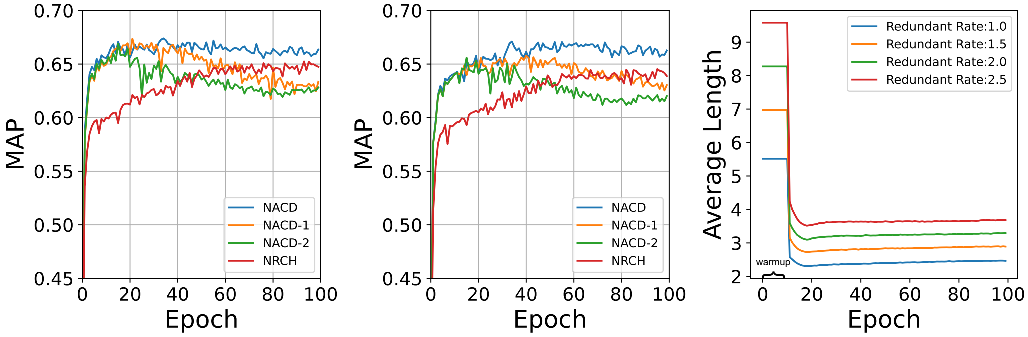
Dataset	Method	Year	1.0			1.5			2.0			2.5		
			32bits	64bits	128bits	32bits	64bits	128bits	32bits	64bits	128bits	32bits	64bits	128bits
Flickr	DJSRH	2019	63.4	64.8	66.5	63.4	64.8	66.5	63.4	64.8	66.5	63.4	64.8	66.5
	DGCPN	2021	69.2	70.5	71.1	69.2	70.5	71.1	69.2	70.5	71.1	69.2	70.5	71.1
	PIP	2021	67.9	68.9	69.8	67.9	68.9	69.8	67.9	68.9	69.8	67.9	68.9	69.8
	CIRH	2022	68.6	69.4	70.1	68.6	69.4	70.1	68.6	69.4	70.1	68.6	69.4	70.1
	UCCH	2023	71.0	71.9	72.0	71.0	71.9	72.0	71.0	71.9	72.0	71.0	71.9	72.0
	CMMQ	2022	63.6	64.1	65.5	66.5	67.6	69.1	63.6	63.9	65.6	63.6	63.8	64.8
	MIAN	2023	71.1	71.8	73.0	67.6	68.4	68.0	61.5	62.8	63.2	61.6	61.3	60.2
	LtCMH	2023	63.2	64.3	65.3	60.0	59.8	60.6	57.9	57.8	58.5	56.9	56.4	57.6
	DHRL	2024	69.5	69.4	69.3	68.8	69.0	68.6	69.2	68.9	68.9	69.0	68.8	68.4
	NRCH	2024	72.9	74.5	74.4	67.9	70.9	69.5	63.0	63.6	65.1	57.7	58.2	63.9
	RSHNL	2025	71.8	72.2	72.5	69.2	69.5	70.5	66.5	67.5	67.9	64.4	65.0	66.0
	NACD	**Ours**	**76.5**	**76.8**	**77.1**	**76.4**	**76.8**	**76.9**	**75.7**	**76.2**	**76.3**	**74.3**	**74.6**	**75.3**
NUS	DJSRH	2019	46.7	49.5	52.1	46.7	49.5	52.1	46.7	49.5	52.1	46.7	49.5	52.1
	DGCPN	2021	60.2	62.4	64.0	60.2	62.4	64.0	60.2	62.4	64.0	60.2	62.4	64.0
	PIP	2021	57.3	59.1	59.9	57.3	59.1	59.9	57.3	59.1	59.9	57.3	59.1	59.9
	CIRH	2022	57.2	59.3	60.4	57.2	59.3	60.4	57.2	59.3	60.4	57.2	59.3	60.4
	UCCH	2023	62.3	63.4	63.9	62.3	63.4	63.9	62.3	63.4	63.9	62.3	63.4	63.9
	CMMQ	2022	57.7	57.7	58.3	52.5	52.5	52.7	44.8	44.2	44.1	36.9	36.4	36.8
	MIAN	2023	63.5	63.6	64.1	58.9	60.2	61.2	57.4	58.1	59.2	57.1	56.5	58.6
	LtCMH	2023	57.0	58.0	59.4	51.7	52.5	53.2	50.6	49.5	49.5	45.3	45.5	46.6
	DHRL	2024	61.0	61.0	60.8	60.3	60.2	58.9	58.7	59.2	57.8	57.7	58.7	57.4
	NRCH	2024	67.9	68.6	69.3	67.4	68.3	68.4	67.0	67.7	68.0	66.6	67.3	67.6
	RSHNL	2025	59.3	59.7	60.3	57.8	57.7	57.5	55.5	55.5	55.7	54.4	54.0	53.0
	NACD	**Ours**	**68.2**	**69.4**	**70.3**	**68.1**	**69.4**	**70.2**	**67.9**	**69.0**	**69.6**	**67.4**	**68.7**	**69.3**
COCO	DJSRH	2019	51.6	54.1	57.0	51.6	54.1	57.0	51.6	54.1	57.0	51.6	54.1	57.0
	DGCPN	2021	63.0	63.5	64.3	63.0	63.5	64.3	63.0	63.5	64.3	63.0	63.5	64.3
	PIP	2021	55.7	57.8	58.2	55.7	57.8	58.2	55.7	57.8	58.2	55.7	57.8	58.2
	CIRH	2022	63.0	63.5	64.2	63.0	63.5	64.2	63.0	63.5	64.2	63.0	63.5	64.2
	UCCH	2023	60.4	61.4	61.8	60.4	61.4	61.8	60.4	61.4	61.8	60.4	61.4	61.8
	CMMQ	2022	43.9	44.1	44.6	39.8	39.8	40.2	35.9	35.5	35.5	33.7	33.7	33.7
	MIAN	2023	61.0	63.8	64.8	60.8	62.9	63.6	59.2	60.2	60.5	57.3	59.4	58.7
	LtCMH	2023	58.1	60.5	62.1	56.3	58.9	60.4	54.7	56.3	57.9	52.3	54.2	55.3
	DHRL	2024	33.4	34.5	61.9	33.3	43.6	60.4	33.2	43.2	61.5	35.2	60.6	59.8
	NRCH	2024	65.9	67.3	67.8	65.6	66.9	67.3	64.9	66.8	67.1	63.2	65.3	66.2
	RSHNL	2025	60.7	60.3	60.2	61.1	61.5	61.7	59.9	59.9	61.7	60.7	59.1	60.3
	NACD	**Ours**	**66.7**	**68.4**	**68.7**	**67.2**	**68.4**	**69.1**	**67.1**	**68.5**	**69.0**	**66.5**	**68.3**	**68.8**

Table 2: The ablation study results on NUS and COCO datasets with 128 bits and across various redundant rates. The highest scores are presented in bold.

Method	NUS				COCO			
	1.0	1.5	2.0	2.5	1.0	1.5	2.0	2.5
only $\mathcal{L}_{nacr}$	34.3	33.5	33.3	36.0	33.3	33.3	33.3	33.3
only $\mathcal{L}_{crch}$	63.3	63.0	63.1	63.2	64.9	64.9	65.6	65.7
NACD w/o CM strategy	69.2	69.1	69.2	69.0	67.2	66.9	67.1	67.1
NACD with t_k remains 0.5	70.0	69.5	69.3	68.9	67.9	68.3	68.3	68.0
NACD w/o $\mathcal{L}_{quant}$	69.0	69.0	68.4	68.1	68.0	67.9	67.9	67.8
NACD w/o warm-up	69.3	69.2	68.6	68.4	68.0	68.1	68.2	68.1
The full NACD	**70.3**	**70.2**	**70.2**	**69.6**	**68.7**	**69.1**	**69.0**	**68.8**

As shown in Fig. 5(a,b): (1) Both NACD-1 and NACD-2 tend to overfit redundant noise, indicating that the CM strategy and class-wise threshold updating are crucial for enhancing the robustness of NACD. (2) Although NRCH exhibits certain robustness under redundant annotations, it still lags significantly behind NACD. This gap widens as noise levels increase, further underscoring NACD's robustness in challenging scenarios.

Disambiguation Analysis. In Fig. 5(c), we analyze the model's ability to disambiguate redundant annotations by computing the average pseudo-label length for all sample pairs in the COCO dataset. Here, the pseudo-label length for the j-th sample pair is defined as the number of positive entries in $\hat{y}_j$, reflecting how many classes the model predicts as positive. As shown in the figure, when the redundant rate decreases, the average pseudo-label length gradually approaches the average number of ground-truth labels in the COCO dataset (2.76). This indicates that NACD can accurately recover the true label subset from redundant annotations, confirming its strong disambiguation capability.



(a) redundant rate of 2.0.

(b) redundant rate of 2.5.

(c) average label length.

Figure 5: The robustness study and disambiguation study results with 32 bits on the COCO dataset.

5 Conclusion

In this paper, we propose a novel Neighbor-aware Contrastive Disambiguation (NACD) framework to address the challenge of redundant annotations in cross-modal hashing. NACD comprises two key modules: Neighbor-aware Confidence Reconstruction (NACR) and Class-aware Robust Contrastive Hashing (CRCH). Specifically, NACR reconstructs label confidence by aggregating information from cross-modal neighbors, thereby distinguishing true labels from ambiguous ones. Meanwhile, CRCH constructs reliable positive and negative pairs based on label confidence, substantially enhancing robustness under noisy supervision. Moreover, a quantization loss is incorporated to reduce the quantization error and enforce binary constraints on the learned hash representations. Extensive experiments on three large-scale multimodal benchmarks demonstrate that NACD consistently outperforms state-of-the-art approaches, showing strong robustness and stable performance for cross-modal hashing with redundant annotations.

Limitation. Although our proposed NACD demonstrates strong performance, there are still some limitations that need to be addressed. In this paper, the "redundant annotations" we study in cross-modal hashing do not account for the inherent similarity between labels. NACD may struggle when noisy labels are highly similar to the ground-truth, such as when the correct label is "car" and the noisy label is "truck". Additionally, we only conduct extensive experiments on image and text modalities to demonstrate NACD's effectiveness. In the future, additional modalities need to be considered to verify the generalization ability of NACD. We encourage further research to better understand and mitigate the limitations and risks of cross-modal hashing with redundant annotations.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62306197, 62372315), Sichuan Science and Technology Planning Project (2025ZNSFSC1507, 2024YFG0007, 2024ZDZX0004, 2024NSFTD0049), China Postdoctoral Science Foundation (2021TQ0223, 2022M712236), Chengdu Science and Technology Project (2023-XT00-00004-GX), Postdoctoral Joint Training Program of Sichuan University (SCDXLHPY2307).

References

[1] Jie Wen, Gehui Xu, Zhanyan Tang, Wei Wang, Lunke Fei, and Yong Xu. Graph regularized and feature aware matrix factorization for robust incomplete multi-view clustering. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(5):3728–3741, 2023.

[2] Chao Su, Zhi Li, Tianyi Lei, Dezhong Peng, and Xu Wang. Metavg: A meta-learning framework for visual grounding. *IEEE Signal Processing Letters*, 31:236–240, 2023.

[3] Song Wu, Yan Zheng, Yazhou Ren, Jing He, Xiaorong Pu, Shudong Huang, Zhifeng Hao, and Lifang He. Self-weighted contrastive fusion for deep multi-view clustering. *IEEE Transactions on Multimedia*, 26:9150–9162, 2024.

[4] Yu Cao, Xu Wang, Qian Wang, Zhong Yuan, Yongguo Shi, and Dezhong Peng. Graph and text multi-modal representation learning with momentum distillation on electronic health records. *Knowledge-Based Systems*, 302:112373, 2024.

[5] Jie Wen, Jiang Long, Xiaohuan Lu, Chengliang Liu, Xiaozhao Fang, and Yong Xu. Partial multiview incomplete multilabel learning via uncertainty-driven reliable dynamic fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[6] Chengliang Liu, Jie Wen, Yong Xu, Bob Zhang, Liqiang Nie, and Min Zhang. Reliable representation learning for incomplete multi-view missing multi-label classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[7] Jianpeng Chen, Yawen Ling, Jie Xu, Yazhou Ren, Shudong Huang, Xiaorong Pu, Zhifeng Hao, Philip S. Yu, and Lifang He. Variational graph generator for multiview graph clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):11078–11091, 2025.

[8] Yanglin Feng, Hongyuan Zhu, Dezhong Peng, Xi Peng, and Peng Hu. Rono: robust discriminative learning with noisy labels for 2d-3d cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11610–11619, 2023.

[9] Haoran Liu, Ying Ma, Ming Yan, Yingke Chen, Dezhong Peng, and Xu Wang. Dida: Disambiguated domain alignment for cross-domain retrieval with partial labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3612–3620, 2024.

[10] Yongxiang Li, Yang Qin, Yuan Sun, Dezhong Peng, Xi Peng, and Peng Hu. Romo: Robust unsupervised multimodal learning with noisy pseudo labels. *IEEE Transactions on Image Processing*, 2024.

[11] Chao Su, Huiming Zheng, Dezhong Peng, and Xu Wang. Dica: Disambiguated contrastive alignment for cross-modal retrieval with partial labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 20610–20618, 2025.

[12] Yongxiang Li, Yuan Sun, Yang Qin, Dezhong Peng, Xi Peng, and Peng Hu. Robust duality learning for unsupervised visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 2025.

[13] Yongxiang Li, Dezhong Peng, Haixiao Huang, Yizhi Liu, Huiming Zheng, and Zheng Liu. Multi-granularity confidence learning for unsupervised text-to-image person re-identification with incomplete modality. *Knowledge-Based Systems*, 315:113304, 2025.

[14] Yanglin Feng, Yang Qin, Dezhong Peng, Hongyuan Zhu, Xi Peng, and Peng Hu. Pointcloud-text matching: Benchmark dataset and baseline. *IEEE Transactions on Multimedia*, 2025.

[15] Ziniu Yin, Yanglin Feng, Ming Yan, Xiaomin Song, Dezhong Peng, and Xu Wang. Roda: Robust domain alignment for cross-domain retrieval against label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9535–9543, 2025.

[16] Ruitao Pu, Yang Qin, Xiaomin Song, Dezhong Peng, Zhenwen Ren, and Yuan Sun. She: Streaming-media hashing retrieval. In *Forty-second International Conference on Machine Learning*, 2025.

[17] Jian Zhang, Yuxin Peng, and Mingkuan Yuan. Unsupervised generative adversarial cross-modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[18] Shupeng Su, Zhisheng Zhong, and Chao Zhang. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3027–3035, 2019.

[19] Dejie Yang, Dayan Wu, Wanqian Zhang, Haisu Zhang, Bo Li, and Weiping Wang. Deep semantic-alignment hashing for unsupervised cross-modal retrieval. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 44–52, 2020.

[20] Jun Yu, Hao Zhou, Yibing Zhan, and Dacheng Tao. Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4626–4634, 2021.

[21] Lei Zhu, Xize Wu, Jingjing Li, Zheng Zhang, Weili Guan, and Heng Tao Shen. Work together: Correlation-identity reconstruction hashing for unsupervised cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8838–8851, 2022.

[22] Peng Hu, Hongyuan Zhu, Jie Lin, Dezhong Peng, Yin-Ping Zhao, and Xi Peng. Unsupervised contrastive cross-modal hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3877–3889, 2022.

[23] Qing-Yuan Jiang and Wu-Jun Li. Deep cross-modal hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3232–3240, 2017.

[24] Xi Zhang, Hanjiang Lai, and Jiashi Feng. Attention-aware deep adversarial hashing for cross-modal retrieval. In *Proceedings of the European Conference on Computer Vision*, pages 591–606, 2018.

[25] De Xie, Cheng Deng, Chao Li, Xianglong Liu, and Dacheng Tao. Multi-task consistency-preserving adversarial hashing for cross-modal retrieval. *IEEE Transactions on Image Processing*, 29:3626–3637, 2020.

[26] Rong-Cheng Tu, Xian-Ling Mao, Bing Ma, Yong Hu, Tan Yan, Wei Wei, and Heyan Huang. Deep cross-modal hashing with hashing functions and unified hash codes jointly learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(2):560–572, 2020.

[27] Zheng Zhang, Haoyang Luo, Lei Zhu, Guangming Lu, and Heng Tao Shen. Modality-invariant asymmetric networks for cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):5091–5104, 2023.

[28] Zijun Gao, Jun Wang, Guoxian Yu, Zhongmin Yan, Carlotta Domeniconi, and Jinglin Zhang. Long-tail cross modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7642–7650, 2023.

[29] Zhenqiu Shu, Yibing Bai, Kailing Yong, and Zhengtao Yu. Deep cross-modal hashing with ranking learning for noisy labels. *IEEE Transactions on Big Data*, 2024.

[30] Longan Wang, Yang Qin, Yuan Sun, Dezhong Peng, Xi Peng, and Peng Hu. Robust contrastive cross-modal hashing with noisy labels. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5752–5760, 2024.

[31] Ruitao Pu, Yuan Sun, Yang Qin, Zhenwen Ren, Xiaomin Song, Huiming Zheng, and Dezhong Peng. Robust self-paced hashing for cross-modal retrieval with noisy labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19969–19977, 2025.

[32] Lijuan Sun, Songhe Feng, Jun Liu, Gengyu Lyu, and Congyan Lang. Global-local label correlation for partial multi-label learning. *IEEE Transactions on Multimedia*, 24:581–593, 2022.

[33] Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning with noisy label identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3676–3687, 2022.

[34] Jun-Yi Hang and Min-Ling Zhang. Partial multi-label learning with probabilistic graphical disambiguation. *Advances in Neural Information Processing Systems*, 36:1339–1351, 2023.

[35] Yu Chen, Yanan Wu, Na Han, Xiaozhao Fang, Bingzhi Chen, and Jie Wen. Partial multi-label learning based on near-far neighborhood label enhancement and nonlinear guidance. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 3722–3731, 2024.

[36] Yuan Sun, Zhenwen Ren, Peng Hu, Dezhong Peng, and Xu Wang. Hierarchical consensus hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*, 26:824–836, 2023.

[37] Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32, Jun 2018.

[38] Gengyu Lyu, Songhe Feng, and Yidong Li. Partial multi-label learning via probabilistic graph matching mechanism. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 105–113, 2020.

[39] Guoxian Yu, Xia Chen, Carlotta Domeniconi, Jun Wang, Zhao Li, Zili Zhang, and Xindong Wu. Feature-induced partial multi-label learning. In *2018 IEEE International Conference on Data Mining*, pages 1398–1403, Nov 2018.

[40] Lijuan Sun, Songhe Feng, Tao Wang, Congyan Lang, and Yi Jin. Partial multi-label learning by low-rank and sparse decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5016–5023, 2019.

[41] Gengyu Lyu, Songhe Feng, Yi Jin, Tao Wang, Congyan Lang, and Yidong Li. Prior knowledge regularized self-representation model for partial multilabel learning. *IEEE Transactions on Cybernetics*, 53(3):1618–1628, Mar 2023.

[42] Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, page 39–43, 2008.

[43] 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*, pages 1–9, 2009.

[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[46] Jey Han Lau and Timothy Baldwin. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 78–86, 2016.

[47] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2):26, 2012.

[48] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

[49] Peng-Fei Zhang, Yang Li, Zi Huang, and Hongzhi Yin. Privacy protection in deep multi-modal retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 634–643, 2021.

[50] Erkun Yang, Dongren Yao, Tongliang Liu, and Cheng Deng. Mutual quantization for cross-modal search with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7551–7560, 2022.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: This paper proposes a novel Neighbor-aware Contrastive Disambiguation (NACD) method for cross-modal hashing with redundant annotations. The effectiveness of this method is thoroughly demonstrated through extensive experiments, which precisely align with the main claims made at the beginning.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in the conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This is not a theoretical paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The architecture of NACD is fully described in the main paper. The experimental setup and implementation details are disclosed in the supplementary sections.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide code and instructions for using the data and code in the supplementary material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We briefly describe the key information of the experimental setting/details in the main paper and provide full details in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The experimental data and comparative experiments in this paper are both very extensive. In addition to the large number of retrieval results, this paper also requires a significant amount of space to analyze the model's performance and influencing factors, which makes the presentation of error bars quite challenging. To enhance the reliability of the experimental evaluation, we conducted multiple experiments and presented the average results. However, this also greatly increased the resources and time consumed in our training.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We detail the computation resources used in our experiments in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We ensure that the research in the paper fully complies with the Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Although our paper addresses the redundant annotations problem, which is related to the robustness of models, it does not have a direct association with societal aspects.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper merely addresses the Redundant Annotations problem in cross-modal hashing and does not involve these risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We used open datasets and correctly referenced the papers.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will release our code, which is well documented, along with a link to the codebase repository.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The LLM is used only for writing.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix

In the following sections, we provide additional details about the optimization process (Sec. A), the datasets (Sec. B), and the experimental results (Sec. C) of our proposed NACD.

A Optimization

To effectively learn under redundant annotation supervision, we design an end-to-end optimization strategy, as summarized in Algorithm 1. During the warm-up phase, we train the model without the CM strategy and the class-wise threshold update. Once the warm-up phase is complete, the stabilized model produces more reliable feature representations, which in turn provide stronger support for updating label confidences through neighborhood-based mechanisms. In addition, before backpropagation, we slightly update the label confidences of the current mini-batch based on the model's latest predictions to refine the supervision signals.

Algorithm 1 Optimization Algorithm for NACD

```
Require: The redundant annotations training set $\mathcal{D}$, the code length $L$, the network $N = \{f_1(\cdot, \Theta_1), f_2(\cdot, \Theta_2)\}$, the learnable matrix $\mathbf{W_1}, \mathbf{W_2}$, the maximal epoch number $T_{\max}$, and the warm-up epoch number $T_{\text{warm}}$;
Randomly initialize network parameters $\{\Theta_i, W_i\}_{i=1}^2$;
Use part of the multi-partial labels of each sample as the initial state of instance confidence, and set the threshold for each class to 0.5.
for $epoch = 1$ to $T_{\max}$ do
    for $\mathcal{D}_n$ in mini-batches sampled from $\mathcal{D}$ do
        Compute $L_{final}$ by Eq. (11);
        Update instance confidence by Eq. (1);
        Optimize network parameters through backpropagation;
    end for
    if $epoch > T_{\text{warm}}$ then
        Select the neighbor update instance label confidence by Eq. (2);
        Update class-aware threshold by Eq. (5);
    end if
end for
Ensure: Network parameters $\{\Theta_i, W_i\}_{i=1}^2$.
```

B Datasets

To verify the effectiveness of our proposed NACD method in addressing the redundant annotations problem, we conduct experiments on the MIRFlickr-25k (Flickr) [42], NUS-WIDE (NUS) [43], and MS-COCO (COCO) [44] datasets. The details are as follows: MIRFlickr-25K [42] contains 25,000 image-text pairs, each belonging to one of 24 categories. In this work, we only select 20,015 pairs that have annotations. NUS-WIDE [43] is a multimodal dataset containing 81 concept categories. In this work, we only consider the subset of data from the most frequent 21 categories, which includes 190,421 image-text pairs. MS-COCO [44] encompasses a vast collection of 123,287 images distributed across 80 diverse categories. Each image is enriched with five detailed textual descriptions. After considering only labeled data, we ultimately select 122,218 image-text pairs.

Following dataset partition strategies adopted in prior works [22, 30], we have structured the datasets accordingly: For MIRFlickr-25K [42], we select 2,000 data points as the test (query) dataset. The remaining data points are used to form the retrieval (database) dataset. From this, we further identified a training subset comprising 10,000 data points. In the case of NUS-WIDE [43], the test (query) dataset is made up of 2,100 data points. The remaining data points form the retrieval (database) dataset. From this dataset, we select 10,500 data points for training. For MS-COCO [44], we extract 5,000 data points to be used for testing. The rest of the data points are pooled into the retrieval (database) dataset. From this, we set aside 10,000 data points specifically for training. Table 3 shows the specific data split information of these three multimodal datasets in our experiments.

Table 3: Data split and basic information for Flickr, NUS, and COCO in our experiments.

Dataset	Test (query)	Database	Train	Average length of GTs	Classes
Flickr	2,000	18,015	10,000	3.78	24
NUS	2,100	188,321	10,500	2.09	21
COCO	5,000	117,218	10,000	2.76	80

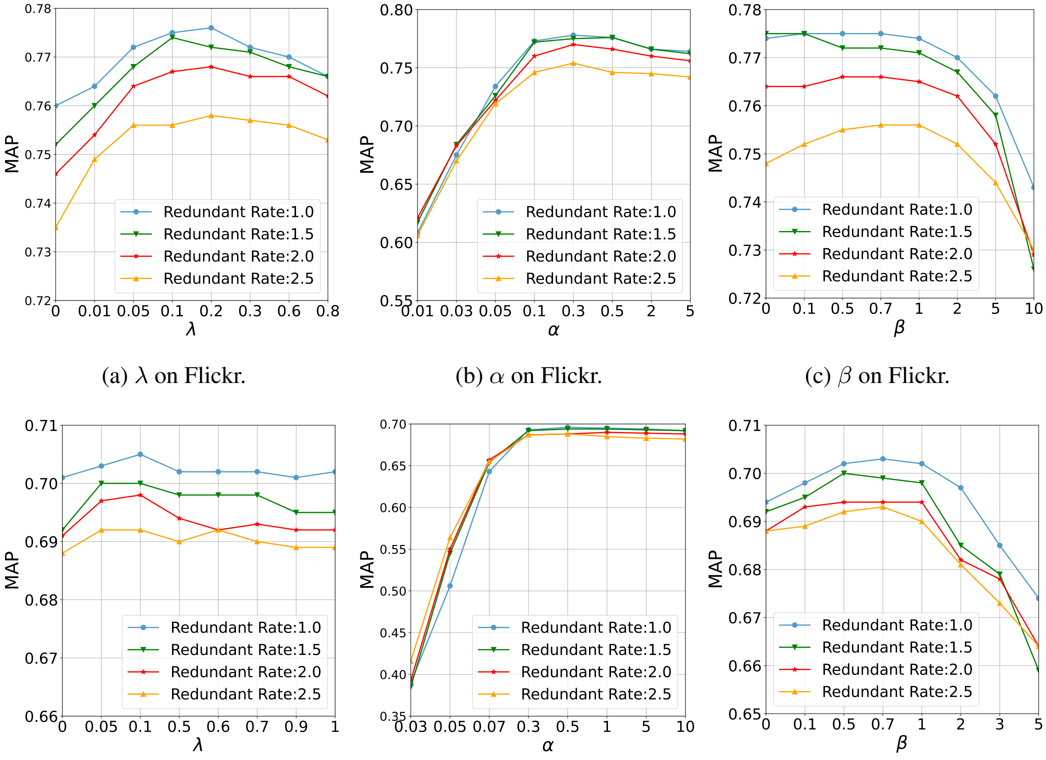
C Experimental Results

C.1 Additional Comparative Experiments

Moreover, to further validate the effectiveness of our method in a more comprehensive manner, we conduct additional experiments using 8-bit and 16-bit hash codes. The results of these experiments are presented in Table 4, demonstrating the stable and competitive performance of our approach under various experimental settings.

C.2 Additional Parameter Analysis

To analyze the impact of the coefficients λ, α, and β in Eq. (3) and Eq. (11), we conduct parameter analysis experiments on the Flickr and NUS datasets under different redundant rates. The results are illustrated in Fig. 6. Specifically, for the Flickr dataset, the model achieves its peak performance when λ is within [0.1, 0.3], α within [0.1, 0.5], and β within [0.5, 1]. This indicates that relatively balanced tuning of these parameters is crucial to optimize the model's performance on this dataset. Similarly, for the NUS dataset, the model exhibits optimal performance when λ is within [0.05, 0.5], α within [0.3, 1], and β within [0.5, 1]. These findings highlight the necessity of dataset-specific parameter tuning to maximize model performance.



(a) λ on Flickr.

(b) α on Flickr.

(c) β on Flickr.

(e) λ on NUS.

(f) α on NUS.

(g) β on NUS.

Figure 6: The performance of NACD in terms of average MAP scores versus different values of λ , β and α on the Flickr and NUS datasets using 128 bits.

Table 4: The performance comparison in terms of average MAP scores (%) of I2T and T2I tasks under different redundant rates and various bit lengths on the MIRFlickr-25K(Flickr), NUS-WIDE(NUS), and MS-COCO(COCO) datasets. The highest and second highest MAP scores among all methods are shown in **bold** and underline respectively.

Dataset	Method	Year	1.0		1.5		2.0		2.5	
			8bits	16bits	8bits	16bits	8bits	16bits	8bits	16bits
Flickr	DJSRH	2019	61.3	61.4	61.3	61.4	61.3	61.4	61.3	61.4
	DGCPN	2021	69.0	67.9	69.0	67.9	69.0	67.9	69.0	67.9
	PIP	2021	66.9	67.4	66.9	67.4	66.9	67.4	66.9	67.4
	CIRH	2022	65.6	66.8	65.6	66.8	65.6	66.8	65.6	66.8
	UCCH	2023	67.5	70.1	67.5	70.1	67.5	70.1	67.5	70.1
	CMMQ	2022	58.1	60.5	61.9	62.7	56.5	61.8	56.5	61.8
	MIAN	2023	65.5	68.6	62.1	64.1	59.7	61.7	58.9	59.9
	LtCMH	2023	60.7	61.4	58.7	57.8	55.6	54.9	56.2	55.0
	DHRL	2024	67.1	68.7	65.7	68.2	65.3	67.6	65.3	66.5
	NRCH	2024	67.0	71.0	53.9	58.5	53.9	57.9	53.9	56.9
	RSHNL	2025	70.3	71.5	68.8	71.0	68.7	70.2	68.9	70.9
	NACD	**Ours**	**73.9**	**75.9**	**73.6**	**75.0**	**72.5**	**74.8**	**70.2**	**73.3**
NUS	DJSRH	2019	42.8	42.9	42.8	42.9	42.8	42.9	42.8	42.9
	DGCPN	2021	55.2	58.6	55.2	58.6	55.2	58.6	55.2	58.6
	PIP	2021	55.7	56.2	55.7	56.2	55.7	56.2	55.7	56.2
	CIRH	2022	54.6	55.5	54.6	55.5	54.6	55.5	54.6	55.5
	UCCH	2023	57.4	60.1	57.4	60.1	57.4	60.1	57.4	60.1
	CMMQ	2022	56.4	57.4	52.1	52.3	46.2	45.5	38.5	37.6
	MIAN	2023	58.9	62.1	54.8	55.2	54.2	54.4	50.0	54.2
	LtCMH	2023	52.2	53.0	36.9	46.1	38.6	45.3	38.0	44.1
	DHRL	2024	58.3	59.3	54.0	58.9	54.0	57.3	52.6	56.8
	NRCH	2024	**63.7**	**65.9**	**61.8**	65.4	61.1	64.9	59.2	64.6
	RSHNL	2025	58.2	61.1	53.9	57.6	54.5	56.0	50.9	55.0
	NACD	**Ours**	62.9	65.0	61.3	**66.1**	**61.3**	**65.8**	**62.1**	**65.2**
COCO	DJSRH	2019	41.9	47.6	41.9	47.6	41.9	47.6	41.9	47.6
	DGCPN	2021	56.6	61.1	56.6	61.1	56.6	61.1	56.6	61.1
	PIP	2021	46.2	52.6	46.2	52.6	46.2	52.6	46.2	52.6
	CIRH	2022	54.2	59.6	54.2	59.6	54.2	59.6	54.2	59.6
	UCCH	2023	55.3	57.0	55.3	57.0	55.3	57.0	55.3	57.0
	CMMQ	2022	45.4	44.2	41.5	39.8	36.1	35.6	33.8	33.7
	MIAN	2023	53.9	58.5	55.0	58.4	53.0	57.4	52.4	56.2
	LtCMH	2023	52.3	54.7	51.2	51.2	50.0	51.5	40.0	49.6
	DHRL	2024	34.4	33.4	33.4	33.8	33.4	33.8	33.4	42.1
	NRCH	2024	59.5	64.2	58.7	63.3	57.5	62.0	56.2	61.0
	RSHNL	2025	56.6	59.3	55.4	59.7	57.6	61.3	55.2	59.3
	NACD	**Ours**	**62.1**	**65.5**	**61.7**	**65.4**	**62.3**	**65.6**	**60.5**	**65.8**