RELIABLE SUPERVISION AND SEMANTIC INITIALIZATION FOR NOISY MULTI-LABEL CROSS-MODAL HASHING RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Supervised multi-label cross-modal hashing relies on label information to construct semantic relations across heterogeneous modalities. However, false-positive and missing-positive labels introduce spurious positive relations and missing associations, respectively, directly distorting cross-modal semantic alignment and degrading binary representation learning. Existing robust methods mainly focus on identifying, reweighting, or correcting noisy supervision, but often underexploit the fine-grained multi-label structure after correction. Consequently, corrected supervision is not fully preserved across continuous semantic relation modeling and discrete hash optimization. To address this issue, we propose Reliability-Based Cross-modal Deep Hashing (RBCDH), which unifies supervision reliability modeling with multi-resolution semantic learning. RBCDH estimates instance–class-level label reliability by integrating intra-modal neighborhood label support with cross-modal consistency, and performs controlled bidirectional correction to suppress false positives while recovering missing positives. To preserve fine-grained semantics after correction, we construct continuous Soft-Jaccard relations from corrected label overlaps, providing a reliability-aware initialization before discrete optimization. We further derive binary reliable relations from the same corrected supervision to constrain the final hash representations, enabling reliable semantics to be consistently exploited across continuous and discrete spaces. Experiments on NUS-WIDE TC-21, and MM-IMDb under different noise levels and code lengths show that RBCDH consistently outperforms the baseline methods in overall cross-modal retrieval performance under noisy supervision.

Keywords: Cross-modal Hashing, Noisy Multi-label Learning, Reliable Supervision Reconstruction, Hash Learning

1 INTRODUCTION

Cross-modal hashing maps heterogeneous modalities into a shared compact binary space, enabling efficient large-scale retrieval with low storage cost and fast Hamming-distance search Ding et al. (2014); Zhou et al. (2014); Yao et al. (2016); Wang et al. (2016). Supervised cross-modal hashing further exploits semantic annotations to establish cross-modal correspondence and learn discriminative hash representations Jiang & Li (2017); Zhong et al. (2018); Li & Sun (2023). In multi-label settings, however, semantic relevance is not naturally binary: two instances may share different subsets of labels and therefore exhibit different degrees of semantic overlap. Recent studies have explored richer semantic structures for multi-label cross-modal retrieval, including label-wise semantic alignment, dual semantic fusion, and multi-label weighted contrastive modeling Li & Sun (2023); Liu et al. (2024); Yi et al. (2024). These methods show that multi-label annotations can describe cross-modal relevance more finely than a simple related/unrelated relation. As illustrated in Figure 1, once such supervision is corrupted, false-positive and missing-positive labels can distort the cross-modal relations induced by label overlap and subsequently interfere with hash representation learning.

Multi-label supervision is often incomplete and noisy in real-world data. A positive label may be incorrectly assigned to an instance, while a valid positive label may be missing Ghiassi et al. (2022);

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

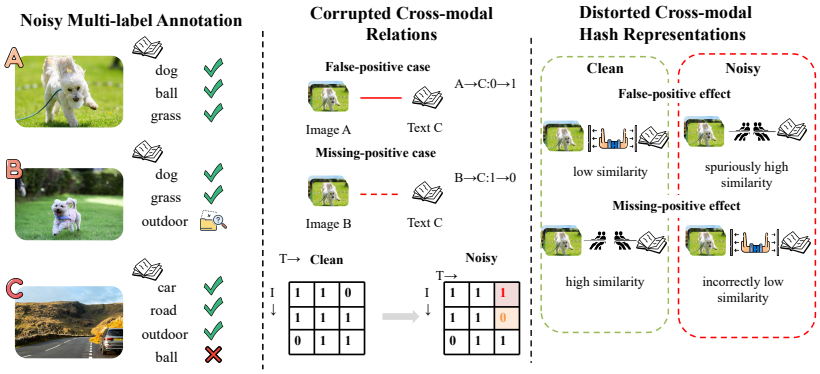


Figure 1: Propagation of supervision corruption in noisy multi-label cross-modal hashing. False-positive and missing-positive labels alter pairwise cross-modal relations, which further distort the learned cross-modal hash representations.

Xia et al. (2023); Xu et al. (2024). These two types of annotation errors affect semantic relations differently: false-positive labels can create spurious associations between irrelevant instances, whereas missing-positive labels can suppress associations that should exist. This issue is particularly important in cross-modal retrieval because label-derived supervision is further used to establish relations between heterogeneous modalities and optimize their shared representations. Early studies on cross-modal retrieval with noisy labels have demonstrated the adverse effects of corrupted supervision on retrieval learning Mandal & Biswas (2020); Hu et al. (2021), while noise-robust cross-modal hashing has explicitly investigated how unreliable supervision affects hash learning Wang et al. (2021).

Existing robust cross-modal hashing methods improve supervision reliability in several ways. Some reduce the influence of unreliable training signals through sample selection or self-paced learning Yang et al. (2022); Pu et al. (2025); Liu et al. (2026), while others assign different importance to training instances or relations according to estimated reliability Tan et al. (2026); Wang et al. (2026a). Early-learning and contrastive regularization have also been used to mitigate the propagation of noisy supervision during representation learning Xu et al. (2022); Wang et al. (2024). Another line of work directly modifies supervision through label correction or explicit relation construction, including LCNME and optimal-transport-based noisy-label learning Okamura et al. (2024); Han et al. (2025). More recently, noisy multi-label cross-modal hashing has incorporated semantic consistency and annotation redundancy into robust representation learning Peng et al. (2026); Su et al. (2025). Despite these advances, how to exploit the multi-label structure exposed after supervision correction remains less explored.

This distinction is important because corrected multi-label supervision contains richer information than the binary relations ultimately used for hash retrieval. Two relevant instance pairs can exhibit substantially different degrees of overlap between their corrected label sets, whereas a binary relation records only whether they are related. Recent multi-label hashing methods have demonstrated the benefit of preserving finer semantic structure Li & Sun (2023); Liu et al. (2024); Yi et al. (2024), and SCBCH further explores semantic consistency under noisy multi-label supervision Peng et al. (2026). However, how to explicitly preserve and reuse corrected multi-label structure across continuous semantic initialization and subsequent binary hash refinement remains underexplored. Directly reducing corrected supervision to binary relations discards the degree of label overlap, while using corrected labels only for noise suppression leaves their fine-grained semantic structure underutilized. This motivates a sequential supervision transformation in which corrected multi-label supervision first provides a graded semantic signal and is then converted into the binary relation required by the final hash space.

To this end, we propose **Reliability-Based Cross-modal Deep Hashing (RBCDH)**, a sequential framework that connects reliable supervision construction, continuous semantic initialization, and binary-relation-guided cross-modal hash refinement. RBCDH first performs Precision First Bidirectional Label Correction (PF-BLC), which estimates instance–class-level reliability by combining intra-modal neighborhood support with cross-modal consistency and applies constrained deletion and addition to obtain fixed corrected multi-label supervision. Rather than treating the corrected

labels as clean ground truth, RBCDH uses them as a common supervision source for subsequent learning stages. Reliable Semantic Initialization converts corrected label overlap into continuous Soft-Jaccard relations, allowing fine-grained multi-label semantics to shape the image and text representations before binary refinement. The initialized networks are then refined using binary reliable relations derived from the same corrected supervision. In this way, RBCDH progressively transforms corrected multi-label supervision from a graded semantic representation into the binary relevance required for final hash retrieval.

The main contributions of this paper are summarized as follows:

- We propose **RBCDH**, a reliability-guided sequential framework that progressively exploits corrected multi-label supervision for semantic initialization and binary cross-modal hash refinement, using the same corrected supervision across continuous and binary learning stages.
- We develop **PF-BLC**, an instance–class-level bidirectional label correction method that integrates intra-modal neighborhood support and cross-modal consistency to remove low-reliability observed positives and supplement high-confidence missing positives under noisy multi-label supervision.
- We introduce **Reliable Semantic Initialization**, which preserves fine-grained multi-label semantics by modeling corrected label overlap with continuous Soft-Jaccard relations before final binary-relation refinement.

2 Related Work

2.1 Supervised Cross-modal Hashing

Early cross-modal hashing jointly learned multimodal features and codes through matrix factorization or shared representations Ding et al. (2014); Zhou et al. (2014) and modeled cross-modal semantic consistency Yao et al. (2016). DCMH Jiang & Li (2017) coupled deep modality representations with pairwise supervision. In multi-label settings, SSAH Li et al. (2018), CPAH Xie et al. (2020), and related methods Li & Sun (2023); Zhong et al. (2018) preserve associations through common semantics, modality consistency, or discrete optimization. Beyond binary relevance, RDCMH Liu et al. (2019) uses semantic ranking, DSFH Liu et al. (2024) combines coarse- and cluster-level fine-grained relations, and MLWCH Yi et al. (2024) uses multi-label similarity and weighted contrastive learning. Robust discrete multimodal hashing also combines robust representation learning with discrete optimization Li et al. (2019). These advances establish multi-label relation learning, but label noise can still distort relations derived from category labels.

2.2 Cross-modal Hashing and Label Correction under Noisy Supervision

Work on noisy cross-modal retrieval Mandal & Biswas (2020) spans reliable sample selection, reweighting, and robust hashing. Mutual Quantization Yang et al. (2022) selects samples from training dynamics, RSHNL Pu et al. (2025) uses self-paced learning, and MGSH Tan et al. (2026) meta-learns sample weights. NRCH Wang et al. (2024) studies robust contrastive hashing, while DHRL Shu et al. (2025) and HRDCH Wang et al. (2026a) model ranking and historical reliability, respectively generative hashing has also addressed noisy retrieval Wang et al. (2026b). For noisy multi-label data, SCBCH Peng et al. (2026) combines sample reliability estimation with continuous semantic relations, weighting supervision from neighborhood labels and modeling multi-label overlap. LCNME Okamura et al. (2024) corrects noisy labels, NACD Su et al. (2025) addresses redundant annotations in cross-modal hashing, NIRNL Liu et al. (2026) refines noisy instances with neighborhoods, and UOT-RCL Han et al. (2025) uses optimal transport. Beyond cross-modal retrieval, HLC Xia et al. (2023) exploits label dependencies for noisy multi-label classification, while iLaCo Xu et al. (2024) selects and corrects instance–label pairs in noisy multi-label text classification. RBCDH instead constructs fixed corrected supervision through instance–class bidirectional correction, then derives continuous Soft-Jaccard and binary reliable relations from it for sequential dual-stream initialization and final refinement.

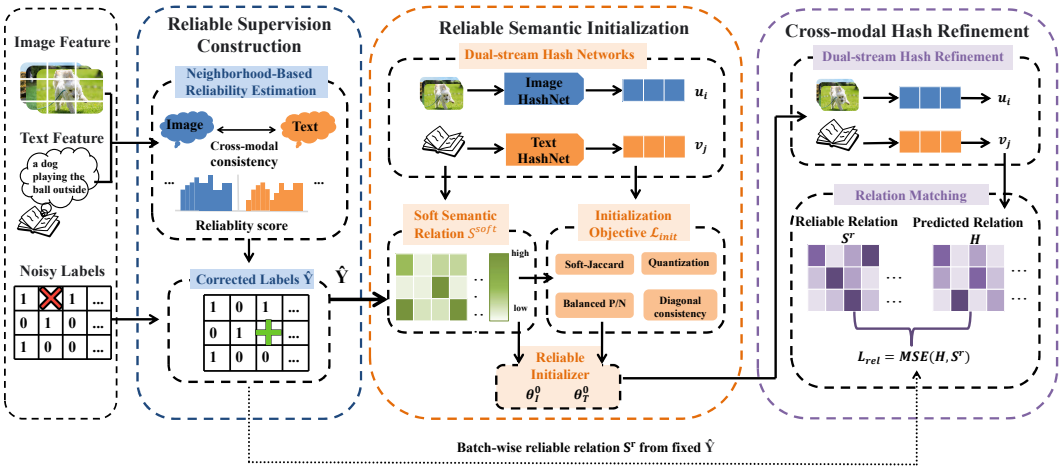


Figure 2: Overall framework of RBCDH. PF-BLC constructs corrected supervision, from which continuous Soft-Jaccard relations and binary reliable relations are sequentially used for semantic initialization and hash refinement.

3 METHOD

3.1 PROBLEM DEFINITION

Supervised multi-label cross-modal hashing learns compact binary representations by using label supervision to construct semantic relations across heterogeneous modalities. Let the training set be:

$$\mathcal{D} = \left\{ \left(\mathbf{x}_i^I, \mathbf{x}_i^T, \tilde{\mathbf{y}}_i \right) \right\}_{i=1}^{N}, \tag{1}$$

where $\mathbf{x}_i^I \in \mathbb{R}^{d_I}$ and $\mathbf{x}_i^T \in \mathbb{R}^{d_T}$ denote the fixed pre-extracted image and text input features, respectively $\tilde{\mathbf{y}}_i \in \{0,1\}^C$ denotes the observed multi-label annotation available during training, and C is the number of semantic categories. All observed labels form $\tilde{\mathbf{Y}} \in \{0,1\}^{N \times C}$.

For any two instances i and j, the initial binary semantic relationship obtained based on the observation tag is defined as:

$$\tilde{S}_{ij} = \mathbb{I}\left[\tilde{\mathbf{y}}_i^\top \tilde{\mathbf{y}}_j > 0 \right], \tag{2}$$

where $\mathbb{I}[\cdot]$ is the indicator function and $\tilde{\mathbf{S}} \in \{0,1\}^{N \times N}$ denotes the relation matrix induced by the observed labels.

However, obtaining more reliable multi-label supervision alone is not sufficient. If the corrected supervision is immediately reduced to a binary related/unrelated relation, different degrees of multi-label overlap are mapped to the same positive value, and the fine-grained semantic structure is lost. Therefore, the problem considered in this work is to construct reliable supervision from noisy multi-label annotations and to use the corrected multi-label structure for both continuous semantic modeling and final binary hash optimization. This motivates RBCDH to first reconstruct reliable supervision and then exploit it at different relation resolutions in the subsequent hashing stages.

3.2 OVERALL FRAMEWORK OF RBCDH

RBCDH addresses two coupled issues: noisy multi-label annotations can corrupt cross-modal relation supervision, while directly reducing corrected multi-label supervision to binary relations may discard fine-grained label overlap. As shown in Figure 2, RBCDH consists of three sequential stages: Reliable Supervision Construction, Reliable Semantic Initialization, and Cross-modal Hash Refinement.

Reliable Supervision Construction uses PF-BLC to combine dual-modal neighborhood support and cross-modal consistency, producing a fixed corrected label matrix $\hat{\mathbf{Y}}$. Reliable Semantic Initialization then starts from a newly initialized dual-stream hashing network and uses continuous Soft-

Jaccard relations derived from $\hat{\mathbf{Y}}$ to obtain (θ_I^0, θ_T^0). Cross-modal Hash Refinement further constructs binary reliable relations $\mathbf{S}^r$ from the same $\hat{\mathbf{Y}}$ and optimizes the final parameters (θ_I^*, θ_T^*).

The auxiliary Neighborhood Encoder is used only for neighborhood evidence construction and is not transferred to the subsequent hashing network. The continuous and binary relations are used sequentially rather than combined into a joint objective: the former preserves graded multi-label semantics during initialization, while the latter guides final hash refinement toward binary relevance.

3.3 Reliable Supervision Construction

Since noisy multi-label annotations directly corrupt relation supervision, RBCDH first corrects labels rather than $\tilde{\mathbf{S}}$. Motivated by neighborhood-based reliability estimation and label correction methods, PF-BLC combines dual-modal neighborhood support and cross-modal consistency to construct a fixed corrected label matrix $\hat{\mathbf{Y}}$.

3.3.1 Instance–Class-Level Label Reliability Estimation

PF-BLC trains an auxiliary Neighborhood Encoder on observed noisy labels. Its normalized outputs are used only for neighborhood construction and are not transferred to subsequent hash learning. The training data are divided into a training subset $\mathcal{D}_{\mathrm{tr}}$ and a held-out validation subset $\mathcal{D}_{\mathrm{val}}$, where the former corresponds to the `Fit` split in our implementation and the latter is used only for model selection.

For instance $i \in \mathcal{D}_{\mathrm{tr}}$ and class c, the neighborhood label support in modality $m \in \{I, T\}$ is

$$q_{ic}^m = \sum_{r=1}^{K_P} w_r \, \tilde{y}_{\mathcal{N}_i^m(r),c}, \qquad m \in \{I, T\}, \tag{3}$$

where $\mathcal{N}_i^m(r)$ is the r-th nearest neighbor in modality m, and w_r is the normalized rank-decay weight. The neighborhood settings are given in Appendix B.

We further define

$$\begin{aligned} Q_{ic} &= \frac{q_{ic}^I + q_{ic}^T}{2}, \\ A_{ic} &= 1 - \left| q_{ic}^I - q_{ic}^T \right|, \\ E_{ic} &= A_{ic} Q_{ic}, \end{aligned} \tag{4}$$

where Q_{ic} measures average dual-modal support, A_{ic} measures cross-modal consistency, and E_{ic} gives the combined semantic evidence at (i, c).

3.3.2 Bidirectional Label Correction

PF-BLC removes weakly supported observed positives and adds strongly supported missing positives. The removal candidates are defined as

$$\begin{aligned} C_{ic}^{\mathrm{remove}} = \mathbb{I}[\tilde{y}_{ic} = 1] \, \mathbb{I}\left[q_{ic}^I \le \tau_{\mathrm{rem}}^q\right] \\ \cdot \, \mathbb{I}\left[q_{ic}^T \le \tau_{\mathrm{rem}}^q\right] \mathbb{I}\left[E_{ic} \le \tau_{\mathrm{rem}}^E\right], \end{aligned} \tag{5}$$

and the addition candidates are defined as

$$\begin{aligned} C_{ic}^{\mathrm{add}} = \mathbb{I}[\tilde{y}_{ic} = 0] \, \mathbb{I}\left[q_{ic}^I \ge \tau_{\mathrm{add}}^q\right] \mathbb{I}\left[q_{ic}^T \ge \tau_{\mathrm{add}}^q\right] \\ \cdot \, \mathbb{I}\left[A_{ic} \ge \tau_{\mathrm{add}}^A\right] \mathbb{I}\left[E_{ic} \ge \tau_{\mathrm{add}}^E\right]. \end{aligned} \tag{6}$$

where the τ terms are reliability thresholds for removal and addition. Their values are reported in Appendix A.

Class- and instance-level budgets further limit the extent of correctionremoval candidates are ranked by ascending E_{ic}, while addition candidates are ranked by descending E_{ic}. The exact quotas and selection procedure are given in Appendix B.

The corrected label matrix is

$$\hat{\mathbf{Y}} = \left(\tilde{\mathbf{Y}} \wedge \neg \mathbf{M}^{\mathrm{remove}}\right) \vee \mathbf{M}^{\mathrm{add}}, \tag{7}$$

where $\mathbf{M}^{\text{remove}}$ and $\mathbf{M}^{\text{add}}$ are the final removal and addition masks. $\hat{\mathbf{Y}}$ is corrected supervision rather than ground-truth clean labels.

3.3.3 Reliable Relation Construction

For instances i and j, the binary reliable relation is

$$S_{ij}^{r} = \mathbb{I}\left[\hat{\mathbf{y}}_i^{\top}\hat{\mathbf{y}}_j > 0\right], \tag{8}$$

where $S_{ij}^{r} \in \{0, 1\}$ indicates whether the two instances share at least one corrected positive label. The mini-batch relation matrix $\mathbf{S}_{\mathcal{B}}^{r}$ is constructed on the fly from fixed $\hat{\mathbf{Y}}$.

3.4 Reliable Semantic Initialization

Binary relations cannot distinguish different degrees of multi-label overlap. RBCDH therefore uses continuous Soft-Jaccard relations for Reliable Semantic Initialization before final binary refinement. This stage starts from a newly initialized dual-stream hashing network and does not inherit Neighborhood Encoder parameters or representations.

For instances i and j,

$$S_{ij}^{\text{soft}} = \frac{\hat{\mathbf{y}}_i^{\top}\hat{\mathbf{y}}_j}{\max\left(\|\hat{\mathbf{y}}_i\|_1 + \|\hat{\mathbf{y}}_j\|_1 - \hat{\mathbf{y}}_i^{\top}\hat{\mathbf{y}}_j, \epsilon\right)}, \tag{9}$$

where $\epsilon > 0$ prevents division by zero.

The positive and negative relation losses $\mathcal{L}_{\text{pos}}$ and $\mathcal{L}_{\text{neg}}$ fit the continuous targets for $S_{ij}^{\text{soft}} > 0$ and $S_{ij}^{\text{soft}} = 0$, respectively. The quantization loss $\mathcal{L}_{\text{q}}$ and diagonal-consistency loss $\mathcal{L}_{\text{diag}}$ provide additional constraints, their definitions are given in Appendix C.

The initialization objective is

$$\mathcal{L}_{\text{init}} = \frac{1}{2}\left(\mathcal{L}_{\text{pos}} + \mathcal{L}_{\text{neg}}\right) + \lambda_{\text{q}}\mathcal{L}_{\text{q}} + \lambda_{\text{diag}}\mathcal{L}_{\text{diag}}, \tag{10}$$

where λ_{q} and λ_{diag} control the two auxiliary terms. The selected state on $\mathcal{D}_{\text{val}}$ provides (θ_I^0, θ_T^0) for refinement.

3.5 Cross-modal Hash Refinement

Reliable Semantic Initialization preserves graded semantic overlap, while the final retrieval stage requires binary relevance supervision. RBCDH therefore refines the initialized networks using $\mathbf{S}^{r}$.

For a mini-batch $\mathcal{B}$ of m pairs, let $\mathbf{U}_{\mathcal{B}}$ and $\mathbf{V}_{\mathcal{B}}$ denote the L_2-normalized image and text representations. The predicted relation matrix is

$$\mathbf{H}_{\mathcal{B}} = \frac{\mathbf{U}_{\mathcal{B}}\mathbf{V}_{\mathcal{B}}^{\top} + \mathbf{1}_{m\times m}}{2}, \tag{11}$$

where $\mathbf{1}_{m\times m}$ is the all-ones matrix.

The relation-matching loss is

$$\mathcal{L}_{\text{rel}} = \frac{1}{m^2}\left\|\mathbf{H}_{\mathcal{B}} - \mathbf{S}_{\mathcal{B}}^{r}\right\|_F^2. \tag{12}$$

Refinement starts from (θ_I^0, θ_T^0) and optimizes only $\mathcal{L}_{\text{rel}}$.

3.6 Overall Optimization and Retrieval

RBCDH adopts sequential rather than joint optimization:

$$\hat{\mathbf{Y}} \xrightarrow{\mathcal{L}_{\text{init}}} (\theta_I^0, \theta_T^0) \xrightarrow{\mathcal{L}_{\text{rel}}} (\theta_I^*, \theta_T^*). \tag{13}$$

The same corrected supervision is therefore used in two forms: continuous relations for semantic initialization and binary relations for final hash refinement.

After training, $\mathbf{b}_i^I = \mathrm{sign}(\mathbf{u}_i^*)$ and $\mathbf{b}_i^T = \mathrm{sign}(\mathbf{v}_i^*)$. Retrieval is performed by Hamming distance

$$d_H(\mathbf{b}_q, \mathbf{b}_d) = \frac{1}{2}\left(b - \mathbf{b}_q^\top \mathbf{b}_d\right)$$

for both I2T and T2I retrieval.

4 EXPERIMENTS

We evaluate RBCDH on NUS-WIDE TC-21 and MM-IMDb using the fixed random seeds 123, 2024, and 3407.

4.1 EXPERIMENTAL SETTINGS

All methods use identical data splits, fixed pre-extracted image and text features, noisy-label instances, and evaluation protocols. NUS-WIDE TC-21 contains 10,500 training, 2,100 query, and 193,734 database instances, while MM-IMDb contains 15,552 training and 7,799 evaluation instances. Only training labels are corrupted at $\rho \in \{0.2, 0.5, 0.8\}$, for each seed, $\lfloor \rho N \rfloor$ instances receive one $1 \to 0$ and one $0 \to 1$ flip. Clean training labels are unavailable for training or model selection. We report Hamming-ranking Avg mAP over I2T and T2I retrieval. Further protocol details are provided in Appendix A.

4.2 BASELINES

We compare RBCDH with NRCH, MGSH, and SCBCH, which represent recent robust learning strategies for cross-modal hashing under noisy supervision. Additional implementation and protocol details are provided in Appendix A.

4.3 OVERALL RETRIEVAL PERFORMANCE

We evaluate all methods at noise rates $\rho \in \{0.2, 0.5, 0.8\}$ and hash-code lengths of 16, 32, 64, and 128 bits on NUS-WIDE TC-21 and MM-IMDb. Each condition is independently evaluated using seeds 123, 2024, and 3407, and Table 1 reports the three-seed mean Avg mAP.

Overall, RBCDH achieves the highest Avg mAP in 23 of the 24 dataset–noise-rate–code-length settings, with an average Avg mAP of 62.57%, compared with 58.35%, 55.98%, and 55.42% for SCBCH, MGSH, and NRCH, respectively. On NUS-WIDE TC-21, RBCDH ranks first under all settings at $\rho = 0.2$ and $\rho = 0.5$. At $\rho = 0.8$, RBCDH remains best at 32, 64, and 128 bits; the only exception is the 16-bit setting, where it obtains 58.535%, slightly below SCBCH's 58.647% by 0.112 percentage points. One possible explanation is that, under severe supervision noise, the limited capacity of very short hash codes makes it more difficult to retain the graded semantic structure established during Soft-Jaccard-based initialization when the representations are further aligned with the final binary retrieval objective. This interpretation is consistent with RBCDH recovering the best performance at longer code lengths under the same noise level, although the current experiments do not establish a causal effect of code length. However, this observation does not establish a causal effect of code length; rather, it is consistent with such an explanation, since RBCDH again ranks first at longer code lengths under the same noise level.

On MM-IMDb, RBCDH obtains the highest Avg mAP in all 12 settings. Taken together, the results show that the advantage of RBCDH is observed across different noise rates and code lengths, while the contributions of its individual stages are further examined in the controlled ablation study.

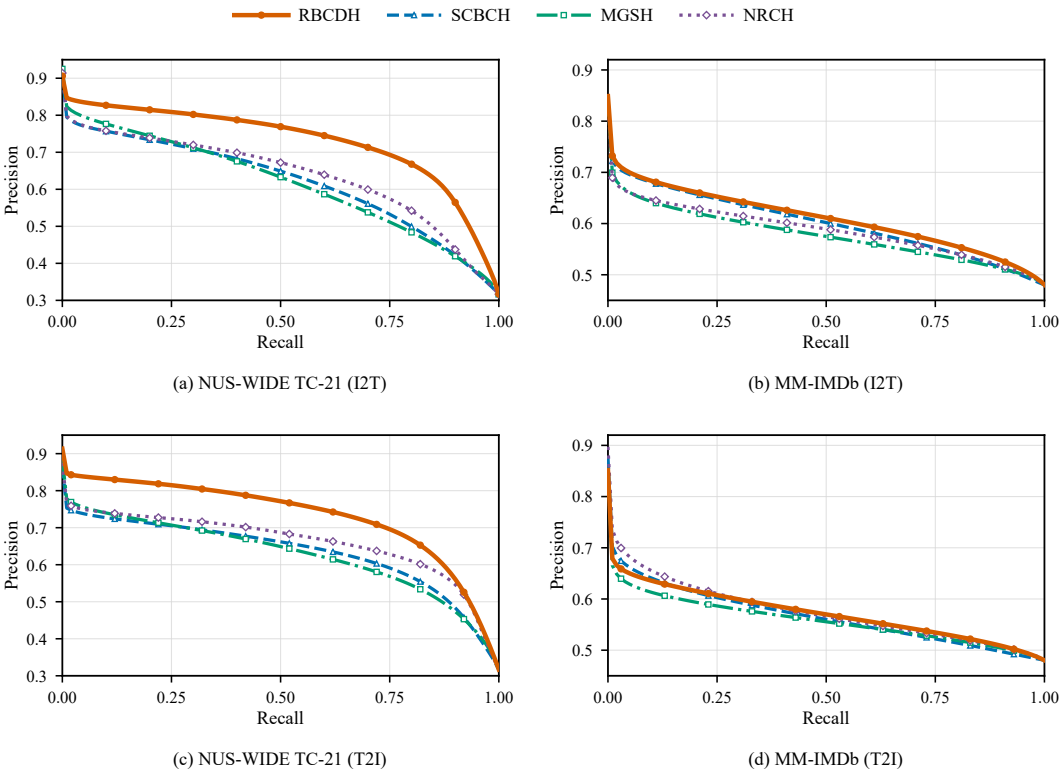
4.4 ABLATION STUDY

In order to analyze the role of reliable supervised construction, continuous semantic initialization and final hash refinement in RBCDH, this paper conducts controlled ablation experiments on NUS-WIDE TC-21 and MM-IMDb. All variants use the same fixed pre-extracted image and text features, data split, noisy-label instances, and evaluation protocol, the fixed noise rate is $rho = 0.5$, the hash code length is 64 bits, and three fixed random seeds are used. Except for the components to be

Table 1: Avg mAP (%) comparison of various methods on NUS-WIDE TC-21 and MM-IMDb under different noise rates and hash code lengths.

Dataset	Method	Year	Noise=0.2				Noise=0.5				Noise=0.8			
			16 bit	32 bit	64 bit	128 bit	16 bit	32 bit	64 bit	128 bit	16 bit	32 bit	64 bit	128 bit
NUS-WIDE TC-21	NRCH	2024	61.369	63.486	63.687	64.831	52.927	60.932	62.268	63.452	40.327	43.064	46.238	50.011
	MGSH	2026	57.014	59.716	61.417	64.353	53.724	56.570	58.267	61.572	52.130	54.479	56.025	59.601
	SCBCH	2026	64.715	65.757	65.684	66.029	61.454	63.419	62.304	61.512	**58.647**	60.150	56.932	60.425
	RBCDH	–	**70.709**	**72.295**	**73.495**	**72.982**	**65.912**	**70.502**	**71.780**	**72.467**	58.535	**62.322**	**64.397**	**65.262**
MM-IMDb	NRCH	2024	53.368	55.039	56.024	57.624	52.708	54.619	55.743	57.061	51.987	53.065	54.542	55.726
	MGSH	2026	51.515	51.440	53.635	54.213	54.013	54.711	55.174	55.642	53.895	53.953	54.886	55.577
	SCBCH	2026	50.827	56.384	57.583	57.985	51.087	55.094	56.410	56.807	46.216	53.721	55.350	55.835
	RBCDH	–	**56.433**	**57.819**	**58.568**	**58.964**	**55.501**	**56.481**	**57.418**	**57.937**	**54.143**	**55.100**	**56.094**	**56.506**

Note: The results in the table are the average value of Avg mAP under the fixed three random seeds and are reported as a percentage. Each method on the same data set adopts consistent data division, noisy tag instance and evaluation protocol. The optimal result is shown in bold, and the suboptimal result is underlined.



(a) NUS-WIDE TC-21 (I2T)

(b) MM-IMDb (I2T)

(c) NUS-WIDE TC-21 (T2I)

(d) MM-IMDb (T2I)

Figure 3: Precision–recall curves on NUS-WIDE TC-21 and MM-IMDb at $\rho = 0.5$ with 128-bit hash codes. Results are obtained under the same fixed three-seed protocol used in the main experiments.

analyzed, the rest of the training settings are consistent, and the results are reported as the standard deviation of the three sub means pm of Avg mAP (%).

Specifically, w/o PF-BLC removes the reliable supervision construction, but retains the following two stages of learning. At this time, the continuous Soft-Jaccard relationship and the final binary relationship are both constructed from the original noisy label to investigate the role of corrected supervision in the complete training path. w/o Semantic Init. and w/Binary Init. keep the PF-BLC and complete two-stage training, but replace the continuous Soft-Jaccard target in the initialization phase with a binary relationship, which is used to compare the difference between the continuous multi label overlap relationship and the binary relationship as the initialization supervision, w/o

Table 2: Results of controlled ablation of RBCDH. All results are reported as mean ± standard deviation of Avg mAP (%) over three fixed random seeds at $\rho = 0.5$ with 64-bit codes.

Configuration	PF-BLC	Init. Target	Hash Refine.	NUS-WIDE TC-21	MM-IMDb
w/o PF-BLC	–	Noisy Soft-Jaccard	✓	71.238 ± 0.408	56.849 ± 0.271
w/o Semantic Init.	✓	–	✓	54.100 ± 0.524	52.654 ± 2.543
w/ Binary Init.	✓	Binary Relation	✓	63.264 ± 1.102	56.900 ± 0.909
w/o Hash Refinement	✓	Soft-Jaccard	–	$\mathbf{72.508 \pm 0.255}$	57.045 ± 0.039
Full RBCDH	✓	Soft-Jaccard	✓	71.751 ± 0.512	$\mathbf{57.553 \pm 0.242}$

Hash Refinement retains the reliable semantic initialization of PF-BLC and Soft-Jaccard, but does not perform the final binary relationship refinement, and directly evaluates the selected initialization model. Full RBCDH executes PF-BLC, Soft-Jaccard reliable semantic initialization and final cross-modal hash refinement in turn.

Reliable Supervision Construction. Compared with w/o PF-BLC, Full RBCDH's Avg mAP on NUS-WIDE TC-21 and MM-IMDb increased by 0.512 and 0.704 percentage points respectively, indicating that when the rest of the training process is consistent, the correction supervision built by PF-BLC can reduce the impact of observation noise on subsequent relationship learning.

Continuous Semantic Initialization. W/Binary Init. only replaces the initialization target from Soft-Jaccard with a binary relationship. Compared to this variant, Full RBCDH increased by 8.487 and 0.653 percentage points on NUS-WIDE TC-21 and MM-IMDb, respectively, indicating that on the basis of corrected supervision, the continuous Soft-Jaccard relationship can more fully utilize multi label overlap information. W/o Semantic Init. directly learning the modified binary relationship from a random initialization state, the results are 54.100 and 52.654, respectively. Due to the fact that this variant simultaneously changes the initialization method and training path, its difference from Full RBCDH is only used as an auxiliary reference and should not be interpreted as an independent gain for reliable semantic initialization.

Final Hash Refinement. Compared with w/o Hash Refinement, Full RBCDH increased by 0.509 percentage points on MM-IMDb and decreased by 0.757 percentage points on NUS-WIDE TC-21. This indicates that the performance changes in the refinement of the final binary relationship are dataset dependent. Therefore, this paper considers it as the alignment stage after continuous semantic initialization towards the final binary correlation target, rather than describing it as a module with consistent independent gains.

5 Conclusion

We proposed RBCDH for noisy multi-label cross-modal hashing, which combines reliable supervision construction with fine-grained semantic initialization and binary hash refinement. PF-BLC constructs corrected supervision by integrating intra-modal neighborhood support and cross-modal consistency, while continuous Soft-Jaccard relations preserve graded multi-label semantics before final binary-relation refinement. Experiments on NUS-WIDE TC-21 and MM-IMDb demonstrate the effectiveness of the proposed framework under different noise rates and hash-code lengths, while controlled ablations further support the roles of corrected supervision and continuous semantic initialization. The effect of final hash refinement is dataset dependent, suggesting that future work may further investigate more effective connections between continuous semantic structure and binary retrieval objectives.

AI Use Statement

In this study, we used ChatGPT solely for language polishing and improving the clarity of the manuscript. The authors reviewed and approved all AI-assisted revisions and take full responsibility for the final content of the paper.

REFERENCES

John Arevalo, Thamar Solorio, Manuel Montes-y Gomez, and Fabio A. Gonzalez. Gated multimodal units for information fusion, 2017. URL https://openreview.net/forum?id=Hy-2G6ile. OpenReview paper and arXiv preprint; source repository: https://github.com/johnarevalo/gmu-mmimdb.

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 8th ACM International Conference on Image and Video Retrieval*, pp. 1–9. ACM, 2009. doi: 10.1145/1646396.1646452.

Guiguang Ding, Yuchen Guo, and Jile Zhou. Collective matrix factorization hashing for multimodal data. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2083–2090, 2014. doi: 10.1109/CVPR.2014.267.

Amirmasoud Ghiassi, Robert Birke, and Lydia Y. Chen. Multi label loss correction against missing and corrupted labels. In *Proceedings of the Asian Conference on Machine Learning*, volume 189 of *Proceedings of Machine Learning Research*, pp. 359–374, 2022.

Haochen Han, Minnan Luo, Huan Liu, Fang Nan, and Jun Liu. A unified optimal transport framework for cross-modal retrieval with noisy labels. *IEEE Transactions on Neural Networks and Learning Systems*, 36(9):16435–16448, 2025. doi: 10.1109/TNNLS.2025.3559533.

Peng Hu, Xi Peng, Hongyuan Zhu, Liangli Zhen, and Jie Lin. Learning cross-modal retrieval with noisy labels. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5399–5409, 2021. doi: 10.1109/CVPR46437.2021.00536.

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

Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao. Self-supervised adversarial hashing networks for cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4242–4251, 2018. doi: 10.1109/CVPR.2018.00446.

Chuan-Xiang Li, Ting-Kun Yan, Xin Luo, Liqiang Nie, and Xin-Shun Xu. Supervised robust discrete multimodal hashing for cross-media retrieval. *IEEE Transactions on Multimedia*, 21(11): 2863–2877, 2019. doi: 10.1109/TMM.2019.2912714.

Liang Li and Weiwei Sun. Label-wise deep semantic-alignment hashing for cross-modal retrieval. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pp. 416–424, 2023. doi: 10.1145/3591106.3592283.

Kaiming Liu, Yunhong Gong, Yu Cao, Zhenwen Ren, Dezhong Peng, and Yuan Sun. Dual semantic fusion hashing for multi-label cross-modal retrieval. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 4569–4577, 2024.

Xuanwu Liu, Guoxian Yu, Carlotta Domeniconi, Jun Wang, Yazhou Ren, and Maozu Guo. Ranking-based deep cross-modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4400–4407, 2019.

Yizhi Liu, Ruitao Pu, Shilin Xu, Yingke Chen, Quan-Hui Liu, and Yuan Sun. Neighbor-aware instance refining with noisy labels for cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23908–23916, 2026. doi: 10.1609/aaai.v40i28.39567.

Devraj Mandal and Soma Biswas. Cross-modal retrieval with noisy labels. In *2020 IEEE International Conference on Image Processing*, pp. 2326–2330, 2020. doi: 10.1109/ICIP40778.2020.9190981.

Daiki Okamura, Ryosuke Harakawa, and Masahiro Iwahashi. LCNME: Label correction using network prediction based on memorization effects for cross-modal retrieval with noisy labels. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(1):590–602, 2024. doi: 10.1109/TCSVT.2023.3286546.

Likang Peng, Chao Su, Wenyuan Wu, Yuan Sun, Dezhong Peng, Xi Peng, and Xu Wang. Semantic-consistent bidirectional contrastive hashing for noisy multi-label cross-modal retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(29):24811–24819, 2026. doi: 10.1609/aaai.v40i29.39667.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(19):19969–19977, 2025. doi: 10.1609/aaai.v39i19.34199.

Zhenqiu Shu, Yibing Bai, Kailing Yong, and Zhengtao Yu. Deep cross-modal hashing with ranking learning for noisy labels. *IEEE Transactions on Big Data*, 11(2):553–565, 2025. doi: 10.1109/TBDATA.2024.3423704.

Chao Su, Likang Peng, Yuan Sun, Dezhong Peng, Xi Peng, and Xu Wang. Neighbor-aware contrastive disambiguation for cross-modal hashing with redundant annotations. In *Advances in Neural Information Processing Systems*, volume 38, pp. 120168–120190, 2025. doi: 10.52202/085713-3620.

Ziang Tan, Weitao An, and Erkun Yang. Meta-guided sample reweighting for robust cross-modal hashing retrieval with noisy labels. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(11):9350–9358, 2026. doi: 10.1609/aaai.v40i11.37894.

Jun Wang, Ting Zhang, Nicu Sebe, and Heng Tao Shen. Learning to hash for indexing big data: A survey. *Proceedings of the IEEE*, 104(1):34–57, 2016.

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*, pp. 5752–5760, 2024. doi: 10.1145/3664647.3680564.

Rui Wang, Haixiao Huang, Likang Peng, Chao Su, Zijie Zhong, Da Rao, Wenyuan Wu, Dezhong Peng, and Xu Wang. Historical reliability-based dual contrastive hashing for robust cross-modal retrieval with noisy labels. *Neurocomputing*, 678:133193, 2026a. doi: 10.1016/j.neucom.2026.133193.

Runmin Wang, Guoxian Yu, Hong Zhang, Maozu Guo, Lizhen Cui, and Xiangliang Zhang. Noise-robust deep cross-modal hashing. *Information Sciences*, 581:136–154, 2021. doi: 10.1016/j.ins.2021.09.030.

Zequn Wang, Tianshi Wang, Fengling Li, Jingjing Li, and Lei Zhu. Noise-robust generative hashing for cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 22(4):1–22, 2026b. doi: 10.1145/3777477.

Xiaobo Xia, Jiankang Deng, Wei Bao, Yuxuan Du, Bo Han, Shiguang Shan, and Tongliang Liu. Holistic label correction for noisy multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1483–1493, 2023.

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. doi: 10.1109/TIP.2020.2963957.

Pengyu Xu, Mingyang Song, Linkaida Liu, Bing Liu, Hongjian Sun, Liping Jing, and Jian Yu. Noisy multi-label text classification via instance-label pair correction. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1446–1458, 2024.

Tianyuan Xu, Xueliang Liu, Zhen Huang, Dan Guo, Richang Hong, and Meng Wang. Early-learning regularized contrastive learning for cross-modal retrieval with noisy labels. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 629–637, 2022. doi: 10.1145/3503161.3548066.

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*, pp. 7541–7550, 2022. doi: 10.1109/CVPR52688.2022.00740.

Tao Yao, Xiangwei Kong, Haiyan Fu, and Qi Tian. Semantic consistency hashing for cross-modal retrieval. *Neurocomputing*, 193:250–259, 2016. doi: 10.1016/j.neucom.2016.02.016.

Zeqian Yi, Xinghui Zhu, Runbing Wu, Zhuoyang Zou, Yi Liu, and Lei Zhu. Multi-label weighted contrastive cross-modal hashing. *Applied Sciences*, 14(1):93, 2024. doi: 10.3390/app14010093.

Fangming Zhong, Zhikui Chen, and Geyong Min. Deep discrete cross-modal hashing for cross-media retrieval. *Pattern Recognition*, 83:64–77, 2018. doi: 10.1016/j.patcog.2018.05.018.

Jile Zhou, Guiguang Ding, and Yuchen Guo. Latent semantic sparse hashing for cross-modal similarity search. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 415–424, 2014. doi: 10.1145/2600428.2609610.

APPENDIX

A IMPLEMENTATION DETAILS AND TRAINING PROTOCOL

RBCDH uses fixed pre-extracted image and text features as network inputs,without updating the upstream modality-specific feature extractors end to end. For NUS-WIDE TC-21, we use the FQ-PNGR-C fused feature protocol, with image/text input dimensions of $4224/1512$. For MM-IMDb, we use the Qwen/ColQwen Full-Fusion feature protocol, with image/text input dimensions of $4224/812$. These fused representations are generated before RBCDH training and remain fixed throughout all subsequent training stages. Both the image and text branches employ a three-layer fully connected architecture of $d \to 1024 \to 512 \to b$. Each of the first two linear transformations is followed by Batch Normalization, ReLU, and Dropout with a probability of 0.2. The output of the final layer is passed through a hyperbolic tangent function and per-instance L_2 normalization to obtain continuous hash representations. The batch size is set to 128 for all training stages. During Reliable Supervision Construction, the base image and text networks are first trained using the observed noisy labels, and their model states are selected according to an internal Validation objective. The image and text representations produced by the selected base models are used only to construct the intra-modal neighborhoods for PF-BLC. They are not used as inputs to Reliable Semantic Initialization, and no network parameters are transferred to subsequent stages. The base dual-stream networks use the Adam optimizer with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}, following the same internal Validation selection and early-stopping rules as Reliable Semantic Initialization.

Unless otherwise stated, the default RBCDH configuration uses $K_P = 30$, $\tau^q_{\mathrm{rem}} = 0.25$, $\tau^E_{\mathrm{rem}} = 0.12$, $\tau^q_{\mathrm{add}} = 0.85$, $\tau^A_{\mathrm{add}} = 0.90$, $\tau^E_{\mathrm{add}} = 0.80$, $\beta_{\mathrm{rem}} = 0.07$, $\beta_{\mathrm{add}} = 0.02$, and $\beta_{\mathrm{inst}} = 0.5$. These default settings are used in the main and ablation experiments. The parameter-sensitivity study in Appendix F varies only K_P or β_{rem} while keeping the remaining settings fixed.

During Reliable Semantic Initialization, the image and text networks are reinitialized and trained using the same fixed pre-extracted modality features and the corrected labels generated by PF-BLC. Reliable Semantic Initialization uses the objective in Eq. equation 10, with ith $\lambda_q = 0.01$ and $\lambda_{\mathrm{diag}} = 0.05$. The Soft-Jaccard numerical-stability constant is fixed to $\epsilon = 10^{-8}$. These values are shared across all formal experiments and are not tuned separately for individual datasets, noise rates, or code lengths. This stage uses the Adam optimizer with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. During training, the complete Reliable Semantic Initialization objective is evaluated on the internal Validation set, and the image and text network states are selected according to this objective. Reliable Semantic Initialization is trained for at most 50 epochs. After at least 10 epochs, training is terminated early if the objective fails to improve by more than 10^{-6} over eight consecutive Validation checks. When objective values are tied, the model state obtained earlier is retained.

During Dual-Stream Cross-Modal Hash Refinement, the complete parameter states of the image and text networks are loaded from the checkpoint selected during Reliable Semantic Initialization. The optimizer state from the preceding stage is not inherited, instead, a new optimizer is constructed to continue training. The only explicit training objective at this stage is the relation matching loss $\mathcal{L}_{\mathrm{rel}}$ defined in Eq. equation 12. The quantization and diagonal-consistency terms used during Reliable Semantic Initialization are no longer included. The Adam optimizer is used with a learning rate of

1×10^{-5}, a weight decay of 1×10^{-5}, a gradient-norm clipping threshold of 5.0, and a maximum of 10 training epochs. The model is selected according to the relation matching objective on the internal Validation set. After model selection, neither the network parameters nor the training configuration is adjusted according to retrieval results on the query or database sets.

A.1 Dataset Sources and Fixed Splits

Experiments are conducted on two multi-label cross-modal datasets:NUS-WIDE TC-21 and MM-IMDb.

NUS-WIDE contains large-scale web images and their associated textual tags Chua et al. (2009). We use its 21-class subset, NUS-WIDE TC-21. Under the fixed protocol, this subset contains 10,500 training instances, 2,100 query instances, and 193,734 database instances. The training set consists of 9,450 Fit instances and 1,050 internal Validation instances. MM-IMDb consists of movie-related visual and textual information, with each instance associated with multiple semantic categories Arevalo et al. (2017). The fixed protocol uses 15,552 training instances, of which 13,969 are assigned to the Fit set and 1,583 to the internal Validation set. An additional 7,799 frozen evaluation instances serve simultaneously as the query and database sets under this protocol.

A.2 Evaluation Protocol

We use mean Average Precision (mAP) as the primary retrieval-performance metric and report results separately for Image-to-Text (I2T) and Text-to-Image (T2I) retrieval.

For a given query instance, its Average Precision (AP) is defined as

$$\mathrm{AP} = \frac{1}{R} \sum_{k=1}^{N} P(k)\delta(k), \tag{14}$$

where N is the total number of retrieved instances in the database, R is the total number of database instances semantically relevant to the current query, and $P(k)$ is the precision among the top-k retrieval results. If the database instance at rank k is semantically relevant to the query, $\delta(k) = 1$, otherwise, $\delta(k) = 0$.

For the Q valid query instances that each have at least one relevant database instance, mAP is defined as

$$\mathrm{mAP} = \frac{1}{Q} \sum_{q=1}^{Q} \mathrm{AP}_q. \tag{15}$$

In the multi-label retrieval setting, a query instance and a database instance are considered semantically relevant if they share at least one ground-truth semantic label. All retrieval results are based on the final binary hash codes and are ranked and evaluated using Hamming distance. For MM-IMDb, retrieval results are ordered by ascending Hamming distance. When multiple database instances have the same Hamming distance, their order is determined by ascending fixed database row index. Under the fixed evaluation protocol for this dataset, the self-match between a query instance and its corresponding database instance is not additionally excluded. We compute mAP separately for I2T and T2I and define their average, Avg mAP, as the primary metric for overall performance comparison:

$$\text{Avg mAP} = \frac{\mathrm{mAP}_{\mathrm{I2T}} + \mathrm{mAP}_{\mathrm{T2I}}}{2}. \tag{16}$$

Comparisons among different methods are based on descriptive statistics, without statistical significance tests between methods.

B Detailed PF-BLC Correction Procedure

B.1 Neighborhood Encoder Representations and Neighborhood Construction

Let the image and text branches of the selected Neighborhood Encoder (NE) be denoted by g_I^{NE} and g_T^{NE}, respectively. For the i-th instance in the Fit set, its base image and text representations are

defined as

$$\mathbf{z}_i^{I,0} = \frac{\tanh\big(g_I^{\mathrm{NE}}(\mathbf{x}_i^I)\big)}{\big\|\tanh\big(g_I^{\mathrm{NE}}(\mathbf{x}_i^I)\big)\big\|_2}, \qquad \mathbf{z}_i^{T,0} = \frac{\tanh\big(g_T^{\mathrm{NE}}(\mathbf{x}_i^T)\big)}{\big\|\tanh\big(g_T^{\mathrm{NE}}(\mathbf{x}_i^T)\big)\big\|_2}. \tag{17}$$

Because these representations are already L_2-normalized, PF-BLC measures the proximity between instances within the same modality using their vector inner products. For each instance, its inner-product similarities with all instances in the same Fit set are computed separately in the image and text spaces, and the Top-(K_P+1) indices are retrieved. The actual implementation discards the first returned entry in each row and uses the remaining K_P entries to construct the image neighborhood $\mathcal{N}_i^I$ and text neighborhood $\mathcal{N}_i^T$. We set $K_P = 30$ throughout all formal experiments. To assign greater label contributions to higher-ranked neighbors, PF-BLC uses the normalized rank-decay weights

$$w_r = \frac{1/\log_2(r+1)}{\sum_{s=1}^{K_P} 1/\log_2(s+1)}, \qquad r = 1, \ldots, K_P, \tag{18}$$

which satisfy

$$w_r \geq 0, \qquad \sum_{r=1}^{K_P} w_r = 1. \tag{19}$$

B.2 Correction Budgets and Label Update

To limit the extent of each label-correction operation, PF-BLC further imposes class-level and instance-level constraints on the candidate sets described above. Let the number of positive labels for class c in the Fit set before correction be

$$n_c^+ = \sum_i \tilde{y}_{ic}. \tag{20}$$

The numbers of labels that may be deleted from and added to this class are then bounded by

$$R_c^{\mathrm{remove}} = \left\lfloor \beta_{\mathrm{rem}}\, n_c^+ \right\rfloor, \qquad R_c^{\mathrm{add}} = \left\lfloor \beta_{\mathrm{add}}\, n_c^+ \right\rfloor. \tag{21}$$

Among the deletion candidates, positions with smaller E_{ic} values are prioritized. Among the addition candidates, positions with larger E_{ic} values are prioritized. Let the number of positive labels associated with instance i before correction be

$$k_i = \sum_c \tilde{y}_{ic}. \tag{22}$$

The maximum number of labels that may be deleted from this instance is

$$R_i^{\mathrm{remove}} = \min\left\{\max(k_i - 1, 0), \lfloor \beta_{\mathrm{inst}} k_i \rfloor\right\}, \tag{23}$$

which limits the removal of existing positive labels from any individual instance. For label addition, at most one class is added to each instance. When multiple addition candidates exist, the candidate with the largest E_{ic} is retained. For an instance that has at least one positive label before correction, the deletion constraint ensures that it does not lose all positive labels through label removal.

Given the actual deletion mask $\mathbf{M}^{\mathrm{remove}}$ and addition mask $\mathbf{M}^{\mathrm{add}}$, the corrected instance–class labels are defined as

$$\hat{y}_{ic} = \begin{cases} 0, & \tilde{y}_{ic} = 1 \text{ and } M_{ic}^{\mathrm{remove}} = 1, \\ 1, & \tilde{y}_{ic} = 0 \text{ and } M_{ic}^{\mathrm{add}} = 1, \\ \tilde{y}_{ic}, & \text{otherwise.} \end{cases} \tag{24}$$

C Reliable Semantic Initialization and Loss Details

C.1 Dual-Stream Representations

Suppose that the current batch contains m paired image–text instances and that the hash-code length is b. During Reliable Semantic Initialization, the continuous representations produced by the image

and text networks are defined as

$$\mathbf{u}_i^0 = \frac{\tanh\big(g_I(\mathbf{x}_i^I;\theta_I)\big)}{\big\|\tanh\big(g_I(\mathbf{x}_i^I;\theta_I)\big)\big\|_2}, \qquad \mathbf{v}_i^0 = \frac{\tanh\big(g_T(\mathbf{x}_i^T;\theta_T)\big)}{\big\|\tanh\big(g_T(\mathbf{x}_i^T;\theta_T)\big)\big\|_2}. \tag{25}$$

For image instance i and text instance j in the batch, the predicted cross-modal relation is

$$H_{ij}^0 = \frac{(\mathbf{u}_i^0)^\top \mathbf{v}_j^0 + 1}{2}. \tag{26}$$

C.2 Positive and Negative Relation Losses

Based on the continuous relations constructed using Eq. equation 9, we define the positive and negative relation sets as

$$\mathcal{P} = \left\{(i,j) \mid S_{ij}^{\text{soft}} > 0\right\}, \qquad \mathcal{N} = \left\{(i,j) \mid S_{ij}^{\text{soft}} = 0\right\}. \tag{27}$$

When the corresponding relation set is nonempty, the positive and negative relation-fitting losses are computed separately as

$$\mathcal{L}_{\text{pos}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \left(H_{ij}^0 - S_{ij}^{\text{soft}}\right)^2, \qquad \mathcal{L}_{\text{neg}} = \frac{1}{|\mathcal{N}|} \sum_{(i,j)\in\mathcal{N}} \left(H_{ij}^0\right)^2. \tag{28}$$

When $\mathcal{P}$ or $\mathcal{N}$ is empty, the corresponding loss term is defined as zero. The positive and negative relation losses are averaged separately so that their contributions to the initialization objective are not directly determined by the numbers of relation pairs in the two sets.

C.3 Quantization and Diagonal Consistency

Because the continuous representations are L_2-normalized, the absolute value of each dimension of a normalized binary code is $1/\sqrt{b}$. The quantization loss is defined as

$$\mathcal{L}_{\text{q}} = \frac{1}{mb} \sum_{i=1}^{m} \sum_{k=1}^{b} \left(|u_{ik}^0| - \frac{1}{\sqrt{b}}\right)^2 + \frac{1}{mb} \sum_{i=1}^{m} \sum_{k=1}^{b} \left(|v_{ik}^0| - \frac{1}{\sqrt{b}}\right)^2. \tag{29}$$

For the image and text representations of the same instance, the diagonal-consistency loss is defined as

$$\mathcal{L}_{\text{diag}} = \frac{1}{m} \sum_{i=1}^{m} \left[1 - (\mathbf{u}_i^0)^\top \mathbf{v}_i^0\right]. \tag{30}$$

The relation-fitting, quantization, and diagonal-consistency constraints jointly constitute the Reliable Semantic Initialization objective defined in Eq. equation 10 in the main text.

D Additional Reliability Diagnostics

As an additional read-only post-hoc reliability diagnostic, we analyze NUS-WIDE TC-21 and MM-IMDb at $\rho = 0.5$ with 128-bit codes over the three fixed model seeds 123, 2024, and 3407. For each dataset and model seed, we independently sample 200,000 ordered off-diagonal pairs with replacement from all formal training rows. The noisy, corrected, and pre-corruption reference labels are evaluated on exactly the same pair indices within each dataset–seed configuration. The

Table 3: Post-hoc reliability diagnostics of PF-BLC at $\rho = 0.5$ with 128-bit codes. Results are averaged over seeds 123, 2024, and 3407.

Dataset	Label Error	Precision	Recall	F1	FPR
NUS-WIDE (Noisy)	0.04762	0.78311	0.63382	0.70060	0.08264
NUS-WIDE (Corrected)	0.04595	0.80451	0.60909	0.69329	0.06968
MM-IMDb (Noisy)	0.04348	0.86661	0.67364	0.75804	0.09765
MM-IMDb (Corrected)	0.04368	0.88124	0.63150	0.73575	0.08015

pre-corruption labels are used only for this post-hoc diagnostic and do not participate in training, hyperparameter selection, or model selection.

On NUS-WIDE TC-21, the mean element-wise label error decreases from 0.04762 to 0.04595. Relation precision increases from 0.78311 to 0.80451, while the false-positive rate decreases from 0.08264 to 0.06968. Meanwhile, recall decreases from 0.63382 to 0.60909 and F1 from 0.70060 to 0.69329.

On MM-IMDb, the mean element-wise label error changes slightly from 0.04348 to 0.04368. Relation precision increases from 0.86661 to 0.88124, and the false-positive rate decreases from 0.09765 to 0.08015, while recall decreases from 0.67364 to 0.63150 and F1 from 0.75804 to 0.73575.

These results show that PF-BLC changes the composition of relation errors in a manner consistent with its precision-first design: it reduces false-positive relations and increases relation precision, while trading off part of the relation recall. The diagnostic does not imply recovery of the ground-truth clean supervision or uniform improvement across all relation-quality metrics.

E Random-Seed Robustness

To further assess run-to-run variability, we evaluate the complete RBCDH model on NUS-WIDE TC-21 using 10 random seeds at a noise rate of 0.5 with 128-bit codes. The seeds are 42, 123, 777, 1001, 2024, 2025, 3407, 27182, 31415, and 65537.

The resulting Avg mAP is 0.724126 ± 0.006179, with values ranging from 0.710602 to 0.733058 and a median of 0.724488.

The 10-seed mean remains close to the result obtained with the original three fixed seeds, indicating that the reported NUS-WIDE TC-21 performance is not determined by a single random seed configuration.

F Parameter Sensitivity Analysis

We evaluate the sensitivity of RBCDH to the neighborhood size $K_P \in \{20, 30, 40\}$ and the class-level removal budget $\beta_{\text{rem}} \in \{0.05, 0.07, 0.10\}$ on NUS-WIDE TC-21 at a noise rate of 0.5 with 128-bit codes. Each configuration is evaluated using seeds 123, 2024, and 3407. The maximum variation in mean Avg mAP across the tested K_P values is 0.051 percentage points. For β_{rem}, the 5% budget yields lower performance, whereas the 7% and 10% settings remain close. The adopted values $K_P = 30$ and $\beta_{\text{rem}} = 0.07$ therefore lie within stable regions of the tested parameter ranges.