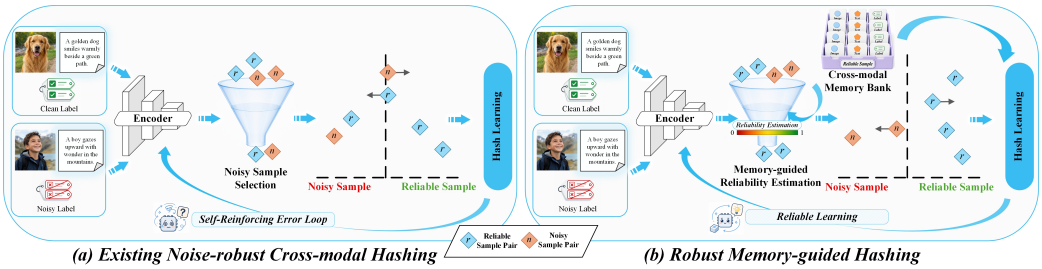
ROBUST MEMORY-GUIDED HASHING FOR CROSS-MODAL RETRIEVAL WITH NOISY LABELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-modal hashing (CMH) provides an efficient solution for large-scale multimodal retrieval, but its performance is vulnerable to noisy labels. Existing noise-robust CMH methods typically estimate sample or label reliability based on signals derived from the current model, such as training losses, prediction confidence, neighborhood relations, or learned semantic representations. However, these signals are themselves affected by noisy supervision, and biases in reliability estimation can be progressively amplified during iterative optimization, leading to noise propagation and error accumulation. To address this issue, we propose Robust Memory-guided Hashing (RMH), which exploits semantic information accumulated from historically reliable samples to provide stable references for current reliability estimation and hash learning. Specifically, Memory-guided Reliability Estimation (MRE) constructs modality-specific historical class centers from historically reliable samples, estimates the reliability of current samples through bidirectional cross-modal matching, and employs a smooth threshold to partition samples into reliable and unreliable subsets while updating the memory only with reliable samples. Furthermore, we develop Reliable Cross-batch Learning (RCL), which jointly exploits in-batch and historical cross-batch semantic relations for learning from reliable samples, while introducing an uncertainty-aware loss to attenuate potentially erroneous supervision associated with unreliable samples, thereby improving data utilization and the robustness of hash representations. Extensive experiments under multiple noise patterns and noise rates demonstrate that RMH achieves consistently competitive retrieval performance.



(a) Existing Noise-robust Cross-modal Hashing *(b) Robust Memory-guided Hashing*

Figure 1: Motivation of the proposed RMH framework. (a) Most existing noise-robust CMH methods rely on current-model signals for reliability estimation. Under noisy supervision, biased representations can lead to erroneous reliability estimation, allowing noisy samples mistakenly retained as reliable to re-enter optimization and further distort the representation space, forming a self-reinforcing error loop. (b) RMH exploits semantic information accumulated from historically reliable samples as a more stable reference for reliability estimation, which further guides robust hash learning to suppress noise propagation and error accumulation.

1 INTRODUCTION

Cross-modal hashing (CMH) has become an important technique for large-scale multimodal retrieval by mapping heterogeneous data into compact binary codes, thereby reducing storage costs and improving similarity search efficiency (Zhu et al., 2023; Wang et al., 2025; Sun et al., 2026a).

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.

Most supervised CMH methods rely on semantic annotations to learn modality-invariant and similarity-preserving representations (Qin et al., 2025b; 2026). In practical applications, however, obtaining complete and accurate annotations is costly, and training labels are often corrupted by annotation errors, semantic ambiguity, or imperfect data collection processes (Song et al., 2022; Peng et al., 2026). Such noisy supervision can distort cross-modal semantic relations and induce incorrect similarity structures, ultimately degrading the quality of hash representations and retrieval performance (Pu et al., 2025; Sun et al., 2026b). Therefore, improving the robustness of CMH under noisy-label supervision remains an important problem for reliable cross-modal retrieval.

To mitigate the adverse effects of noisy labels, existing noise-robust CMH methods mainly adopt three strategies: robust objectives (Pu et al., 2025), label correction (Yang et al., 2025), and sample selection (Wang et al., 2024; Liu et al., 2026; Peng et al., 2026). Robust objectives reduce the influence of potentially corrupted supervision on optimization; label-correction methods aim to recover more reliable semantic annotations; and sample-selection methods identify trustworthy samples to provide relatively reliable supervision for subsequent model training. However, when explicitly estimating sample or label reliability, existing methods often rely on signals derived from the current model, including training losses, prediction confidence, neighborhood relations, and learned semantic representations. Since these signals are themselves obtained under noisy supervision, their biases tend to accumulate as the model progressively fits noisy labels, thereby degrading the accuracy and stability of reliability estimation (Wang et al., 2020; Liu et al., 2021).

As illustrated in Figure 1 (a), this dependency can induce a ***self-reinforcing error loop***. Noisy labels bias the semantic representation space, causing clean samples to be filtered out and noisy samples to be mistakenly retained. The retained noisy samples further distort the representation space, aggravating reliability estimation bias and leading to progressive noise propagation and error accumulation. Therefore, reliability estimation should not rely solely on the current model state. In contrast, semantic information accumulated from historically reliable samples provides a more stable reference for assessing current samples. As illustrated in Figure 1 (b), we therefore exploit such historically reliable semantics to guide reliability estimation and subsequent robust hash learning.

Building on this idea, we propose ***Robust Memory-guided Hashing (RMH)***, which leverages historically reliable cross-modal memory to provide temporally stable semantic references for both reliability estimation and robust hash learning, thereby mitigating the propagation and accumulation of noisy supervision. Figure 2 illustrates the architecture of RMH. Specifically, Memory-guided Reliability Estimation (MRE) constructs modality-specific class centers from previously selected reliable samples and estimates the reliability of each sample pair through bidirectional matching between current representations and historical heterogeneous-modal centers. A smoothed threshold is then used to partition samples into reliable and unreliable subsets, while the reliable subset is used to update the memory bank. Based on this partition, Reliable Cross-batch Learning (RCL) jointly exploits in-batch and historical cross-batch semantic relations for reliable samples, while retaining unreliable samples through an uncertainty-aware loss to reduce the influence of potentially erroneous supervision. In general, the main contributions of this paper are as follows:

- Firstly, we propose RMH, which leverages a cross-modal memory of historically reliable samples to provide stable semantic references for reliability estimation and cross-batch representation learning, thereby suppressing the propagation and accumulation of noisy supervision during iterative optimization.
- Secondly, we propose ***Memory-guided Reliability Estimation (MRE)***, which constructs cross-modal class centers from historically reliable samples and estimates pairwise reliability by matching current representations with historical heterogeneous-modal centers to assess the semantic support for observed labels. A smooth threshold is then applied to partition samples into reliable and unreliable subsets, while only reliable samples update the memory bank to preserve reliable historical semantic references.
- Thirdly, we design a ***Reliable Cross-batch Learning (RCL)*** approach that jointly exploits in-batch and historical cross-batch semantic relations to learn from reliable samples, while employing an uncertainty-aware loss to weaken noisy supervision from unreliable samples, thereby improving data utilization and the robustness of hash codes.
- Finally, RMH can be flexibly integrated with various CMH methods to enhance robustness under noisy labels. Extensive experiments demonstrate that RMH achieves superior retrieval performance and robustness across diverse noisy-label scenarios.

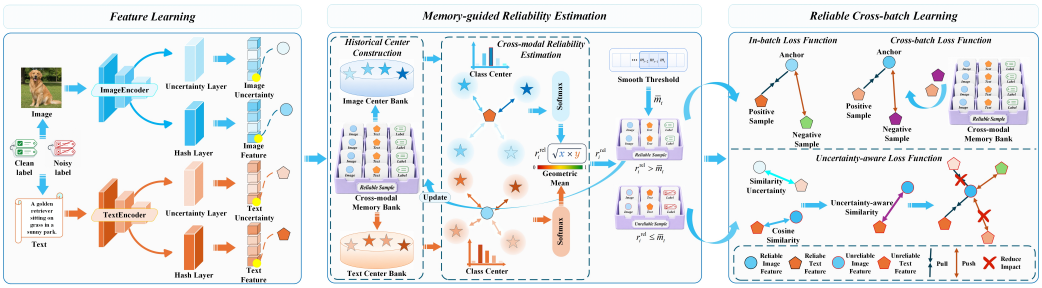


Figure 2: Overview of the RMH framework. It consists of three components: (1) Feature Learning: CLIP-based encoders extract modality-specific hash and uncertainty representations. (2) Memory-guided Reliability Estimation: historically reliable samples are used to construct cross-modal class centers, and pairwise reliability is estimated by matching current representations with historical heterogeneous-modal class centers to assess the semantic support for labels. Smooth thresholding then separates reliable from unreliable samples, with reliable samples updating the memory bank. (3) Reliable Cross-batch Learning: reliable samples are optimized with in-batch and cross-batch losses to exploit current and historical semantic relations, while unreliable samples are trained with an uncertainty-aware loss to attenuate noisy supervision and improve hash-code robustness.

2 Related Work

Cross-modal hashing maps heterogeneous data into a shared Hamming space, enabling efficient large-scale cross-modal retrieval with compact binary codes (Huo et al., 2024; Cheng et al., 2026a). Recent advances in semantic alignment (Qin et al., 2025a), structure-preserving learning (Qin et al., 2023), and uncertainty modeling (Cheng et al., 2025) have further improved the discriminability and semantic consistency of hash representations. However, most supervised CMH methods construct semantic relations from given annotations and are therefore highly sensitive to supervision quality. When training labels are corrupted, erroneous semantic relations can interfere with similarity learning, distort the cross-modal representation space, and ultimately degrade retrieval performance.

To mitigate the effects of noisy supervision, existing ***noise-robust cross-modal hashing*** methods mainly adopt three strategies: robust objectives (Pu et al., 2025), label correction (Yang et al., 2025), and sample selection (Wang et al., 2024; Liu et al., 2026; Peng et al., 2026). Robust objectives reduce the influence of noisy supervision through noise-tolerant losses or regularization, label-correction methods aim to recover more reliable annotations, and sample-selection methods identify trustworthy samples for optimization. However, when assessing sample or label reliability, these methods often rely on signals derived from the current model, such as losses, predictions, or semantic representations. As training progresses, such signals can become increasingly coupled with noise-induced representation bias, thereby undermining the stability of reliability estimation.

Memory-based learning preserves historical representations beyond the current mini-batch to enrich relation modeling (Wang et al., 2020). Similar memory structures have also been explored in cross-modal hashing; for example, FPO-CMH employs a multimodal dual-tier anchor bank to support online cross-modal learning (Li et al., 2025). Different from memory mechanisms primarily designed to enlarge the available relation set, RMH selectively maintains historical samples and assigns the resulting memory two coupled roles: providing cross-modal semantic references for reliability estimation and supplying reliable cross-batch relations for robust hash learning. A more detailed discussion of related work is provided in Appendix A.

3 Methodology

3.1 Notation and Feature Learning

Given a training set $\mathcal{D} = \{(x_i^{\mathcal{I}}, x_i^{\mathcal{T}}, \tilde{\mathbf{y}}_i)\}_{i=1}^{N}$ of N image–text pairs, $x_i^{\mathcal{I}}$ and $x_i^{\mathcal{T}}$ denote the image and text of the i-th pair, respectively, and $\tilde{\mathbf{y}}_i \in \{0,1\}^C$ is its potentially corrupted multi-label annotation over C categories. Two instances are considered semantically related if their observed annotations share at least one positive label, i.e., $r_{ij} = \mathbb{I}(\tilde{\mathbf{y}}_i^\top \tilde{\mathbf{y}}_j > 0)$. Thus, r_{ij} represents the observed seman-

tic relation induced by potentially corrupted annotations and may differ from the underlying clean relation. RMH employs CLIP-based encoders followed by separate hash and uncertainty mappings to obtain modality-specific relaxed hash and uncertainty representations:

$$\mathbf{h}_i^m = \tanh(H^m(E^m(x_i^m))), \qquad \mathbf{u}_i^m = U^m(E^m(x_i^m)), \quad m \in \{\mathcal{I}, \mathcal{T}\} \tag{1}$$

Here, E^m denotes the encoder for modality m, and H^m and U^m denote the hash and uncertainty mappings, respectively. $\mathbf{h}_i^m \in (-1, 1)^K$ is the relaxed hash representation, where K denotes the hash-code length, while $\mathbf{u}_i^m$ is the learned uncertainty representation used for uncertainty-aware optimization. During training, the relaxed hash representations are optimized directly; at inference, they are binarized as $\mathbf{b}_i^m = \text{sign}(\mathbf{h}_i^m) \in \{-1, +1\}^K$.

3.2 Memory-guided Reliability Estimation

Most existing noise-robust CMH methods estimate sample or label reliability using signals produced by the current model. However, these signals are themselves learned under noisy supervision, and their biases can accumulate as the model progressively fits noisy labels, thereby degrading reliability estimation and amplifying errors. To reduce this dependence on the current model state, we propose Memory-guided Reliability Estimation (MRE). MRE maintains a fixed-capacity first-in-first-out (FIFO) Cross-modal Memory Bank $\mathcal{M} = \{(\mathbf{v}_j^{\mathcal{I}}, \mathbf{v}_j^{\mathcal{T}}, \tilde{\mathbf{y}}_j)\}$, where $\mathbf{v}_j^m$ denotes a detached copy of the hash representation $\mathbf{h}_j^m$ stored at the time of insertion. During a short warm-up stage, before MRE produces reliability partitions, the uncertainty-aware loss is applied to all samples, and representations initialize the memory. MRE is activated once the memory reaches capacity. Thereafter, only samples identified as reliable are admitted into the memory, while the oldest entries are removed in FIFO order to reduce contamination of historical semantic references by noisy labels.

Historical Center Construction. To obtain compact historical semantic references, we aggregate the reliable representations stored in the Cross-modal Memory Bank according to their semantic categories and construct modality-specific historical class centers. For category c, let $\mathcal{M}_c = \{j \mid \tilde{y}_{jc} = 1\}$ denote the set of memory samples whose observed labels contain category c. The historical class center of modality $m \in \{\mathcal{I}, \mathcal{T}\}$ is defined as:

$$\mathbf{c}_c^m = \frac{1}{|\mathcal{M}_c|} \sum_{j \in \mathcal{M}_c} \frac{\mathbf{v}_j^m}{\|\mathbf{v}_j^m\|_2} \tag{2}$$

Here, $\mathbf{c}_c^m$ denotes the modality-specific historical class center of category c. For a multi-label sample, its representation may contribute to multiple class centers. These historical class centers provide compact semantic references for subsequent cross-modal reliability estimation.

Cross-modal Reliability Estimation. Given the historical center banks, MRE evaluates sample reliability according to how well the observed labels of a current image-text pair are supported by historical semantics from the heterogeneous modality. Specifically, the current image representation is matched with historical text centers, and the similarities are normalized by softmax to obtain:

$$p_{ic}^{\mathcal{I}\to\mathcal{T}} = \frac{\exp\big(\text{sim}(\mathbf{h}_i^{\mathcal{I}}, \mathbf{c}_c^{\mathcal{T}})/\tau_r\big)}{\sum_{k \in \mathcal{C}_{\mathcal{M}}} \exp\big(\text{sim}(\mathbf{h}_i^{\mathcal{I}}, \mathbf{c}_k^{\mathcal{T}})/\tau_r\big)}, \qquad c \in \mathcal{C}_{\mathcal{M}} \tag{3}$$

where $\mathcal{C}_{\mathcal{M}}$ denotes the set of categories with established historical centers, τ_r is the reliability temperature, and $\text{sim}(\cdot, \cdot)$ denotes cosine similarity. For multi-label supervision, we measure the semantic support for the observed positive labels by their total probability mass, $s_i^{\mathcal{I}\to\mathcal{T}} = \sum_{c \in \mathcal{C}_{\mathcal{M}}, \tilde{y}_{ic}=1} p_{ic}^{\mathcal{I}\to\mathcal{T}}$, which avoids penalizing samples whose semantics are distributed across multiple valid categories. Similarly, $s_i^{\mathcal{T}\to\mathcal{I}}$ is obtained by matching the current text representation with the historical image centers. The reliability score of the i-th image–text pair is then defined as the geometric mean of the bidirectional supports, $r_i^{\text{rel}} = \sqrt{s_i^{\mathcal{I}\to\mathcal{T}} s_i^{\mathcal{T}\to\mathcal{I}}}$, which favors consistent cross-modal semantic support and reduces the influence of unilateral high confidence.

To reduce batch-wise fluctuations in reliability partitioning, we employ a temporally smoothed threshold. Given an estimated noise rate $\hat{\rho}$, the instantaneous threshold for the t-th mini-batch $\mathcal{B}_t$ is defined as $m_t = Q_{\hat{\rho}}(\{r_i^{\text{rel}}\}_{i \in \mathcal{B}_t})$, where $Q_{\hat{\rho}}(\cdot)$ denotes the $\hat{\rho}$-th quantile of the reliability scores. The threshold is then smoothed over the most recent τ mini-batches:

$$\bar{m}_t = \frac{1}{\tau_t} \sum_{j=t-\tau_t+1}^{t} m_j, \qquad \tau_t = \min(t, \tau) \tag{4}$$

Samples with $r_i^{\text{rel}} > \bar{m}_t$ are assigned to the reliable subset $\mathcal{B}_r$, and the rest to the unreliable subset $\mathcal{B}_u$. After warm-up, only samples in $\mathcal{B}_r$ update the Cross-modal Memory Bank and historical class centers, providing reliable semantic references for subsequent cross-batch learning.

3.3 Reliable Cross-batch Learning

After MRE partitions the training samples by reliability, the reliable subset provides trustworthy supervision. However, reliability filtering reduces the number of usable relations within each mini-batch under high noise rates, while discarding unreliable samples would waste useful information. To address both issues, we propose Reliable Cross-batch Learning (RCL). RCL extends reliable supervision beyond the mini-batch by jointly exploiting in-batch relations and historical reliable representations stored in the Cross-modal Memory Bank, while retaining unreliable samples through an uncertainty-aware loss that reduces the influence of potentially erroneous supervision. In this work, RCL is instantiated with the triplet loss. Given an anchor a, a semantic positive sample p, and a semantic negative sample n, the triplet loss for reliable samples is defined as:

$$\ell_{\text{tri}}(a, p, n) = [\mu + \text{sim}(a, n) - \text{sim}(a, p)]_+ \tag{5}$$

where μ is the margin and $\text{sim}(\cdot, \cdot)$ denotes cosine similarity. A sample is considered positive if its observed annotation shares at least one positive label with that of the anchor, and negative otherwise. Cross-modal triplets are constructed using semi-hard negative mining as in (Schroff et al., 2015). For the reliable subset selected by MRE, the in-batch loss is defined as:

$$\mathcal{L}_{\text{in}} = \frac{1}{|\mathcal{T}_{\text{in}}^r|} \sum_{(a,p,n)\in\mathcal{T}_{\text{in}}^r} \ell_{\text{tri}}(a, p, n) \tag{6}$$

where $\mathcal{T}_{\text{in}}^r$ denotes the set of cross-modal triplets constructed from the reliable samples in the current mini-batch. This objective pulls semantically related cross-modal representations closer than unrelated ones. To further expand the reliable relation set, RCL leverages historical reliable representations stored in the Cross-modal Memory Bank. Current reliable samples are used as anchors, while positives and negatives are drawn from the historical heterogeneous-modal memory to form cross-batch triplets. The corresponding loss is defined as:

$$\mathcal{L}_{\text{mem}} = \frac{1}{|\mathcal{T}_{\text{mem}}^r|} \sum_{(a,p,n)\in\mathcal{T}_{\text{mem}}^r} \ell_{\text{tri}}(a, p, n) \tag{7}$$

where $\mathcal{T}_{\text{mem}}^r$ denotes the set of cross-batch triplets formed between current reliable samples and historical heterogeneous-modal representations. Together with in-batch relations, these cross-batch relations provide more diverse semantic constraints for reliable representation learning.

For unreliable samples, conventional metric constraints may induce excessive gradients from potentially erroneous semantic relations. We therefore introduce an Uncertainty-aware Loss Function that adaptively modulates pairwise similarity based on semantic discrepancy and uncertainty. Given two representations $(\mathbf{h}_i, \mathbf{u}_i)$ and $(\mathbf{h}_j, \mathbf{u}_j)$, we define $\alpha_{ij} = \|\hat{\mathbf{h}}_i - \hat{\mathbf{h}}_j\|_2$ and $\beta_{ij} = \|\mathbf{u}_i + \mathbf{u}_j\|_2$, respectively, where $\hat{\mathbf{h}} = \mathbf{h}/\|\mathbf{h}\|_2$. The resulting Uncertainty-aware Similarity is formulated as:

$$S_u(i, j) = 1 - (1 - \text{sim}(\mathbf{h}_i, \mathbf{h}_j)) \exp\left(-\frac{\beta_{ij} + \gamma}{\tau_u \max(\alpha_{ij}, \epsilon_s)}\right) \tag{8}$$

where γ and τ_u control the uncertainty modulation, and ϵ_s ensures numerical stability. This transformation is applied only to unreliable samples, where larger uncertainty attenuates the gradient contribution of uncertain semantic relations. The uncertainty-aware loss is then defined as:

$$\mathcal{L}_{\text{unc}} = \frac{1}{|\mathcal{T}_u|} \sum_{(a,p,n)\in\mathcal{T}_u} [\mu + S_u(a, n) - S_u(a, p)]_+ \tag{9}$$

where $\mathcal{T}_u$ denotes the set of cross-modal triplets involving unreliable samples. Thus, unreliable samples can contribute useful information while exerting reduced influence on optimization. They are excluded from updating the Cross-modal Memory Bank to prevent unreliable supervision from contaminating historical references. Appendix B analyzes how reliable-memory updating constrains supervision propagation and how uncertainty-aware similarity attenuates uncertain gradients, providing a theoretical interpretation of RMH's robustness to noise propagation and error accumulation.

Table 1: Average mAP results (%) on four datasets under different symmetric noise rates. The best and second-best results are highlighted in boldface and underlined, respectively.

Dataset	Method	0.0				0.2				0.5				0.8			
		16	32	64	128	16	32	64	128	16	32	64	128	16	32	64	128
MIRFlickr -25K	DCHMT	80.92	81.51	81.27	81.65	72.53	74.20	76.27	76.48	69.08	72.70	73.64	74.66	68.71	69.57	68.72	72.63
	DDWSH	77.32	82.03	83.37	84.49	72.84	75.06	77.20	79.73	62.28	71.36	73.54	75.39	61.40	67.66	70.56	72.80
	DSSH	82.44	82.40	83.22	83.34	80.22	81.28	82.29	82.51	78.55	79.57	80.40	80.82	76.26	77.13	77.52	77.64
	CMMQ	77.65	79.29	80.59	81.17	76.55	77.99	78.93	79.91	70.10	75.29	76.46	77.53	66.02	68.33	69.97	73.22
	NRCH	78.46	79.66	81.02	80.99	77.07	79.11	79.25	78.97	75.81	77.58	77.78	77.50	75.50	76.52	75.94	76.33
	NIRNL	77.61	78.87	79.36	79.18	76.96	78.23	77.85	77.66	75.79	76.83	76.54	76.98	75.12	75.79	75.78	76.78
	RFCMH	80.69	79.68	80.31	81.13	78.77	79.20	79.20	79.51	78.20	78.30	78.82	79.18	77.10	77.79	78.07	77.91
	RMH	**82.77**	**84.09**	**85.17**	**85.69**	**80.99**	**82.22**	**83.65**	**84.56**	**79.40**	**80.75**	**81.74**	**82.43**	**77.26**	**78.60**	**79.62**	**80.34**
NUS-WIDE	DCHMT	66.79	67.72	69.03	70.35	61.31	65.63	66.84	66.44	51.67	54.58	56.94	56.99	44.02	45.78	45.68	45.69
	DDWSH	66.88	71.25	72.10	72.83	58.29	59.85	63.69	66.52	47.44	47.64	48.56	51.57	37.86	42.34	44.95	46.80
	DSSH	70.94	71.57	72.85	73.63	68.03	70.40	71.32	72.76	63.48	65.91	67.68	68.83	53.16	53.83	54.66	56.47
	CMMQ	67.91	70.55	72.13	72.49	66.32	68.90	70.34	70.93	61.29	65.73	67.29	65.83	36.33	47.06	46.51	47.93
	NRCH	67.18	70.19	69.95	70.57	66.72	66.20	66.20	65.82	64.41	62.14	62.13	63.29	62.03	61.41	61.38	61.77
	NIRNL	65.94	65.19	66.12	68.74	65.70	66.30	66.65	68.04	65.88	68.37	68.94	69.17	62.55	66.42	67.82	67.29
	RFCMH	69.40	70.43	71.07	71.50	69.53	70.57	70.90	71.68	67.06	69.21	69.74	70.30	62.26	66.36	67.42	68.57
	RMH	**72.12**	**72.50**	**73.86**	**74.78**	**69.59**	**71.37**	**73.16**	**74.01**	**68.18**	**70.67**	**72.13**	**73.74**	**66.32**	**68.67**	**68.79**	**68.75**
MS COCO	DCHMT	62.75	62.14	65.20	67.80	58.70	59.89	60.35	61.38	57.66	58.67	59.77	58.93	55.60	57.55	58.21	57.88
	DDWSH	66.71	70.63	73.55	75.55	60.73	66.82	70.04	73.21	50.89	60.09	63.27	67.51	47.67	55.84	58.82	61.22
	DSSH	69.47	71.03	73.39	74.73	66.46	70.47	72.60	74.08	64.35	64.90	68.45	70.40	62.20	66.10	68.41	70.25
	CMMQ	60.77	69.43	75.12	76.76	58.72	66.08	70.70	71.23	48.26	56.64	60.28	61.45	41.54	41.47	47.59	47.94
	NRCH	67.08	72.24	74.74	75.85	68.34	70.54	71.52	72.11	64.81	67.01	67.77	69.12	64.09	65.88	65.90	67.16
	NIRNL	60.12	61.88	64.14	64.78	59.69	61.50	62.78	64.12	58.40	61.04	60.21	63.38	55.57	58.36	58.57	60.97
	RFCMH	67.50	68.75	70.69	71.13	69.21	70.56	71.88	72.03	**68.47**	70.51	71.36	71.15	**66.46**	68.68	70.21	70.98
	RMH	**69.95**	**73.68**	**75.51**	**77.06**	**69.42**	**72.52**	**74.97**	**76.57**	68.00	**71.10**	**73.39**	**75.29**	66.39	**68.86**	**70.64**	**72.48**
IAPR TC-12	DCHMT	60.45	62.41	63.21	62.91	52.67	55.54	56.72	58.55	51.71	54.47	56.41	56.86	50.97	54.23	55.41	57.00
	DDWSH	50.19	61.42	64.40	66.92	47.92	54.17	59.02	61.96	44.47	52.17	56.91	59.74	41.11	50.93	56.02	58.89
	DSSH	63.29	66.04	68.51	70.11	61.00	65.06	67.93	67.97	60.24	63.76	66.63	68.41	59.13	63.04	65.86	66.62
	CMMQ	45.37	50.67	56.28	62.60	44.36	50.00	53.27	59.05	42.84	48.79	49.59	51.58	41.54	43.95	45.34	43.60
	NRCH	52.22	56.74	59.57	60.94	51.66	56.74	57.48	58.38	51.62	55.34	56.44	56.96	49.88	54.57	55.72	55.85
	NIRNL	50.84	52.02	52.64	53.09	49.87	51.61	52.57	52.64	50.84	51.86	52.39	52.95	50.38	52.00	52.87	52.84
	RFCMH	55.07	57.47	59.28	59.21	55.50	57.77	59.46	59.78	55.37	57.01	58.45	59.59	54.19	56.49	58.44	58.80
	RMH	**63.71**	**67.15**	**70.33**	**72.76**	**62.17**	**65.55**	**68.78**	**70.40**	**61.47**	**64.45**	**67.38**	**69.31**	**59.48**	**63.38**	**66.12**	**67.58**

Table 2: Average mAP results (%) on NUS-WIDE under different asymmetric noise rates.

Dataset	Method	0.0				0.2				0.5				0.8			
		16	32	64	128	16	32	64	128	16	32	64	128	16	32	64	128
NUS-WIDE	DCHMT	66.79	67.72	69.03	70.35	61.50	65.60	66.91	67.95	57.52	62.34	64.94	64.54	58.78	64.07	64.97	65.59
	DDWSH	66.88	71.25	72.10	72.83	57.11	57.62	61.68	65.85	54.12	51.70	53.03	58.11	54.42	55.22	54.78	58.84
	DSSH	70.94	71.57	72.85	73.63	67.40	69.58	71.76	72.97	64.59	67.77	68.34	69.95	65.73	65.91	68.12	69.15
	CMMQ	67.91	70.55	72.13	72.49	66.14	68.33	70.11	70.72	59.44	63.88	66.40	66.52	43.20	45.37	51.48	51.28
	NRCH	67.18	70.19	69.95	70.57	66.51	66.26	66.34	66.34	64.63	63.46	63.37	64.14	64.58	62.91	63.62	63.69
	NIRNL	65.94	65.19	66.12	68.74	66.06	67.14	67.94	68.68	66.63	66.72	68.41	68.27	65.96	66.53	68.25	68.12
	RFCMH	69.40	70.43	71.07	71.50	69.46	**71.13**	70.79	71.57	67.88	69.50	70.30	70.69	66.78	68.10	**69.42**	69.41
	RMH	**72.12**	**72.50**	**73.86**	**74.78**	**69.64**	71.12	**71.92**	**73.01**	**68.50**	**69.71**	**70.93**	**72.01**	**67.54**	**68.41**	68.99	**70.10**

3.4 Overall Objective

To regularize the continuous hash representations with respect to their binarized counterparts, we employ the following quantization loss:

$$\mathcal{L}_q = \frac{1}{|\mathcal{B}|K} \sum_{i \in \mathcal{B}} \left(\left\| \hat{\mathbf{h}}_i^{\mathcal{I}} - \operatorname{sign}(\hat{\mathbf{h}}_i^{\mathcal{I}}) \right\|_1 + \left\| \hat{\mathbf{h}}_i^{\mathcal{T}} - \operatorname{sign}(\hat{\mathbf{h}}_i^{\mathcal{T}}) \right\|_1 \right) \tag{10}$$

where $\hat{\mathbf{h}}_i^m = \mathbf{h}_i^m / \|\mathbf{h}_i^m\|_2$. The overall objective of RMH jointly optimizes the in-batch loss, cross-batch loss, uncertainty-aware loss, and quantization loss:

$$\mathcal{L}_{\text{RMH}} = \mathcal{L}_{\text{mem}} + \mathcal{L}_{\text{in}} + \mathcal{L}_{\text{unc}} + \lambda_q \mathcal{L}_q \tag{11}$$

where λ_q controls the contribution of the quantization loss.

4 Experiments

4.1 Benchmark Datasets & Baseline Methods

To evaluate the retrieval performance of RMH, we conduct experiments on four benchmark datasets: MIRFlickr-25K (Huiskes & Lew, 2008), NUS-WIDE (Chua et al., 2009), MS COCO (Lin et al., 2014), and IAPR TC-12 (Escalante et al., 2010). Further details are provided in Appendix C.

To verify the effectiveness of RMH, we compare it with seven representative cross-modal hashing methods, including DCHMT (Tu et al., 2022), DDWSH (Cheng et al., 2026b), DSSH (Qin et al., 2026), CMMQ (Yang et al., 2022), NRCH (Wang et al., 2024), NIRNL (Liu et al., 2026), and

Table 3: Average mAP results (%) on IAPR TC-12 under different semantic confusion noise rates.

Dataset	Method	0.0				0.2				0.5				0.8			
		16	32	64	128	16	32	64	128	16	32	64	128	16	32	64	128
IAPR TC-12	DCHMT	60.45	62.41	63.21	62.91	53.13	55.51	57.67	58.69	52.16	54.74	58.01	58.49	51.94	55.41	57.28	58.30
	DDWSH	50.19	61.42	64.40	66.92	48.92	54.20	57.97	62.99	47.87	53.92	58.18	60.99	45.92	52.10	56.94	60.53
	DSSH	63.29	66.04	68.51	70.11	59.35	63.95	**66.48**	67.87	58.84	63.06	65.84	67.62	58.54	62.38	64.75	66.93
	CMMQ	45.37	50.67	56.28	62.60	44.58	50.14	53.58	59.16	43.14	48.54	49.26	51.99	41.83	45.33	43.88	44.27
	NRCH	52.22	56.74	59.57	60.94	51.50	56.38	57.86	58.75	50.67	54.73	56.08	57.29	52.59	54.21	55.35	55.86
	NIRNL	50.84	52.02	52.64	53.09	50.32	51.83	52.60	52.97	50.31	51.68	52.93	52.33	50.20	51.40	52.56	53.10
	RFCMH	55.07	57.47	59.28	59.21	55.40	57.64	58.90	59.81	54.58	57.87	58.55	59.37	53.73	56.75	58.03	56.73
	RMH	**63.71**	**67.15**	**70.33**	**72.76**	**60.35**	**64.10**	66.40	**68.97**	**59.42**	**63.35**	**66.37**	**68.14**	**58.93**	**62.70**	**64.97**	**67.16**

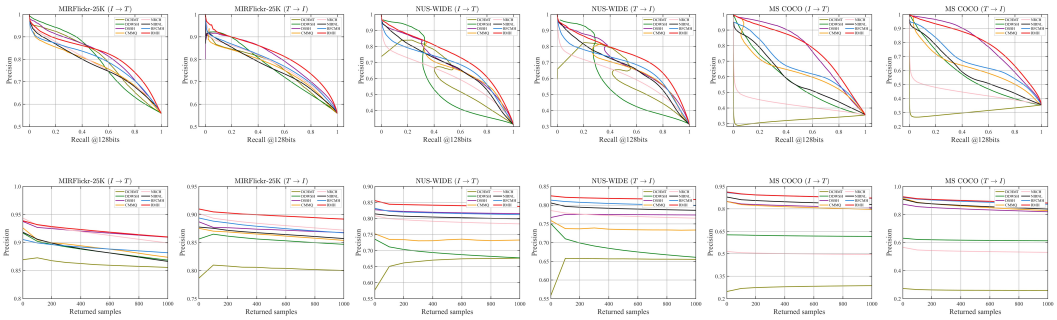


Figure 3: Results of Precision-Recall and Top-N Precision curves on three benchmark datasets with 128 bits under 50% symmetric noise.

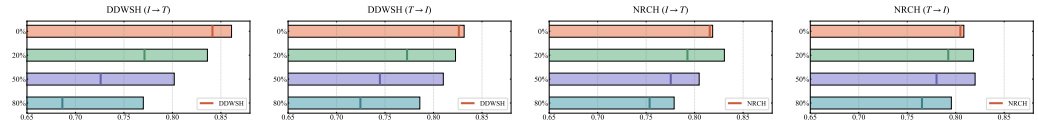


Figure 4: Bullet chart visualization on MIRFlickr-25K with 64 bits under symmetric noise. The target markers indicate the baselines, and the bars show performance after incorporating RMH.

RFCMH (Sun et al., 2026c), among which NRCH, NIRNL, and RFCMH are specifically designed for robust learning under noisy labels. For fair comparison, all methods adopt CLIP ViT-B/32 as the common backbone and use identical training, query, and retrieval sets as well as the same evaluation protocol. Method-specific hyperparameters follow the settings reported in the original papers. For each dataset, noise type, and noise rate, all methods are trained with the same pre-generated and fixed noisy labels, while the query and retrieval sets are always evaluated using the original annotations.

4.2 Evaluation Metrics & Implementation Details

We evaluate cross-modal retrieval performance in image-to-text (I2T) and text-to-image (T2I) directions using mean average precision (mAP), Precision-Recall (PR) curves, and Top-N Precision as primary retrieval metrics (Sun et al., 2026a; Cheng et al., 2026b). To assess robustness across noise patterns, we consider symmetric label noise (Wang et al., 2024), asymmetric label noise (Pu et al., 2025), and semantic confusion noise at multiple noise rates. Semantic confusion noise generates semantically plausible label corruption by replacing labels with semantically related categories. In addition, we employ AUROC, Reliable Precision, False-retention Rate, and Memory Purity to evaluate sample reliability estimation, measuring the separability between clean and noisy samples, the purity of the selected reliable subset, the proportion of noisy samples incorrectly identified as reliable, and the reliability of historical samples stored in the Cross-modal Memory Bank, respectively.

All experiments are implemented in PyTorch and conducted on two NVIDIA A100 GPUs. Models are trained for 100 epochs with a batch size of 128. For RMH, the capacity of the Cross-modal Memory Bank is set to 640, the reliability temperature is set to $\tau_r = 0.2$, and the smoothing window of the Smooth Threshold is set to 40. The parameters of the Uncertainty-aware Loss Function are set to $\tau_u = 9$, $\gamma = 4$, and $\epsilon_s = 10^{-12}$. Detailed definitions of the evaluation metrics, noise construction protocols, and experimental settings are provided in Appendix D.

4.3 Performance Comparison

To evaluate RMH under noisy supervision, we compare it with baselines under symmetric, asymmetric, and semantic confusion noise. The average mAP results are reported in Tables 1–3. (1) As the noise rate increases, most methods show performance degradation, indicating interference from corrupted labels. In contrast, RMH remains comparatively stable across noise types and rates, even under severe corruption. (2) Semantic confusion noise is challenging because labels are replaced with semantically related categories, making corrupted annotations semantically plausible. RMH remains highly competitive under this setting, supporting the use of historically reliable semantics as stable references when noisy labels retain substantial semantic similarity to the original labels. Overall, RMH shows consistent robustness across different noise patterns. This benefits from the complementary roles of memory-guided reliability estimation, reliable cross-batch learning, and uncertainty-aware optimization, which respectively provide stable semantic references, extend trustworthy relations beyond the current mini-batch, and suppress potentially erroneous supervision. Together, they reduce the accumulation of unreliable semantic constraints and improve robust cross-modal hash learning. Complete I2T and T2I mAP results are provided in Appendix E.6.

Figure 3 further compares ranking quality using Precision-Recall and Top-N Precision curves. RMH maintains higher precision over a broad recall range and stable Top-N Precision across retrieval depths, indicating better preservation of semantic relevance and ranking quality. To examine whether RMH depends on a specific hashing objective, we integrate it into DDWSH and NRCH while keeping other training settings unchanged, as shown in Figure 4. Both baselines improve after incorporating RMH, demonstrating its compatibility with different hashing objectives and its ability to enhance robustness through historical reliability estimation and reliable cross-batch learning.

4.4 Reliability Estimation Analysis

To analyze the effect of semantic references on reliability estimation, we construct three controlled mechanisms. Current-batch builds class centers from the current mini-batch and adopts a leave-one-out strategy to prevent each sample from contributing to its own reference; Historical-all maintains a fixed-capacity FIFO historical memory, where all current samples are admitted and used to construct historical class centers; RMH, in contrast, allows only samples identified as reliable to update the historical memory. The three mechanisms share the same current representations, reliability scoring function, and thresholding strategy, and differ only in their semantic references and memory update rules. As shown in Figure 5, Current-batch is more susceptible to local noise in the current mini-batch; Historical-all improves estimation stability by incorporating historical semantics, but unfiltered historical samples can still contaminate the semantic reference. RMH suppresses such contamination through reliability-driven memory updates, yielding more stable reliability estimation in terms of AUROC, Reliable Precision, and FRR, while maintaining higher Memory Purity. For the first three metrics, we further compare RMH with CMMQ and NRCH. The results further support the effectiveness of reliable historical semantics for stable reliability estimation.

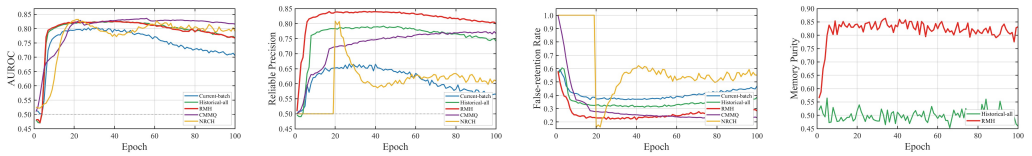


Figure 5: Reliability estimation analysis on NUS-WIDE with 64 bits under 50% symmetric noise.

Table 4: Ablation study results of RMH under 50% symmetric noise. MRE, CB, and UA denote Memory-guided Reliability Estimation, Cross-batch Loss, and Uncertainty-aware Loss, respectively. Results are reported as the average mAP (%) over I2T and T2I retrieval tasks.

Variant	MRE	CB	UA	MIRFlickr-25K				NUS-WIDE				MS COCO				IAPR TC-12			
				16	*32*	*64*	*128*	*16*	*32*	*64*	*128*	*16*	*32*	*64*	*128*	*16*	*32*	*64*	*128*
Base	–	–	–	76.79	78.08	79.70	79.89	62.17	66.36	67.69	69.78	65.60	68.50	70.37	70.95	59.13	63.46	66.36	67.95
Base + MRE	✓	–	–	76.29	78.66	80.26	81.14	68.39	70.66	71.59	72.14	66.50	68.61	70.84	71.01	59.81	63.57	65.83	67.53
Base + MRE + CB	✓	✓	–	77.03	79.65	80.34	81.51	**69.94**	70.36	72.07	73.04	66.58	69.30	71.81	73.21	60.78	64.28	66.04	68.64
Base + MRE + UA	✓	–	✓	77.93	80.25	81.22	82.01	69.29	70.33	71.88	73.36	67.35	69.88	73.11	74.90	61.19	**64.47**	67.27	69.29
RMH	✓	✓	✓	**79.40**	**80.75**	**81.74**	**82.43**	68.18	**70.67**	**72.13**	**73.74**	**68.00**	**71.10**	**73.39**	**75.29**	**61.47**	64.45	**67.38**	**69.31**

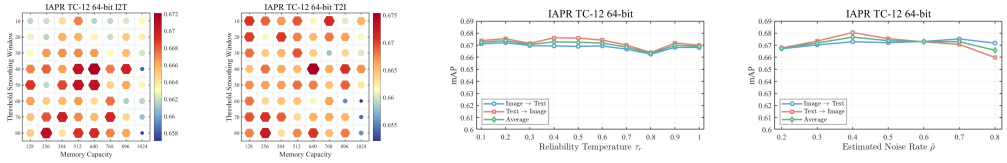


Figure 6: Parameter analysis on IAPR TC-12 with 64 bits under 50% symmetric noise.

4.5 Ablation and Sensitivity Analysis

1) Ablation Study. To assess the contribution of each component, we conduct ablation experiments under identical settings, with results reported in Table 4. MRE leverages historically reliable semantics to improve current-sample reliability estimation; CB further extends reliable supervision across batches using historical relations; and UA reduces the influence of potentially erroneous supervision from unreliable samples. The full RMH achieves the best performance in most settings, demonstrating the complementary roles of the three components in reliability estimation, reliable-relation expansion, and robust utilization of unreliable samples.

2) Parameter Analysis. We analyze the effects of Memory Capacity, Threshold Smoothing Window, Reliability Temperature, and the estimated noise rate, as shown in Figure 6. RMH remains stable across a broad range of settings. Memory Capacity controls the scale and recency of historical semantics, Threshold Smoothing Window smooths the reliability threshold, and Reliability Temperature controls the sharpness of the matching distribution. Based on overall performance, they are set to 640, 40, and 0.2. With the true noise rate fixed at $\rho = 0.5$, varying $\hat{\rho}$ from 0.2 to 0.8 causes only minor performance fluctuations, indicating robustness to noise-rate estimation errors. Efficiency results for parameter count, FLOPs, and training/encoding time are provided in Appendix E.8.

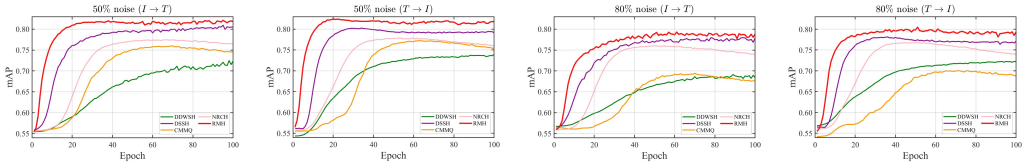


Figure 7: mAP versus training epochs on MIRFlickr-25K using 128-bit hash codes under 50% and 80% symmetric label noise.

4.6 Robustness Analysis

To evaluate RMH's robustness throughout training, we compare the epoch-wise mAP of different methods, as shown in Figure 7. RMH remains relatively stable after convergence across different noise levels, whereas several baselines exhibit increasing fluctuations or late-stage degradation as the noise rate rises. This stability is consistent with the design of RMH: historically reliable semantics provide stable references for reliability estimation, while reliable cross-batch learning and uncertainty-aware optimization mitigate the influence of noisy supervision on representation learning. Together, these mechanisms alleviate noise propagation and error accumulation during training.

5 Conclusion

This paper proposes a Robust Memory-guided Hashing (RMH) framework for noisy-label cross-modal retrieval, which introduces historically reliable semantics as stable references to alleviate reliability estimation bias and error accumulation caused by noisy supervision. Specifically, MRE constructs cross-modal semantic references from historically reliable samples to improve the accuracy of reliability estimation for current samples while limiting the persistent contamination of historical memory. RCL further exploits reliable semantic relations from both the current mini-batch and historical memory, while employing uncertainty-aware learning to reduce the influence of potentially erroneous supervision from unreliable samples. Extensive experiments demonstrate that RMH achieves superior retrieval performance across various noise types and noise rates, supporting the effectiveness of RMH in mitigating noise propagation and progressive error accumulation.

AI USE STATEMENT

Large Language Models (LLMs) were used solely to assist with language editing during manuscript preparation, including grammatical correction, sentence polishing, and improving readability. LLMs were not used for research ideation, development of the methodology, design of experiments, implementation of algorithms, data processing or analysis, interpretation of experimental results, generation of mathematical claims or proofs, or drafting substantive sections of the manuscript. All scientific contributions, including the proposed method, theoretical analysis, experimental design, results, and conclusions, were developed and verified by the authors. The authors have reviewed all AI-assisted edits and take full responsibility for the final content of this work.

ETHICS STATEMENT

This work does not involve human subjects, user studies, or the collection of any new personal data. All experiments are conducted on publicly available benchmark datasets that are widely used in the cross-modal retrieval community, and we follow the corresponding dataset licenses and usage guidelines. The proposed method aims to improve the robustness of cross-modal retrieval systems under noisy supervision and does not introduce mechanisms intended to cause harm, discriminate against individuals or groups, or compromise user privacy. We acknowledge that machine learning models may inherit biases from their training data; however, this work focuses on algorithmic robustness and does not make decisions affecting individuals or sensitive applications. We have made efforts to ensure scientific integrity by providing detailed descriptions of the proposed method, experimental protocols, and analysis. All reported results and conclusions are the responsibility of the authors, and we have followed standard research practices regarding attribution, reproducibility, and transparent reporting.

REFERENCES

Kun Cheng, Qibing Qin, Wenfeng Zhang, Lei Huang, and Jie Nie. Deep probabilistic binary embedding via learning reliable uncertainty for cross-modal retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 6393–6402, 2025.

Kun Cheng, Qibing Qin, Wenfeng Zhang, Lei Huang, and Jie Nie. Deep global-sense hard-negative discriminative generation hashing for cross-modal retrieval. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Kun Cheng, Qibing Qin, Wenfeng Zhang, Lei Huang, and Jie Nie. Deep distance weighted sampling hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*, 2026b.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pp. 1–9, 2009.

Hugo Jair Escalante, Carlos A Hernández, Jesus A Gonzalez, Aurelio López-López, Manuel Montes, Eduardo F Morales, L Enrique Sucar, Luis Villasenor, and Michael Grubinger. The segmented and annotated iapr tc-12 benchmark. *Computer vision and image understanding*, 114 (4):419–428, 2010.

Jingzhong Fang, Zidong Wang, Weibo Liu, Yong Zhang, Yimeng He, and Xiaohui Liu. A comprehensive review on deep learning with noisy labels: Challenges, solutions, and frontiers. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2026.

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.

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

Yadong Huo, Qibing Qin, Jiangyan Dai, Lei Wang, Wenfeng Zhang, Lei Huang, and Chengduan Wang. Deep semantic-aware proxy hashing for multi-label cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(1):576–589, 2023.

Yadong Huo, Qibing Qin, Wenfeng Zhang, Lei Huang, and Jie Nie. Deep hierarchy-aware proxy hashing with self-paced learning for cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Fengling Li, Yang Sun, Tianshi Wang, Lei Zhu, and Xiaojun Chang. Fast partial-modal online cross-modal hashing. *IEEE Transactions on Image Processing*, 34:4440–4455, 2025.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

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*, pp. 740–755. Springer, 2014.

Chang Liu, Han Yu, Boyang Li, Zhiqi Shen, Zhanning Gao, Peiran Ren, Xuansong Xie, Lizhen Cui, and Chunyan Miao. Noise-resistant deep metric learning with ranking-based instance selection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6807–6816. IEEE, 2021.

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.

Junyu Luo, Jinsheng Huang, Yang Xu, Lutong Zou, Xiao Luo, Bohan Wu, Yifan Wang, Wei Ju, and Ming Zhang. Conformalized hierarchical calibration for uncertainty-aware adaptive hashing. In *International Conference on Learning Representations*, volume 2026, pp. 105223–105246, 2026.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24811–24819, 2026.

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, pp. 19969–19977, 2025.

Qibing Qin, Kezhen Xie, Wenfeng Zhang, Chengduan Wang, and Lei Huang. Deep neighborhood structure-preserving hashing for large-scale image retrieval. *IEEE Transactions on Multimedia*, 26:1881–1893, 2023.

Qibing Qin, Yadong Huo, Wenfeng Zhang, Lei Huang, and Jie Nie. Deep discriminative boundary hashing for cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025a.

Qibing Qin, Lei Wu, Wenfeng Zhang, Lei Huang, and Jie Nie. Deep semantic-consistent penalizing hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*, 2025b.

Qibing Qin, Meiling Ge, Wenfeng Zhang, Lei Huang, and Jie Nie. Deep stochastic spherical hashing with von mises-fisher distributions for cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Selfie: Refurbishing unclean samples for robust deep learning. In *International conference on machine learning*, pp. 5907–5915. PMLR, 2019.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 34(11):8135–8153, 2022.

Hao Sun, Yadong Huo, Qibing Qin, Wenfeng Zhang, and Lei Huang. Intra-class distribution-guided generative hashing with neighbor refinement for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2671–2681, 2026a.

Hao Sun, Qibing Qin, and Lei Huang. Robust self-reflective hashing for cross-modal retrieval with noisy label. In *Forty-third International Conference on Machine Learning*, 2026b.

Xincheng Sun, Ruitao Pu, Guangsi Shi, Zhenwen Ren, Peng Hu, and Yuan Sun. Learning with admissibility: Robust fuzzy hashing for cross-modal retrieval with noisy labels. In *Forty-third International Conference on Machine Learning*, 2026c.

Yuan Sun, Jian Dai, Zhenwen Ren, Yingke Chen, Dezhong Peng, and Peng Hu. Dual self-paced cross-modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15184–15192, 2024.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Junfeng Tu, Xueliang Liu, Zongxiang Lin, Richang Hong, and Meng Wang. Differentiable cross-modal hashing via multimodal transformers. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 453–461, 2022.

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.

Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. Cross-modal retrieval: a systematic review of methods and future directions. *Proceedings of the IEEE*, 2025.

Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R Scott. Cross-batch memory for embedding learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6387–6396. IEEE, 2020.

Qi Wei, Haoliang Sun, Xiankai Lu, and Yilong Yin. Self-filtering: A noise-aware sample selection for label noise with confidence penalization. In *European Conference on Computer Vision*, pp. 516–532. Springer, 2022.

Zhiping Xiao, Junyu Luo, Hang Zhou, Yusheng Zhao, Xiao Luo, Pengyun Wang, Wei Ju, Siyu Heng, and Ming Zhang. Hierarchical encoding tree with modality mixup for cross-modal hashing. In *International Conference on Learning Representations*, volume 2026, pp. 126617–126641, 2026.

Ao Yang, Yanglin Feng, Yuan Sun, Dezhong Peng, Guiduo Duan, and Yang Qin. Noise-robust cross-modal learning for reliable 2d-3d retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 8154–8163, 2025.

Erkun Yang, Dongren Yao, Tongliang Liu, and Cheng Deng. Mutual quantization for cross-modal search with noisy labels. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7541–7550. IEEE, 2022.

Lei Zhu, Chaoqun Zheng, Weili Guan, Jingjing Li, Yang Yang, and Heng Tao Shen. Multi-modal hashing for efficient multimedia retrieval: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(1):239–260, 2023.

CONTENTS OF APPENDIX

A Related Work **14**

A.1 Deep Cross-modal Hashing 14

A.2 Noise-robust Cross-modal Hashing 14

A.3 Memory-based Learning 15

B Theoretical Analysis of Robustness Mechanisms in RMH **16**

B.1 Stability of Historical Semantic References 16

B.2 Local Gradient Modulation of Uncertainty-aware Learning 19

B.3 Overall Interpretation 20

C Datasets **21**

D Experimental Settings **21**

D.1 Implementation Details 21

D.2 Noise Construction Protocols 22

D.3 Evaluation Metrics 22

D.3.1 Mean Average Precision 23

D.3.2 Precision–Recall Curve 23

D.3.3 Top-N Precision 23

D.3.4 Normalized Discounted Cumulative Gain 23

D.3.5 Reliability Estimation Metrics 24

E Additional Experiments **24**

E.1 Historical Semantic Reference Strategy 24

E.2 Reliability Aggregation Strategy 25

E.3 Positive-label Support Aggregation 25

E.4 Uncertainty-aware Optimization and Gradient Stability 26

E.5 Additional Hyperparameter Analysis 26

E.6 Complete Retrieval Results 28

E.7 NDCG@1000 Results 29

E.8 Efficiency Analysis 29

E.9 Backbone Generalization 30

E.10 Statistical Significance 30

E.11 Visualization 31

E.11.1 Visualization of Reliability-Score Distributions 31

E.11.2 Top-10 Retrieval Visualization 31

A Related Work

This section provides a more detailed discussion of the related studies briefly introduced in Section 2, including deep cross-modal hashing, noise-robust cross-modal hashing, and memory-based learning.

A.1 Deep Cross-modal Hashing

Deep cross-modal hashing aims to map heterogeneous data, such as images and text, into a shared Hamming space while preserving cross-modal semantic relationships and generating compact binary representations (Hu et al., 2022; Sun et al., 2024; Xiao et al., 2026; Luo et al., 2026). With the development of deep representation learning, existing studies have mainly improved the discriminability of hash representations through cross-modal semantic alignment, structure preservation, sample selection, uncertainty modeling, and training-data augmentation. Tu *et al.* proposed Differentiable Cross-modal Hashing via Multimodal Transformers (DCHMT), which employs multimodal Transformers to model cross-modal semantics and introduces a differentiable selection mechanism to alleviate the difficulty of discrete binary optimization (Tu et al., 2022). Huo *et al.* proposed Deep Hierarchy-aware Proxy Hashing (DHaPH), which exploits learnable shared hierarchical proxies to capture latent semantic hierarchies and incorporates multimodal self-paced learning to dynamically regulate the learning process of sample pairs (Huo et al., 2024). Cheng *et al.* proposed Deep Distance Weighted Sampling Hashing (DDWSH), which performs distance-weighted sampling over different distance ranges to select more informative training samples for enhanced cross-modal similarity learning (Cheng et al., 2026b). Beyond deterministic representations and sample selection, recent studies have further investigated complex semantic variations and representation uncertainty. Qin *et al.* proposed Deep Stochastic Spherical Hashing (DSSH), which models spherical embeddings with the von Mises–Fisher distribution and exploits uncertainty information to guide discriminative hash-code learning (Qin et al., 2026). Sun *et al.* proposed Intra-class Distribution-guided Generative Hashing (IDGH), which models intra-class semantic variations through class-specific distribution estimation and refines unreliable distribution estimates using statistical information from neighboring classes. Based on the refined distributions, semantically diverse training samples are adaptively generated according to class-specific distribution patterns to enhance similarity learning (Sun et al., 2026a). These methods have substantially improved the discriminability of cross-modal hashing from the perspectives of representation modeling, semantic structure, sample selection, and data augmentation. However, they mainly focus on learning more effective cross-modal representations from the provided supervision and generally assume that the semantic annotations are sufficiently reliable. When training labels are corrupted, the semantic relations induced by erroneous annotations may further misguide the optimization of the Hamming space, motivating dedicated studies on robust hash learning under noisy supervision.

A.2 Noise-robust Cross-modal Hashing

Annotations of real-world multimodal data are often affected by human errors, non-expert labeling, and semantic ambiguity, making the semantic relations constructed from labels potentially unreliable (Song et al., 2019; 2022; Wei et al., 2022; Fang et al., 2026). To mitigate this problem, existing studies mainly reduce the influence of erroneous supervision through robust objectives, label correction, sample selection, and self-paced learning. Yang *et al.* proposed Cross-Modal Mutual Quantization (CMMQ), which establishes semantic correlations across modalities through proxy-based contrastive learning, preferentially selects more trustworthy samples according to the small-loss criterion, and further improves cross-modal output consistency and sample-selection quality through mutual quantization (Yang et al., 2022). Wang *et al.* proposed Robust Contrastive Cross-modal Hashing with Noisy Labels (NRCH), which introduces a robust contrastive hashing objective to reduce the influence of unreliable semantic relations and employs a dynamic noise separator to model the loss distribution for distinguishing reliable and noisy samples (Wang et al., 2024). Compared with fixed small-loss selection, dynamic separation can adapt the reliability judgment during training, but the selection criterion still depends strongly on loss information produced by the current model. Pu *et al.* proposed Robust Self-Paced Hashing with Noisy Labels (RSHNL), which enhances cross-modal consistency and intra-class compactness through contrastive hash learning and center aggregation, and further employs noise-tolerant self-paced learning to identify noisy samples according to learning difficulty, allowing reliable samples to be learned progressively from easy to

hard (Pu et al., 2025). This learning strategy can reduce the interference of difficult noisy samples during training, but both sample difficulty and reliability are still estimated according to the current learning state of the model. Liu *et al.* proposed Neighbor-aware Instance Refining with Noisy Labels (NIRNL), which generates soft labels using cross-modal neighborhood consistency, partitions training samples into pure, hard, and noisy subsets, and applies corresponding optimization strategies according to their reliability levels (Liu et al., 2026). This method exploits neighborhood information to provide richer semantic evidence than individual sample losses, but the neighborhood relations are still constructed from the currently learned representation space and may therefore be affected by noisy supervision. Peng *et al.* proposed Semantic-Consistent Bidirectional Contrastive Hashing (SCBCH) for noisy multi-label scenarios. Its Cross-modal Semantic-Consistent Classification module adaptively adjusts sample weights according to cross-modal neighborhood semantic consistency, while Bidirectional Soft Contrastive Hashing constructs soft contrastive relations based on partial semantic overlap among multi-label samples, thereby modeling their semantic relations in a more fine-grained manner (Peng et al., 2026). Yang *et al.* further investigated noisy cross-modal learning from the perspective of label correction, improving the reliability of cross-modal representation learning by correcting potentially erroneous supervision (Yang et al., 2025). Sun *et al.* proposed Robust Fuzzy Cross-modal Hashing (RFCMH), which introduces fuzzy set theory to model the admissibility of noisy labels and employs Fuzzy Admissibility Refinement to dynamically calibrate supervision signals, while Dual-Granularity Structural Alignment preserves cross-modal consistency and inter-instance discriminative structure (Sun et al., 2026c). Such continuous supervision or supervision-refinement strategies can reduce the information loss caused by hard sample partitioning, but their supervision adjustment still generally relies on predictions or semantic responses progressively learned during training. Overall, existing noise-robust CMH methods have substantially improved robustness under noisy-label supervision through strategies such as small-loss selection, dynamic noise separation, self-paced learning, neighborhood consistency, and supervision refinement. However, when explicitly estimating sample or label reliability, these methods largely rely on signals produced by the current model, including training losses, prediction results, and semantic representations. As the model progressively fits corrupted labels, biases in these signals may accumulate over training, thereby further degrading the accuracy and stability of subsequent reliability estimation. In contrast, RMH constructs historical cross-modal semantic references from previously identified reliable samples and uses them to assess the reliability of current samples, thereby reducing the direct dependence of reliability estimation on currently noise-corrupted representations.

A.3 Memory-based Learning

Mini-batch training limits the range of sample relations available in a single optimization step. Memory-based learning therefore preserves representations or semantic information from previous iterations, enabling the current batch to exploit sample relations beyond the scope of the current mini-batch. Wang *et al.* proposed Cross-Batch Memory (XBM), which caches feature representations from previous mini-batches and reuses them as additional samples in subsequent metric learning, thereby extending the range of sample relations available for relation modeling (Wang et al., 2020). Similar memory structures have also been introduced into cross-modal hashing. Li *et al.* proposed Fast Partial-Modal Online Cross-Modal Hashing (FPO-CMH), which employs a multimodal dual-tier anchor bank to preserve and organize historical cross-modal information for online cross-modal hash learning (Li et al., 2025). These studies demonstrate that historical information can effectively overcome the relational limitations of a single mini-batch and provide richer semantic constraints for current optimization. However, under noisy-label settings, the reliability of historical information itself also becomes important. If unreliable samples are continuously stored in memory, their erroneous semantic relations may be repeatedly reused in subsequent training.

Motivated by this consideration, RMH updates the Cross-modal Memory Bank only with previously identified reliable samples and assigns historical memory two closely related roles. On the one hand, historical reliable samples are aggregated into cross-modal class centers to provide relatively stable semantic references for estimating the reliability of current samples. On the other hand, historical reliable representations are further used for cross-batch relation learning, providing current reliable samples with semantic constraints beyond the current mini-batch. Therefore, unlike conventional memory mechanisms that mainly expand the available training relations, historical memory in RMH jointly supports reliability estimation and robust cross-batch hash learning.

B THEORETICAL ANALYSIS OF ROBUSTNESS MECHANISMS IN RMH

We provide an analytical interpretation of two robustness mechanisms in RMH. First, we analyze the initialization and subsequent evolution of the historical semantic reference, including warm-up memory contamination, its effect on historical class centers, and the post-warm-up FIFO update dynamics under explicit sufficient conditions. Second, we analyze the local gradient modulation induced by the uncertainty-aware similarity for unreliable samples. These results characterize the mechanisms underlying RMH rather than provide a global convergence guarantee for the full non-convex optimization procedure.

B.1 STABILITY OF HISTORICAL SEMANTIC REFERENCES

We analyze the historical semantic reference in two stages. We first characterize the contamination of the memory when it is initially populated during warm-up, before reliability-based admission becomes available. We then examine how this initialized reference affects the historical class centers and how its contamination evolves after MRE is activated.

Warm-up Initialization. Before MRE is activated, the Cross-modal Memory Bank is populated during the memory warm-up stage without reliability-based admission. Therefore, this stage does not guarantee a clean initial memory at the sample level. However, the sample-level corruption rate does not directly equal the contamination level of the class-specific semantic references used by MRE.

Let $\mathcal{M}_0$ denote the memory when it first reaches capacity M, and let $\mathbf{y}_j$ denote the original clean annotation of sample j. The label cardinality is preserved under the noise construction protocols considered in this work, i.e., $L_j = \|\mathbf{y}_j\|_0 = \|\tilde{\mathbf{y}}_j\|_0$. Specifically, each corrupted sample replaces one positive label with one absent category, thereby introducing one false-positive class assignment while maintaining the same number of positive labels.

Let $Z_j \in \{0, 1\}$ indicate whether sample j is corrupted. The total number of false-positive class assignments in the initial memory is therefore

$$F_0 = \sum_{j \in \mathcal{M}_0} Z_j, \tag{12}$$

while the total number of observed positive class assignments is

$$A_0 = \sum_{j \in \mathcal{M}_0} L_j. \tag{13}$$

We define the assignment-level contamination of the initialized memory as

$$\chi_0 = \frac{F_0}{A_0}. \tag{14}$$

If corrupted samples are selected independently of the warm-up memory with sample-level rate ρ, then, conditioned on the samples contained in $\mathcal{M}_0$,

$$\mathbb{E}[\chi_0 \mid \mathcal{M}_0] = \frac{\rho M}{\sum_{j \in \mathcal{M}_0} L_j} = \frac{\rho}{\bar{L}_0}, \qquad \bar{L}_0 = \frac{1}{M} \sum_{j \in \mathcal{M}_0} L_j. \tag{15}$$

For a fixed-size corruption set, the same expression holds with ρ replaced by its exact sampling fraction.

This distinction is important in the multi-label setting: a sample that is corrupted at the sample level can still retain multiple correct positive class assignments. Hence, a high sample-level noise rate does not imply an equally high contamination rate for the class assignments from which historical semantic centers are constructed.

Let $\zeta_{c,0}$ denote the category-specific contamination ratio defined below. Since every false-positive assignment contributes to exactly one category,

$$\sum_c \frac{|\mathcal{M}_{c,0}|}{A_0} \zeta_{c,0} = \chi_0. \tag{16}$$

Thus, Equation 14 characterizes the weighted average contamination of the class-specific references at the moment MRE is activated.

This analysis does not imply that every category center is clean after warm-up; category-specific contamination may vary with class frequency and the corruption target distribution. Rather, it shows why sample-level corruption and semantic-reference contamination should be distinguished. Once MRE is activated, the subsequent evolution of memory contamination is governed by the selective FIFO update analyzed below.

Effect of class-center contamination. For category c at training step t, let

$$\mathcal{M}_{c,t} = \{j \in \mathcal{M}_t \mid \tilde{y}_{jc} = 1\}$$

denote the memory entries assigned to category c according to their observed labels. For analysis only, let $\mathbf{y}_j$ denote the corresponding clean annotation. We partition $\mathcal{M}_{c,t}$ into

$$\mathcal{M}^+_{c,t} = \{j \in \mathcal{M}_{c,t} \mid y_{jc} = 1\}, \qquad \mathcal{M}^-_{c,t} = \{j \in \mathcal{M}_{c,t} \mid y_{jc} = 0\},$$

where $\mathcal{M}^-_{c,t}$ contains entries whose observed annotation incorrectly assigns them to category c. Define the category-specific contamination ratio as

$$\zeta_{c,t} = \frac{|\mathcal{M}^-_{c,t}|}{|\mathcal{M}_{c,t}|}. \tag{17}$$

Let

$$\bar{\mathbf{v}}^m_j = \frac{\mathbf{v}^m_j}{\|\mathbf{v}^m_j\|_2}$$

denote a normalized historical representation. The historical class center in Equation 2 can be decomposed as

$$\mathbf{c}^m_{c,t} = (1 - \zeta_{c,t})\bar{\mathbf{c}}^{m,+}_{c,t} + \zeta_{c,t}\bar{\mathbf{c}}^{m,-}_{c,t}, \tag{18}$$

where $\bar{\mathbf{c}}^{m,+}_{c,t}$ and $\bar{\mathbf{c}}^{m,-}_{c,t}$ denote the mean normalized representations of the semantically valid and contaminating subsets, respectively.

Let

$$\boldsymbol{\mu}^m_c = \mathbb{E}[\bar{\mathbf{v}}^m \mid y_c = 1] \tag{19}$$

denote the clean reference center of category c in modality m, and assume $\nu^m_c = \|\boldsymbol{\mu}^m_c\|_2 > 0$. Define

$$\delta^m_{c,t} = \left\|\bar{\mathbf{c}}^{m,+}_{c,t} - \boldsymbol{\mu}^m_c\right\|_2,$$

which captures the deviation of the valid historical subset from the clean reference, including finite-sample variation and temporal representation drift.

Because normalized sample representations have unit norm, $\|\bar{\mathbf{c}}^{m,-}_{c,t}\|_2 \le 1$ and $\|\boldsymbol{\mu}^m_c\|_2 \le 1$. Therefore,

$$\left\|\bar{\mathbf{c}}^{m,-}_{c,t} - \boldsymbol{\mu}^m_c\right\|_2 \le 2.$$

Using Equation 18 and the triangle inequality gives

$$\left\|\mathbf{c}^m_{c,t} - \boldsymbol{\mu}^m_c\right\|_2 \le (1 - \zeta_{c,t})\delta^m_{c,t} + 2\zeta_{c,t} \le \delta^m_{c,t} + 2\zeta_{c,t}. \tag{20}$$

MRE uses cosine similarity and therefore depends on the direction of the historical center. Let

$$\hat{\mathbf{c}}^m_{c,t} = \frac{\mathbf{c}^m_{c,t}}{\|\mathbf{c}^m_{c,t}\|_2}, \qquad \hat{\boldsymbol{\mu}}^m_c = \frac{\boldsymbol{\mu}^m_c}{\|\boldsymbol{\mu}^m_c\|_2},$$

for nonzero established centers. For any nonzero vectors $\mathbf{x}$ and $\mathbf{y}$,

$$\left\|\frac{\mathbf{x}}{\|\mathbf{x}\|_2} - \frac{\mathbf{y}}{\|\mathbf{y}\|_2}\right\|_2 \le \frac{2\|\mathbf{x} - \mathbf{y}\|_2}{\|\mathbf{y}\|_2}.$$

Hence,

$$\left\|\hat{\mathbf{c}}^m_{c,t} - \hat{\boldsymbol{\mu}}^m_c\right\|_2 \le \frac{2}{\nu^m_c}\left(\delta^m_{c,t} + 2\zeta_{c,t}\right). \tag{21}$$

For any normalized current representation $\hat{\mathbf{h}}_i$, the corresponding cosine-similarity deviation is therefore bounded by

$$\begin{aligned}\left|\mathrm{sim}(\hat{\mathbf{h}}_i, \mathbf{c}_{c,t}^m) - \mathrm{sim}(\hat{\mathbf{h}}_i, \boldsymbol{\mu}_c^m)\right| \\ \leq \left\|\hat{\mathbf{c}}_{c,t}^m - \hat{\boldsymbol{\mu}}_c^m\right\|_2 \\ \leq \frac{2}{\nu_c^m}\left(\delta_{c,t}^m + 2\zeta_{c,t}\right).\end{aligned} \tag{22}$$

Equation 22 establishes a direct class-level relationship between contamination of the historical reference and the semantic similarity used by MRE. Lower category-specific contamination yields a tighter approximation to the corresponding clean semantic reference, provided that the deviation of the valid historical subset remains controlled.

Post-warm-up FIFO memory dynamics. After the memory first reaches capacity, MRE is activated and only samples classified as reliable are admitted into the Cross-modal Memory Bank. We now analyze the sample-level contamination dynamics during this post-warm-up stage. This quantity is distinct from the category-specific contamination ratio $\zeta_{c,t}$ analyzed above. Let the fixed memory capacity be M, and let

$$\eta_t = \frac{|\{j \in \mathcal{M}_t : \tilde{\mathbf{y}}_j \neq \mathbf{y}_j\}|}{M}$$

denote the fraction of globally corrupted samples stored in the memory after it has reached capacity.

Suppose that K_t newly selected samples are admitted at update t, causing the same number of oldest entries to be removed by FIFO. Let

$$\omega_t = \frac{K_t}{M},$$

and let a_t and d_t denote the corruption ratios among the newly admitted and evicted samples, respectively. The FIFO update then satisfies the exact identity

$$\eta_{t+1} = \eta_t + \omega_t(a_t - d_t). \tag{23}$$

Thus, memory contamination decreases whenever the newly admitted subset is cleaner than the subset removed from memory, i.e., $a_t < d_t$, and it does not increase when $a_t \leq d_t$. Importantly, Equation 23 is an exact consequence of the fixed-capacity FIFO update and does not require random replacement.

A contraction-type bound can be obtained under explicit sufficient conditions. Let $\mathcal{F}_t$ denote the training history before the t-th memory update. Assume that

$$\mathbb{E}[d_t \mid \mathcal{F}_t] \geq \eta_t, \qquad \mathbb{E}[a_t \mid \mathcal{F}_t] \leq \kappa\eta_t + \xi_t, \qquad 0 \leq \kappa < 1, \tag{24}$$

where $\xi_t \geq 0$ represents residual reliability-estimation error. The first condition requires the evicted block to be no cleaner, in expectation, than the current memory average, while the second requires the contamination of newly admitted samples to grow sublinearly with the current memory contamination.

Taking conditional expectation in Equation 23 yields

$$\begin{aligned}\mathbb{E}[\eta_{t+1} \mid \mathcal{F}_t] &= \eta_t + \omega_t\left(\mathbb{E}[a_t \mid \mathcal{F}_t] - \mathbb{E}[d_t \mid \mathcal{F}_t]\right) \\ &\leq \left[1 - \omega_t(1-\kappa)\right]\eta_t + \omega_t\xi_t.\end{aligned} \tag{25}$$

For any update with $\omega_t > 0$, the coefficient $1-\omega_t(1-\kappa)$ lies in $[0, 1)$. Therefore, under the sufficient conditions in Equation 24, contamination inherited from the previous memory state is contractive up to the residual term ξ_t. If $\xi_t \leq \bar{\xi}$, the recursion is driven toward a neighborhood characterized by $\bar{\xi}/(1-\kappa)$; if ξ_t decreases toward zero and memory updates persist, the residual neighborhood correspondingly shrinks.

This result should be interpreted conditionally rather than as an unconditional guarantee of monotonic memory purification. Its implication is that RMH restricts persistent contamination to samples

that first pass the reliability filter and are subsequently admitted into the memory. The empirical Memory Purity results evaluate how well this condition is realized in practice.

The temporally smoothed threshold in Equation 4 further reduces the sensitivity of the decision boundary to an individual mini-batch threshold, although this smoothing mechanism is not required for the bounds above.

B.2 Local Gradient Modulation of Uncertainty-aware Learning

MRE prevents samples classified as unreliable from entering the historical memory, but discarding them entirely would remove potentially useful training information. RMH therefore retains these samples through the uncertainty-aware similarity in Equation 8. We analyze its local sensitivity with respect to the hash cosine similarity.

For a pair (i, j), define

$$c_{ij} = \mathrm{sim}(\mathbf{h}_i, \mathbf{h}_j). \tag{26}$$

Because $\hat{\mathbf{h}}_i$ and $\hat{\mathbf{h}}_j$ are L_2-normalized,

$$\alpha_{ij} = \|\hat{\mathbf{h}}_i - \hat{\mathbf{h}}_j\|_2 = \sqrt{2(1 - c_{ij})}. \tag{27}$$

Consider first the branch $\alpha_{ij} > \epsilon_s$. For fixed β_{ij}, define

$$x_{ij} = \frac{\beta_{ij} + \gamma}{\tau_u \alpha_{ij}} = \frac{\beta_{ij} + \gamma}{\tau_u \sqrt{2(1 - c_{ij})}}. \tag{28}$$

Since $\beta_{ij} \geq 0$, $\gamma \geq 0$, and $\tau_u > 0$, we have $x_{ij} \geq 0$. Equation 8 can then be rewritten as

$$S_u(i, j) = 1 - (1 - c_{ij}) e^{-x_{ij}}. \tag{29}$$

Differentiating with respect to c_{ij} while holding β_{ij} fixed gives

$$\frac{\partial S_u(i, j)}{\partial c_{ij}} = e^{-x_{ij}} \left(1 + \frac{x_{ij}}{2}\right). \tag{30}$$

Define

$$G(x) = e^{-x} \left(1 + \frac{x}{2}\right). \tag{31}$$

Since

$$G'(x) = -\frac{1 + x}{2} e^{-x} \leq 0, \qquad G(0) = 1,$$

we obtain

$$0 < G(x) \leq 1, \qquad x \geq 0. \tag{32}$$

Moreover, for fixed c_{ij}, and hence fixed α_{ij},

$$\frac{\partial x_{ij}}{\partial \beta_{ij}} = \frac{1}{\tau_u \alpha_{ij}} > 0.$$

Together with $G'(x) \leq 0$, this implies

$$\frac{\partial G(x_{ij})}{\partial \beta_{ij}} \leq 0. \tag{33}$$

Thus, along the hash-similarity pathway, the local sensitivity of S_u to c_{ij} is bounded by one and decreases as the uncertainty relation β_{ij} increases.

For the clipped branch $\alpha_{ij} < \epsilon_s$, the denominator in Equation 8 is constant with respect to c_{ij}, yielding

$$\frac{\partial S_u(i, j)}{\partial c_{ij}} = \exp\left(-\frac{\beta_{ij} + \gamma}{\tau_u \epsilon_s}\right) \in (0, 1]. \tag{34}$$

At the boundary $\alpha_{ij} = \epsilon_s$, the $\max$ operation is generally non-differentiable. The one-sided derivatives are

$$\exp\left(-\frac{\beta_{ij}+\gamma}{\tau_u \epsilon_s}\right)\left(1+\frac{\beta_{ij}+\gamma}{2\tau_u \epsilon_s}\right)$$

for the unclipped branch and

$$\exp\left(-\frac{\beta_{ij}+\gamma}{\tau_u \epsilon_s}\right)$$

for the clipped branch, both of which are bounded by one. Hence, the same local non-amplification property holds on both sides of the clipping boundary.

Consider the uncertainty-aware triplet term

$$\ell_{\mathrm{unc}} = [\mu + S_u(a,n) - S_u(a,p)]_+ . \tag{35}$$

Conditional on the hinge being active and holding the uncertainty relations fixed, for the unclipped branch we have

$$\frac{\partial \ell_{\mathrm{unc}}}{\partial c_{an}} = G(x_{an}), \qquad \frac{\partial \ell_{\mathrm{unc}}}{\partial c_{ap}} = -G(x_{ap}), \tag{36}$$

with the corresponding clipped factors given by Equation 34. When the hinge is inactive, both derivatives are zero. Therefore, at every differentiable point,

$$\left|\frac{\partial \ell_{\mathrm{unc}}}{\partial c_{an}}\right| \le 1, \qquad \left|\frac{\partial \ell_{\mathrm{unc}}}{\partial c_{ap}}\right| \le 1. \tag{37}$$

For comparison, an active conventional cosine-triplet loss has derivative magnitude one with respect to its positive and negative cosine similarities. Hence, the uncertainty-aware transformation does not amplify the local gradient transmitted through the cosine-similarity pathway and attenuates this pathway more strongly as β_{ij} increases.

This local result does not imply an analogous bound on the complete network gradient. At differentiable points, if the hash and uncertainty mappings share lower-level parameters θ, then

$$\nabla_\theta \ell_{\mathrm{unc}} = \sum_{(i,j)\in\{(a,n),(a,p)\}} \frac{\partial \ell_{\mathrm{unc}}}{\partial c_{ij}} \nabla_\theta c_{ij} + \sum_{(i,j)\in\{(a,n),(a,p)\}} \frac{\partial \ell_{\mathrm{unc}}}{\partial \beta_{ij}} \nabla_\theta \beta_{ij}. \tag{38}$$

The bounds above control only the first term, namely the gradient pathway transmitted through hash similarity, and do not imply that the norm of the full parameter gradient is bounded by that of a conventional triplet loss. Since $\beta_{ij} = \|\mathbf{u}_i + \mathbf{u}_j\|_2$ is non-differentiable at $\mathbf{u}_i + \mathbf{u}_j = \mathbf{0}$, the corresponding full-gradient statement is understood at differentiable points; a subgradient treatment is required at this degenerate case.

B.3 Overall Interpretation

The two analyses characterize complementary robustness mechanisms in RMH.

First, the warm-up analysis distinguishes sample-level corruption from the assignment-level contamination that directly affects class-center construction, thereby characterizing the initial contamination state from which MRE starts. Equation 22 then relates category-specific contamination to the deviation of the semantic reference used by MRE. After MRE is activated, Equation 23 describes the exact FIFO contamination dynamics, while Equation 25 establishes a contraction-type bound under explicit sufficient conditions.

Second, Equations 32–37 establish a local non-amplification property for uncertainty-aware learning along the hash-similarity pathway. Relations associated with larger β_{ij} receive weaker similarity-path gradients, while no corresponding scalar bound is claimed for the full network gradient.

Together, these results provide a conditional theoretical interpretation of RMH. After activation, MRE restricts the persistent contamination of historical semantic references through reliability-based memory admission, while uncertainty-aware learning limits the local influence of uncertain relations along the hash-similarity optimization pathway. These mechanisms constrain two pathways through which noisy supervision can affect subsequent training, without assuming an initially clean memory or claiming global convergence or complete elimination of label noise.

C DATASETS

We conduct experiments on four widely used image–text benchmark datasets. Their detailed descriptions and experimental splits are provided below, with the dataset statistics summarized in Table 5.

MIRFlickr-25K: MIRFlickr-25K contains 24,581 image-text pairs collected from Flickr (Huiskes & Lew, 2008). Each pair consists of an image and its associated textual information, and is annotated with one or more labels from 24 semantic categories.

NUS-WIDE: NUS-WIDE contains 269,648 image-text pairs covering 81 semantic categories (Chua et al., 2009). Following common practice, we retain the 21 most frequent categories and remove sparsely represented categories, resulting in 195,834 image–text pairs for evaluation.

MS COCO: MS COCO is a widely used benchmark for visual recognition and cross-modal retrieval (Lin et al., 2014). It contains 82,785 training images and 40,504 validation images, each associated with textual descriptions and semantic annotations from 80 categories. We combine the training and validation sets in our experiments.

IAPR TC-12: IAPR TC-12 contains 20,000 image–text pairs collected from a wide range of real-world scenes, including people, animals, and landscapes (Escalante et al., 2010). The dataset covers 291 semantic categories, and each image is accompanied by textual descriptions in multiple languages. We use the English captions in all experiments.

Following a unified experimental protocol (Huo et al., 2023; 2024), 10,000 image-text pairs are randomly sampled for training and 5,000 pairs are used as the query set. The retrieval database consists of all non-query samples, with the training samples included in the retrieval set. The same data splits are used for all compared methods to ensure a fair evaluation.

Table 5: Statistics of the four benchmark datasets. "Train / Query / Retrieval" denotes the numbers of image–text pairs used for training, query, and retrieval, respectively, and C denotes the number of semantic categories.

Dataset	Train / Query / Retrieval	C
MIRFlickr-25K (Huiskes & Lew, 2008)	10,000 / 5,000 / 19,581	24
NUS-WIDE (Chua et al., 2009)	10,000 / 5,000 / 190,834	21
MS COCO (Lin et al., 2014)	10,000 / 5,000 / 118,289	80
IAPR TC-12 (Escalante et al., 2010)	10,000 / 5,000 / 15,000	291

D EXPERIMENTAL SETTINGS

D.1 IMPLEMENTATION DETAILS

Experimental Setup. All experiments are implemented in PyTorch 2.3.0 (Paszke et al., 2019) and trained on two NVIDIA A100 GPUs. Following the main experiments, all methods adopt CLIP ViT-B/32 (Radford et al., 2021) as the common backbone. Images are resized to 224×224, while textual inputs are truncated or padded to a maximum length of 32 tokens. We evaluate hash codes of 16, 32, 64, and 128 bits.

Training Configuration. All models are trained for 100 epochs with a batch size of 128. For RMH, the capacity of the Cross-modal Memory Bank is set to 640, the reliability temperature is $\tau_r = 0.2$, and the smoothing window of the Smooth Threshold is set to 40. For uncertainty-aware learning, τ_u and γ are set to 9 and 4, respectively. Unless otherwise specified, the same data splits and fixed noisy-label realizations are used for all compared methods.

Optimization Details. All methods are optimized with BertAdam using the same optimizer configuration and a cosine learning-rate schedule with linear warm-up. The initial learning rates are set to 1×10^{-5} for the CLIP backbone and 1×10^{-3} for the task-specific mappings. We use $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon_{\text{Adam}} = 10^{-6}$, a weight decay of 10^{-4}, and gradient clipping with a maximum norm of 1.0. The warm-up is defined over optimization steps with a proportion of 0.1. For RMH, the

triplet margin is fixed to $\mu = 0.25$. The loss terms $\mathcal{L}_{\text{mem}}$, $\mathcal{L}_{\text{in}}$, and $\mathcal{L}_{\text{unc}}$ are equally weighted, while the quantization loss weight is set to $\lambda_q = 0.8$. Method-specific hyperparameters of the compared baselines otherwise follow their original settings.

Memory Initialization and Update. The Cross-modal Memory Bank is initially empty. Before the memory reaches its predefined capacity, MRE remains inactive, and RMH performs warm-up by applying the uncertainty-aware loss to all training samples while progressively populating the memory. Once the memory reaches capacity, MRE is activated and constructs cross-modal class centers from the historical memory for reliability estimation. Thereafter, each mini-batch is first partitioned into reliable and unreliable subsets, and only reliable samples are inserted into the memory. The memory is therefore updated at the mini-batch level using a fixed-capacity FIFO strategy, where the oldest entries are removed when new samples exceed the capacity. The image representations, text representations, and observed labels stored in the memory are detached from the current computation graph before insertion. Historical class centers are constructed by aggregating these detached representations according to their observed labels. Consequently, the memory serves only as a non-gradient historical semantic reference, and no gradients are propagated across mini-batches through the stored representations.

D.2 Noise Construction Protocols

To systematically evaluate robustness to noisy supervision, we construct symmetric label noise (Wang et al., 2024), asymmetric label noise (Pu et al., 2025), and semantic confusion noise. Let $\mathbf{y}_i \in \{0,1\}^C$ denote the original multi-label annotation of the i-th training sample and ρ denote the sample-level noise rate. We consider $\rho \in \{0.2, 0.5, 0.8\}$, while $\rho = 0$ corresponds to clean supervision. All noisy labels are generated before training and remain fixed throughout optimization.

For **symmetric label noise**, we randomly select $\lfloor \rho N \rfloor$ training samples. For each selected sample, one positive label is randomly changed from 1 to 0 and one negative label is changed from 0 to 1. Thus, each corrupted sample contains exactly one removed positive label and one introduced false-positive label, while its label cardinality remains unchanged.

For **asymmetric label noise**, corruption follows a predefined class-dependent mapping. For each selected sample, one positive source category c is replaced by its mapped target category $\phi(c)$, provided that the target category is absent from the original annotation. Each corrupted sample therefore undergoes exactly one valid source-to-target replacement without changing its label cardinality.

For **semantic confusion noise**, we construct semantically plausible label perturbations using a frozen external text encoder that is independent of the CLIP backbone used by RMH. Specifically, we employ `sentence-transformers/all-mpnet-base-v2` to encode each category name with the unified prompt "a photo containing {class name}", where {class name} is replaced by the corresponding category name. The resulting category embeddings are L_2-normalized, and their semantic similarity is measured by the inner product of normalized embeddings, which is equivalent to cosine similarity. For each selected training sample, one source category is uniformly sampled from its positive labels, and the absent category with the highest semantic similarity to the source category is selected as the target category. The source category is then replaced by the target category, preserving the number of positive labels. In the case of tied maximum similarities, the first candidate according to the dataset category index is selected deterministically. The same encoder, prompt, normalization, and similarity computation protocol are used across all datasets under this noise setting.

Only the training annotations are corrupted. The original annotations of the query and retrieval sets are retained for evaluation.

D.3 Evaluation Metrics

We evaluate cross-modal retrieval in both image-to-text (I2T) and text-to-image (T2I) directions. Two instances are regarded as relevant if they share at least one semantic label. Specifically, for a query i and a retrieval sample j, their relevance is defined as $g_{ij} = \mathbb{I}(\mathbf{y}_i^\top \mathbf{y}_j > 0)$. The retrieval results are ranked according to the Hamming distance between their binary codes. We employ mean average precision (mAP), Precision–Recall (PR) curves, and Top-N Precision as the primary

retrieval metrics (Sun et al., 2026a; Cheng et al., 2026b), and additionally report NDCG@1000 to evaluate ranking quality.

D.3.1 MEAN AVERAGE PRECISION

For the i-th query, let $\pi_i(k)$ denote the retrieval sample ranked at position k, and let

$$P_i(k) = \frac{1}{k} \sum_{j=1}^{k} g_{i\pi_i(j)} \tag{39}$$

denote the precision at rank k. The Average Precision (AP) of query i is defined as:

$$AP_i = \frac{1}{R_i} \sum_{k=1}^{M} P_i(k)\, g_{i\pi_i(k)}, \tag{40}$$

where $R_i = \sum_{j=1}^{M} g_{ij}$ is the number of relevant samples in the retrieval set and M is the retrieval-set size. The mAP is then computed as:

$$\mathrm{mAP} = \frac{1}{n} \sum_{i=1}^{n} AP_i, \tag{41}$$

where n denotes the number of queries. In our experiments, the entire retrieval set is considered when computing mAP.

D.3.2 PRECISION–RECALL CURVE

The PR curve characterizes the trade-off between precision and recall under different retrieval thresholds. For a given retrieved set, precision and recall are defined as:

$$\mathrm{Precision} = \frac{TP}{TP + FP}, \qquad \mathrm{Recall} = \frac{TP}{TP + FN}, \tag{42}$$

where TP, FP, and FN denote the numbers of relevant retrieved samples, irrelevant retrieved samples, and relevant but unretrieved samples, respectively. A PR curve that remains higher over a broad recall range indicates more accurate semantic ranking.

D.3.3 TOP-N PRECISION

Top-N Precision measures retrieval accuracy at different ranking depths. For query i, it is defined as:

$$P_i@N = \frac{1}{N} \sum_{k=1}^{N} g_{i\pi_i(k)}. \tag{43}$$

The final Top-N Precision is obtained by averaging $P_i@N$ over all queries. By varying N, this metric reflects how well semantic relevance is preserved from the top-ranked results to deeper retrieval positions.

D.3.4 NORMALIZED DISCOUNTED CUMULATIVE GAIN

We further employ Normalized Discounted Cumulative Gain (NDCG) to evaluate the ranking quality of retrieved samples. For query i, the Discounted Cumulative Gain at rank K_r is defined as:

$$DCG_i@K_r = \sum_{k=1}^{K_r} \frac{g_{i\pi_i(k)}}{\log_2(k+1)}. \tag{44}$$

where K_r denotes the ranking cutoff. Let $IDCG_i@K_r$ denote the corresponding DCG under the ideal ranking. The normalized score is:

$$NDCG@K_r = \frac{1}{n} \sum_{i=1}^{n} \frac{DCG_i@K_r}{IDCG_i@K_r}. \tag{45}$$

NDCG assigns larger weights to relevant samples appearing at higher ranks. We report NDCG@1000 in our experiments.

D.3.5 RELIABILITY ESTIMATION METRICS

To directly evaluate the reliability estimation capability of MRE, we use the known corruption masks of the synthetically generated noisy labels only for post-hoc evaluation. These ground-truth masks are not involved in reliability-score computation, thresholding, sample partitioning, memory updating decisions, or model optimization.

Let $\mathcal{C}$ and $\mathcal{N}$ denote the sets of truly clean and corrupted training samples, respectively, and let $\mathcal{R}_t$ denote the subset predicted as reliable at training step t. We evaluate reliability estimation using AUROC, Reliable Precision, False-retention Rate, and Memory Purity.

AUROC. We use the reliability score r_i^{rel} defined in Section 3.2 as the scoring function, with clean samples treated as the positive class. AUROC measures the probability that a randomly selected clean sample receives a higher reliability score than a randomly selected noisy sample:

$$\mathrm{AUROC} = \Pr\left(r_c^{\mathrm{rel}} > r_n^{\mathrm{rel}}\right) + \frac{1}{2}\Pr\left(r_c^{\mathrm{rel}} = r_n^{\mathrm{rel}}\right), \tag{46}$$

where r_c^{rel} and r_n^{rel} denote the reliability scores of randomly selected clean and noisy samples, respectively. A higher AUROC indicates better threshold-independent separability between clean and noisy samples.

Reliable Precision. Reliable Precision measures the proportion of truly clean samples among those selected as reliable:

$$\mathrm{RP}_t = \frac{|\mathcal{R}_t \cap \mathcal{C}|}{|\mathcal{R}_t|}. \tag{47}$$

A higher value indicates that the selected reliable subset contains less noisy supervision.

False-retention Rate. False-retention Rate measures the proportion of truly noisy samples that are incorrectly retained as reliable:

$$\mathrm{FRR}_t = \frac{|\mathcal{R}_t \cap \mathcal{N}|}{|\mathcal{N}|}. \tag{48}$$

A lower FRR indicates that fewer corrupted samples are allowed to enter subsequent reliable-sample learning and memory updates.

Memory Purity. For the Cross-modal Memory Bank $\mathcal{M}_t$ at training step t, Memory Purity is defined as:

$$\mathrm{MP}_t = \frac{|\mathcal{M}_t \cap \mathcal{C}|}{|\mathcal{M}_t|}. \tag{49}$$

This metric directly measures the proportion of truly clean samples retained in the historical memory. Higher Memory Purity indicates that the memory provides a more reliable historical semantic reference for subsequent reliability estimation and cross-batch learning.

E ADDITIONAL EXPERIMENTS

E.1 HISTORICAL SEMANTIC REFERENCE STRATEGY

To investigate how different semantic reference construction strategies affect reliability estimation, we compare four variants on MIRFlickr-25K with 64-bit hash codes under 50% symmetric label noise, as reported in Table 6. Current-batch does not use historical semantic information; instead, it constructs modality-specific class centers from the current mini-batch and adopts a leave-one-out scheme to exclude the evaluated sample itself from the corresponding class centers. Periodic Re-computation periodically reconstructs historical class centers from reliable samples stored in the FIFO memory and keeps the centers fixed between consecutive updates. EMA Center does not retain sample-level historical memory for MRE, but instead updates the historical center of each class and modality through an exponential moving average of the reliable samples selected at each training step. FIFO Memory maintains a fixed-capacity set of reliable historical samples and reconstructs historical class centers from the current memory at every training step. Since current samples are admitted into the memory only after their reliability has been determined, the semantic references used for current reliability estimation are strictly derived from previous training steps. Except for

the semantic reference construction used by MRE, all other training settings are kept unchanged, and all four variants use the same historical memory mechanism for RCL cross-batch learning.

The results show that all three strategies using historical semantic references outperform Current-batch, which relies only on the current mini-batch, indicating that cross-batch historical semantics provide more effective references for estimating the reliability of current samples. Among them, FIFO Memory achieves the best I2T, T2I, and average mAP, further outperforming both Periodic Re-computation and EMA Center. These results support continuously maintaining reliable historical samples and dynamically reconstructing class centers from them as an effective semantic reference strategy for MRE.

Table 6: Comparison of different semantic reference strategies on MIRFlickr-25K with 64-bit hash codes under 50% symmetric label noise. Results are reported as mAP (%).

Reference Strategy	I2T	T2I	Average
Current-batch	79.90	79.09	79.50
Periodic Re-computation	81.16	81.72	81.44
EMA Center	<u>81.30</u>	<u>81.77</u>	<u>81.54</u>
FIFO Memory	**81.52**	**81.95**	**81.74**

E.2 Reliability Aggregation Strategy

To investigate how bidirectional cross-modal semantic support should be aggregated, we compare four reliability estimation strategies on MIRFlickr-25K with 64-bit hash codes under 50% symmetric label noise: using only the $I \rightarrow T$ support, using only the $T \rightarrow I$ support, and aggregating the bidirectional supports by either the arithmetic mean or the geometric mean. The results are reported in Table 7. Jointly exploiting bidirectional semantic support generally outperforms relying on a single direction, with the geometric mean achieving the best performance in I2T, T2I, and average mAP. These results indicate that combining evidence from both cross-modal directions provides more informative reliability cues, while the geometric mean further suppresses reliability estimates dominated by high confidence in only one direction, thereby favoring samples with more consistent bidirectional semantic support.

Table 7: Comparison of different reliability aggregation strategies on MIRFlickr-25K with 64-bit hash codes under 50% symmetric label noise. Results are reported as mAP (%).

Aggregation Strategy	I2T	T2I	Average
$I \rightarrow T$	80.76	81.28	81.02
$T \rightarrow I$	80.31	81.73	81.02
Arithmetic Mean	80.86	81.79	81.33
Geometric Mean	**81.52**	**81.95**	**81.74**

E.3 Positive-label Support Aggregation

To examine how positive-label semantic support should be aggregated in multi-label settings, we compare sum support and mean support on MIRFlickr-25K with 128-bit hash codes under 50% and 80% symmetric label noise, as reported in Table 8. For each cross-modal direction, sum support directly accumulates the matching probabilities assigned to all observed positive labels, whereas mean support further divides this quantity by the number of valid positive labels; all other training settings remain unchanged. The two strategies achieve comparable retrieval performance: sum support performs slightly better at 50% noise, while mean support has a small advantage at 80% noise, indicating that RMH is not sensitive to this aggregation choice. We adopt sum support because it directly measures the total probability mass assigned to all observed positive labels and avoids additionally scaling the semantic support by the number of positive labels, which better matches the multi-label semantic setting.

Table 8: Comparison of positive-label support aggregation strategies on MIRFlickr-25K with 128-bit hash codes under different symmetric noise rates. Results are reported as mAP (%).

Aggregation Strategy	50% Noise			80% Noise		
	I2T	T2I	Average	I2T	T2I	Average
Mean Support	81.62	**82.87**	82.25	79.79	**81.18**	**80.49**
Sum Support	**82.11**	82.74	**82.43**	**80.56**	80.12	80.34

E.4 Uncertainty-aware Optimization and Gradient Stability

To evaluate the effectiveness and optimization stability of uncertainty-aware learning, we compare four strategies for optimizing unreliable samples on MIRFlickr-25K with 64-bit hash codes under 50% and 80% symmetric label noise. Fixed Weight does not use learned uncertainty and instead assigns the same fixed coefficient to all unreliable relations, optimizing them with conventional cosine-based triplet constraints so that all relations contribute with equal strength. Learned Weight converts the learned relation uncertainty into a relation-specific scalar weight that directly scales the corresponding relation loss, while jointly optimizing the uncertainty and hash representations, thereby providing a direct learnable reweighting mechanism. Proposed-UA (Detached) retains the proposed uncertainty-aware similarity but stops gradient propagation through the uncertainty representation for this objective. Proposed-UA (Joint) uses the complete uncertainty-aware similarity and jointly optimizes the hash and uncertainty representations. All other training settings are kept unchanged across the four variants.

As shown in Table 9, Proposed-UA (Joint) achieves the best I2T, T2I, and average mAP under both noise rates. Its advantage over the detached variant becomes more pronounced under 80% noise, indicating that structured uncertainty-aware modulation together with joint uncertainty optimization is particularly beneficial when unreliable supervision becomes more severe.

Figure 8 further illustrates the optimization dynamics under 80% noise. For Learned Weight, the relation uncertainty continuously increases, while both the effective gradient-scaling factor and the direct gradient of the uncertainty-aware objective with respect to the hash representation decay toward zero. This behavior indicates that direct multiplicative reweighting can admit a degenerate optimization path in which increasing uncertainty progressively suppresses the corresponding supervision signal. In contrast, Proposed-UA (Joint) maintains relation uncertainty within a finite range, keeps the gradient modulation factor away from zero, and preserves non-vanishing gradients along the hash-similarity pathway.

Taken together, the retrieval results and optimization dynamics show that Fixed Weight lacks relation-adaptive modulation, Learned Weight is susceptible to gradient collapse, and the detached variant prevents the uncertainty representation from being updated through the uncertainty-aware objective. In contrast, Proposed-UA (Joint) maintains effective adaptive gradient modulation without exhibiting self-gating collapse under the evaluated settings, while achieving the strongest retrieval performance. These results support the proposed uncertainty-aware similarity and joint optimization design.

Table 9: Comparison of different uncertainty optimization strategies on MIRFlickr-25K with 64-bit hash codes under 50% and 80% symmetric label noise. Results are reported as mAP (%).

Strategy	50% Noise			80% Noise		
	I2T	T2I	Average	I2T	T2I	Average
Fixed Weight	<u>80.08</u>	81.18	80.63	<u>76.79</u>	77.28	77.04
Learned Weight	77.88	80.01	78.95	71.12	74.03	72.58
Proposed-UA (Detached)	79.92	<u>81.46</u>	<u>80.69</u>	76.75	<u>77.62</u>	<u>77.19</u>
Proposed-UA (Joint)	**81.52**	**81.95**	**81.74**	**79.60**	**79.63**	**79.62**

E.5 Additional Hyperparameter Analysis

To further examine the robustness of RMH to hyperparameter selection, we conduct additional sensitivity analyses on several key parameters, including the uncertainty-aware parameters τ_u and γ,

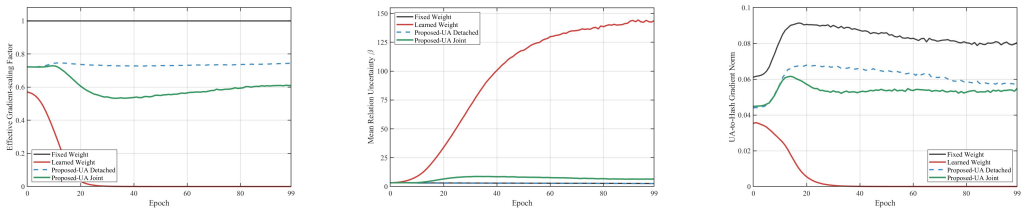


Figure 8: Training dynamics of different uncertainty optimization strategies on MIRFlickr-25K with 64-bit hash codes under 80% symmetric label noise. (a) Effective gradient-scaling factor along the hash-similarity pathway. (b) Mean relation uncertainty β. (c) Direct gradient norm of the uncertainty-aware objective with respect to the hash representations. Learned Weight exhibits progressive uncertainty inflation together with vanishing effective scaling and hash-path gradients, whereas Proposed-UA (Joint) maintains bounded uncertainty modulation and non-vanishing gradients throughout training.

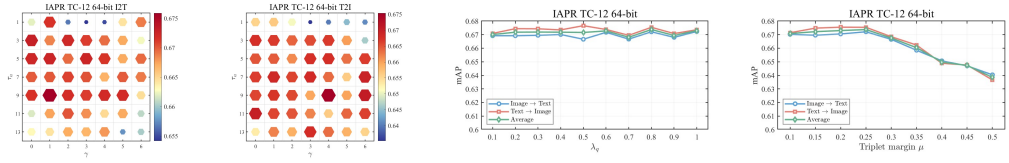


Figure 9: Sensitivity analysis of hyperparameters τ_u, γ, λ_q, and μ on IAPR TC-12 with 64 bits under 50% symmetric label noise.

the quantization-loss weight λ_q, and the triplet margin μ. All experiments are conducted on IAPR TC-12 using 64-bit hash codes under 50% symmetric label noise.

Figure 9 presents the corresponding sensitivity results. For the uncertainty-aware parameters τ_u and γ, RMH maintains relatively stable retrieval performance over a broad range of values, with the highest average mAP achieved at $(\tau_u, \gamma) = (9, 4)$. For the quantization-loss weight λ_q, the performance varies only moderately across the evaluated range, and $\lambda_q = 0.8$ yields the highest average mAP. For the triplet margin μ, the best average performance is obtained at $\mu = 0.25$, whereas further increasing the margin leads to a noticeable performance degradation. Based on these results, we set $\tau_u = 9$, $\gamma = 4$, $\lambda_q = 0.8$, and $\mu = 0.25$ as the default hyperparameter configuration in our experiments.

Table 10: mAP results (%) of RMH and baselines under different symmetric noise rates on MIRFlickr-25K with four hash-code lengths.

Task	Method	16 bits				32 bits				64 bits				128 bits			
		0.0	*0.2*	*0.5*	*0.8*	*0.0*	*0.2*	*0.5*	*0.8*	*0.0*	*0.2*	*0.5*	*0.8*	*0.0*	*0.2*	*0.5*	*0.8*
Image → Text	DCHMT	82.01	73.33	68.99	68.80	82.53	74.60	72.79	69.34	82.22	76.27	73.47	67.94	82.59	77.15	74.81	72.25
	DDWSH	77.26	72.01	60.55	58.90	83.20	74.65	70.16	65.77	84.13	77.12	72.60	68.65	85.89	80.11	74.87	71.35
	DSSH	83.51	81.20	79.06	76.66	83.56	82.33	79.90	76.59	84.41	83.22	80.87	76.87	84.51	83.28	81.02	76.80
	CMMQ	77.74	76.24	69.85	66.02	79.27	77.62	74.83	67.95	80.45	78.65	75.89	69.95	81.06	79.56	76.77	72.64
	NRCH	79.03	77.01	75.85	75.46	80.06	79.16	77.22	76.18	81.56	79.26	77.54	75.36	81.32	78.71	77.01	75.87
	NIRNL	77.94	76.92	75.64	74.84	79.71	78.63	76.51	75.29	79.86	77.67	76.19	75.30	79.74	77.46	76.43	76.26
	RFCMH	80.74	79.18	78.51	76.76	80.33	79.66	78.52	77.43	80.66	79.64	79.00	77.85	81.33	79.88	79.17	77.48
	RMH	**83.94**	**81.53**	**79.14**	**77.06**	**85.46**	**82.65**	**80.54**	**78.31**	**86.61**	**84.27**	**81.52**	**79.60**	**87.14**	**85.38**	**82.11**	**80.56**
Text → Image	DCHMT	79.83	71.73	69.17	68.62	80.48	73.79	72.61	69.80	80.31	76.27	73.81	69.50	80.70	75.80	74.50	73.01
	DDWSH	77.37	73.67	64.00	63.89	80.86	75.47	72.56	69.55	82.61	77.27	74.47	72.46	83.08	79.34	75.91	74.24
	DSSH	81.36	79.24	78.03	75.86	81.24	80.22	79.23	77.66	82.02	81.36	79.93	78.16	82.17	81.73	80.61	78.48
	CMMQ	77.56	76.86	70.35	66.02	79.31	78.36	75.75	68.70	80.72	79.21	77.03	69.98	81.28	80.26	78.29	73.80
	NRCH	77.88	77.13	75.77	75.54	79.25	79.06	77.94	76.85	80.48	79.23	78.02	76.52	80.65	79.22	77.98	76.79
	NIRNL	77.28	77.00	75.93	75.40	78.03	77.83	77.15	76.28	78.86	78.03	76.89	76.25	78.61	77.86	77.52	77.30
	RFCMH	80.63	78.36	77.89	77.43	79.02	78.74	78.07	78.14	79.96	78.76	78.63	78.28	80.92	79.14	79.19	78.33
	RMH	**81.59**	**80.44**	**79.65**	**77.45**	**82.72**	**81.79**	**80.96**	**78.88**	**83.72**	**83.02**	**81.95**	**79.63**	**84.23**	**83.73**	**82.74**	**80.12**

Table 11: mAP results (%) of RMH and baselines under different symmetric noise rates on NUS-WIDE with four hash-code lengths.

Task	Method	16 bits				32 bits				64 bits				128 bits			
		0.0	*0.2*	*0.5*	*0.8*	*0.0*	*0.2*	*0.5*	*0.8*	*0.0*	*0.2*	*0.5*	*0.8*	*0.0*	*0.2*	*0.5*	*0.8*
Image → Text	DCHMT	65.96	60.42	50.66	43.20	67.06	64.94	53.74	44.71	68.63	66.19	56.15	44.50	69.83	65.68	56.48	44.60
	DDWSH	66.09	57.32	46.31	37.15	70.76	59.27	46.31	40.33	71.67	63.76	48.02	43.67	72.47	66.17	51.27	45.79
	DSSH	70.53	67.30	62.98	52.45	71.07	69.79	64.59	52.67	72.53	71.22	67.33	53.39	73.48	72.56	68.82	55.70
	CMMQ	66.86	65.24	60.52	35.80	69.51	67.82	64.81	46.61	71.35	69.29	66.40	46.02	71.69	69.91	64.75	48.24
	NRCH	65.99	66.09	63.75	61.35	69.79	65.76	61.54	60.61	69.40	65.74	61.72	60.91	70.19	65.32	62.98	61.36
	NIRNL	65.36	65.04	65.19	61.98	64.34	65.31	67.84	65.97	65.19	65.98	68.38	67.50	67.84	67.01	68.64	67.01
	RFCMH	68.91	68.82	66.57	61.84	69.80	69.87	68.60	65.81	70.26	70.24	69.29	67.02	71.50	71.23	69.82	**68.11**
	RMH	**71.42**	**68.94**	**66.89**	**65.22**	**72.05**	**70.85**	**69.64**	**67.64**	**73.46**	**72.60**	**71.42**	**67.92**	**74.50**	**73.46**	**72.92**	68.01
Text → Image	DCHMT	67.61	62.19	52.68	44.83	68.37	66.31	55.41	46.85	69.43	67.48	57.72	46.86	70.87	67.19	57.50	46.78
	DDWSH	67.66	59.25	48.56	38.57	71.73	60.43	48.96	44.35	72.53	63.61	49.09	46.22	73.19	66.87	51.87	47.81
	DSSH	71.35	68.75	63.98	53.87	72.06	71.01	67.23	54.98	73.16	71.42	68.03	55.93	73.77	72.95	68.83	57.24
	CMMQ	68.95	67.39	62.05	36.85	71.58	69.98	66.65	47.51	72.91	71.39	68.18	47.00	73.29	71.94	66.90	47.62
	NRCH	68.37	67.34	65.07	62.71	70.59	66.63	62.74	62.21	70.49	66.66	62.53	61.85	70.94	66.31	63.59	62.18
	NIRNL	66.51	66.36	66.56	63.11	66.04	67.29	68.90	66.87	67.05	67.32	69.50	68.13	69.63	69.07	69.70	67.57
	RFCMH	69.89	**70.23**	67.54	62.68	71.06	71.27	69.81	66.91	71.87	71.56	70.19	67.82	71.50	72.13	70.77	69.02
	RMH	**72.81**	**70.23**	**69.46**	**67.41**	**72.95**	**71.88**	**71.69**	**69.69**	**74.25**	**73.71**	**72.83**	**69.66**	**75.05**	**74.55**	**74.56**	**69.48**

Table 12: mAP results (%) of RMH and baselines under different symmetric noise rates on MS COCO with four hash-code lengths.

Task	Method	16 bits				32 bits				64 bits				128 bits			
		0.0	*0.2*	*0.5*	*0.8*	*0.0*	*0.2*	*0.5*	*0.8*	*0.0*	*0.2*	*0.5*	*0.8*	*0.0*	*0.2*	*0.5*	*0.8*
Image → Text	DCHMT	63.09	59.05	57.97	55.94	62.16	60.11	59.08	57.55	65.53	60.75	60.48	58.81	67.82	61.84	58.72	58.01
	DDWSH	67.18	61.44	49.92	47.27	70.91	67.86	60.80	55.85	73.42	72.01	64.78	59.10	75.51	74.59	69.78	61.97
	DSSH	70.02	66.80	64.57	62.07	71.54	70.75	64.78	65.99	73.68	72.84	68.34	68.30	75.61	74.29	70.43	70.44
	CMMQ	61.20	59.65	49.56	43.91	69.66	66.39	57.21	42.56	75.21	71.26	60.79	47.36	76.61	71.97	61.86	47.85
	NRCH	67.38	68.25	64.89	64.25	72.15	70.68	67.02	65.73	74.80	71.52	67.61	65.69	75.70	71.96	69.02	67.00
	NIRNL	60.15	59.82	58.41	55.69	61.78	62.00	61.08	58.34	64.23	62.98	60.16	58.42	64.93	64.02	63.27	61.10
	RFCMH	67.60	69.45	**68.33**	**66.96**	68.75	70.60	70.24	68.55	70.84	71.71	71.19	70.02	71.35	71.97	71.17	70.85
	RMH	**70.12**	**69.51**	67.77	65.96	**73.65**	**72.57**	**70.74**	**68.80**	**75.23**	**74.95**	**73.05**	**70.19**	**76.77**	**76.12**	**74.83**	**72.10**
Text → Image	DCHMT	62.41	58.35	57.34	55.26	62.12	59.66	58.26	57.55	64.86	59.94	59.06	57.60	67.78	60.91	59.13	57.74
	DDWSH	66.24	60.01	51.85	48.06	70.35	65.78	59.38	55.83	73.67	68.06	61.76	58.53	75.58	71.83	65.24	60.47
	DSSH	68.92	66.11	64.12	62.32	70.52	70.18	65.02	66.21	73.10	72.36	68.55	68.51	73.84	73.87	70.36	70.05
	CMMQ	60.33	57.78	46.95	39.16	69.19	65.77	56.06	40.37	75.03	70.13	59.76	47.82	76.90	70.49	61.04	48.03
	NRCH	66.77	68.43	64.73	63.93	72.33	70.40	67.00	66.02	74.67	71.52	67.92	66.10	76.00	72.26	69.22	67.31
	NIRNL	60.08	59.55	58.39	55.45	61.98	60.99	61.00	58.38	64.05	62.58	60.26	58.71	64.63	64.22	63.49	60.84
	RFCMH	67.39	68.97	**68.60**	65.95	68.75	70.52	70.77	68.80	70.54	72.05	71.52	70.40	70.90	72.08	71.13	71.11
	RMH	**69.77**	**69.32**	68.23	**66.82**	**73.71**	**72.47**	**71.46**	**68.91**	**75.78**	**74.98**	**73.73**	**71.09**	**77.35**	**77.02**	**75.74**	**72.85**

Table 13: mAP results (%) of RMH and baselines under different symmetric noise rates on IAPR TC-12 with four hash-code lengths.

Task	Method	16 bits				32 bits				64 bits				128 bits			
		0.0	*0.2*	*0.5*	*0.8*	*0.0*	*0.2*	*0.5*	*0.8*	*0.0*	*0.2*	*0.5*	*0.8*	*0.0*	*0.2*	*0.5*	*0.8*
Image → Text	DCHMT	60.21	54.51	53.64	53.42	61.96	56.96	55.72	56.23	62.54	57.66	56.27	55.94	62.69	59.34	58.20	58.09
	DDWSH	51.96	50.40	46.98	43.35	61.70	55.64	53.57	52.92	64.41	59.71	57.78	57.33	66.71	61.64	60.52	60.05
	DSSH	64.20	62.16	61.35	60.44	67.03	**65.75**	**64.60**	**64.08**	69.15	68.60	**67.55**	66.68	71.00	68.93	**69.01**	67.65
	CMMQ	45.34	44.15	42.96	41.74	51.59	50.78	49.51	44.65	57.06	53.72	50.45	45.42	62.86	59.02	51.75	43.65
	NRCH	52.71	52.36	52.08	50.30	56.95	56.88	55.88	54.89	59.66	57.60	56.65	55.93	61.41	58.66	57.42	56.13
	NIRNL	50.96	50.01	50.88	50.41	52.06	51.67	52.02	52.04	52.71	52.63	52.55	53.11	53.28	52.83	53.16	53.00
	RFCMH	55.39	55.65	55.81	54.30	58.09	58.20	57.25	56.62	60.05	60.07	58.64	58.65	59.45	60.61	59.89	58.99
	RMH	**64.52**	**62.52**	**62.44**	**60.65**	**67.08**	65.32	64.41	63.43	**70.05**	**68.98**	67.21	**66.80**	**72.21**	**70.34**	68.89	**67.79**
Text → Image	DCHMT	60.68	50.82	49.78	48.52	62.86	54.11	53.22	52.23	63.87	55.77	56.54	54.88	63.13	57.75	55.51	55.90
	DDWSH	48.42	45.44	41.96	38.87	61.13	52.69	50.76	48.94	64.38	58.33	56.04	54.70	67.12	62.28	58.95	57.73
	DSSH	62.37	59.83	59.12	57.81	65.04	64.37	62.91	62.00	67.86	67.26	65.70	65.03	69.21	67.00	67.80	65.58
	CMMQ	45.40	44.56	42.71	41.34	49.74	49.21	48.06	43.24	55.49	52.82	48.73	45.26	62.34	59.07	51.41	43.55
	NRCH	51.73	50.95	51.16	49.45	56.52	56.59	54.80	54.25	59.47	57.35	56.22	55.51	60.46	58.10	56.50	55.56
	NIRNL	50.72	49.72	50.80	50.34	51.97	51.54	51.70	51.95	52.56	52.50	52.23	52.63	52.90	52.44	52.73	52.67
	RFCMH	54.74	55.35	54.93	54.07	56.84	57.33	56.76	56.36	58.51	58.84	58.26	58.22	58.97	58.94	59.28	58.60
	RMH	**62.89**	**61.82**	**60.49**	**58.30**	**67.21**	**65.77**	**64.48**	**63.33**	**70.61**	**68.57**	**67.54**	**65.44**	**73.30**	**70.45**	**69.72**	**67.37**

E.6 Complete Retrieval Results

The main paper reports the average mAP over the I2T and T2I retrieval directions to provide a compact comparison under different noisy-label settings. Here, we further report the complete mAP results for the two retrieval directions separately. Specifically, the results under symmetric label noise on MIRFlickr-25K, NUS-WIDE, MS COCO, and IAPR TC-12 are presented in Tables 10, 11, 12, and 13, respectively. The corresponding results under asymmetric label noise on NUS-WIDE and semantic confusion noise on IAPR TC-12 are reported in Tables 14 and 15, respectively.

Across both retrieval directions, RMH exhibits performance trends consistent with the average mAP results reported in the main paper. In particular, its advantages remain evident as the noise rate increases, indicating that the robustness of RMH is not specific to either I2T or T2I retrieval. These

Table 14: mAP results (%) of RMH and baselines under different asymmetric noise rates on NUS-WIDE with four hash-code lengths.

Task	Method	16 bits				32 bits				64 bits				128 bits			
		0.0	*0.2*	*0.5*	*0.8*	*0.0*	*0.2*	*0.5*	*0.8*	*0.0*	*0.2*	*0.5*	*0.8*	*0.0*	*0.2*	*0.5*	*0.8*
Image → Text	DCHMT	65.96	60.52	56.46	57.50	67.06	65.10	61.49	63.41	68.63	66.21	64.34	64.15	69.83	67.27	63.95	64.61
	DDWSH	66.09	56.25	52.61	53.27	70.76	57.93	51.01	54.28	71.67	62.01	52.70	54.64	72.47	65.44	57.84	58.67
	DSSH	70.53	66.38	63.89	64.55	71.07	68.94	67.12	65.28	72.53	**71.33**	67.82	67.65	73.48	**72.70**	69.55	68.67
	CMMQ	66.86	65.13	58.52	43.12	69.51	67.19	63.11	43.22	71.35	69.23	65.57	50.78	71.69	69.81	65.68	50.98
	NRCH	65.99	65.72	64.08	63.78	69.79	65.80	63.00	62.31	69.40	65.57	62.63	63.17	70.19	65.85	63.78	63.14
	NIRNL	65.36	65.23	66.05	65.40	64.34	66.55	66.09	65.53	65.19	67.34	67.71	67.81	67.84	68.07	68.03	67.51
	RFCMH	68.91	68.64	67.29	66.01	69.80	70.60	**68.94**	67.00	70.26	70.23	69.74	68.44	71.50	71.00	70.02	68.88
	RMH	**71.42**	**68.98**	**67.65**	**66.24**	**72.05**	**70.82**	68.92	**67.21**	**73.46**	71.09	**70.14**	**68.74**	**74.50**	72.53	**71.14**	**69.22**
Text → Image	DCHMT	67.61	62.47	58.58	60.06	68.37	66.10	63.19	64.72	69.43	67.61	65.54	65.79	70.87	68.63	65.13	66.57
	DDWSH	67.66	57.96	55.62	55.57	71.73	57.30	52.39	56.15	72.53	61.35	53.35	54.92	73.19	66.26	58.38	59.01
	DSSH	71.35	68.41	65.29	66.90	72.06	70.22	68.42	66.53	73.16	72.19	68.85	68.59	73.77	73.23	70.34	69.63
	CMMQ	68.95	67.14	60.35	43.27	71.58	69.47	64.64	47.52	72.91	70.98	67.23	52.18	73.29	71.63	67.36	51.57
	NRCH	68.37	67.29	65.18	65.38	70.59	66.71	63.92	63.50	70.49	67.11	64.10	64.06	70.94	66.83	64.49	64.23
	NIRNL	66.51	66.89	67.21	66.51	66.04	67.72	67.35	67.53	67.05	68.54	69.11	68.69	69.63	69.28	68.50	68.72
	RFCMH	69.89	70.27	68.46	67.54	71.06	**71.66**	70.06	69.20	71.87	71.35	70.85	**70.40**	71.50	72.13	71.36	69.93
	RMH	**72.81**	**70.30**	**69.34**	**68.84**	**72.95**	71.41	**70.50**	**69.60**	**74.25**	**72.74**	**71.72**	69.24	**75.05**	**73.49**	**72.87**	**70.98**

Table 15: mAP results (%) of RMH and baselines under different semantic confusion noise rates on IAPR TC-12 with four hash-code lengths.

Task	Method	16 bits				32 bits				64 bits				128 bits			
		0.0	*0.2*	*0.5*	*0.8*	*0.0*	*0.2*	*0.5*	*0.8*	*0.0*	*0.2*	*0.5*	*0.8*	*0.0*	*0.2*	*0.5*	*0.8*
Image → Text	DCHMT	60.21	54.94	54.10	53.76	61.96	56.47	56.31	56.85	62.54	58.59	58.95	58.41	62.69	59.66	59.64	59.63
	DDWSH	51.96	51.64	50.10	48.06	61.70	55.41	55.38	53.50	64.41	58.55	59.24	58.10	66.71	63.04	61.76	61.72
	DSSH	64.20	60.23	60.03	**59.54**	67.03	**65.11**	**64.46**	**63.78**	69.15	**67.49**	**67.00**	**66.29**	71.00	69.00	**68.76**	**68.02**
	CMMQ	45.34	44.58	43.30	42.30	51.59	51.11	49.27	46.17	57.06	54.14	49.40	43.80	62.86	59.48	51.64	44.64
	NRCH	52.71	51.73	50.86	52.97	56.95	56.45	54.99	54.22	59.66	57.82	56.06	55.75	61.41	58.97	57.70	55.96
	NIRNL	50.96	50.40	50.44	50.28	52.06	51.95	51.98	51.67	52.71	52.66	53.09	52.80	53.28	53.17	52.68	53.25
	RFCMH	55.39	55.80	54.83	53.86	58.09	58.14	58.24	56.89	60.05	59.76	58.66	58.63	59.45	60.29	59.45	57.05
	RMH	**64.52**	**60.80**	**60.25**	59.39	**67.08**	64.41	63.94	63.69	**70.05**	66.92	66.67	65.85	**72.21**	**69.56**	68.34	67.92
Text → Image	DCHMT	60.68	51.31	50.22	50.11	62.86	54.54	53.16	53.97	63.87	56.75	57.06	56.15	63.13	57.72	57.34	56.96
	DDWSH	48.42	46.19	45.63	43.77	61.13	52.99	52.46	50.70	64.38	57.39	57.11	55.77	67.12	62.93	60.22	59.33
	DSSH	62.37	58.46	57.64	57.54	65.04	62.78	61.66	60.97	67.86	65.46	64.67	63.20	69.21	66.73	66.47	65.83
	CMMQ	45.40	44.57	42.98	41.35	49.74	49.16	47.80	44.49	55.49	53.01	49.12	43.96	62.34	58.83	52.33	43.90
	NRCH	51.73	51.27	50.47	52.21	56.52	56.31	54.46	54.20	59.47	57.90	56.09	54.94	60.46	58.53	56.88	55.76
	NIRNL	50.72	50.23	50.17	50.11	51.97	51.71	51.37	51.12	52.56	52.54	52.76	52.32	52.90	52.77	51.97	52.94
	RFCMH	54.74	55.00	54.33	53.59	56.84	57.13	57.50	56.60	58.51	58.03	58.44	57.43	58.97	59.33	59.29	56.41
	RMH	**62.89**	**59.90**	**58.58**	**58.46**	**67.21**	**63.78**	**62.75**	**61.70**	**70.61**	**65.87**	**66.06**	**64.09**	**73.30**	**68.38**	**67.94**	**66.39**

results further demonstrate that RMH can consistently preserve cross-modal semantic relationships in both retrieval directions under noisy supervision.

E.7 NDCG@1000 Results

To further assess the ranking quality of the retrieval results, we report NDCG@1000, which assigns larger weights to relevant samples appearing at higher ranks. As shown in Figure 10, RMH consistently achieves the highest NDCG@1000 across all three datasets and both retrieval directions, indicating its ability to rank semantically relevant cross-modal instances more accurately under noisy supervision. This advantage can be attributed to the reliable semantic relationships preserved by RMH. Historical reliable memory reduces the influence of corrupted supervision, while uncertainty-aware learning further suppresses unreliable semantic relations. As a result, RMH learns more robust semantic structures and produces higher-quality retrieval rankings.

E.8 Efficiency Analysis

To further evaluate the computational efficiency of RMH, we compare it with representative baselines in terms of retrieval performance, parameter count, FLOPs, training time, and encoding latency. The experiments are conducted on MIRFlickr-25K under 80% symmetric label noise with 32-bit hash codes, and the results are reported in Table 16. FLOPs are measured for encoding one paired image–text query, while Encoding Time denotes the average latency per query pair.

RMH achieves the best retrieval performance while maintaining computational costs comparable to the competing methods. Its modest increase in parameter count mainly arises from the uncertainty mappings introduced for robust training. These mappings, together with the memory mechanism, reliability estimation, and uncertainty-aware optimization, are not required during retrieval. Consequently, RMH retains the encoder and hash mappings as its deployment path and exhibits FLOPs

and encoding latency comparable to the baselines. These results indicate that the robustness improvement of RMH is achieved without introducing substantial additional inference cost.

Table 16: Efficiency comparison on MIRFlickr-25K under 80% symmetric label noise with 32-bit hash codes. Training time denotes the estimated total time for 100 training epochs. Encoding time is measured as the average latency for encoding one image-text query pair.

Method	FLOPs (G)	Params. (M)	Training Time (h)	Encoding Time (ms)	mAP (%)
DCHMT	11.16	151.43	1.27	1.285	69.57
DSSH	11.15	151.31	0.95	1.271	77.13
NIRNL	11.15	151.31	0.78	1.262	75.79
RFCMH	11.15	151.31	0.89	1.274	77.79
RMH	11.15	151.34	0.88	1.270	78.60

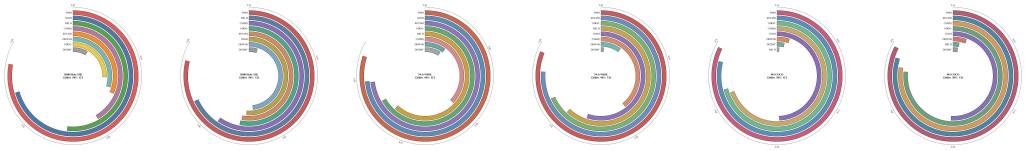


Figure 10: NDCG@1000 results (%) of RMH and the baselines on three benchmark datasets with 128 bits under 50% symmetric noise.

E.9 Backbone Generalization

To evaluate the generalization capability of RMH across different vision-language pretrained backbones, we further conduct experiments with a variety of CLIP, BLIP, and SigLIP2 architectures (Li et al., 2022; Tschannen et al., 2025). All experiments are performed on the MIRFlickr-25K dataset with 32-bit hash codes under 50% symmetric noise, where only the backbone is replaced while keeping the remaining training strategies unchanged. As shown in Table 17, we report the I2T and T2I mAP (%) results, along with other metrics, including the number of parameters in each backbone (in millions), total training time for 100-epoch training (in hours), encoding time for query dataset encoding (in milliseconds), and the best epoch achieving the highest performance within 100 epochs.

RMH achieves stable cross-modal retrieval performance across different architectures, with all backbones obtaining an average mAP above 80%. Among them, SigLIP2-L/16-256 achieves the highest performance with an average mAP of 82.20%, while CLIP-ViT-B/32 with a relatively smaller parameter count still obtains an average mAP of 80.75%. These results indicate that RMH does not rely on a specific large-scale backbone and can effectively exploit the cross-modal semantic representations provided by diverse vision-language models. Furthermore, the results show that model size does not exhibit a strictly positive correlation with retrieval performance, suggesting that the proposed historical semantic-guided reliability estimation and noise-robust learning mechanisms remain effective across different backbone architectures. Although different backbones introduce varying computational costs, all models achieve stable retrieval performance within the training schedule, further demonstrating the adaptability of RMH to diverse vision-language representation models.

E.10 Statistical Significance

To further assess the statistical reliability of the retrieval improvements, we conduct significance tests over 10 independent runs under 50% symmetric label noise using 64-bit hash codes. Specifically, paired one-sided t-tests are performed between RMH and each baseline on both I2T and T2I retrieval tasks over MIRFlickr-25K and IAPR TC-12. As shown in Table 18, RMH consistently achieves higher mAP than the compared methods, and all 12 comparisons are statistically significant ($p \leq 1.5 \times 10^{-4}$). These results further support the statistical reliability of the retrieval improvements achieved by RMH across different datasets and retrieval directions.

Table 17: Various metrics of RMH with different backbones on MIRFlickr-25K with 32 bits under 50% symmetric noise.

Metric	CLIP			BLIP		SigLIP2		
	ViT-B/32	*ViT-L/14*	*RN50x16*	*I-C-Base*	*I-C-Large*	*Base-16/224*	*Base-32/256*	*Large-16/256*
Parameters (M)	151.3	427.7	291.1	223.8	446.2	375.3	377.0	881.7
Image→Text	80.54	81.75	80.55	79.07	80.87	79.02	79.21	82.20
Text→Image	80.96	81.93	80.76	82.02	81.63	81.51	81.29	82.21
Average mAP	80.75	81.84	80.65	80.54	81.25	80.26	80.25	82.20
Training Time (h)	0.88	8.97	5.55	8.11	21.45	1.66	1.78	8.18
Encoding Time (ms)	4809.3	51137.0	24769.3	39426.2	105127.2	11616.7	4920.4	44930.6
Best Epoch	35	90	25	25	35	40	40	75

Table 18: Paired one-sided t-tests over 10 independent runs under 50% symmetric label noise using 64-bit hash codes. The mAP (%) values are reported as mean $\pm$ standard deviation. The p-values are raw one-sided values for testing whether RMH outperforms each baseline. All 12 comparisons are statistically significant at $\alpha = 0.05$.

	MIRFlickr-25K				IAPR TC-12			
Method	**I2T**	p	**T2I**	p	**I2T**	p	**T2I**	p
DSSH	79.26±0.18	1.3×10^{-8}	79.78±0.15	1.8×10^{-11}	66.71±0.19	1.5×10^{-4}	64.42±0.34	3.4×10^{-9}
RFCMH	78.85±0.16	2.1×10^{-9}	78.84±0.23	2.4×10^{-11}	59.00±0.32	3.3×10^{-15}	58.20±0.25	3.7×10^{-14}
NRCH	77.43±0.24	2.6×10^{-11}	78.01±0.29	5.0×10^{-11}	56.62±0.38	3.9×10^{-14}	56.01±0.36	6.1×10^{-14}
RMH	**81.10**±0.15	–	**81.89**±0.12	–	**67.01**±0.16	–	**67.33**±0.26	–

E.11 VISUALIZATION

E.11.1 VISUALIZATION OF RELIABILITY-SCORE DISTRIBUTIONS

As shown in Figure 11, we compare the reliability-score distributions obtained using different semantic references under the same current representations. Current-batch constructs semantic centers from the current mini-batch, Historical-all constructs historical semantic references from a fixed-capacity FIFO memory updated with all current samples without reliability filtering, while RMH uses the same memory mechanism but admits only samples identified as reliable. The clean and noisy samples exhibit substantial overlap under Current-batch. After introducing historical information, Historical-all shows a clearer separation between the two groups. RMH further promotes the concentration of clean samples in the high-reliability region and noisy samples in the low-reliability region through reliability-driven memory updates, indicating that historically reliable semantics can reduce noise contamination in the semantic references and improve the discriminability of reliability estimation.

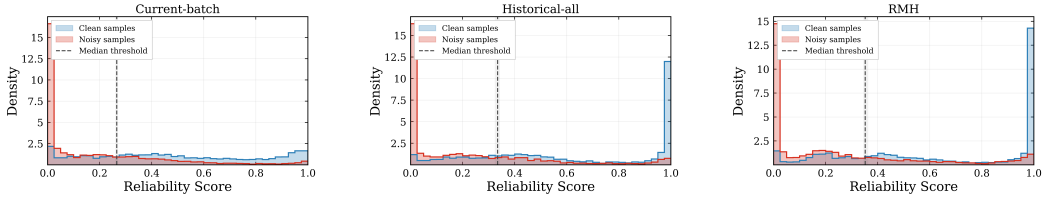


Figure 11: Reliability-score distributions of clean and noisy samples on NUS-WIDE with 64 bits under 50% symmetric noise at the 50th training epoch. Current-batch, Historical-all, and RMH share the same current representations and differ only in the references used for reliability estimation.

E.11.2 TOP-10 RETRIEVAL VISUALIZATION

On MS COCO with 64-bit hash codes and 80% symmetric label noise, we further visualize the top-10 cross-modal retrieval results of RMH to qualitatively assess its semantic retrieval capability under severe label corruption. As shown in Figure 12, despite the presence of a large proportion of incorrect labels in the training data, the retrieved samples exhibit a high degree of semantic

consistency with the query samples. The top-ranked results show clear relevance in terms of objects, scenes, and actions, indicating that the learned hash representations can preserve a stable cross-modal semantic structure even under extremely noisy conditions.

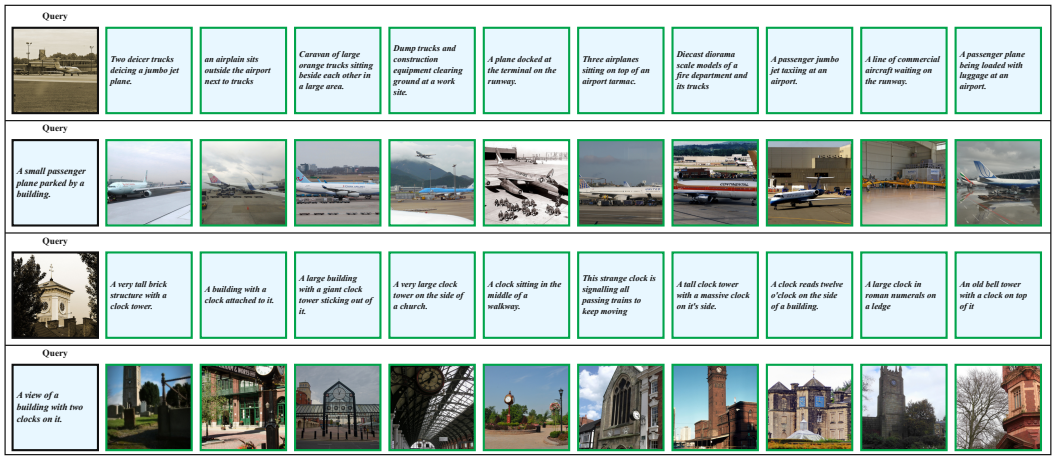


Figure 12: Top-10 retrieval results of RMH on MS COCO with 64 bits under 80% symmetric noise.