REVISITING NOISY-LABEL CROSS-MODAL HASHING WITH DUAL-SOURCE SEMANTIC CONSENSUS

Anonymous authors
Paper under double-blind review

ABSTRACT

Most cross-modal hashing (CMH) methods encourage semantic consistency across modalities, implicitly assuming that cross-modal agreement is uniformly beneficial. Under noisy labels, however, different modalities can respond differently to the same supervision, making indiscriminate semantic alignment susceptible to propagating unstable predictions across modalities. This motivates a key question: how can cross-modal semantic interaction be conditionally regulated when modality-specific predictions exhibit heterogeneous reliability? To address this issue, we propose a dual-source semantic consensus framework, termed DUET, which regulates cross-modal semantic interaction from complementary temporal and instantaneous perspectives. Specifically, Temporal Consensus Learning (TCL) aggregates predictions across training epochs to characterize their relative temporal stability, while Modality Consensus Learning (MCL) exploits the concentration of current prediction distributions to adaptively determine each modality's contribution to the semantic consensus and the corresponding consensus constraint. By jointly considering these complementary cues, DUET avoids treating all modality predictions as equally influential and enables conditional semantic interaction during cross-modal learning. Extensive experiments on three benchmark datasets demonstrate that DUET consistently improves retrieval performance across a broad range of noise levels and exhibits effective adaptive behavior under heterogeneous modality-specific responses to noisy supervision.

1 INTRODUCTION

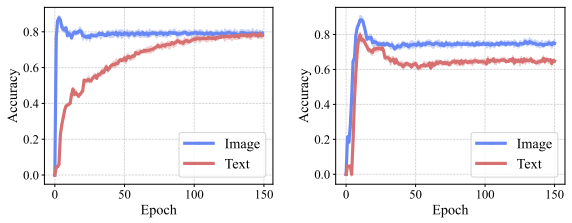
With the rapid growth of multimodal data, cross-modal retrieval increasingly depends on learning a shared semantic space from heterogeneous observations (Tu et al., 2025; Liu et al., 2025; Liang et al., 2024). Cross-modal hashing (CMH) addresses this challenge by mapping different modalities into a unified binary Hamming space, enabling efficient retrieval of semantically related instances across modalities (Qin et al., 2022; Hu et al., 2021b; Tu et al., 2023; Zhu et al., 2022; Wang et al., 2025; 2024; Hu et al., 2024). The central premise of such a shared space is that different modalities can be optimized toward a common semantic target. In practice, however, this target is often constructed from imperfect labels. Semantic ambiguity, subjective annotation, and automatically generated supervision inevitably introduce label noise (Zhang et al., 2022a; Hu et al., 2021a; Patel & Sastry, 2023; Li et al., 2025b). Consequently, corrupted semantic supervision may affect modality-specific representations differently and make indiscriminate cross-modal semantic alignment potentially harmful.

To address CMH learning under noisy supervision, most existing studies treat label noise as a modality-agnostic perturbation and focus on suppressing its adverse effects through robust loss functions, regularization constraints, or sample selection strategies (Li et al., 2024; Hu et al., 2021a; Li et al., 2025a; Pu et al., 2025; Feng et al., 2023). Under this paradigm, noisy annotations are implicitly regarded as a common semantic reference across modalities, while their influence is controlled through a unified optimization mechanism. However, such a treatment overlooks an important characteristic of multimodal learning: the same noisy label does not necessarily affect different modalities to the same extent. Owing to differences in representation capacity, semantic separability, and information redundancy, modalities may exhibit substantially different responses to identical label corruption. Some modalities may maintain relatively stable semantic predictions, whereas others may undergo pronounced fluctuations or systematic deviations. In such cases, enforcing uniform

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.

cross-modal consistency can couple semantically heterogeneous predictions and potentially transfer unstable information from one modality to another. Therefore, the key issue is not simply how to enforce stronger cross-modal agreement, but when cross-modal consistency is beneficial and how its influence should be conditionally regulated according to modality-specific prediction behavior.

Here, an empirical study is conducted to examine whether different modalities exhibit distinct responses to the same noisy labels. As shown in Figure 1, the image and text modalities exhibit distinct classification-accuracy trajectories under identical noise rates during independent training, revealing different degrees of overfitting to corrupted labels. When cross-modal alignment is introduced, such modality-specific overfitting behaviors become coupled through shared semantic constraints, potentially transferring noisy or unstable predictions across modalities. This observation suggests that uniform cross-modal interaction may be suboptimal when modalities differ in their susceptibility to noisy-label overfitting. Cross-modal learning should therefore account for modality-specific prediction states and adapt the strength of semantic interaction accordingly. This motivates the following question: how can semantic consensus be formed while adaptively regulating cross-modal interaction according to the evolving prediction behavior of each modality?



(a) Unimodal Training. (b) Joint Training.

Figure 1: Illustration of modality-dependent responses to label noise and their influence on cross-modal learning. XMedia with a noise rate of 0.4 and 32-bit hash codes is used as an example. (a) **Unimodal Training:** Under independent training, the image and text modalities exhibit different prediction dynamics and performance trajectories under the same noisy supervision. (b) **Joint Training:** Introducing cross-modal alignment further couples the prediction behaviors of the two modalities, leading to different effects compared with independent unimodal training.

Revisiting noisy-label CMH from this perspective, we propose a dual-source semantic consensus framework, termed DUET, as shown in Figure 2. Rather than treating all modality-specific predictions as equally informative, DUET exploits two complementary sources of semantic evidence to regulate cross-modal interaction: temporal stability accumulated during training and the concentration of the current prediction. Temporal Consensus Learning (TCL) tracks the evolution of modality-specific predictions across training epochs and aggregates persistent semantic patterns to attenuate temporally varying fluctuations. The resulting temporal stability provides a relative cue for characterizing the consistency of modality-specific predictions over time. Modality Consensus Learning (MCL) then examines the current prediction distributions and uses their relative concentration to determine each modality's contribution to the cross-modal semantic consensus. Less concentrated predictions receive stronger consensus constraints, whereas more concentrated predictions contribute more to the aggregated semantic state. These two complementary cues regulate both the relative contribution of each modality and the strength of the consensus imposed on it, thereby avoiding indiscriminate cross-modal alignment and enabling conditional semantic interaction.

The main contributions of this work are summarized as follows:

- We revisit noisy-label CMH from the perspective of modality-dependent semantic behavior and show that identical label corruption can induce substantially different prediction dynamics across modalities. This observation reveals a potential limitation of uniform cross-modal consistency, which may become harmful when modality-specific predictions exhibit heterogeneous responses to noisy supervision.
- We propose a dual-source semantic consensus framework, termed DUET, for conditionally regulating cross-modal semantic interaction. TCL exploits temporal prediction stability to attenuate transient fluctuations, while MCL uses current prediction concentration to adapt the contribution and consensus constraint of each modality.
- Extensive experiments on three benchmark datasets demonstrate that DUET consistently improves cross-modal retrieval performance across a broad range of noise levels. Further ablation and reliability analyses validate the contribution of the dual-source consensus mechanism and its adaptive modulation of cross-modal semantic interaction.

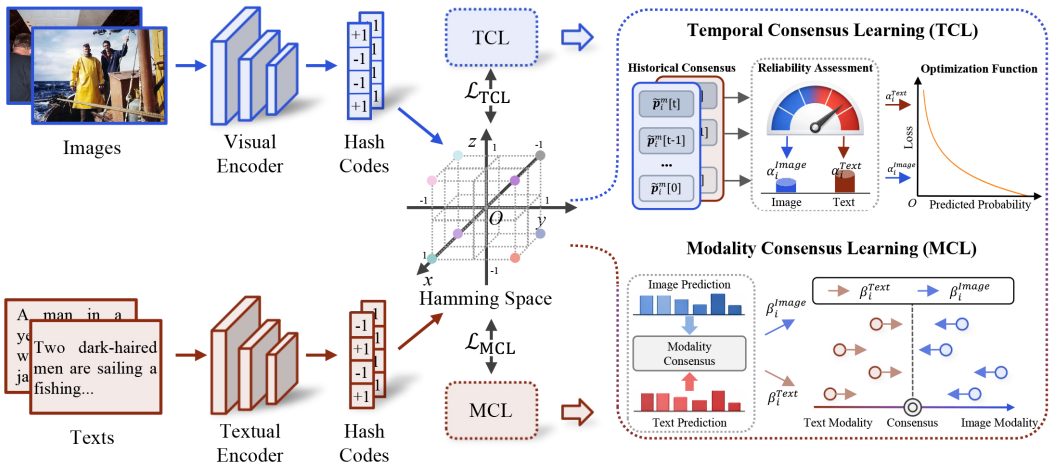


Figure 2: Framework of the proposed DUET for cross-modal hashing with noisy labels. DUET regulates cross-modal semantic interaction using two complementary prediction cues: TCL captures temporal prediction stability within each modality, while MCL exploits current prediction concentration to adaptively regulate modality contributions and consensus constraints.

2 Methodology

2.1 Problem Formulation

To establish a rigorous formulation, we consider the problem of CMH under noisy supervision. Let $\mathcal{D} = \{(\{\boldsymbol{x}_i^m\}_{m=1}^M, \tilde{y}_i)\}_{i=1}^N$ denote a multimodal training dataset containing N instances observed from M heterogeneous modalities, where $\boldsymbol{x}_i^m \in \mathcal{X}^m$ represents the observation of the i-th instance in the m-th modality and $\mathcal{X}^m$ denotes the corresponding modality-specific input space. Each instance is associated with an only underlying ground-truth semantic category $y_i \in \{1, \ldots, K\}$, which is not directly accessible during training. Instead, only a noisy label $\tilde{y}_i \in \{1, \ldots, K\}$ is available, where $\tilde{y}_i$ may differ from y_i due to annotation errors, semantic ambiguity, or automated data collection. Given the noisy multimodal training data, the objective of CMH is to learn modality-specific hash functions that map heterogeneous observations into a common Hamming space, such that semantically related instances are effectively aligned across modalities while remaining robust to corrupted semantic supervision and heterogeneous modality-specific sensitivity to label noise.

2.2 Cross-modal Hashing Representations

Building upon the above problem formulation, we learn a modality-specific hashing network $\mathcal{H}^m(\cdot;\theta^m)$ for each modality m, where θ^m denotes the learnable parameters of the corresponding network. Given an input $\boldsymbol{x}_i^m$, the network produces an L-dimensional continuous representation:

$$\boldsymbol{h}_i^m = \mathcal{H}^m(\boldsymbol{x}_i^m;\theta^m) \in \mathbb{R}^L. \tag{1}$$

The corresponding binary hash code is defined as:

$$\boldsymbol{b}_i^m = \operatorname{sign}(\boldsymbol{h}_i^m) \in \{-1,+1\}^L. \tag{2}$$

These binary codes provide a common representation for cross-modal retrieval, where semantic similarity can be efficiently measured in the Hamming space. Since the $\operatorname{sign}(\cdot)$ function in Equation (2) is non-differentiable, directly optimizing the binary codes would lead to a discrete optimization problem. We therefore adopt a differentiable relaxation during training:

$$\hat{\boldsymbol{b}}_i^m = \tanh(\boldsymbol{h}_i^m) \in (-1,1)^L. \tag{3}$$

Here, $\hat{\boldsymbol{b}}_i^m$ denotes the continuous hash representation used for gradient-based optimization, whereas $\boldsymbol{b}_i^m$ denotes the final binary hash code used for retrieval. Accordingly, the training objectives are formulated on $\hat{\boldsymbol{b}}_i^m$, while the $\operatorname{sign}(\cdot)$ operation is applied to $\boldsymbol{h}_i^m$ only during inference to obtain the final binary codes.

2.3 TEMPORAL CONSENSUS LEARNING

Given that noisy labels can affect different modalities to different extents, the semantic predictions learned from different modalities may exhibit distinct temporal behaviors during optimization. We therefore establish modality-specific semantic evidence from the temporal evolution of each modality. Rather than relying solely on the prediction at the current training epoch, TCL maintains a historical semantic state to accumulate persistent prediction patterns.

To obtain a semantic representation that can be tracked throughout training, we introduce a set of learnable class prototypes $\boldsymbol{\mathcal{C}} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k, \ldots, \boldsymbol{c}_K\}$, where $\boldsymbol{c}_k \in \mathbb{R}^L$ represents the prototype of the k-th semantic category in the continuous hash space. Each prototype is ℓ_2-normalized to provide a comparable semantic reference across categories:

$$\bar{\boldsymbol{c}}_k = \boldsymbol{c}_k / \|\boldsymbol{c}_k\|_2. \tag{4}$$

The prototypes are jointly optimized with the hashing networks, allowing the semantic references to adapt to the evolving distribution of the learned representations. Given the continuous hash representation $\hat{\boldsymbol{b}}_i^m$ obtained from Equation (3), the semantic affinity between instance i from modality m and category k is defined as

$$s_{i,k}^m = (\hat{\boldsymbol{b}}_i^m)^\top \bar{\boldsymbol{c}}_k / \tau, \tag{5}$$

where $\tau > 0$ controls the sharpness of the semantic prediction. Collecting the category-wise affinity scores gives the score vector $\boldsymbol{s}_i^m = [s_{i,1}^m, \ldots, s_{i,K}^m]^\top \in \mathbb{R}^K$. The semantic prediction is then represented by a soft probability distribution over the K semantic categories:

$$\boldsymbol{p}_i^m = \mathrm{softmax}\,(\boldsymbol{s}_i^m) \in (0,1)^K, \tag{6}$$

where the probability assigned to category k is

$$p_{i,k}^m = \frac{\exp(s_{i,k}^m)}{\sum_{j=1}^K \exp(s_{i,j}^m)}, \tag{7}$$

and $\sum_{k=1}^K p_{i,k}^m = 1$. Here, $p_{i,k}^m$ denotes the predicted probability assigned to category k, while $\boldsymbol{p}_i^m$ collects the predicted probabilities over all K categories. Thus, $\boldsymbol{s}_i^m$ represents the unnormalized semantic affinity scores, whereas $\boldsymbol{p}_i^m$ represents their normalized probabilistic form. A hard semantic prediction can be obtained as $\hat{y}_i^m = \arg\max_k p_{i,k}^m$ for prediction analysis, but it is not involved in the differentiable training objectives.

The semantic prediction evolves continuously as the hashing network is optimized. Here, we denote the prediction of instance i from modality m at epoch t by $\boldsymbol{p}_i^m[t]$. To retain semantic evidence that persists across epochs, TCL maintains a temporal consensus state $\tilde{\boldsymbol{p}}_i^m[t]$. Since the observed noisy label may itself be incorrect, initializing this state from $\tilde{y}_i$ could introduce an undesirable bias into the subsequent temporal aggregation. Instead, the temporal state is initialized from the model prediction at the beginning of training:

$$\tilde{\boldsymbol{p}}_i^m[0] = \boldsymbol{p}_i^m[0]. \tag{8}$$

For subsequent epochs, the state is updated as

$$\tilde{\boldsymbol{p}}_i^m[t] = \eta \tilde{\boldsymbol{p}}_i^m[t-1] + (1-\eta)\boldsymbol{p}_i^m[t], \tag{9}$$

where $\eta \in [0,1)$ controls the contribution of historical predictions. During implementation, the historical state is detached from the computation graph and treated as a fixed reference for optimization. Consequently, semantic evidence that remains consistent across training epochs is progressively accumulated, while short-term prediction fluctuations are attenuated. The accumulated temporal state can further serve as a self-supervisory target for stabilizing the current semantic prediction. At epoch t, the temporal consensus loss for modality m is defined as

$$\ell_i^m[t] = -\sum_{k=1}^K \tilde{p}_{i,k}^m[t-1] \log p_{i,k}^m[t]. \tag{10}$$

Here, $\tilde{\boldsymbol{p}}_i^m[t-1]$ is treated as a fixed historical target, while $\boldsymbol{p}_i^m[t]$ is the current prediction being optimized. This objective encourages the current prediction to remain consistent with the accumulated

intra-modal semantic evidence, thereby reducing unnecessary temporal fluctuations without directly forcing the model to reproduce the potentially corrupted noisy label.

The temporal evolution of the semantic prediction also provides a basis for characterizing its stability. We measure the deviation of the current prediction from its historical state by

$$\Delta_i^m[t] = \|\boldsymbol{p}_i^m[t] - \tilde{\boldsymbol{p}}_i^m[t-1]\|_2^2. \tag{11}$$

A smaller $\Delta_i^m[t]$ indicates that the current prediction remains closer to its historical semantic state, whereas a larger value indicates a stronger temporal change. We convert this discrepancy into a temporal stability score:

$$r_i^m[t] = \exp\left(-\frac{\Delta_i^m[t]}{\sigma}\right), \tag{12}$$

where $\sigma > 0$ controls the sensitivity of the stability score to temporal changes. A larger $r_i^m[t]$ therefore indicates stronger temporal consistency between the current prediction and its historical state. Because the same instance may exhibit different temporal stability across modalities, the stability scores are further normalized to characterize their relative temporal stability:

$$\alpha_i^m[t] = \frac{\exp\left(r_i^m[t]/\tau\right)}{\sum_{m'=1}^{M} \exp\left(r_i^{m'}[t]/\tau\right)}. \tag{13}$$

where the resulting $\alpha_i^m[t]$ reflects the relative temporal stability of modality m for instance i at epoch t. These weights are then used to adapt the contribution of different modalities:

$$\mathcal{L}_{\mathrm{TCL}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M} \alpha_i^m[t]\ell_i^m[t]. \tag{14}$$

TCL thus establishes a temporal semantic reference for each modality. The historical state $\tilde{\boldsymbol{p}}_i^m[t]$ consolidates persistent semantic evidence, while $\alpha_i^m[t]$ quantifies cross-modal relative temporal stability. Importantly, temporal stability serves only as a reliability cue rather than a guarantee of semantic correctness: a modality may remain temporally stable even under noisy-induced systematic prediction bias. Nevertheless, following the neural network memorization effect, deep networks prioritize true patterns before fitting label noise; temporally-stable predictions are thus empirically more likely semantically correct.

2.4 Modality Consensus Learning

Although TCL characterizes the temporal stability of modality-specific semantic predictions, temporal stability alone does not describe the current predictive state of each modality. MCL therefore considers the uncertainty of the current predictions and uses their relative concentration to determine the contribution of each modality to the cross-modal semantic consensus.

At epoch t, the predictive uncertainty of modality m for instance i is quantified by the entropy of its soft prediction:

$$H_i^m[t] = -\sum_{k=1}^{K} p_{i,k}^m[t] \log p_{i,k}^m[t]. \tag{15}$$

A lower entropy indicates a more concentrated prediction distribution, meaning that the modality exhibits a stronger preference toward a particular semantic category for the current instance. Such prediction concentration provides a relative signal for determining the contribution of each modality to the semantic consensus. Accordingly, the normalized consensus weight is defined as

$$\beta_i^m[t] = \frac{\exp\left(-H_i^m[t]/\tau\right)}{\sum_{m'=1}^{M} \exp\left(-H_i^{m'}[t]/\tau\right)}, \tag{16}$$

where a larger $\beta_i^m[t]$ indicates that modality m has a more concentrated prediction relative to the other modalities for the current instance. Importantly, this relative weighting reflects the current predictive state rather than semantic correctness.

The modality-specific predictions are then aggregated according to their relative weights to form a semantic consensus:

$$\boldsymbol{q}_i[t] = \sum_{m=1}^{M} \beta_i^m[t]\boldsymbol{p}_i^m[t]. \tag{17}$$

Since $\sum_{m=1}^{M} \beta_i^m[t] = 1$ and each $\boldsymbol{p}_i^m[t]$ is a valid probability distribution, $\boldsymbol{q}_i[t]$ also constitutes a valid probability distribution. The consensus is therefore dominated by modalities whose current predictions are more concentrated, while the contribution of less concentrated predictions is correspondingly reduced.

To further regulate the influence of modalities with different degrees of prediction concentration, MCL imposes stronger consensus constraints on modalities with smaller consensus weights. Specifically, the constraint strength for modality m is set to $1 - \beta_i^m[t]$, leading to the following loss:

$$\mathcal{L}_{\mathrm{MCL}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{m=1}^{M} (1 - \beta_i^m[t]) \, \mathrm{KL}\left(\mathrm{sg}\left(\boldsymbol{q}_i[t]\right) | \boldsymbol{p}_i^m[t]\right), \tag{18}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation, which prevents gradients from propagating through the consensus distribution. Consequently, a modality with a larger $\beta_i^m[t]$ contributes more strongly to the consensus while receiving a weaker consensus constraint, whereas a modality with a smaller $\beta_i^m[t]$ contributes less to the consensus and is subject to stronger guidance toward the aggregated semantic state. This asymmetric formulation prevents less concentrated predictions from exerting the same influence as more concentrated predictions while encouraging cross-modal consistency.

2.5 Training Objective

The optimization of DUET consists of two stages. Let t denote the training epoch, E_{warm} the number of warm-up epochs, and E_{all} the total number of training epochs. During the warm-up stage ($t < E_{warm}$), only TCL is activated to establish temporally persistent semantic evidence within each modality. After the warm-up stage, MCL is introduced to further exploit the current predictive state for adaptive cross-modal consensus learning. The overall objective is formulated as

$$\mathcal{L} = \begin{cases} \mathcal{L}_{\mathrm{TCL}}, & \text{if } t < E_{warm}, \\ \mathcal{L}_{\mathrm{TCL}} + \lambda \mathcal{L}_{\mathrm{MCL}}, & \text{if } E_{warm} \le t < E_{all}, \end{cases} \tag{19}$$

where λ controls the contribution of modality consensus learning. TCL exploits the temporal evolution of modality-specific predictions to preserve persistent semantic evidence, while MCL uses the current prediction concentration to determine the relative contribution of each modality to the cross-modal consensus and adapt the corresponding consensus constraint.

2.6 Theoretical Analysis

To further characterize the properties of DUET under noisy supervision, we establish two theoretical results concerning the temporal aggregation in TCL and the weighted consensus in MCL.

Theorem 1. *Assume that the semantic prediction of modality m at epoch t can be modeled as $\hat{\boldsymbol{p}}_i^m[t] = \boldsymbol{y}_i^* + \boldsymbol{\epsilon}_i^m[t]$, where $\boldsymbol{y}_i^*$ denotes the underlying semantic representation and $\boldsymbol{\epsilon}_i^m[t]$ denotes a temporally varying prediction perturbation. Assume that the perturbations are zero-mean, $\mathbb{E}[\boldsymbol{\epsilon}_i^m[t]] = 0$, temporally uncorrelated, i.e., $\mathbb{E}[(\boldsymbol{\epsilon}_i^m[t])^\top \boldsymbol{\epsilon}_i^m[s]] = 0$ for $t \neq s$, and have bounded second moments $\mathbb{E}[\|\boldsymbol{\epsilon}_i^m[t]\|_2^2] \le \sigma^2$. Assume further that the initial temporal state is fixed. For the temporal consensus state updated by TCL as $\tilde{\boldsymbol{p}}_i^m[t] = \eta \tilde{\boldsymbol{p}}_i^m[t-1] + (1-\eta)\hat{\boldsymbol{p}}_i^m[t]$, where $0 < \eta < 1$, the semantic deviation satisfies*

$$\mathbb{E}\left[\|\tilde{\boldsymbol{p}}_i^m[t] - \boldsymbol{y}_i^*\|_2^2\right] \le \eta^{2t} \|\tilde{\boldsymbol{p}}_i^m[0] - \boldsymbol{y}_i^*\|_2^2 + \frac{(1-\eta)^2(1-\eta^{2t})}{1-\eta^2}\sigma^2. \tag{20}$$

As $t \to \infty$, we have $\lim_{t\to\infty} \mathbb{E}\left[\|\tilde{\boldsymbol{p}}_i^m[t] - \boldsymbol{y}_i^*\|_2^2\right] \le \frac{1-\eta}{1+\eta}\sigma^2$. Therefore, under the assumed zero-mean and temporally uncorrelated perturbations, temporal aggregation attenuates temporally varying prediction fluctuations.

Theorem 2. *Assume that the prediction of modality m can be expressed as $\hat{\boldsymbol{p}}_i^m = \boldsymbol{y}_i^* + \boldsymbol{\iota}_i^m$, where $\boldsymbol{\iota}_i^m$ denotes the residual modality-specific error satisfying $\|\boldsymbol{\iota}_i^m\|_2 \le \delta_m$. Given the MCL consensus $\hat{\boldsymbol{q}}_i = \sum_{m=1}^{M} \beta_i^m[t]\hat{\boldsymbol{p}}_i^m$, where $\beta_i^m[t] \ge 0$ and $\sum_{m=1}^{M} \beta_i^m[t] = 1$, the consensus error satisfies*

$$\|\hat{\boldsymbol{q}}_i - \boldsymbol{y}_i^*\|_2 \le \sum_{m=1}^{M} \beta_i^m[t]\delta_m. \tag{21}$$

Theorem 2 shows that the consensus error is bounded by the weighted residual errors of the participating modalities. While assigning larger weights to modalities with smaller residual errors can tighten this bound, the entropy-based weights $\beta_i^m[t]$ only characterize the relative contribution of each modality according to its current prediction concentration. MCL further assigns stronger consensus constraints to lower-weight modalities through $1 - \beta_i^m[t]$, encouraging them to conform to the aggregated semantic state.

3 Experiments and Analysis

3.1 Datasets and Evaluation Metric

To comprehensively evaluate the effectiveness and robustness of the proposed method, we conduct extensive experiments on three representative cross-modal retrieval benchmark datasets, including XMedia (Peng et al., 2015), XMediaNet (Peng et al., 2018), and INRIA-Websearch (Krapac et al., 2010). These datasets cover diverse multimodal retrieval scenarios and exhibit substantial variations in data scale, modality characteristics, and semantic categories. Detailed descriptions of the datasets, data preprocessing procedures, and implementation settings are provided in the Appendix Section E. Following previous cross-modal retrieval studies (Li et al., 2024), we adopt mean Average Precision over all retrieved samples (mAP@All) as the primary evaluation metric, which measures the overall ranking quality of the retrieved results without restricting the evaluation to a predefined number of top-ranked samples. To comprehensively assess bidirectional retrieval capability, we report the final performance by averaging the mAP@All scores of the two retrieval directions, i.e., image-to-text and text-to-image retrieval. The symbol "/" denotes that the corresponding method cannot be evaluated under the specified setting due to out-of-memory issues.

3.2 Implementation Details

All experiments are implemented in PyTorch and conducted on an NVIDIA RTX 3090 GPU. For each modality, the backbone network is followed by a lightweight projection head consisting of three fully connected layers to obtain compact hash representations. The first two layers perform feature transformation, while the last layer projects the features into an L-dimensional hashing space. A $\tanh$ activation function is applied to generate bounded continuous hash codes, which are subsequently binarized for retrieval. The model is trained for $E_{all} = 150$ epochs, with an initial warm-up stage of $E_{warm} = 3$ epochs. The learning rate is fixed at $lr = 1 \times 10^{-4}$ throughout training. The temperature parameter is set to $\tau = 0.5$, the stability scale is set to $\sigma = 0.1$, and the batch size is fixed at $B = 256$ for all experiments. Additional analyses on the sensitivity of the hyperparameters η and λ are provided in the Appendix Section F.5.

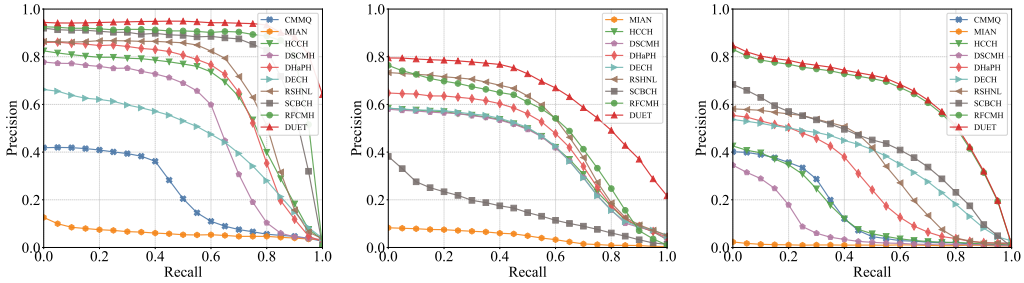
3.3 Experiments and Analysis

Comparison with the State-of-the-Art Methods: To evaluate the robustness of DUET under noisy supervision, we compare it with nine representative CMH methods, including CMMQ (Yang et al., 2022), MIAN (Zhang et al., 2022b), HCCH (Sun et al., 2023), DSCMH (Sun et al., 2024), DHaPH (Huo et al., 2024), DECH (Li et al., 2025c), RSHNL (Pu et al., 2025), SCBCH (Peng et al., 2026), and RFCMH (Sun et al., 2026). The mAP (%) results on XMedia, XMediaNet, and INRIA-Websearch are reported in Table 1, with the corresponding precision-recall curves shown in Figure 3. Several observations can be drawn: 1) Most methods exhibit noticeable performance degradation as the noise rate increases, whereas DUET maintains relatively stable retrieval performance across different noise levels, demonstrating its robustness to increasingly corrupted supervision. 2) Longer hash codes generally yield better retrieval performance by providing greater capacity for semantic representation, although slight non-monotonic variations are observed in some settings. 3) DUET achieves the best or highly competitive performance in most settings, generally outperforming RFCMH on XMedia and XMediaNet while remaining competitive on INRIA-Websearch, demonstrating the effectiveness and generalization of the proposed semantic consensus strategy under diverse noisy-label conditions.

Ablation Analysis: To disentangle the contributions of temporal stability, modality-dependent weighting, and adaptive consensus constraints, we evaluate five DUET variants on three benchmark datasets under different noise rates, as shown in Table 2.

Table 1: The mAP scores (%) with different bits under different noise rates on XMedia, XMediaNet, and INRIA-Websearch datasets. The best and second best results are marked in **bold** and underline.

	Methods	Venues	Noise Rate: 0.2				Noise Rate: 0.4				Noise Rate: 0.6				Noise Rate: 0.8			
			16	32	64	128	16	32	64	128	16	32	64	128	16	32	64	128
XMedia	CMMQ (Yang et al., 2022)	CVPR'22	85.8	85.6	84.7	82.4	72.8	75.9	71.0	68.2	48.4	39.9	36.0	41.1	16.4	16.2	13.4	11.4
	MIAN (Zhang et al., 2022b)	TKDE'22	16.8	25.8	31.5	32.2	11.7	15.2	17.0	15.6	9.6	9.8	11.2	10.9	7.3	7.5	7.0	7.8
	HCCH (Sun et al., 2023)	TMM'23	70.3	81.3	81.0	84.2	70.8	76.2	77.4	80.3	59.2	60.8	60.0	72.6	26.2	41.0	41.2	48.4
	DSCMH (Sun et al., 2024)	AAAI'24	80.8	85.3	87.1	88.0	78.7	82.6	84.4	85.3	63.3	70.6	74.2	77.4	30.0	34.0	42.4	46.1
	DHaPH (Huo et al., 2024)	TKDE'24	79.1	84.9	87.6	89.2	69.6	78.0	83.2	85.3	51.9	65.4	72.8	80.2	42.5	48.8	60.3	70.6
	DECH (Li et al., 2025c)	AAAI'25	72.2	77.3	78.4	80.4	62.3	63.1	67.6	68.7	52.9	54.2	54.5	53.8	49.5	49.2	49.6	50.1
	RSHNL (Pu et al., 2025)	AAAI'25	88.0	89.7	91.0	90.7	88.6	90.2	90.7	90.0	82.8	87.6	89.0	88.6	80.7	87.4	87.2	85.0
	SCBCH (Peng et al., 2026)	AAAI'26	35.4	75.0	82.8	86.0	33.4	66.9	83.5	85.6	38.5	75.4	82.9	84.0	31.0	69.8	79.3	84.2
	RFCMH (Sun et al., 2026)	ICML'26	88.6	90.4	91.7	91.2	88.8	90.1	91.0	91.0	88.5	89.4	90.2	91.1	87.1	89.4	89.4	89.9
	DUET (Ours)	-	**90.2**	**91.6**	**92.2**	**92.6**	**89.6**	**91.4**	**92.0**	**91.8**	**89.7**	**91.3**	**92.0**	**92.2**	**87.9**	**90.9**	**91.6**	**92.0**
XMediaNet	CMMQ (Yang et al., 2022)	CVPR'22	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
	MIAN (Zhang et al., 2022b)	TKDE'22	0.8	1.4	1.5	2.2	0.8	1.1	1.2	1.6	0.7	0.9	0.9	1.1	0.7	0.8	0.8	0.9
	HCCH (Sun et al., 2023)	TMM'23	1.7	1.8	3.4	12.0	1.4	1.4	2.6	12.1	1.4	1.2	1.8	3.8	0.8	0.9	1.0	1.6
	DSCMH (Sun et al., 2024)	AAAI'24	4.7	9.2	17.3	28.1	4.2	7.8	14.2	24.2	3.0	4.9	9.8	18.0	1.6	2.2	3.6	6.2
	DHaPH (Huo et al., 2024)	TKDE'24	10.4	17.0	24.8	30.6	9.8	16.6	24.0	29.6	9.8	16.0	23.4	29.4	9.2	16.2	22.8	29.0
	DECH (Li et al., 2025c)	AAAI'25	12.7	13.5	14.4	13.6	13.5	13.4	10.6	11.2	3.7	5.0	5.6	5.0	0.4	0.6	0.6	1.0
	RSHNL (Pu et al., 2025)	AAAI'25	35.3	43.0	47.0	48.2	35.2	41.8	47.0	46.2	33.2	41.6	46.3	45.6	28.8	39.8	45.0	39.8
	SCBCH (Peng et al., 2026)	AAAI'26	3.6	10.7	15.8	27.8	0.7	7.7	14.2	24.8	0.7	6.0	17.6	24.6	0.7	9.8	14.8	26.7
	RFCMH (Sun et al., 2026)	ICML'26	40.8	54.2	53.0	55.4	41.0	**50.5**	53.4	54.8	39.2	49.5	53.5	54.6	31.1	49.0	50.8	54.0
	DUET (Ours)	-	**41.6**	**54.6**	**54.4**	**56.4**	**41.5**	49.6	**54.8**	**56.7**	**41.0**	**50.3**	**55.1**	**57.3**	**41.9**	**51.5**	**54.9**	**55.7**
INRIA-Websearch	CMMQ (Yang et al., 2022)	CVPR'22	31.2	36.0	38.4	39.3	27.6	32.2	34.4	35.0	16.6	20.9	27.6	30.4	4.2	3.6	11.2	9.2
	MIAN (Zhang et al., 2022b)	TKDE'22	2.0	2.0	1.7	2.0	2.0	1.4	2.0	2.0	2.0	1.8	1.4	1.4	2.0	1.8	1.5	1.3
	HCCH (Sun et al., 2023)	TMM'23	12.2	22.0	31.8	40.0	9.1	15.0	25.6	36.4	5.6	9.6	16.2	25.8	3.2	3.4	6.3	11.6
	DSCMH (Sun et al., 2024)	AAAI'24	19.0	27.3	34.0	39.4	18.2	24.8	32.2	37.6	13.7	19.0	26.0	32.4	5.6	9.2	16.2	20.3
	DHaPH (Huo et al., 2024)	TKDE'24	23.2	32.8	40.1	45.0	21.6	29.3	37.3	42.6	19.2	28.2	35.0	40.3	19.0	25.8	33.7	39.0
	DECH (Li et al., 2025c)	AAAI'25	27.8	32.8	34.6	36.1	17.8	18.2	23.3	24.1	18.4	19.2	20.4	19.2	15.2	14.7	15.6	15.4
	RSHNL (Pu et al., 2025)	AAAI'25	38.8	48.0	52.0	52.8	37.0	45.8	50.3	52.0	30.7	38.0	47.8	49.4	27.7	38.1	41.9	42.9
	SCBCH (Peng et al., 2026)	AAAI'26	6.6	20.8	29.2	41.7	5.2	11.4	31.5	37.7	5.6	15.1	35.0	36.2	5.4	22.8	30.9	36.8
	RFCMH (Sun et al., 2026)	ICML'26	**45.6**	51.4	54.5	55.4	45.5	51.0	53.6	54.9	42.9	49.6	53.0	55.4	40.2	47.4	50.7	52.0
	DUET (Ours)	-	44.5	**52.2**	**56.8**	**56.7**	**46.8**	**51.4**	**55.3**	**56.6**	**43.6**	**50.9**	**53.2**	**55.8**	**41.6**	**47.9**	**52.2**	**53.8**



(a) XMedia. (b) XMediaNet. (c) INRIA-Websearch.

Figure 3: Precision-recall curves of different methods with 128-bit hash codes under a 0.6 noise rate.

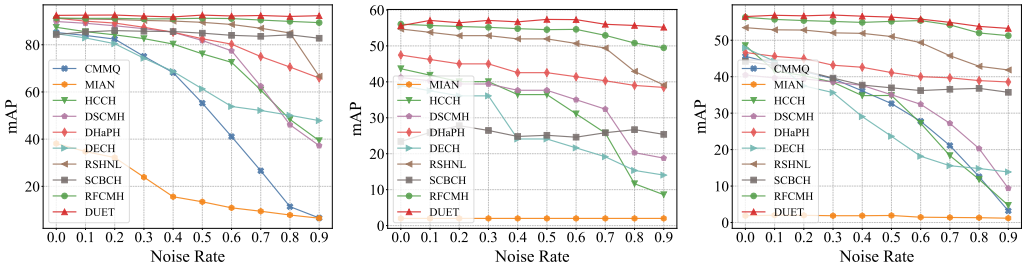
Table 2: Ablation study of different DUET variants under different noise rates with 64-bit hash codes.

Methods	XMedia				XMediaNet				INRIA-Websearch			
	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
DUET-1	88.3	88.0	88.0	87.8	51.5	51.8	51.8	51.7	47.3	46.8	45.6	46.2
DUET-2	79.9	79.8	79.9	79.8	19.5	19.6	19.5	19.6	32.3	32.2	32.3	32.2
DUET-3	89.7	89.4	89.2	88.9	52.6	52.8	52.9	52.6	50.8	49.6	48.7	47.9
DUET-4	90.5	90.3	90.1	89.7	53.3	53.5	53.8	53.4	52.3	51.0	49.7	49.0
DUET-5	**92.2**	**92.0**	**92.0**	**91.6**	**54.4**	**54.8**	**55.1**	**54.9**	**56.8**	**55.3**	**53.2**	**52.2**

DUET-1 uses only TCL and provides a strong baseline with relatively stable performance across noise levels. DUET-2 uses only MCL and gives substantially lower but identical results across noise rates, since MCL does not involve the noisy labels and therefore its objective is unaffected by the noise rate. DUET-3 replaces modality-dependent weighting with uniform weighting while retaining TCL, showing that cross-modal consensus improves the baseline, while adaptive weighting provides further gains. DUET-4 uses the temporal stability weights α_i^m for consensus and consistently outperforms DUET-3, confirming that temporal prediction behavior provides useful modality-dependent information. Finally, DUET-5 further employs the current concentration weights $\beta_i^m[t]$ and the adaptive constraint strength $(1-\beta_i^m[t])$, achieving the best results across all three datasets and noise

levels. These results verify the complementary roles of temporal stability and current prediction concentration in regulating cross-modal semantic consensus.

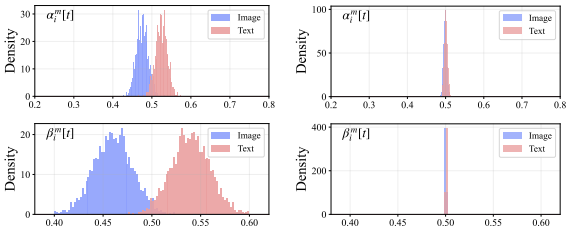
Robustness Analysis: To evaluate robustness under noisy supervision, we compare retrieval performance across noise rates from 0.0 to 0.9 in Figure 4. Most baseline methods degrade substantially as the noise rate increases, demonstrating their vulnerability to unreliable semantic supervision. This degradation stems from their inability to distinguish reliable modality cues from noise-induced signals, allowing unreliable semantics to propagate across modalities and impair cross-modal alignment. In contrast, DUET maintains relatively stable performance even under severe noise by modeling modality-aware reliability and regulating cross-modal interactions, thereby preserving reliable semantics while suppressing noisy signals.



(a) XMedia. (b) XMediaNet. (c) INRIA-Websearch.

Figure 4: Performance comparison of mAP versus noise rate using 128-bit hash codes.

Modality-wise Prediction Analysis: To examine how DUET responds to heterogeneous modality-wise prediction behavior under noisy labels, we analyze the distributions of the temporal-stability weights $\alpha_i^m[t]$ and entropy-based consensus weights $\beta_i^m[t]$ at different training stages in Figure 5. At the early stage (Figure 5 (a)), the text modality exhibits higher $\alpha_i^m[t]$ than the image modality, indicating more stable temporal predictions. This observation is consistent with the modality-dependent noise sensitivity shown in Figure 1, where the image modality is more susceptible to noisy-label fitting and consequently exhibits larger prediction fluctuations. The text modality also shows higher $\beta_i^m[t]$ at this stage, indicating more concentrated current predictions. Accordingly, the text modality contributes more strongly to the consensus prediction, while the image modality receives a stronger consensus constraint through the $1 - \beta_i^m[t]$ term in MCL. As training proceeds (Figure 5 (b)), the gap in $\alpha_i^m[t]$ becomes smaller, while the $\beta_i^m[t]$ values gradually become more balanced across modalities, indicating increasingly similar temporal stability and current prediction concentration.



(a) Distribution of epoch 1. (b) Distribution of epoch 150.

Figure 5: Distributions of temporal-stability weights $\alpha_i^m[t]$ and cross-modal consensus weights $\beta_i^m[t]$ at different stages.

4 Conclusion

This paper investigates the heterogeneous effects of noisy labels across modalities in cross-modal hashing, where uniform semantic alignment may propagate unreliable information and degrade cross-modal retrieval. We propose DUET, a dual-source semantic consensus learning framework that combines temporal prediction stability with current prediction concentration to regulate cross-modal semantic interaction. By controlling the contribution and constraint strength of different modalities, DUET learns more robust cross-modal hash representations under varying levels of label noise. Future work will focus on more reliable semantic correctness estimation and self-adaptive consensus mechanisms for complex and dynamically evolving noisy environments.

REFERENCES

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

Peng Hu, Xi Peng, Hongyuan Zhu, Liangli Zhen, and Jie Lin. Learning cross-modal retrieval with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5403–5413, 2021a.

Yupeng Hu, Meng Liu, Xiaobin Su, Zan Gao, and Liqiang Nie. Video moment localization via deep cross-modal hashing. *IEEE Transactions on Image Processing*, 30:4667–4677, 2021b.

Zhikai Hu, Yiu-Ming Cheung, Mengke Li, and Weichao Lan. Cross-modal hashing method with properties of hamming space: A new perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):7636–7650, 2024.

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*, 36(11):5926–5939, 2024.

Longlong Jing, Elahe Vahdani, Jiaxing Tan, and Yingli Tian. Cross-modal center loss for 3d cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3142–3151, 2021.

Josip Krapac, Moray Allan, Jakob Verbeek, and Frédéric Juried. Improving web image search results using query-relative classifiers. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1094–1101, 2010.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Yongxiang Li, Yang Qin, Yuan Sun, Dezhong Peng, Xi Peng, and Peng Hu. Romo: Robust unsupervised multimodal learning with noisy pseudo labels. *IEEE Transactions on Image Processing*, 33:5086–5097, 2024.

Yongxiang Li, Yanglin Feng, Yuan Sun, Dezhong Peng, Xi Peng, and Peng Hu. Learning source-free domain adaptation for visible-infrared person re-identification. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Yongxiang Li, Yuan Sun, Yang Qin, Dezhong Peng, Xi Peng, and Peng Hu. Robust duality learning for unsupervised visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 20:1937–1948, 2025b.

Yuan Li, Liangli Zhen, Yuan Sun, Dezhong Peng, Xi Peng, and Peng Hu. Deep evidential hashing for trustworthy cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18566–18574, 2025c.

Xiao Liang, Erkun Yang, Yanhua Yang, and Cheng Deng. Multi-relational deep hashing for cross-modal search. *IEEE Transactions on Image Processing*, 33:3009–3020, 2024.

Xiaoqing Liu, Huanqiang Zeng, Yifan Shi, Jianqing Zhu, Chih-Hsien Hsia, and Kai-Kuang Ma. Deep cross-modal hashing based on semantic consistent ranking. *IEEE Transactions on Multimedia*, 25:9530–9542, 2023a.

Xiaoqing Liu, Zhiwen Yu, Kaixiang Yang, Jun Yu, Huanqiang Zeng, and CL Philip Chen. Enhanced cross-modal hashing via hybrid distillation and structural refinement. *IEEE Transactions on Image Processing*, 34:7138–7151, 2025.

Yang Liu, Hao Cheng, and Kun Zhang. Identifiability of label noise transition matrix. In *International Conference on Machine Learning*, pp. 21475–21496. PMLR, 2023b.

Deep Patel and PS Sastry. Adaptive sample selection for robust learning under label noise. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3932–3942, 2023.

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.

Yuxin Peng, Xiaohua Zhai, Yunzhen Zhao, and Xin Huang. Semi-supervised cross-media feature learning with unified patch graph regularization. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(3):583–596, 2015.

Yuxin Peng, Xin Huang, and Yunzhen Zhao. An overview of cross-media retrieval: Concepts, methodologies, benchmarks and challenges. *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 2372–2385, Sep 2018.

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.

Jianyang Qin, Lunke Fei, Zheng Zhang, Jie Wen, Yong Xu, and David Zhang. Joint specifics and consistency hash learning for large-scale cross-modal retrieval. *IEEE Transactions on Image Processing*, 31:5343–5358, 2022.

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.

Zhenqiu Shu, Yibing Bai, Kailing Yong, and Zhengtao Yu. Deep cross-modal hashing with ranking learning for noisy labels. *IEEE Transactions on Big Data*, 11(2):553–565, 2024.

Chao Su, Likang Peng, Yuan Sun, Dezhong Peng, Xi Peng, and Xu Wang. Neighbor-aware contrastive disambiguation for cross-modal hashing with redundant annotations. *Advances in Neural Information Processing Systems*, 38:108581–108603, 2026.

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*, 2026.

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

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.

Wentao Tan, Lei Zhu, Jingjing Li, Huaxiang Zhang, and Junwei Han. Teacher-student learning: Efficient hierarchical message aggregation hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*, 25:4520–4532, 2022.

Junfeng Tu, Xueliang Liu, Zhen Huang, Yanbin Hao, Richang Hong, and Meng Wang. Cross-modal hashing via diverse instances matching. *IEEE Transactions on Image Processing*, 34:2737–2749, 2025.

Rong-Cheng Tu, Jie Jiang, Qinghong Lin, Chengfei Cai, Shangxuan Tian, Hongfa Wang, and Wei Liu. Unsupervised cross-modal hashing with modality-interaction. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):5296–5308, 2023.

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 ACM International Conference on Multimedia*, pp. 5752–5760, 2024.

Rui Wang, Haixiao Huang, Yanglin Feng, Dezhong Peng, Peng Hu, and Yongxiang Li. Rise: Semantics-guided discriminative and robust learning for unsupervised cross-modal hashing. *Knowledge-Based Systems*, pp. 115242, 2025.

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

Guoxian Yu, Xuanwu Liu, Jun Wang, Carlotta Domeniconi, and Xiangliang Zhang. Flexible cross-modal hashing. *IEEE Transactions on Neural Networks and Learning Systems*, 33(1):304–314, 2020.

Chao Zhang, Huaxiong Li, Yang Gao, and Chunlin Chen. Weakly-supervised enhanced semantic-aware hashing for cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 35(6):6475–6488, 2022a.

Zheng Zhang, Haoyang Luo, Lei Zhu, Guangming Lu, and Heng Tao Shen. Modality-invariant asymmetric networks for cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):5091–5104, 2022b.

Xiong Zhou, Xianming Liu, Junjun Jiang, Xin Gao, and Xiangyang Ji. Asymmetric loss functions for learning with noisy labels. In *International Conference on Machine Learning*, pp. 12846–12856. PMLR, 2021.

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

APPENDIX

A The Use of Large Language Models (LLMs) **14**

B Related Work **14**

B.1 Cross-modal Hashing 14

B.2 Learning with Noisy Labels 14

C Proofs of Theorems **15**

C.1 Proof of Theorem 1 15

C.2 Proof of Theorem 2 16

D Optimization Pipeline **17**

E More Implementation Details **18**

E.1 Datasets 18

E.2 Feature Extraction 18

F Additional Experimental Results **18**

F.1 Additional Image-to-text and Text-to-image Retrieval Results 18

F.2 Results under Clean Labels 18

F.3 Statistical Performance Analysis 20

F.4 More Ablation Analysis 20

F.5 Parameter Analysis 21

F.6 Convergence Analysis 21

F.7 Additional Experiments with Various Backbones 22

F.8 Additional Experiments on Tri-modal Datasets 23

F.9 Visualization Analysis 23

F.10 Limitation Statement 24

A THE USE OF LARGE LANGUAGE MODELS (LLMS)

Yes, we utilize Large Language Models (LLMs) to assist in manuscript composition and language polishing. In particular, LLMs are adopted to refine wording, fix grammatical issues, and enhance readability across multiple sections including the Introduction, Method, and Experiments. It is worth clarifying that the LLMs were only used for linguistic optimization rather than generating our core ideas, algorithm designs, experimental analyses, or conclusions. All contents produced or revised with the assistance of LLMs have been fully checked, edited, and validated by the authors to guarantee correctness and consistency.

B RELATED WORK

B.1 CROSS-MODAL HASHING

Over the past decade, cross-modal hashing (CMH) has emerged as an effective paradigm for large-scale multimodal retrieval. Existing CMH methods mainly follow unsupervised or supervised paradigms. Unsupervised methods exploit structural or distributional regularities to learn a shared Hamming space without semantic annotations (Tu et al., 2023; Zhu et al., 2022; Wang et al., 2025; 2024; Hu et al., 2024; Liu et al., 2023a), but often struggle to capture fine-grained semantics. Supervised methods therefore leverage semantic labels to achieve more discriminative hash representations (Tan et al., 2022; Sun et al., 2024; Li et al., 2025c; Tu et al., 2023). However, their effectiveness relies heavily on reliable annotations, which are frequently corrupted by subjective bias, semantic ambiguity, and limited labeling resources. Noisy labels can distort the shared semantic structure, induce spurious cross-modal associations, and encourage models to memorize erroneous semantics (Zhang et al., 2022a; Yang et al., 2022; Yu et al., 2020). Although several noise-robust CMH methods have been proposed (Yang et al., 2022; Shu et al., 2024), most treat label noise as a static and modality-agnostic perturbation, overlooking the fact that heterogeneous modalities differ in semantic expressiveness, noise sensitivity, and learning dynamics. As a result, noisy supervision may cause different modalities to form conflicting semantic judgments, while uniform cross-modal alignment can further propagate unreliable evidence across modalities. Therefore, beyond simply suppressing noisy labels, a key challenge is to identify consistent semantic evidence, resolve cross-modal conflicts, and progressively establish a reliable semantic consensus.

B.2 LEARNING WITH NOISY LABELS

In real-world applications, training labels are inevitably corrupted by diverse noise sources, making direct reliance on noisy supervision prone to overfitting (Zhou et al., 2021). Existing noise-robust learning methods have explored sample selection (Patel & Sastry, 2023; Li et al., 2025b), noise transition modeling (Liu et al., 2023b), and robust loss design (Hu et al., 2021a). Despite their effectiveness, these approaches primarily focus on estimating label reliability or reducing the influence of noisy labels, while largely treating supervision as a fixed and externally provided signal. Such a static view becomes particularly limiting in cross-modal learning, where heterogeneous modalities may capture different aspects of the same instance and thus provide complementary but potentially inconsistent semantic evidence. Under noisy supervision, these discrepancies can be further amplified, causing different modalities to form conflicting semantic judgments and making direct information aggregation susceptible to unreliable semantic propagation. Meanwhile, as training progresses, modality-specific predictions continuously evolve and their agreement or disagreement provides valuable internal evidence about the reliability of the underlying semantics. Consistent predictions across modalities can serve as a stronger indication of semantic reliability, whereas persistent disagreements may reveal ambiguous or corrupted supervision. However, existing noise-robust methods rarely exploit such evolving cross-modal interactions to explicitly identify and resolve semantic conflicts. Consequently, how to transform cross-modal agreement into reliable semantic consensus while suppressing conflicting evidence remains largely underexplored, highlighting the need for a conflict-aware learning paradigm.

C PROOFS OF THEOREMS

C.1 PROOF OF THEOREM 1

Theorem 1. *Assume that the semantic prediction of modality m at epoch t can be modeled as $\hat{\boldsymbol{p}}_i^m[t] = \boldsymbol{y}_i^* + \boldsymbol{\epsilon}_i^m[t]$, where $\boldsymbol{y}_i^*$ denotes the underlying semantic representation and $\boldsymbol{\epsilon}_i^m[t]$ denotes a temporally varying prediction perturbation. Assume that the perturbations are zero-mean, $\mathbb{E}[\boldsymbol{\epsilon}_i^m[t]] = 0$, temporally uncorrelated, i.e., $\mathbb{E}[(\boldsymbol{\epsilon}_i^m[t])^\top \boldsymbol{\epsilon}_i^m[s]] = 0$ for $t \neq s$, and have bounded second moments $\mathbb{E}[\|\boldsymbol{\epsilon}_i^m[t]\|_2^2] \leq \sigma^2$. Assume further that the initial temporal state is fixed. For the temporal consensus state updated by TCL as $\tilde{\boldsymbol{p}}_i^m[t] = \eta\tilde{\boldsymbol{p}}_i^m[t-1] + (1-\eta)\hat{\boldsymbol{p}}_i^m[t]$, where $0 < \eta < 1$, the semantic deviation satisfies*

$$\mathbb{E}\left[\|\tilde{\boldsymbol{p}}_i^m[t] - \boldsymbol{y}_i^*\|_2^2\right] \leq \eta^{2t}\|\tilde{\boldsymbol{p}}_i^m[0] - \boldsymbol{y}_i^*\|_2^2 + \frac{(1-\eta)^2(1-\eta^{2t})}{1-\eta^2}\sigma^2. \tag{A-1}$$

Proof. Define the semantic deviation of the temporal consensus as

$$\boldsymbol{e}_i^m[t] = \tilde{\boldsymbol{p}}_i^m[t] - \boldsymbol{y}_i^*. \tag{A-2}$$

According to the TCL update rule, we have

$$\begin{aligned}\boldsymbol{e}_i^m[t] &= \eta\tilde{\boldsymbol{p}}_i^m[t-1] + (1-\eta)\hat{\boldsymbol{p}}_i^m[t] - \boldsymbol{y}_i^* \\ &= \eta(\tilde{\boldsymbol{p}}_i^m[t-1] - \boldsymbol{y}_i^*) + (1-\eta)(\hat{\boldsymbol{p}}_i^m[t] - \boldsymbol{y}_i^*) \\ &= \eta\boldsymbol{e}_i^m[t-1] + (1-\eta)\boldsymbol{\epsilon}_i^m[t].\end{aligned} \tag{A-3}$$

By repeatedly substituting the recursive formulation, we obtain

$$\begin{aligned}\boldsymbol{e}_i^m[t] &= \eta\left(\eta\boldsymbol{e}_i^m[t-2] + (1-\eta)\boldsymbol{\epsilon}_i^m[t-1]\right) + (1-\eta)\boldsymbol{\epsilon}_i^m[t] \\ &= \eta^2\boldsymbol{e}_i^m[t-2] + (1-\eta)\left(\eta\boldsymbol{\epsilon}_i^m[t-1] + \boldsymbol{\epsilon}_i^m[t]\right) \\ &\cdots \\ &= \eta^t\boldsymbol{e}_i^m[0] + (1-\eta)\sum_{\tau=1}^{t}\eta^{t-\tau}\boldsymbol{\epsilon}_i^m[\tau].\end{aligned} \tag{A-4}$$

Therefore, the expected squared semantic deviation can be expressed as

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{e}_i^m[t]\|_2^2\right] =& \mathbb{E}\left[\left\|\eta^t\boldsymbol{e}_i^m[0] + (1-\eta)\sum_{\tau=1}^{t}\eta^{t-\tau}\boldsymbol{\epsilon}_i^m[\tau]\right\|_2^2\right] \\ =& \eta^{2t}\|\boldsymbol{e}_i^m[0]\|_2^2 + (1-\eta)^2\mathbb{E}\left[\left\|\sum_{\tau=1}^{t}\eta^{t-\tau}\boldsymbol{\epsilon}_i^m[\tau]\right\|_2^2\right] + \\ & 2\eta^t(1-\eta)\mathbb{E}\left[(\boldsymbol{e}_i^m[0])^\top\sum_{\tau=1}^{t}\eta^{t-\tau}\boldsymbol{\epsilon}_i^m[\tau]\right].\end{aligned} \tag{A-5}$$

Since the perturbations are unbiased, i.e., $\mathbb{E}[\boldsymbol{\epsilon}_i^m[t]] = 0$, the third term becomes

$$\mathbb{E}\left[(\boldsymbol{e}_i^m[0])^\top\sum_{\tau=1}^{t}\eta^{t-\tau}\boldsymbol{\epsilon}_i^m[\tau]\right] = \sum_{\tau=1}^{t}\eta^{t-\tau}(\boldsymbol{e}_i^m[0])^\top\mathbb{E}[\boldsymbol{\epsilon}_i^m[\tau]] = 0. \tag{A-6}$$

Next, consider the accumulated perturbation term. Expanding its squared norm gives

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{\tau=1}^{t}\eta^{t-\tau}\boldsymbol{\epsilon}_i^m[\tau]\right\|_2^2\right] &= \mathbb{E}\left[\sum_{\tau=1}^{t}\sum_{s=1}^{t}\eta^{t-\tau}\eta^{t-s}(\boldsymbol{\epsilon}_i^m[\tau])^\top\boldsymbol{\epsilon}_i^m[s]\right] \\ &= \sum_{\tau=1}^{t}\sum_{s=1}^{t}\eta^{t-\tau+t-s}\mathbb{E}\left[(\boldsymbol{\epsilon}_i^m[\tau])^\top\boldsymbol{\epsilon}_i^m[s]\right].\end{aligned} \tag{A-7}$$

According to the temporal uncorrelated assumption,

$$\mathbb{E}\left[(\boldsymbol{\epsilon}_i^m[\tau])^\top \boldsymbol{\epsilon}_i^m[s]\right] = 0, \quad \tau \neq s, \tag{A-8}$$

therefore all cross-epoch covariance terms vanish. Only the diagonal terms remain:

$$\mathbb{E}\left[\left\|\sum_{\tau=1}^{t} \eta^{t-\tau} \boldsymbol{\epsilon}_i^m[\tau]\right\|_2^2\right] = \sum_{\tau=1}^{t} \eta^{2(t-\tau)} \mathbb{E}\left[\|\boldsymbol{\epsilon}_i^m[\tau]\|_2^2\right]. \tag{A-9}$$

Using the bounded perturbation assumption $\mathbb{E}\left[\|\boldsymbol{\epsilon}_i^m[\tau]\|_2^2\right] \leq \sigma^2$, we obtain

$$\mathbb{E}\left[\left\|\sum_{\tau=1}^{t} \eta^{t-\tau} \boldsymbol{\epsilon}_i^m[\tau]\right\|_2^2\right] \leq \sigma^2 \sum_{\tau=1}^{t} \eta^{2(t-\tau)}. \tag{A-10}$$

The summation is a finite geometric series:

$$\sum_{\tau=1}^{t} \eta^{2(t-\tau)} = \sum_{k=0}^{t-1} \eta^{2k} = \frac{1-\eta^{2t}}{1-\eta^2}. \tag{A-11}$$

Consequently,

$$\mathbb{E}\left[\|\boldsymbol{e}_i^m[t]\|_2^2\right] \leq \eta^{2t} \|\boldsymbol{e}_i^m[0]\|_2^2 + (1-\eta)^2 \frac{1-\eta^{2t}}{1-\eta^2} \sigma^2. \tag{A-12}$$

Substituting $\boldsymbol{e}_i^m[t] = \tilde{\boldsymbol{p}}_i^m[t] - \boldsymbol{y}_i^*$ and $\boldsymbol{e}_i^m[0] = \tilde{\boldsymbol{p}}_i^m[0] - \boldsymbol{y}_i^*$ into the above inequality yields

$$\mathbb{E}\left[\|\tilde{\boldsymbol{p}}_i^m[t] - \boldsymbol{y}_i^*\|_2^2\right] \leq \eta^{2t} \|\tilde{\boldsymbol{p}}_i^m[0] - \boldsymbol{y}_i^*\|_2^2 + \frac{(1-\eta)^2(1-\eta^{2t})}{1-\eta^2} \sigma^2. \tag{A-13}$$

Hence, the theorem is proved. □

C.2 Proof of Theorem 2

Theorem 2. *Assume that the prediction of modality m can be expressed as $\hat{\boldsymbol{p}}_i^m = \boldsymbol{y}_i^* + \boldsymbol{\iota}_i^m$, where $\boldsymbol{\iota}_i^m$ denotes the residual modality-specific error satisfying $\|\boldsymbol{\iota}_i^m\|_2 \leq \delta_m$. Given the MCL consensus $\hat{\boldsymbol{q}}_i = \sum_{m=1}^{M} \beta_i^m \hat{\boldsymbol{p}}_i^m$, where $\beta_i^m \geq 0$ and $\sum_{m=1}^{M} \beta_i^m = 1$, the consensus error satisfies*

$$\|\hat{\boldsymbol{q}}_i - \boldsymbol{y}_i^*\|_2 \leq \sum_{m=1}^{M} \beta_i^m \delta_m. \tag{A-14}$$

Proof. According to the assumption, the semantic prediction of modality m can be decomposed into the underlying semantic representation and a modality-specific residual error $\hat{\boldsymbol{p}}_i^m[t] = \boldsymbol{y}_i^* + \boldsymbol{\iota}_i^m$. The consensus representation generated by MCL is obtained through the reliability-weighted aggregation of modality-specific predictions $\hat{\boldsymbol{q}}_i[t] = \sum_m \beta_i^m \hat{\boldsymbol{p}}_i^m[t]$, where β_i^m denotes the aggregation weight of modality m and satisfies $\sum_m \beta_i^m = 1$, $\beta_i^m \geq 0$.

Substituting the semantic prediction into the consensus representation gives

$$\begin{aligned} \hat{\boldsymbol{q}}_i[t] &= \sum_m \beta_i^m (\boldsymbol{y}_i^* + \boldsymbol{\iota}_i^m) \\ &= \sum_m \beta_i^m \boldsymbol{y}_i^* + \sum_m \beta_i^m \boldsymbol{\iota}_i^m. \end{aligned} \tag{A-15}$$

Since the aggregation weights form a normalized distribution, i.e., $\sum_m \beta_i^m = 1$, we have

$$\sum_m \beta_i^m \boldsymbol{y}_i^* = \boldsymbol{y}_i^*. \tag{A-16}$$

Therefore, the deviation between the consensus representation and the underlying semantic representation can be written as

$$\hat{\boldsymbol{q}}_i[t] - \boldsymbol{y}_i^* = \sum_m \beta_i^m \boldsymbol{\iota}_i^m. \tag{A-17}$$

Taking the ℓ_2 norm on both sides yields

$$\|\hat{\boldsymbol{q}}_i[t] - \boldsymbol{y}_i^*\|_2 = \left\| \sum_m \beta_i^m \boldsymbol{\iota}_i^m \right\|_2. \tag{A-18}$$

By applying the triangle inequality, we have

$$\left\| \sum_m \beta_i^m \boldsymbol{\iota}_i^m \right\|_2 \leq \sum_m \|\beta_i^m \boldsymbol{\iota}_i^m\|_2 = \sum_m \beta_i^m \|\boldsymbol{\iota}_i^m\|_2 . \tag{A-19}$$

According to the residual error assumption, $\|\boldsymbol{\iota}_i^m\|_2 \leq \delta_m$, we further obtain

$$\|\hat{\boldsymbol{q}}_i[t] - \boldsymbol{y}_i^*\|_2 \leq \sum_m \beta_i^m \delta_m. \tag{A-20}$$

This indicates that the consensus error is bounded by the weighted residual error of all modalities.

Hence, the theorem is proved. □

D Optimization Pipeline

The complete training procedure of DUET is summarized in Algorithm 1. The training process consists of two stages: TCL is first optimized to establish temporally stable intra-modal semantic representations, followed by MCL to perform reliability-aware cross-modal consensus learning.

Algorithm 1 Training procedure of the proposed DUET.

Input: Training dataset $\mathcal{D}$, batch size B, warm-up epoch E_{warm}, total epoch E_{all}, learning rate lr, and hyper-parameters $L, \tau, \sigma, \eta, \lambda$.

Randomly initialize the hashing networks $\{\mathcal{H}^m(\cdot)\}_{m=1}^M$ and learnable class prototypes $\boldsymbol{\mathcal{C}}$;
for $t = 1$ **to** E_{all} **do**
  repeat
    Randomly sample a mini-batch with B instances from $\mathcal{D}$;
    Generate continuous hash representations $\hat{\boldsymbol{b}}_i^m$ via Equation (3);
    Compute modality-specific semantic predictions $\boldsymbol{p}_i^m[t]$ via Equations (5) to (7);
    Initialize or update temporal consensus states $\tilde{\boldsymbol{p}}_i^m[t]$ via Equations (8) and (9);
    Compute temporal prediction deviations $\Delta_i^m[t]$ and stability scores $r_i^m[t]$ via Equations (11) and (12);
    Compute normalized temporal stability weights $\alpha_i^m[t]$ via Equation (13);
    Compute TCL loss $\mathcal{L}_{\mathrm{TCL}}$ via Equations (10) and (14);
    if $t < E_{warm}$ **then**
      Optimize $\{\mathcal{H}^m(\cdot)\}_{m=1}^M$ and $\boldsymbol{\mathcal{C}}$ by minimizing $\mathcal{L}_{\mathrm{TCL}}$;
    end
    else
      Compute entropy $H_i^m[t]$ and consensus weights $\beta_i^m[t]$ via Equations (15) and (16);
      Construct the cross-modal semantic consensus $\boldsymbol{q}_i[t]$ via Equation (17);
      Compute MCL loss $\mathcal{L}_{\mathrm{MCL}}$ via Equation (18);
      Optimize $\{\mathcal{H}^m(\cdot)\}_{m=1}^M$ and $\boldsymbol{\mathcal{C}}$ by minimizing $\mathcal{L}_{\mathrm{TCL}} + \lambda\mathcal{L}_{\mathrm{MCL}}$;
    end
  until *All training samples are traversed*;
end

Output: Optimized hashing networks $\{\mathcal{H}^m(\cdot)\}_{m=1}^M$.

E More Implementation Details

E.1 Datasets

To assess the effectiveness of the proposed method, extensive experiments are conducted on three representative benchmark datasets, including XMedia (Peng et al., 2015), XMediaNet (Peng et al., 2018) and INRIA-Websearch (Krapac et al., 2010). The details of each dataset are introduced below:

- **XMedia** (Peng et al., 2015) is a heterogeneous multi-modal dataset that contains five modalities annotated with 20 semantic categories. Following common practice, experiments are restricted to the image and text modalities. The dataset is randomly partitioned into 4,000 image–text pairs for training, 500 pairs for validation, and 500 pairs for testing. Owing to its moderate scale, XMedia provides an appropriate testbed for evaluating cross-modal retrieval methods in relatively constrained data regimes.
- **XMediaNet** (Peng et al., 2018) is a large-scale multi-modal benchmark consisting of five modalities annotated with 200 semantic categories. Similar to XMedia, only the image and text modalities are considered in this work. The dataset is randomly divided into 32,000 training pairs, 4,000 validation pairs, and 4,000 test pairs. Its large scale and fine-grained semantic annotations impose significant challenges on learning modality-invariant representations.
- **INRIA-Websearch** (Krapac et al., 2010) is a large-scale dataset comprising over 70,000 image–text pairs associated with 353 semantic labels, exhibiting substantial semantic diversity and intra-class variation. In our experiments, a widely used subset of 14,698 pairs covering 100 categories is adopted. This subset is randomly divided into 9,000 training pairs, 1,332 validation pairs, and 4,366 test pairs.

E.2 Feature Extraction

To ensure a fair comparison with existing methods, we follow the standard feature settings adopted by the respective datasets and prior CMH studies. Specifically, for XMedia and XMediaNet, we directly use the official pre-extracted features provided by the dataset authors. For INRIA-Websearch, image representations are extracted using a pre-trained AlexNet (Krizhevsky et al., 2017), resulting in 4,096-dimensional visual features, while 1,000-dimensional textual features are generated using Latent Dirichlet Allocation (LDA) (Hu et al., 2021a). Following these standard feature settings reduces potential confounding effects introduced by different feature extraction strategies and allows the comparison to focus on the effectiveness of the learning methods themselves. To further examine the adaptability of DUET to different feature extraction strategies, we conduct additional experiments using CLIP and BLIP as general-purpose vision-language feature extractors. Instead of relying on the provided pre-extracted features, these experiments directly extract representations from the raw data and retrain the models accordingly. The results are reported in Appendix Section F.7. DUET remains effective with these general-purpose pre-trained representations, indicating that its performance is not restricted to a specific dataset-dependent feature configuration.

F Additional Experimental Results

F.1 Additional Image-to-text and Text-to-image Retrieval Results

In the main paper, the retrieval performance is summarized using the average mAP of *Image to Text* and *Text to Image* retrieval tasks for concise comparison. This section further reports the detailed results of each retrieval direction under different noise levels and hash code lengths. As shown in Tables 3 and 4, DUET consistently achieves competitive performance across various settings, demonstrating its robustness against noisy supervision and effectiveness in learning reliable cross-modal hash representations.

F.2 Results under Clean Labels

To further evaluate the retrieval performance of RFCMH under the clean setting, we conduct experiments on three benchmark datasets, i.e., XMedia, INRIA-Websearch, and XMediaNet. Table 5

Table 3: Image-to-text retrieval performance (mAP%) with different hash code lengths under various noise ratios on the XMedia, XMediaNet, and INRIA-Websearch datasets. The best and second-best results are highlighted in **bold** and underline, respectively.

	Methods	Venues	Noise Rate: 0.2				Noise Rate: 0.4				Noise Rate: 0.6				Noise Rate: 0.8			
			16	32	64	128	16	32	64	128	16	32	64	128	16	32	64	128
XMedia	CMMQ (Yang et al., 2022)	CVPR'22	86.6	87.9	87.4	86.6	74.8	77.4	74.5	72.0	51.5	43.8	38.9	38.7	18.3	18.5	13.7	12.6
	MIAN (Zhang et al., 2022b)	TKDE'22	18.1	29.6	34.5	35.4	12.4	16.2	17.7	16.2	10.7	9.9	11.5	11.1	7.6	7.7	7.1	8.0
	HCCH (Sun et al., 2023)	TMM'23	71.1	81.9	82.1	84.4	71.6	76.8	78.2	80.6	60.3	61.5	61.5	72.5	26.3	41.2	41.1	48.5
	DSCMH (Sun et al., 2024)	AAAI'24	81.2	85.6	87.9	88.2	79.4	83.6	85.4	85.8	63.1	70.4	74.4	77.5	31.0	35.1	43.7	47.2
	DHaPH (Huo et al., 2024)	TKDE'24	79.3	84.9	86.9	88.6	69.6	78.2	82.6	84.6	52.3	66.0	72.8	79.8	42.7	48.3	60.6	70.2
	DECH (Li et al., 2025c)	AAAI'25	72.9	78.5	80.1	81.0	62.4	63.2	68.0	69.1	52.3	53.7	53.4	52.8	47.9	47.6	48.1	48.5
	RSHNL (Pu et al., 2025)	AAAI'25	89.5	90.0	90.9	90.8	89.7	90.2	91.1	89.8	83.0	87.7	88.9	88.4	81.7	88.1	87.6	84.7
	SCBCH (Peng et al., 2026)	AAAI'26	34.6	74.4	80.5	85.5	33.4	66.9	81.2	84.8	38.6	75.1	81.8	83.1	29.7	67.7	77.9	83.4
	RFCMH (Sun et al., 2026)	ICML'26	89.5	91.4	**92.2**	91.3	89.8	90.7	91.4	90.7	89.3	90.1	90.5	91.2	88.1	89.7	89.9	90.2
	DUET (Ours)	-	**91.1**	**91.7**	92.1	**92.5**	**90.1**	**91.7**	**92.1**	**91.9**	**90.8**	**91.5**	**91.6**	**92.3**	**89.2**	**91.2**	**92.1**	**92.0**
XMediaNet	CMMQ (Yang et al., 2022)	CVPR'22	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
	MIAN (Zhang et al., 2022b)	TKDE'22	0.8	1.5	1.8	2.5	0.8	1.2	1.3	1.8	0.7	0.9	1.0	1.2	0.7	0.8	0.8	0.9
	HCCH (Sun et al., 2023)	TMM'23	1.6	2.0	4.8	15.0	1.4	1.5	3.4	12.1	1.3	1.4	2.2	5.8	0.8	0.9	1.1	1.9
	DSCMH (Sun et al., 2024)	AAAI'24	5.2	9.9	18.1	28.7	4.7	8.5	14.9	24.7	3.4	5.4	10.3	18.5	1.8	2.3	3.9	6.6
	DHaPH (Huo et al., 2024)	TKDE'24	9.8	15.9	23.4	28.5	9.1	15.8	22.6	27.6	9.2	15.1	21.9	27.5	8.8	15.2	21.5	27.3
	DECH (Li et al., 2025c)	AAAI'25	13.8	11.6	12.7	13.9	15.1	16.2	10.5	11.6	4.9	6.1	6.6	6.0	0.4	0.6	0.8	1.1
	RSHNL (Pu et al., 2025)	AAAI'25	86.5	89.4	91.0	90.6	87.4	90.2	90.3	90.3	82.6	87.4	89.1	88.8	79.7	86.8	86.7	85.3
	SCBCH (Peng et al., 2026)	AAAI'26	3.8	11.4	16.3	28.0	0.7	8.3	14.4	25.0	0.7	7.5	17.8	24.6	0.7	10.1	15.1	26.8
	RFCMH (Sun et al., 2026)	ICML'26	40.9	54.4	52.7	55.7	41.2	**50.5**	52.3	54.5	39.5	49.5	53.3	54.2	30.2	48.9	50.4	54.0
	DUET (Ours)	-	**42.0**	**55.4**	**55.1**	**56.5**	**42.1**	50.4	**55.6**	**56.7**	**41.4**	**51.2**	**55.8**	**57.4**	**42.6**	**52.4**	**55.2**	**55.7**
INRIA-Websearch	CMMQ (Yang et al., 2022)	CVPR'22	31.1	35.5	38.3	39.6	27.8	32.1	34.4	35.8	17.0	21.2	27.9	30.4	4.1	3.9	11.6	9.8
	MIAN (Zhang et al., 2022b)	TKDE'22	2.7	2.7	2.2	2.8	2.7	1.7	2.8	2.8	2.8	2.4	1.7	1.7	2.8	2.3	1.8	1.4
	HCCH (Sun et al., 2023)	TMM'23	11.8	20.5	28.5	37.6	8.8	13.1	21.9	33.9	5.4	8.3	12.8	22.2	3.3	3.2	5.5	10.2
	DSCMH (Sun et al., 2024)	AAAI'24	18.8	27.6	35.2	41.2	17.9	25.0	33.4	39.6	13.5	19.4	27.7	34.1	5.6	9.7	16.7	21.6
	DHaPH (Huo et al., 2024)	TKDE'24	24.0	32.8	39.5	44.3	22.3	29.3	37.0	42.0	19.9	28.1	34.7	39.8	19.5	26.1	33.6	38.5
	DECH (Li et al., 2025c)	AAAI'25	30.1	36.2	37.8	38.6	19.9	20.6	25.5	26.7	19.5	21.0	20.8	20.3	15.6	15.2	15.7	16.0
	RSHNL (Pu et al., 2025)	AAAI'25	39.3	48.0	51.9	52.4	37.9	45.8	50.3	51.6	31.2	38.3	47.6	49.0	28.3	38.2	41.8	42.9
	SCBCH (Peng et al., 2026)	AAAI'26	6.7	20.0	28.6	40.5	5.1	11.4	31.4	38.6	5.8	14.7	35.0	35.5	5.0	22.3	29.8	36.1
	RFCMH (Sun et al., 2026)	ICML'26	**46.2**	51.9	54.5	55.2	45.9	**51.2**	53.4	54.7	**43.2**	49.7	**53.3**	**54.9**	41.2	**47.7**	50.6	52.0
	DUET (Ours)	-	45.3	**52.3**	**57.6**	**57.3**	**47.5**	48.5	**55.0**	**57.2**	42.9	**51.5**	**53.3**	**54.9**	**42.9**	47.5	**52.7**	**54.8**

Table 4: Text-to-image retrieval performance (mAP%) with different hash code lengths under various noise ratios on the XMedia, XMediaNet, and INRIA-Websearch datasets. The best and second-best results are highlighted in **bold** and underline, respectively.

	Methods	Venues	Noise Rate: 0.2				Noise Rate: 0.4				Noise Rate: 0.6				Noise Rate: 0.8			
			16	32	64	128	16	32	64	128	16	32	64	128	16	32	64	128
XMedia	CMMQ (Yang et al., 2022)	CVPR'22	85.1	83.4	82.0	78.2	70.8	74.4	67.6	64.4	45.2	36.0	33.2	43.5	14.4	14.0	13.2	10.2
	MIAN (Zhang et al., 2022b)	TKDE'22	15.5	22.1	28.5	29.0	11.0	14.3	16.2	15.0	8.6	9.6	10.9	10.7	7.0	7.3	7.0	7.7
	HCCH (Sun et al., 2023)	TMM'23	69.5	80.7	80.0	84.0	70.0	75.7	76.7	80.0	58.1	60.2	58.5	72.6	26.2	40.8	41.2	48.2
	DSCMH (Sun et al., 2024)	AAAI'24	80.5	85.0	86.3	87.9	78.0	81.6	83.5	84.8	63.5	70.9	73.9	77.4	29.1	33.0	41.0	45.0
	DHaPH (Huo et al., 2024)	TKDE'24	78.9	84.9	88.2	89.8	69.5	77.9	83.9	86.0	51.5	64.9	72.9	80.7	42.3	49.2	60.0	70.9
	DECH (Li et al., 2025c)	AAAI'25	8.1	73.6	79.9	80.7	8.1	55.7	57.8	65.7	9.3	28.5	34.4	37.0	6.1	8.4	11.5	23.6
	RSHNL (Pu et al., 2025)	AAAI'25	86.5	89.4	91.0	90.6	87.4	90.2	90.3	90.3	82.6	87.4	89.1	88.8	79.7	86.8	86.7	85.3
	SCBCH (Peng et al., 2026)	AAAI'26	36.1	75.6	85.1	86.5	33.4	66.9	85.8	86.3	38.4	75.8	84.0	84.8	32.4	71.9	80.7	85.1
	RFCMH (Sun et al., 2026)	ICML'26	87.6	89.3	91.1	91.1	87.8	89.5	90.5	91.4	87.7	88.8	89.8	91.0	86.1	89.1	88.9	89.6
	DUET (Ours)	-	**89.2**	**91.4**	**92.2**	**92.6**	**89.0**	**91.1**	**91.9**	**91.8**	**88.5**	**91.1**	**92.4**	**92.2**	**86.6**	**90.6**	**91.1**	**92.1**
XMediaNet	CMMQ (Yang et al., 2022)	CVPR'22	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
	MIAN (Zhang et al., 2022b)	TKDE'22	0.8	1.2	1.2	1.8	0.7	1.0	1.0	1.3	0.7	0.8	0.8	1.0	0.7	0.7	0.7	0.8
	HCCH (Sun et al., 2023)	TMM'23	1.8	1.5	2.1	9.0	1.4	1.2	1.7	12.1	1.5	1.1	1.3	1.8	0.8	0.9	1.0	1.2
	DSCMH (Sun et al., 2024)	AAAI'24	4.1	8.4	16.5	27.5	3.6	7.0	13.5	23.8	2.5	4.4	9.3	17.5	1.4	2.0	3.4	5.9
	DHaPH (Huo et al., 2024)	TKDE'24	11.0	18.0	26.3	32.6	10.5	17.3	25.3	31.6	10.3	16.8	24.8	31.2	9.6	17.1	24.0	30.7
	DECH (Li et al., 2025c)	AAAI'25	11.6	15.4	16.1	13.3	11.9	10.7	10.6	10.8	2.5	3.9	4.7	4.0	0.3	0.6	0.5	0.9
	RSHNL (Pu et al., 2025)	AAAI'25	35.3	42.5	46.5	47.8	35.0	41.6	46.9	46.6	33.3	41.3	46.4	46.1	29.3	39.9	45.1	40.6
	SCBCH (Peng et al., 2026)	AAAI'26	3.3	9.9	15.4	27.6	0.7	7.0	14.1	24.6	0.7	4.4	17.4	24.7	0.7	9.5	14.5	26.6
	RFCMH (Sun et al., 2026)	ICML'26	40.6	**53.9**	53.2	55.2	40.8	**50.5**	**54.5**	55.2	38.9	**49.5**	53.7	54.9	32.0	49.2	51.2	54.1
	DUET (Ours)	-	**41.2**	53.8	**53.8**	**56.3**	**40.9**	48.8	54.0	**56.7**	**40.6**	49.4	**54.4**	**57.2**	**41.2**	**50.6**	**54.6**	**55.7**
INRIA-Websearch	CMMQ (Yang et al., 2022)	CVPR'22	31.4	36.6	38.5	39.0	27.3	32.3	34.5	34.1	16.2	20.6	27.3	30.4	4.3	3.4	10.7	8.5
	MIAN (Zhang et al., 2022b)	TKDE'22	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2
	HCCH (Sun et al., 2023)	TMM'23	12.7	23.6	35.0	42.5	9.4	17.0	29.4	39.0	5.9	10.8	19.7	29.3	3.0	3.6	7.1	13.1
	DSCMH (Sun et al., 2024)	AAAI'24	19.1	27.0	32.8	37.6	18.5	24.6	31.1	35.7	13.9	18.5	24.4	30.7	5.6	8.7	15.6	19.0
	DHaPH (Huo et al., 2024)	TKDE'24	22.4	32.8	40.7	45.7	21.0	29.3	37.6	43.1	18.6	28.4	35.4	40.8	18.6	25.6	33.8	39.5
	DECH (Li et al., 2025c)	AAAI'25	25.4	29.5	31.3	33.6	15.8	15.9	21.1	21.5	17.3	17.4	20.0	18.0	14.9	14.2	15.4	14.7
	RSHNL (Pu et al., 2025)	AAAI'25	38.2	47.9	52.1	53.3	36.2	45.9	50.3	52.3	30.2	37.7	47.9	49.8	27.1	38.0	42.0	42.9
	SCBCH (Peng et al., 2026)	AAAI'26	6.5	21.5	29.9	42.9	5.4	11.5	31.6	36.8	5.4	15.5	35.1	36.8	5.8	23.2	32.0	37.5
	RFCMH (Sun et al., 2026)	ICML'26	**45.1**	50.8	54.5	55.6	45.1	50.9	53.7	55.2	42.6	49.5	52.7	56.0	39.1	47.0	50.8	51.9
	DUET (Ours)	-	43.7	**52.1**	**56.0**	**56.1**	**46.1**	**54.3**	**55.6**	**56.0**	**44.3**	**50.3**	**53.1**	**56.7**	**40.3**	**48.3**	**51.7**	**52.8**

reports the mAP results under the clean setting on both *Image to Text* and *Text to Image* tasks. The results show that DUET achieves stable and competitive performance across different datasets and hash code lengths.

Table 5: The mAP scores on three datasets with clean labels. The best results are highlighted in **boldface** and the second-best are underlined. The column Sum indicates the total score of all hash lengths.

Task	Methods	Venues	Xmedia					INRIA-Websearch					XmediaNet				
			16	32	64	128	Sum	16	32	64	128	Sum	16	32	64	128	Sum
Image to Text	HMAH (Tan et al., 2022)	TMM'22	86.0	89.2	90.0	90.3	355.5	33.6	43.1	47.6	46.2	170.5	7.8	13.0	18.5	25.7	65.0
	CMMQ (Yang et al., 2022)	CVPR'22	81.6	86.0	90.1	90.3	348.0	32.1	32.4	46.0	48.7	159.2	3.2	3.8	7.9	17.6	32.5
	WASH (Zhang et al., 2022a)	TKDE'23	88.2	87.4	89.8	90.0	355.4	33.4	40.1	45.6	48.0	167.1	8.6	16.3	24.9	35.5	85.3
	HCCH (Sun et al., 2023)	TMM'24	79.1	78.2	52.0	77.5	286.8	14.2	21.7	31.6	30.7	98.2	2.2	3.3	6.8	18.6	30.9
	NRCH (Wang et al., 2024)	MM'24	86.5	89.5	90.1	91.2	357.3	30.6	37.0	40.7	46.1	154.4	7.4	12.9	16.7	18.7	55.7
	DHRL (Shu et al., 2024)	TBD'24	65.4	75.2	83.2	84.5	308.3	2.7	3.5	37.8	43.1	87.1	1.1	20.1	26.0	34.1	81.3
	DHaPH (Huo et al., 2024)	TKDE'24	88.1	90.4	91.8	**93.2**	363.5	24.8	34.3	41.2	45.8	146.1	6.1	10.7	16.3	22.5	55.6
	RSHNL (Pu et al., 2025)	AAAI'25	90.5	91.2	92.3	91.9	365.9	36.2	45.2	50.0	51.9	183.3	32.4	41.4	47.6	48.6	170.0
	DECH (Li et al., 2025c)	AAAI'25	85.5	89.7	90.9	91.8	357.9	30.9	44.5	50.1	52.0	177.5	6.9	25.1	42.3	49.9	124.2
	NACD (Su et al., 2026)	NeurIPS'25	85.2	89.4	90.9	90.7	356.2	29.0	35.1	39.1	41.0	144.2	2.1	7.5	8.9	12.3	30.8
	SCBCH (Peng et al., 2026)	AAAI'26	41.3	67.9	82.9	82.3	274.4	5.8	18.5	31.9	42.6	98.8	1.1	5.8	13.6	23.5	44.0
	RFCMH (Sun et al., 2026)	ICML'26	90.2	91.3	92.1	91.4	365.0	**48.4**	**52.0**	**54.0**	**55.8**	**210.2**	**43.7**	**55.5**	55.9	56.0	**211.1**
	DUET(Ours)	-	**91.4**	**92.0**	**92.5**	92.3	**368.2**	46.8	50.3	52.4	54.1	203.6	41.8	53.6	**56.1**	**56.4**	207.9
Text to Image	HMAH (Tan et al., 2022)	TMM'22	85.3	88.9	89.6	90.5	354.3	33.3	44.3	49.6	48.2	175.4	9.1	14.5	20.0	27.2	70.8
	CMMQ (Yang et al., 2022)	CVPR'22	74.1	80.4	90.6	90.6	335.7	30.9	30.4	48.6	51.6	161.5	4.7	6.2	10.4	20.2	41.5
	WASH (Zhang et al., 2022a)	TKDE'23	87.4	88.4	89.0	89.2	354.0	32.5	41.2	47.6	50.5	171.8	10.7	18.6	26.8	36.8	92.9
	HCCH (Sun et al., 2023)	TMM'24	79.7	77.4	56.8	74.5	288.4	13.2	24.9	37.5	26.7	102.3	2.5	2.4	3.8	15.9	24.6
	NRCH (Wang et al., 2024)	MM'24	87.7	89.9	90.0	90.2	357.8	31.9	35.7	37.0	40.0	144.6	6.3	10.6	15.1	16.9	48.9
	DHRL (Shu et al., 2024)	TBD'24	65.0	78.7	83.6	85.2	312.5	2.9	2.6	38.3	44.2	88.0	1.0	18.7	25.9	33.6	79.2
	DHaPH (Huo et al., 2024)	TKDE'24	89.5	91.6	91.8	**92.4**	365.3	23.6	34.5	42.7	47.2	148.0	6.6	11.6	17.6	24.0	59.8
	RSHNL (Pu et al., 2025)	AAAI'25	88.9	90.5	91.6	91.9	362.9	35.6	45.7	51.1	53.5	185.9	33.3	41.6	47.7	49.3	171.9
	DECH (Li et al., 2025c)	AAAI'25	82.0	88.8	90.5	90.8	352.1	30.0	44.8	52.0	53.3	180.1	9.6	26.3	42.8	51.1	129.8
	NACD (Su et al., 2026)	NeurIPS'25	84.1	89.1	91.1	91.3	355.6	27.3	35.8	40.0	42.0	145.1	2.1	8.3	10.2	13.4	34.0
	SCBCH (Peng et al., 2026)	AAAI'26	33.3	71.4	86.5	86.4	277.6	7.0	18.9	34.0	45.8	105.7	1.2	6.2	12.5	23.3	43.2
	RFCMH (Sun et al., 2026)	ICML'26	87.9	89.4	90.8	91.2	359.3	**45.9**	**51.3**	**54.2**	**56.8**	**208.2**	**43.3**	**54.6**	55.5	56.1	**209.5**
	DUET(Ours)	-	**90.8**	**91.7**	**92.2**	92.1	**366.8**	45.2	51.0	53.7	55.6	205.5	41.2	53.1	**55.8**	**56.6**	206.7

F.3 Statistical Performance Analysis

To verify the statistical reliability of DUET, we repeat each experiment using three independent random seeds and report the mean and standard deviation of the retrieval performance. This evaluation aims to investigate whether the performance improvements observed in the main manuscript are consistently reproducible rather than resulting from favorable random initialization. Table 6 summarizes the statistical results on XMedia, INRIA-Websearch, and XMediaNet under different hash code lengths and label noise rates. Across all datasets, DUET consistently achieves competitive mean retrieval performance while maintaining very small standard deviations. Although increasing label noise inevitably introduces slightly larger performance variations, the fluctuations remain limited across all experimental settings. Overall, the statistical analysis confirms that the performance gains reported in the main manuscript are robust and highly reproducible.

Table 6: Statistical retrieval performance of DUET (mean $\pm$ standard deviation over three independent runs).

Dataset	Task	Noise 0.2				Noise 0.4				Noise 0.6				Noise 0.8			
		16	32	64	128	16	32	64	128	16	32	64	128	16	32	64	128
XMedia	mean	90.2	91.6	92.2	92.6	89.6	91.4	92.0	91.8	89.7	91.3	92.0	92.2	87.9	90.9	91.6	92.0
	std	0.31	0.27	0.24	0.22	0.35	0.30	0.26	0.28	0.38	0.32	0.27	0.25	0.46	0.39	0.34	0.31
XMediaNet	mean	41.6	54.6	54.4	56.4	41.5	49.6	54.8	56.7	41.0	50.3	55.1	57.3	41.9	51.5	54.9	55.7
	std	0.42	0.38	0.35	0.32	0.45	0.40	0.36	0.33	0.48	0.43	0.39	0.36	0.52	0.47	0.44	0.41
INRIA-Websearch	mean	44.5	52.2	56.8	56.7	46.8	51.4	55.3	56.6	43.6	50.9	53.2	55.8	41.6	47.9	52.2	53.8
	std	0.36	0.31	0.28	0.30	0.41	0.35	0.32	0.29	0.45	0.39	0.34	0.31	0.51	0.46	0.40	0.37

F.4 More Ablation Analysis

To examine the potential coupling caused by shared class prototypes during warm-up, we compare shared and modality-specific prototypes under otherwise identical settings. As shown in Table 7, the

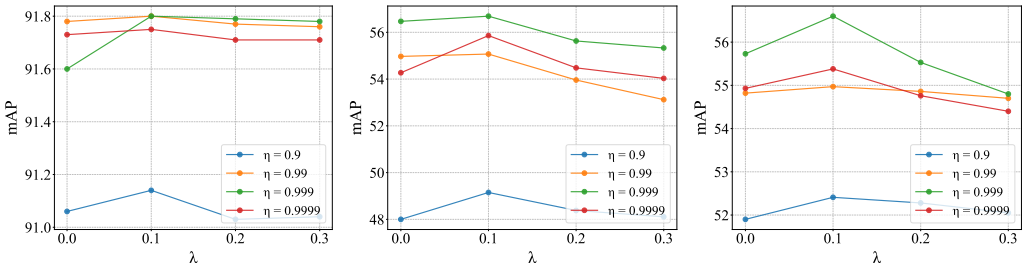
two configurations achieve comparable performance at the end of warm-up, with only minor differences under both noise rates. This suggests that prototype sharing does not substantially affect the initial modality-specific learning. Nevertheless, shared prototypes introduce parameter-level coupling by providing a common semantic space across modalities, whereas MCL introduces explicit prediction-level coupling through cross-modal consensus. The comparable warm-up performance therefore indicates that the former does not dominate the reliability established before explicit consensus learning.

Table 7: Effect of shared and modality-specific class prototypes on DUET performance at different training stages and noise rates on XMedia with 64-bit hash codes.

Stage	Noise Rate = 0.2		Noise Rate = 0.4		Noise Rate = 0.6		Noise Rate = 0.8	
	Shared	Specific	Shared	Specific	Shared	Specific	Shared	Specific
End of Warm-up	88.2	88.0	87.6	87.4	87.3	87.0	86.9	86.6
Final	92.2	91.9	92.0	91.7	92.0	91.7	91.6	91.3

F.5 Parameter Analysis

We investigate the sensitivity of DUET to two key hyper-parameters, η and λ, on three benchmark datasets, as shown in Figure 6. The parameter η controls the temporal authority of historical predictions in TCL. Increasing η from 0.9 to 0.999 generally improves performance, indicating that stronger reliance on stable historical evidence helps suppress transient noise from corrupted labels. Further increasing η to 0.9999 leads to marginal gains or slight degradation, suggesting that excessive historical reliance may hinder adaptation to newly emerging reliable signals. The parameter λ controls the contribution of MCL. When $\lambda = 0$, cross-modal consensus learning is removed, resulting in inferior performance. Increasing λ initially improves performance, confirming the benefit of reliability-aware cross-modal consensus. However, overly large λ degrades performance by imposing excessive agreement and suppressing informative modality-specific semantics. Overall, moderate values of η and λ provide a better balance between intra-modal stability and cross-modal consensus.



(a) XMedia. (b) XMediaNet. (c) INRIA-Websearch.

Figure 6: Sensitivity analysis of parameters η and λ with 128-bit hash codes under 0.4 noise rate.

F.6 Convergence Analysis

To further examine the optimization behavior of DUET, we track the training objective under different hash code lengths. As shown in Figure 7, the training loss decreases rapidly during the initial epochs and gradually converges to a stable level across all datasets and code lengths. Although longer hash codes may require slightly more iterations to reach convergence, all configurations exhibit smooth optimization without evident oscillations or abrupt fluctuations. This stable behavior indicates that the temporally stabilized semantic states provided by TCL can offer reliable intra-modal supervision, while MCL further incorporates cross-modal semantic consensus without destabilizing the optimization process. The consistent convergence patterns across different datasets and code lengths demonstrate that DUET maintains a stable optimization process under noisy-label supervision.

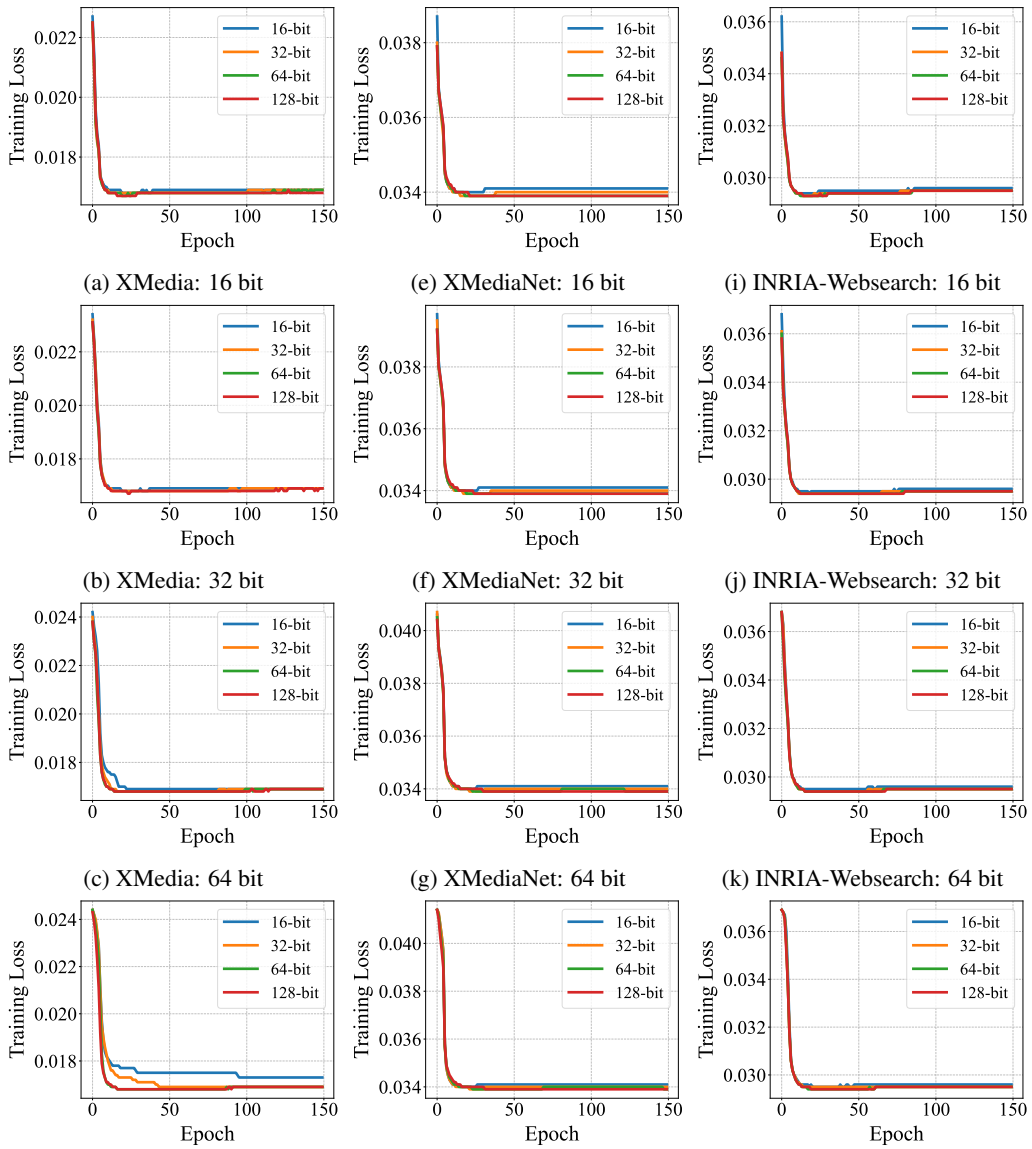


Figure 7: Convergence curves of the total training loss with different hash code lengths on three benchmark datasets.

F.7 Additional Experiments with Various Backbones

To further investigate the dependence of DUET on the underlying feature extractor, we conduct additional experiments with different backbone configurations on XMedia and XMediaNet. The original dataset features are first adopted as the default setting, while CLIP (Radford et al., 2021) is further considered with both frozen and fine-tuned encoders. The results are summarized in Table 8. Several observations can be made from the results. First, DUET remains effective when the feature extractor is replaced, although the absolute retrieval performance varies across different backbone configurations. This variation is expected because different feature extractors provide representations with different levels of semantic separability and cross-modal alignment. Second, the effect of CLIP is not strictly consistent across noise levels or hash code lengths: its stronger pre-trained representations can provide gains in some settings, whereas the improvement becomes less pronounced under severe label noise or certain code lengths. Fine-tuning exhibits a similar behavior, with its effect depending on both the quality of the initial representation and the degree of noise in the su-

pervision. More importantly, the relative robustness of DUET is preserved across these feature configurations. Regardless of whether conventional pre-extracted features or foundation-model-based representations are used, DUET explicitly estimates modality-specific semantic reliability and constructs cross-modal consensus from more reliable predictions rather than directly enforcing uniform agreement. Therefore, the proposed consensus learning strategy is complementary to the choice of feature extractor. These results suggest that the effectiveness of DUET is not tied to a particular backbone, but can be retained under substantially different representation learning settings.

Table 8: Effect of different feature extractors on DUET performance. "Original Features" denote the pre-extracted features provided by the datasets, while "CLIP (Frozen)" and "CLIP (Fine-tuned)" denote features extracted using frozen and fine-tuned CLIP encoders, respectively.

Bits	Configuration	XMedia								XMediaNet							
		Image to Text				*Text to Image*				*Image to Text*				*Text to Image*			
	Noise	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
16	Original Features	91.1	90.1	90.8	89.2	89.2	89.0	88.5	86.6	42.0	42.1	41.4	42.6	41.2	40.9	40.6	41.2
	CLIP (Frozen)	90.8	90.5	89.7	89.4	89.6	88.7	88.9	86.9	41.6	41.0	40.8	42.0	40.8	40.5	41.1	40.7
	CLIP (Fine-tuned)	91.6	90.8	90.4	89.0	89.8	89.3	88.1	87.2	42.4	41.8	41.0	41.5	41.5	40.6	40.2	41.4
32	Original Features	91.7	91.7	91.5	91.2	91.4	91.1	91.1	90.6	55.4	50.4	51.2	52.4	53.8	48.8	49.4	50.6
	CLIP (Frozen)	92.0	91.3	91.7	90.8	91.0	90.8	90.4	90.2	54.1	51.2	50.7	51.8	52.6	49.1	49.8	50.0
	CLIP (Fine-tuned)	92.4	91.8	91.2	91.0	91.6	91.0	90.7	89.8	55.8	50.9	51.6	52.0	53.2	49.5	49.0	50.9
64	Original Features	92.1	92.1	91.6	92.1	92.2	91.9	92.4	91.1	55.1	55.6	55.8	55.2	53.8	54.0	54.4	54.6
	CLIP (Frozen)	92.4	91.8	91.3	91.7	91.7	91.5	91.9	90.8	55.8	54.9	55.1	54.6	54.1	53.6	54.0	54.2
	CLIP (Fine-tuned)	92.0	92.4	91.8	91.3	92.0	91.6	92.1	90.9	56.2	55.3	55.5	54.9	54.5	53.8	54.7	54.0
128	Original Features	92.5	92.5	92.3	92.0	92.6	92.1	92.2	91.8	56.5	56.7	57.4	55.7	56.3	56.7	57.2	55.7
	CLIP (Frozen)	92.7	92.1	92.0	91.7	92.3	91.8	91.9	91.5	56.0	56.4	56.8	55.2	55.8	56.1	56.9	55.4
	CLIP (Fine-tuned)	92.9	92.6	92.1	92.2	92.5	91.9	92.4	91.4	57.1	56.2	57.0	55.9	56.5	56.0	57.4	55.8

F.8 Additional Experiments on Tri-modal Datasets

To further evaluate the generalization capability of DUET, we conduct experiments on a three-modal retrieval benchmark involving images, meshes, and point clouds. Following RONO (Feng et al., 2023), we adopt the same experimental protocol, including feature extraction, symmetric label-noise construction, and evaluation settings. DUET is compared with CLF (Jing et al., 2021), RONO (Feng et al., 2023), RSHNL (Pu et al., 2025) and RFCMH (Sun et al., 2026). For a fair comparison, CLF and RONO are equipped with the same hashing layer to enable retrieval in the Hamming space. Table 9 reports the complete retrieval results across all six retrieval directions, including Image→Mesh (I→M), Image→Point Cloud (I→P), Mesh→Image (M→I), Mesh→Point Cloud (M→P), Point Cloud→Mesh (P→M), and Point Cloud→Image (P→I). DUET generally achieves the best or highly competitive performance across different retrieval directions, hash code lengths, and label noise rates. These results demonstrate that the proposed conflict-aware semantic consensus learning strategy can generalize beyond conventional image-text retrieval and remain effective under heterogeneous three-modal retrieval with noisy labels. The consistent performance across image, mesh, and point-cloud modalities further supports the robustness of DUET to diverse multimodal data and heterogeneous semantic relationships.

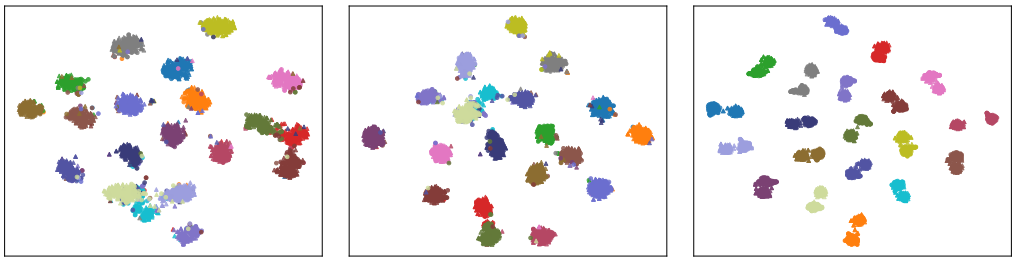
F.9 Visualization Analysis

To further examine the learned representations under noisy supervision, we employ t-SNE visualization to compare the feature spaces produced by different methods across modalities, as shown in Figure 8. DUET produces relatively compact and well-separated semantic clusters, with reduced overlap among samples from different categories and modalities. This suggests that the conflict-aware semantic consensus strategy facilitates more reliable cross-modal semantic alignment while preserving intra-modal discriminability. In contrast, existing methods designed for noisy supervision, such as RSHNL (Pu et al., 2025) and RFCMH (Sun et al., 2026), exhibit more pronounced cross-category and cross-modal overlap. Such entanglement indicates that noisy supervision can introduce inconsistent semantic associations across modalities, making it more difficult to preserve a coherent shared feature space. By accounting for modality-specific reliability during semantic consensus formation, DUET reduces the influence of potentially unreliable modality-specific rep-

Table 9: Retrieval performance on the three-modal ModelNet10 benchmark under different hash code lengths and label noise rates.

Bits		16				32				64				128			
Noise		0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
I → M	CLF (Jing et al., 2021)	81.2	73.6	54.8	18.5	81.8	74.2	55.3	19.0	82.8	74.8	56.0	20.0	83.4	75.6	56.8	20.9
	RONO (Feng et al., 2023)	85.8	84.9	82.0	81.2	86.4	85.1	82.6	81.8	87.0	86.2	83.0	82.1	87.6	86.6	83.1	82.3
	RSHNL (Pu et al., 2025)	85.5	84.2	82.8	82.5	86.2	85.8	83.4	82.9	86.8	86.0	83.6	83.3	87.2	86.8	84.2	83.8
	RFCMH (Sun et al., 2026)	86.8	85.4	83.8	82.9	87.6	86.7	85.1	84.4	88.4	87.6	86.2	85.5	89.3	88.5	87.4	86.6
	DUET	87.5	86.2	85.4	84.8	88.3	87.6	86.5	85.9	89.1	88.5	87.3	86.8	90.2	89.4	88.4	87.6
I → P	CLF (Jing et al., 2021)	81.0	74.2	59.0	20.5	81.6	74.8	59.8	21.2	82.8	75.8	61.0	21.8	83.4	77.8	62.3	22.2
	RONO (Feng et al., 2023)	86.2	85.5	84.2	83.8	86.8	86.2	85.0	84.6	87.4	86.8	85.8	85.1	87.3	86.8	85.8	85.5
	RSHNL (Pu et al., 2025)	86.5	85.8	84.5	84.2	87.2	86.6	85.2	84.8	87.0	87.0	86.0	85.6	87.5	87.1	86.2	85.9
	RFCMH (Sun et al., 2026)	87.0	86.2	84.8	84.4	88.0	87.1	85.7	85.0	88.8	88.0	86.9	86.2	89.8	89.0	87.9	87.2
	DUET	87.9	87.0	86.1	85.5	88.7	87.9	86.8	86.1	89.5	88.8	87.7	87.0	90.7	89.9	88.8	88.1
M → I	CLF (Jing et al., 2021)	82.0	74.8	56.0	21.0	82.6	75.5	57.0	21.8	83.2	76.0	58.0	22.5	84.4	77.1	58.7	24.8
	RONO (Feng et al., 2023)	87.0	86.2	84.8	84.2	87.6	86.8	85.4	84.8	88.0	87.2	85.8	85.2	88.1	87.3	85.8	84.7
	RSHNL (Pu et al., 2025)	87.2	86.4	85.0	84.6	87.8	87.0	85.6	85.0	88.4	87.6	86.2	85.8	88.8	88.0	86.6	86.2
	RFCMH (Sun et al., 2026)	87.7	86.8	85.5	85.1	88.6	87.9	86.5	85.8	89.6	88.9	87.5	86.9	90.7	89.9	88.7	88.1
	DUET	88.5	87.5	86.6	86.0	89.3	88.6	87.5	87.0	90.4	89.7	88.6	88.0	91.5	90.7	89.5	88.9
M → P	CLF (Jing et al., 2021)	80.6	73.2	54.0	18.0	81.2	73.8	54.8	18.6	82.2	74.4	55.6	19.2	83.0	75.2	56.2	20.0
	RONO (Feng et al., 2023)	85.4	84.6	81.8	81.2	86.0	85.2	82.5	82.0	86.6	85.8	83.2	82.6	87.2	86.4	83.6	83.0
	RSHNL (Pu et al., 2025)	85.2	84.5	82.9	83.2	85.8	85.6	83.4	83.8	86.4	86.0	83.8	84.1	87.0	86.8	84.2	84.5
	RFCMH (Sun et al., 2026)	86.5	85.7	84.0	83.8	87.8	86.9	85.4	84.9	88.2	87.5	86.2	85.6	89.6	88.8	87.7	87.0
	DUET	87.3	86.5	85.6	84.9	88.5	87.7	86.6	85.8	89.0	88.3	87.2	86.6	90.4	89.6	88.5	87.8
P → M	CLF (Jing et al., 2021)	81.5	74.2	55.5	20.0	82.0	75.0	56.5	21.0	82.8	75.8	58.0	22.0	83.5	76.5	59.0	23.0
	RONO (Feng et al., 2023)	86.0	85.2	83.8	83.2	86.8	86.0	84.6	84.0	87.2	86.5	85.0	84.5	87.6	86.8	85.5	85.0
	RSHNL (Pu et al., 2025)	86.2	85.5	84.0	83.5	87.0	86.2	84.8	84.2	87.6	86.8	85.4	84.8	87.8	87.2	86.0	85.6
	RFCMH (Sun et al., 2026)	87.0	86.1	84.9	84.4	88.1	87.3	86.0	85.4	88.9	88.1	87.0	86.3	90.0	89.2	88.1	87.4
	DUET	87.8	86.8	85.9	85.2	88.7	88.0	86.9	86.3	89.6	88.8	87.8	87.1	90.8	90.0	88.9	88.2
P → I	CLF (Jing et al., 2021)	81.2	74.6	58.5	20.8	81.8	75.2	59.5	21.5	82.5	76.0	60.0	22.0	83.5	77.4	60.0	24.1
	RONO (Feng et al., 2023)	85.5	84.8	83.5	83.0	86.0	85.5	84.2	83.8	85.8	85.9	86.1	84.4	86.3	85.9	86.1	84.4
	RSHNL (Pu et al., 2025)	85.8	85.2	83.8	83.2	86.5	85.8	84.5	84.0	87.2	86.5	85.2	84.8	87.6	87.0	85.6	85.2
	RFCMH (Sun et al., 2026)	86.9	85.9	84.6	84.0	87.7	86.9	85.5	84.9	88.5	87.7	86.6	85.9	89.7	88.9	87.8	87.1
	DUET	87.4	86.3	85.3	84.7	88.2	87.5	86.4	85.7	89.2	88.4	87.4	86.7	90.5	89.6	88.5	87.9

resentations and yields more structured feature distributions. These qualitative observations are consistent with the quantitative retrieval results reported above.



(a) RSHNL. (b) RFCMH. (c) DUET.

Figure 8: t-SNE visualization on XMedia under a noise rate of 0.8. Different colors indicate distinct semantic categories, where ● denotes image features and ▲ denotes text features.

F.10 Limitation Statement

Although DUET achieves promising performance under label-noise settings, two limitations remain. First, the temporal-stability weight α_i^m and entropy-based consensus weight β_i^m are empirical cues rather than rigorous indicators of ground-truth correctness. Under severe noise and prolonged training, modalities may memorize corrupted labels and produce stable, high-confidence but incorrect predictions. When such errors are shared across modalities, consensus may instead reinforce them. Second, DUET relies on several manually specified hyper-parameters, whose optimal values may vary across datasets and noise conditions.

These limitations motivate our future research in two directions: developing stronger semantic correctness estimation by incorporating complementary evidence beyond prediction stability and confidence, and designing self-adaptive consensus mechanisms that automatically adjust their interaction strength and hyper-parameters according to the observed training dynamics. These directions will be a primary focus of our future work.