ENHash: Error Notebook-Guided Fine-Grained Learning for Unsupervised Cross-Modal Hashing

Hao Fu[1], Zebing Yao[1], Chuangchuang Tan[2], Guanghua Gu[1]*

[1]School of Information Science and Engineering, Yanshan University
[2]Institute of Information Science, Beijing Jiaotong University
guguanghua@ysu.edu.cn

Abstract

Without manual annotations, unsupervised cross-modal hashing (UCMH) aims to achieve efficient clustering and retrieval by leveraging data interrelationships. However, the retrieval accuracy is constrained by two main aspects: 1) insufficient exploration of data relationships; 2) existing knowledge mining strategies are not well aligned with the architectural properties of multilayer perceptrons. Through summary and error analysis, the human brain is able to achieve fast learning through experience and minimal data. Inspired by this cognitive process, we propose a novel Error Notebook strategy, named ENHash, to more effectively capture similarity information between multi-modal data for fine-grained unsupervised clustering. Firstly, simulating the human process of summarizing experiences, ENHash gradually integrates the information from each batch into a global clustering representation. Secondly, drawing upon human error analysis capabilities, ENHash utilizes the summarized experiences to identify and record incorrectly predicted hash codes. Finally, by leveraging the knowledge derived from this analysis, ENHash guides the hash function to learn fine-grained patterns from the errors. To the best of our knowledge, ENHash represents the first attempt at integrating cognitively-inspired mechanisms into fine-grained UCMH optimization paradigms. We evaluate the proposed ENHash against eight state-of-the-art methods on three widely used datasets and one fine-grained cross-modal dataset. Experimental results show that ENHash achieves substantial improvements over existing approaches.

Introduction

Due to the inherent distributional differences between heterogeneous data types, direct measurement of their similarity relationships remains a challenging task. To address this challenge, cross-modal retrieval (CMR) methods (Cheng et al. 2022; Li et al. 2024) have been proposed to bridge the heterogeneity gap by mapping these modalities into a shared representation space. Similar to CMR approaches, cross-modal hashing (CMH) methods (Yu et al. 2020; Zheng et al. 2023) aimed to facilitate more efficient retrieval processes by enabling rapid similarity comparisons through Hamming distance computation.

*Corresponding author

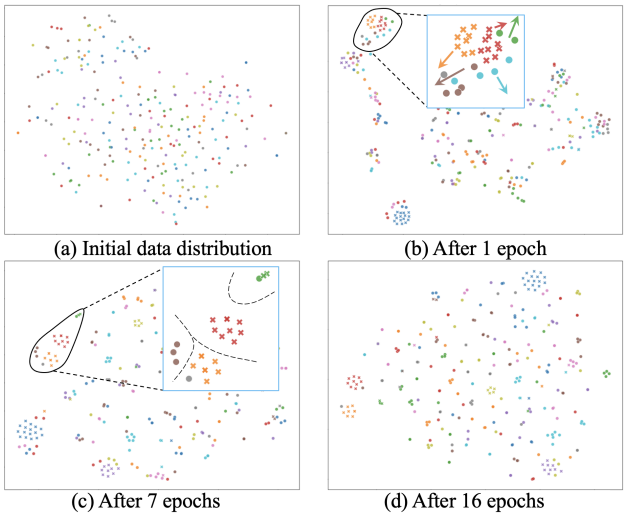

Figure 1: The t-SNE visualization illustrating how the proposed ENHash framework identifies and mitigates model errors on the FG-XMedia dataset. Each colored dot represents an image sample from a distinct category, while cross marks indicate misclassified instances by the hashing model. Colored arrows depict the trajectory of cluster centroids associated with each category across training iterations.

Without extensive annotation, unsupervised deep methods (Zhang et al. 2021; Hu et al. 2022) have garnered significant attention in recent years. However, existing deep methods mostly follow the concept of supervised learning, such as leveraging auxiliary networks to learn label distributions, and fail to fully exploit the latent information embedded within the data. Moreover, current deep UCMH networks generally adopt fully connected layers as the backbone and rely on backpropagation strategies to map features into hash codes. Building upon insights from (Zhou et al. 2025; Zheng et al. 2025), we find that underrepresented samples tend to be forgotten by the model during training. Although the more frequent samples may better represent the main content of the dataset, we argue that fine-grained optimization of hard-to-learn and minority samples can indirectly influence the model's overall learning performance on the entire dataset.

Reflection, feedback, and assessment are crucial components of the learning processes in both the human brain and deep learning systems (Kudithipudi et al. 2022; Yan et al. 2024). As one of the important methods for providing feedback and assessing learning, error analysis is a widely utilized technique in cognitive psychology to investigate the brain's encoding mechanisms (Bayne et al. 2019). This principle extends to human learning. Through induction and summarization, humans refine error patterns and use these errors to reinforce the knowledge they have acquired. This error-correction mechanism is particularly relevant to unsupervised learning, where the absence of annotated information often leads to higher error rates (Zhang and Peng 2019). In other words, if the similarity knowledge within errors can be effectively utilized, unsupervised methods will achieve more comprehensive information extraction from the data. Therefore, inspired by human learning paradigms, this paper proposes a new form of dynamic supervisory feedback, named Error Notebook (ENHash).

By simulating the process of experience summarization and error analysis in human cognition, ENHash seeks to leverage prediction errors for self-optimization of the hashing model, thereby facilitating more comprehensive information extraction from unsupervised data. Since the training set is shuffled during each iteration, to better illustrate this effect, we randomly sample and save one batch of training data to visualize the model's dynamic changes on this batch at the end of each epoch. Figure 1 visualizes the optimization trajectory of misclassified samples using our proposed method on the FG-XMedia (He, Peng, and Xie 2019) dataset across multiple training stages. As depicted in Figure 1(a), initial data distributions are unstructured; however, during early iterations such as that shown in (b), preliminary clustering behavior emerges despite unreliable feature representations at this stage of training. Accordingly, ENHash aims to identify these incorrectly clustered instances and utilize them to refine hash function learning progressively. By examining enlarged views presented in (b) and (c), it becomes evident that ENHash promotes separation among samples with different colors while encouraging proximity among those with the same color. Through iterative refinement, ENHash gradually achieves unsupervised fine-grained clustering performance over time. As illustrated in Figure 1(d), distinct inter-class boundaries emerge between differently colored clusters, with limited cross-category contamination within each cluster.

To achieve the aforementioned effect, ENHash first summarizes the distribution of data within each batch using an incremental iterative approach, representing the summarized knowledge through pseudo-labels. Next, by comparing the predicted values to these pseudo-labels, ENHash filters and records the incorrect hash codes, which are designated as errors. Finally, for the hard-to-learn errors, ENHash introduces a novel multi-cluster boundary loss that drives the model toward regions with the highest correct probabilities, thus ensuring accurate predictions while enabling fine-grained clustering of unsupervised samples.

In summary, the main contributions are:

- We propose ENHash, a novel unsupervised cross-modal hashing framework designed to emulate human cognitive learning mechanisms for inductively summarizing fine-grained unlabeled data. To the best of our knowledge, ENHash represents the first attempt to address fine-grained learning in the UCMH domain.
- The proposed approach designs a novel multi-cluster boundary loss. The primary objective of this loss is to encourage challenging samples to align more closely with the correct distribution, thereby maximizing the probability of accurate predictions.
- By simulating the process of human learning and cognition, ENHash enables more precise self-optimization and fine-grained knowledge acquisition through both the features and the model itself. To be specific, ENHash reaches a MAP@ALL of 0.610 on 128-bit $I \rightarrow T$ task of the FG-XMedia dataset. For the larger MS COCO dataset, ENHash achieves the best performance with a significant improvement of 3.0% on 16-bit $I \rightarrow T$ task.

Related Work

Human-like Learning Paradigm

The definition of artificial cognition is the use of psychology to understand, evaluate, and interpret machine learning algorithms based on human performance (Ritter et al. 2017). Regardless of the cognitive theory it is based on—whether it involves computational theories using abstract symbols or theories simulating brain functions—it contributes to the development of artificial intelligence (Marcus and Davis 2019). In recent years, many outstanding works have emerged by incorporating mechanisms that simulate the human brain's learning process. By simulating the attention mechanism of the human brain, transformer (Vaswani 2017) was proposed to enable the model to dynamically focus on important parts of the data during processing, allowing it to capture long-term dependencies between words in a sentence. Furthermore, vision transformer (ViT) (Dosovitskiy 2020) was applied this mechanism to computer vision tasks, achieving performance superior to CNNs in large-scale data training. In the domain of large-scale models with even greater parameter counts, GPT (Radford et al. 2018) was designed to simulate human-like capabilities in language generation and task understanding. Additionally, due to its extensive learning of linguistic knowledge, GPT has demonstrated a certain degree of reasoning ability that resembles aspects of human thinking. Moreover, the introduction of CLIP (Radford et al. 2021) further enriched the exploration of deep learning in multi-modal zero-shot tasks.

Unsupervised Cross-modal Hashing

The main idea of unsupervised cross-modal retrieval methods is to guide the model to learn high-quality hash functions through the intrinsic relationships within the data. By mitigating the impact of false negative pairs (FNPs) on deep learning models, UCCH (Hu et al. 2022) proposed a novel momentum-based hashing optimizer to enable instance-level contrastive learning within a unified hash

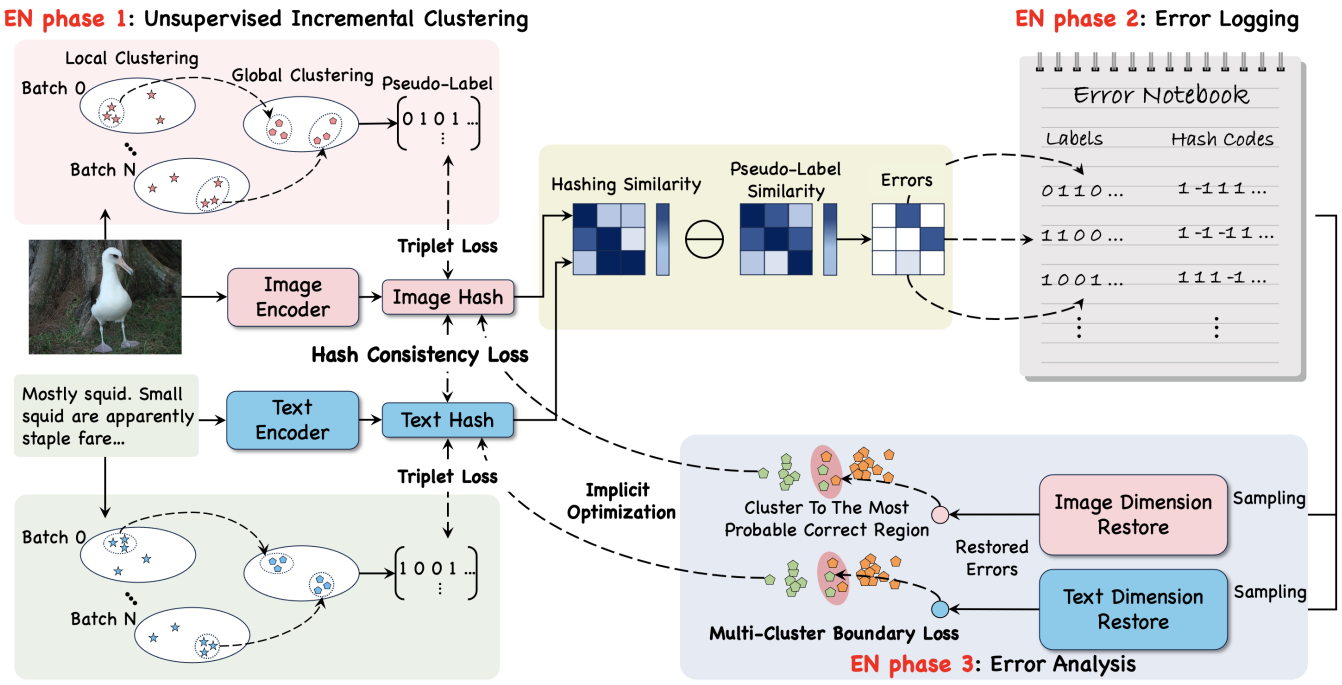

Figure 2: The proposed ENHash framework integrates an unsupervised cross-modal hashing network with the three-phase EN strategy. Specifically, EN consists of an unsupervised incremental clustering strategy, an error logging and analysis process. Through unsupervised incremental clustering, the hashing model is able to acquire similarity knowledge from the training data, while more fine-grained optimization information is obtained through error logging and error analysis.

dictionary. Through knowledge distillation, UKD (Hu et al. 2020) and SKDCH (Su et al. 2021) enabled supervised training using the outputs of an unsupervised and a semi-supervised teacher model, respectively. In addition to optimizing the performance at the model level, recent approaches have also focused on the mathematical statistics of the unsupervised samples themselves. AGCH (Zhang et al. 2021) attempted to use an affinity graph construction approach to uncover the latent semantic structure between the data. UDDH (Zhang et al. 2024) achieved outstanding retrieval performance by fitting cross-modal semantic information and multi-modal content information through training of the head and tail codes, respectively. Although the above methods have achieved promising results, none of them have realized the simultaneous optimization at both the model level and the data level within a generalized network.

In general, the above studies mainly focus on exploring semantic relationships between data at the feature level, such as through labels or graph-based relationships, often overlooking the hashing function's capacity to learn from these high-dimensional representations. As a result, these approaches fail to fully capture the latent similarity information between modalities. Furthermore, in large-scale unsupervised datasets, samples frequently exhibit one-to-many relationships, posing an additional challenge for effectively learning fine-grained semantic correspondences within the network. Consequently, it is crucial to develop a generalized learning paradigm capable of addressing the complexities of unsupervised fine-grained cross-modal tasks.

Methodology

The proposed ENHash aims to enhance the discriminability of hash codes in fine-grained tasks while maintaining consistency. As depicted in Figure 2, our approach comprises two main components: the proposed error notebook and the unsupervised cross-modal hashing network.

Problem Formulation

$D = \{d_j \mid d_j = (I_j, T_j), j = 1, 2, \ldots, N\}$ is a given unsupervised multi-modal dataset, where (I_j, T_j) denotes the j-th paired tuple of an image and text from the dataset. The goal of ENHash is to learn the mapping functions $H^I : F^I \rightarrow B^H$ and $H^T : F^T \rightarrow B^H$ to minimize the distance between image and text modalities in the common Hamming space, where F^I and F^T represent the features of image and text samples, respectively. $B^H \in \{-1, 1\}^P$ denotes the P-bit hash codes. After obtaining the unified hash representations, queries can efficiently retrieve semantically similar samples from another modality via Hamming distance.

Error Notebook

The main idea of the proposed ENHash module is to provide global feedback to the cross-modal hashing model during its training process. As shown in Figure 2, ENHash is composed of three processes: unsupervised incremental clustering, error logging, and error analysis.

Unsupervised Incremental Clustering Currently, the primary solution to unsupervised cross-modal hashing is

based on contrastive learning (Hadsell, Chopra, and LeCun 2006), which aims to reduce the semantic distance between similar samples while increasing the distance between dissimilar ones. Because the data contains only pairs of semantic information and lacks category and label details, it is difficult to accurately identify the negative samples required for contrastive learning. Similar to (Yang, Hu, and Hu 2025; Wang et al. 2021b), this paper adopts an unsupervised hierarchical clustering method for batch-sized data to avoid a significant loss of useful information. Specifically, when the input arrives, the task of the local cluster can be formulated as follows. Without losing commonality, the following content will use the image modal I as an example.

$$DIST\left(C_m, C_n\right) = \left\|U_m - U_n\right\|_2^2, \tag{1}$$

$$U_m = \frac{1}{|C_m|} \sum_{x_l \in C_m} x_l, \tag{2}$$

$$C_{new} = CUP_{min}\left(DIST\left(C_m, C_n\right)\right). \tag{3}$$

where $m \neq n$ and $m, n \in N$. N is the size of the dataset and x_l denotes a sample belonging to C_m. Through dividing by the number of sample points, C_m represents the m-th cluster and U_m represents the center of C_m. The same applies to C_n and U_n. As described in Equations 1, 2 and 3, this process first calculates the distance between clusters and identifies the closest pair of clusters. Then, via $CUP_{min}\left(\cdot\right)$, the two closest clusters are merged into a new cluster C_{new}. The clustering process will iterate and terminate when the number of clusters is equal to the number of categories, thereby achieving sample clustering through similarity measurement. The main idea of unsupervised hierarchical clustering is to partition samples into different clusters.

Nonetheless, due to the large number of samples in the dataset, the model cannot feasibly perform clustering and training on the entire dataset. Instead, a mini-batch training strategy is employed in this paper. Because of the randomness in the samples of each batch, the performance of unsupervised hierarchical clustering will be significantly affected. To tackle this problem, this paper further proposes an unsupervised incremental clustering (UIC) method, which can be defined as follows:

$$U_{cur}^I = CUP_{min}\left(DIST\left(U_{temp}^I, U_{pre}^I\right)\right). \tag{4}$$

here U_{temp}^I is an intermediate state representing the unsupervised hierarchical cluster result of the current batch. U_{cur}^I denotes the current cluster center, while U_{pre}^I represents the previous ones. For each batch, UIC compares the current batch's cluster centers with the previously recorded cluster centers. By repeatedly calculating the squared error between the two sets of cluster centers, the clustering results are refined. The updated clustering results are then recorded for use in the next batch of training. In this way, by the end of an epoch, UIC can estimate the global cluster centers based on local information.

$$B_j = \begin{cases} 1, & \text{if } \left\|I_j - U_{cur}^I\right\|^2 < \bar{d} \\ 0, & \text{otherwise} \end{cases}. \tag{5}$$

When obtaining stable cluster centers, binary pseudo-labels B_j will be assigned for d_j. As shown in Equation 5, $\bar{d} = \frac{\gamma}{N} \sum_j^N \left\|I_j - U_{cur}^I\right\|^2$ denotes the average distance between each sample and cluster center. And γ is the noise suppression factor used to reduce the impact of clustering noise. In addition, to further improve the accuracy of the pseudo-binary label, we adopt the intersection of image and text labels as the final pseudo-label B^{pse}. Compared to the ground truth of the MS COCO and MIRFlickr datasets, the accuracy of B^{pse} can reach 96% and 81%, respectively.

Error Logging To simulate the human process of self-optimization through errors, this paper considers the distribution knowledge from UIC as the learning target. By comparing the outputs of the hash function with this target, the required optimization for the model can be determined. Subsequent ablation experiments demonstrate that this approach can not only enhance the ability to optimize the objective function, but also compensate for information loss caused by nonlinear transformations within the network. The error logging process can be described as follows:

$$Err = TopK\left(S_{B^{pse}} - S_{B^H}\right). \tag{6}$$

where $S_{B^{pse}}$ is the similarity matrix of the pseudo-label B^{pse} and S_{B^H} is the similarity matrix of the hash codes.

Error Analysis To better analyze the underlying distribution of errors, during this stage, ENHash module first randomly selects a portion of the recorded errors based on a percentage of the batch size. By leveraging errors recorded across multiple batches, ENHash module can offer more generalized knowledge, thereby preventing the model from overfitting to the sample distribution of the current batch.

Subsequently, to make the information contained in the binary codes easier to learn, ENHash module utilizes a hash reconstruction network to restore the hash codes of images and texts to their respective feature dimensions. The hash reconstruction network consists of three fully connected layers and is designed to identify shortcomings in the hash function by leveraging a shared subspace. Meanwhile, by modulating latent dimensionality within the shared subspace, our approach enables effective alignment among hash codes of differing lengths and mitigates discrepancies arising from modality-specific feature representations. This design choice further alleviates the risk of overfitting or underfitting that may occur when directly optimizing parameters in the primary hashing network. It is worth noting that, to maintain the consistency of the hash function, the hash codes of images and texts are cross-utilized to learn information from texts and images, respectively.

For the logged errors, most are hard-to-learn samples. Common contrastive loss lacks the ability to effectively learn from such hard examples, while triplet loss struggles to construct negative samples under unsupervised conditions. To address this challenge, we propose a multi-cluster boundary loss. As illustrated in Figure 3, unlike single-label and supervised multi-label tasks, the proposed loss function does not aim to minimize the semantic distance between samples

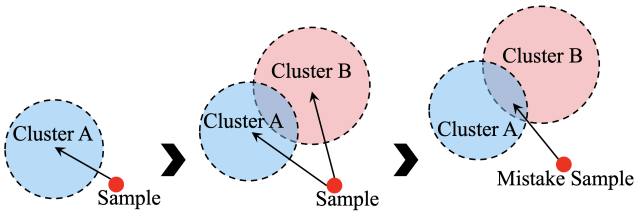

Figure 3: The comparison between common multi-label loss and the proposed multi-cluster boundary loss.

and their corresponding clusters, but rather pushes errors toward the region with the highest probability of correctness. Furthermore, since this stage does not directly measure hash codes but instead adopts an implicit optimization approach, it enhances the learning capability of the hashing model.

At this point, the proposed multi-cluster boundary loss of this error analysis process can be formalized as follows:

$$L_{mcb} = \frac{1}{M}\sum_{j=1}^{M}\left\|F_j^{RI} - S_j^I\right\|_2^2 + \frac{1}{M}\sum_{j=1}^{M}\left\|F_j^{RT} - S_j^T\right\|_2^2. \tag{7}$$

here F^{RI} and F^{RT} represent the reconstructed image and text hashing representation, respectively. S^I denotes the dot product similarity between B^{pse} and U_{cur}^I, and likewise for S^T. M is the number of the selected errors.

Cross-modal Hashing As shown in Figure 2, the proposed cross-modal hashing network comprises an image branch and a text branch, each consisting of an encoder followed by a multi-layer perceptron. Unlike traditional tokenizations, the proposed hashing network first forms a sequence of input image features $Image^s = \{\mathcal{F}_1^I, \ldots, \mathcal{F}_{batch}^I\}$ into the multi-head attention module. Through this serialized input, the transformer encoder can focus on the feature differences between different samples within the same modality and discard redundant information.

The goal of the proposed hashing network is to jointly learn modality-specific discriminative hash functions and modality-invariant consistent representations. Given that high-quality binary pseudo-labels B^{pse} are generated by the ENHash module, we leverage a triplet loss function to enhance the model's ability to capture semantic discrimination across modalities, which is defined as:

$$\begin{aligned} L_{dis} &= \tfrac{1}{2N}\sum_{\substack{m=1\\n=1}}^{N}\max\left(0, d_{I2T}^{+}[m] - d_{I2T}^{-}[m,n] + M_{dis}\right) \\ &+ \tfrac{1}{2N}\sum_{\substack{m=1\\n=1}}^{N}\max\left(0, d_{T2I}^{+}[m] - d_{T2I}^{-}[m,n] + M_{dis}\right) \end{aligned}. \tag{8}$$

where $m \neq n$ and $m, n \in N$. d^+ denotes the l_2-norms between the query and positive samples, and d^- denotes the l_2-norms between the query and negative samples. To reduce the influence of pseudo label noise, four negative pairs are randomly sampled from all possible negatives for L_{dis}. By using the margin parameter M_{dis}, L_{dis} is able to ensure the semantic similarity of non-corresponding sample pairs is more discriminative than that between corresponding sample pairs. Then, the hashing network further ensures the consistency of the hashing functions through a cross-modal attraction loss, which is formulated as:

$$\begin{aligned} L_{att} &= \frac{1}{N}\sum_{m=1}^{N}\max\left(0, d_{I2T}[m] - M_{att}\right) \\ &+ \frac{1}{N}\sum_{m=1}^{N}\max\left(0, d_{T2I}[m] - M_{att}\right) \end{aligned}. \tag{9}$$

here M_{att} is the margin parameter used to constrain the semantic distance between the hash codes of image and text modalities.

Finally, the overall objective function L_{ENHash} of the proposed approach is:

$$L_{ENHash} = \alpha L_{dis} + \beta L_{att} + (1-\alpha-\beta)L_{mcb}. \tag{10}$$

where α and β are the hyperparameters.

Experiments

Datasets and Evaluation Metrics

MS COCO (Lin et al. 2014) comprises 123,287 images, each annotated with five textual descriptions and labeled across 80 object categories. We randomly select 5,000 samples to form the training set, while the remainder serve as the retrieval database. For MIRFlickr (Huiskes and Lew 2008), we adopt a subset comprising 2,000 image-text pairs, which are randomly designated as queries, while all other instances constitute the retrieval database. Following (Hu et al. 2022), we select a subset of NUS-WIDE (Rasiwasia et al. 2010) to the ten most frequently occurring classes. FG-XMedia (Peng, Huang, and Zhao 2017; He, Peng, and Xie 2019) consists of 200 fine-grained bird categories. Its image modality contains 5,994 training samples and 5,794 testing samples. The text modality contains 4,000 samples for both training and testing sets. Furthermore, for unsupervised experiments, we adopt the same selection strategy as MIRFlickr to construct image-text pairs.

In this paper, we employ the mean average precision (MAP) to compute all retrieved results (MAP@ALL).

Comparison Methods

In this paper, we evaluate 8 state-of-the-art cross-modal approaches as baselines, including five UCMH methods—CIRH (Zhu et al. 2022), UCMFH (Xia et al. 2023), UCCH (Hu et al. 2022), and SACH (Cui et al. 2024), and four supervised fine-grained real-valued cross-modal retrieval approaches-FGCrossNet (He, Peng, and Xie 2019), SAFGCM (Wang et al. 2021a), CMCM (Shen et al. 2022), and FGAN (Ge et al. 2025).

TASK	METHOD	MIRFLICKR				NUS-WIDE				MS COCO			
		16BITS	32BITS	64BITS	128BITS	16BITS	32BITS	64BITS	128BITS	16BITS	32BITS	64BITS	128BITS
I→T	CIRH	0.697	0.704	0.717	0.726	0.573	0.602	0.606	0.623	0.552	0.590	0.602	0.596
	UCMFH	0.706	0.732	0.728	0.730	0.585	0.613	0.624	0.630	0.565	0.587	0.601	0.613
	UCCH	**0.739**	0.744	0.754	0.760	0.698	0.708	**0.737**	**0.742**	0.605	0.645	0.655	0.665
	SACH	0.724	0.726	0.732	0.734	0.612	0.624	0.627	0.633	/	/	/	/
	ENHASH	0.738	**0.751**	**0.756**	**0.763**	**0.701**	**0.735**	0.734	0.733	**0.635**	**0.652**	**0.669**	**0.670**
T→I	CIRH	0.699	0.712	0.722	0.735	0.579	0.607	0.614	0.628	0.564	0.590	0.597	0.597
	UCMFH	0.708	0.735	0.729	0.734	0.596	0.622	0.637	0.640	0.567	0.592	0.598	0.610
	UCCH	0.725	0.725	0.743	0.747	0.701	0.724	0.745	0.750	0.610	0.655	0.666	**0.677**
	SACH	0.727	**0.736**	0.740	0.743	0.614	0.621	0.625	0.630	/	/	/	/
	ENHASH	**0.733**	0.732	**0.747**	**0.748**	**0.735**	**0.759**	**0.755**	**0.758**	**0.636**	**0.659**	**0.667**	0.673

Table 1: MAP@ALL comparison of different methods on three datasets under I→T and T→I tasks.

METHODS	$I \to T$	$T \to I$	AVG
FGCROSSNET	0.210	0.255	0.233
SAFGCM	0.293	0.335	0.314
CMCM	0.349	**0.353**	0.351
FGAN	0.161	0.147	0.154
ENHASH(128BITS)	**0.610**	0.138	**0.374**

Table 2: MAP@ALL comparison on FG-XMedia dataset.

MS COCO (123,287 images in total)			
Category	**Freq.**	**Category**	**Freq.**
$[Elephant]$	1094	$[Person, Elephant]$	427
$[Bear]$	770	$[Person, Bear]$	45
$[Cow]$	958	$[Person, Cow]$	185
$[Sheep]$	844	$[Person, Sheep]$	133
$[Toothbrush]$	56	$[Person, Toothbrush]$	226

Table 3: Frequency counts of the hard-to-learn categories.

Implementation Details

To ensure reproducibility, we fix the random seed at 123 during training. The hyperparameters α and β in Equation 10 are set to 0.6 and 0.2, respectively. A batch size of 256 and a learning rate of 1×10^{-4} are used across all experiments, which are conducted on an NVIDIA H800 GPU.

Experimental Results and Analysis

Table 1 shows the MAP@ALL scores of eight cross-modal retrieval tasks on three datasets and the MAP@ALL results are displayed in Table 2. $I \to T$ represents using image as query to retrieve text, and $T \to I$ represents using text as query to retrieve image. From these two tables, it can be observed that our proposed ENHash method achieves outstanding retrieval accuracy.

For MIRFlickr dataset in Table 1, compared with the latest SACH method, the highest improvement appears at the 128-bit $I \to T$ task of 2.9%. For the comparisons on datasets with more categories and a larger number of samples, the proposed ENHash outperforms UCCH by 3.4% and 3.5% in 16-bit and 32-bit $T \to I$ tasks on NUS-WIDE dataset, and 3.0% in 16-bit $I \to T$ task on MS COCO dataset, respectively. Both SACH and UCCH are based on contrastive learning, these improvements demonstrate that the proposed human-like learning paradigm gains a better capability for information extraction from unsupervised paired data.

We adopt 128-bit results for compare with the four real-valued methods. Although accurate annotations are unavailable, the proposed ENHash achieves notable improvements on the $I \to T$ task. As presented in Table 2, ENHash surpasses FGAN, CMCM, SAFGCM and FGCrossNet by 44.9%, 26.1%, 31.7% and 40.0%, respectively. Moreover, ENHash also gains the highest average MAP@ALL scores which outperform the second-best CMCM by 2.3%. Although ENHash achieves notable improvements in fine-grained image clustering for the $I \to T$ task on the FG-XMedia dataset, its performance on the reverse $T \to I$ task remains limited. Further analysis reveals that the dataset's textual samples are characterized by a certain extent of ambiguous semantics. For instance, a sample from the *American_Crow* category states: "Frequently forages at landfills and other areas with garbage." The inherently abstract nature of such descriptions—especially within unsupervised settings like ENHash—introduces additional challenges and highlights an important direction for future research.

Fine-grained Visual Analysis

To further assess ENHash's performance on fine-grained clustering, we select five infrequent labels from MS COCO dataset along with their semantically similar counterparts due to limited sample availability and semantic overlap. The frequency distributions of these samples within the training set are summarized in Table 3.

From the t-SNE results in Figure 4, it is evident that UCCH tends to cluster multiple sub-groups for samples of the same color. For example, in the first column, two green clusters of $[Person, Elephant]$ are separated by multiple red clusters of $[Elephant]$. Similarly, in the fourth column, the green cluster of $[Person, Sheep]$ is surrounded by multiple red clusters of $[Sheep]$. In contrast, ENHash produces more compact and distinct red and green clusters. Notably, as shown in the last column of Figure 4 for $[Toothbrush]$ and $[Person, Toothbrush]$, the distance between the red

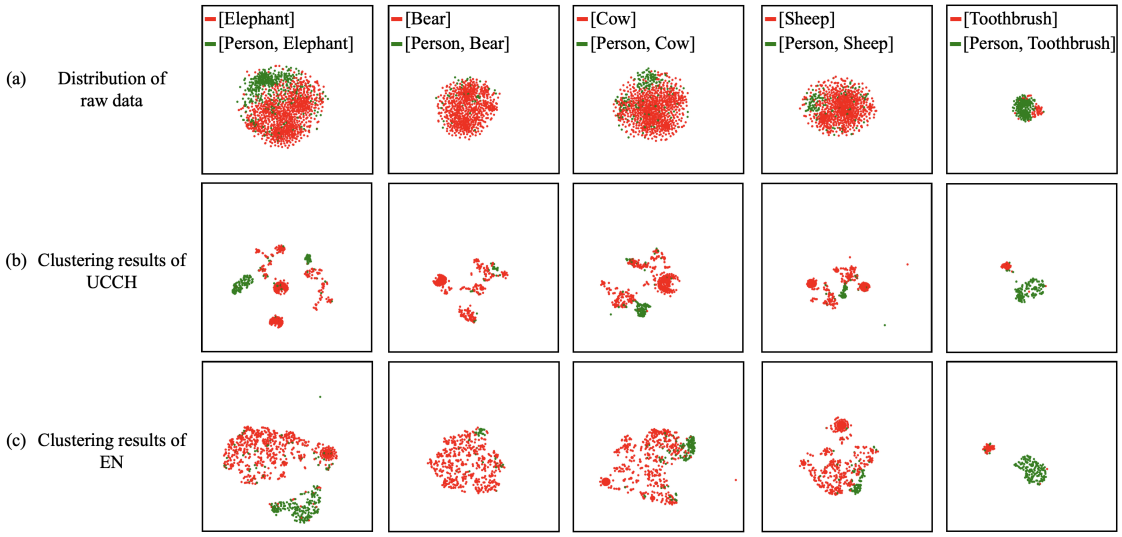

Figure 4: Comparison of t-SNE visualizations corresponding to explicit and implicit optimization strategies under the 64-bit configuration on the MS COCO dataset. Each column denotes a specific group comprising semantically similar yet challenging categories selected from the dataset.

Method(64-bit)	I→T	T→I
ENHash with ordinary hierarchical clustering	0.621	0.627
ENHash-$L_{dis+att}$	0.658	0.662
ENHash-$L_{dis+mcb}$	0.659	0.662
ENHash-$L_{att+mcb}$	0.369	0.433
ENHash-FULL ($L_{dis+att+mcb}$)	**0.669**	**0.667**

Table 4: Ablation study in terms of MAP@ALL.

and green clusters computed by ENHash is significantly larger than that of UCCH. Furthermore, ENHash enlarges the inter-class distances among semantically similar categories when compared to their original distribution representations. These results show that our proposed L_{mcb} achieves superior intra-class compactness and inter-class separability compared to the ranking loss employed in UCCH.

Ablation Study

As shown in Table 4, ablation studies are carried out on the large-scale MS COCO dataset. Based on the three components of the proposed method, UIC, L_{dis}, L_{att}, and L_{mcb}, we design five ablation settings covering the primary combinations on MS COCO utilizing 64-bit hash codes.

We first compare the proposed UIC against the conventional unsupervised hierarchical clustering method. As reported in Table 4, our incremental iterative strategy yields performance gains of 4.8% on $I \rightarrow T$ retrieval and 4.0% on $T \rightarrow I$ retrieval. Such results shows that the UIC facilitates improved compatibility with mini-batch training strategy and enhances the quality of discovered cluster structures in an unsupervised setting.

From Table 4, it can be observed that the proposed ENHash method outperforms all other combinations. Therefore, the mutual collaboration among L_{dis}, L_{att} and L_{mcb} contributes to improving the performance of unsupervised cross-modal retrieval. Benefiting from richer optimization knowledge about errors, ENHash-FULL achieves higher accuracies than ENHash-$L_{dis+att}$, with improvements of 1.1% and 0.5% on the $I \rightarrow T$ and $T \rightarrow I$ tasks, respectively. Meanwhile, by leveraging cross-modal consistency knowledge, ENHash-FULL outperforms ENHash-$L_{dis+mcb}$ by 1.0% and 0.5% on the $I \rightarrow T$ and $T \rightarrow I$ tasks, respectively. Furthermore, serving as an implicit optimization mechanism, ENHash-$L_{dis+mcb}$ outperforms its counterpart ENHash-$L_{dis+att}$—which employs a comparatively robust and simplified optimization scheme—by 0.1% on the $I \rightarrow T$ task, while maintaining comparable performance on the $T \rightarrow I$ task. These results highlight the effectiveness of the proposed L_{mcb} in utilizing prediction errors as complementary supervisory signals during training, and additionally confirm the high reliability of the pseudo-labels produced by our method.

Conclusion

This work introduces a novel unsupervised fine-grained learning framework inspired by human cognitive processes, referred to as Error Notebook (ENHash). ENHash initially aggregates intra-batch data distributions into a unified global cluster representation. It then identifies and excludes erroneously predicted hash codes to emulate analytical correction mechanisms. Finally, a newly designed multi-cluster boundary loss facilitates fine-grained feature discrimination and enhances information extraction from unlabeled samples. Showcasing a notable average retrieval improvement of +0.4%, +1.3%, +1.1% and +2.3%, the proposed method demonstrates its effectiveness through comprehensive experimental comparisons with the state-of-the-art method on three widely used datasets and one fine-grained dataset.

Acknowledgements

This work was partly supported by National Natural Science Foundation of China (No.62072394), and Natural Science Foundation of Hebei province (F2024203049).

References

Bayne, T.; Brainard, D.; Byrne, R. W.; Chittka, L.; Clayton, N.; Heyes, C.; Mather, J.; Ölveczky, B.; Shadlen, M.; Suddendorf, T.; et al. 2019. What is cognition? *Current Biology*, 29(13): R608–R615.

Cheng, Y.; Zhu, X.; Qian, J.; Wen, F.; and Liu, P. 2022. Cross-modal graph matching network for image-text retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(4): 1–23.

Cui, J.; He, Z.; Huang, Q.; Fu, Y.; Li, Y.; and Wen, J. 2024. Structure-aware contrastive hashing for unsupervised cross-modal retrieval. *Neural Networks*, 174: 106211.

Dosovitskiy, A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Ge, S.; Jiang, Z.; Yin, Y.; Wang, C.; Cheng, Z.; and Gu, Q. 2025. Fine-Grained Alignment Network for Zero-Shot Cross-Modal Retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, 1735–1742. IEEE.

He, X.; Peng, Y.; and Xie, L. 2019. A new benchmark and approach for fine-grained cross-media retrieval. In *Proceedings of the 27th ACM international conference on multimedia*, 1740–1748.

Hu, H.; Xie, L.; Hong, R.; and Tian, Q. 2020. Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3123–3132.

Hu, P.; Zhu, H.; Lin, J.; Peng, D.; Zhao, Y.-P.; and Peng, X. 2022. Unsupervised contrastive cross-modal hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3877–3889.

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

Kudithipudi, D.; Aguilar-Simon, M.; Babb, J.; Bazhenov, M.; Blackiston, D.; Bongard, J.; Brna, A. P.; Chakravarthi Raja, S.; Cheney, N.; Clune, J.; et al. 2022. Biological underpinnings for lifelong learning machines. *Nature Machine Intelligence*, 4(3): 196–210.

Li, Y.; Tang, X.; Lu, J.; and Huang, Y. 2024. Dual graph-structured semantics multi-subspace learning for cross-modal retrieval. *Multimedia Systems*, 30(5): 1–14.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Marcus, G.; and Davis, E. 2019. *Rebooting AI: Building artificial intelligence we can trust*. Vintage.

Peng, Y.; Huang, X.; and Zhao, Y. 2017. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. *IEEE Transactions on circuits and systems for video technology*, 28(9): 2372–2385.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training.

Rasiwasia, N.; Costa Pereira, J.; Coviello, E.; Doyle, G.; Lanckriet, G. R.; Levy, R.; and Vasconcelos, N. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, 251–260.

Ritter, S.; Barrett, D. G.; Santoro, A.; and Botvinick, M. M. 2017. Cognitive psychology for deep neural networks: A shape bias case study. In *International conference on machine learning*, 2940–2949. PMLR.

Shen, Y.; Sun, X.; Wei, X.-S.; Hu, H.; and Chen, Z. 2022. A channel mix method for fine-grained cross-modal retrieval. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 01–06. IEEE.

Su, M.; Gu, G.; Ren, X.; Fu, H.; and Zhao, Y. 2021. Semi-supervised knowledge distillation for cross-modal hashing. *IEEE Transactions on Multimedia*, 25: 662–675.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Wang, C.; Yao, Y.; Wang, Q.; and Tang, Z. 2021a. Local self-attention on fine-grained cross-media retrieval. In *Proceedings of the 3rd ACM International Conference on Multimedia in Asia*, 1–7.

Wang, L.; Yang, J.; Zareapoor, M.; and Zheng, Z. 2021b. Cluster-wise unsupervised hashing for cross-modal similarity search. *Pattern Recognition*, 111: 107732.

Xia, X.; Dong, G.; Li, F.; Zhu, L.; and Ying, X. 2023. When CLIP meets cross-modal hashing retrieval: A new strong baseline. *Information Fusion*, 100: 101968.

Yan, L.; Greiff, S.; Teuber, Z.; and Gašević, D. 2024. Promises and challenges of generative artificial intelligence for human learning. *Nature Human Behaviour*, 8(10): 1839–1850.

Yang, Y.; Hu, W.; and Hu, H. 2025. Progressive Cross-modal Association Learning for Unsupervised Visible-Infrared Person Re-Identification. *IEEE Transactions on Information Forensics and Security*.

Yu, G.; Liu, X.; Wang, J.; Domeniconi, C.; and Zhang, X. 2020. Flexible cross-modal hashing. *IEEE Transactions on Neural Networks and Learning Systems*, 33(1): 304–314.

Zhang, B.; Zhang, Y.; Li, J.; Chen, J.; Akutsu, T.; Cheung, Y.-m.; and Cai, H. 2024. Unsupervised Dual Deep Hashing with Semantic-Index and Content-Code for Cross-Modal Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, J.; and Peng, Y. 2019. Multi-pathway generative adversarial hashing for unsupervised cross-modal retrieval. *IEEE Transactions on Multimedia*, 22(1): 174–187.

Zhang, P.-F.; Li, Y.; Huang, Z.; and Xu, X.-S. 2021. Aggregation-based graph convolutional hashing for unsupervised cross-modal retrieval. *IEEE Transactions on Multimedia*, 24: 466–479.

Zheng, H.; Wang, J.; Zhen, X.; Song, J.; Zheng, F.; Lu, K.; and Qi, G.-J. 2023. Continuous cross-modal hashing. *Pattern Recognition*, 142: 109662.

Zheng, J.; Cai, X.; Qiu, S.; and Ma, Q. 2025. Spurious Forgetting in Continual Learning of Language Models. In *The Thirteenth International Conference on Learning Representations*.

Zhou, D.-W.; Cai, Z.-W.; Ye, H.-J.; Zhang, L.; and Zhan, D.-C. 2025. Dual consolidation for pre-trained model-based domain-incremental learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 20547–20557.

Zhu, L.; Wu, X.; Li, J.; Zhang, Z.; Guan, W.; and Shen, H. T. 2022. Work together: Correlation-identity reconstruction hashing for unsupervised cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 35(9): 8838–8851.