Intra-class Distribution-guided Generative Hashing with Neighbor Refinement for Cross-modal Retrieval

Supplementary Material

In this supplementary material, we provide additional information concerning both theoretical and experimental aspects. Specifically, Appendix A supplements the theoretical basis for the distribution-aware adaptive generation approach. Appendix B provides additional experimental setups and results. More specifically, Appendix B.1 outlines the datasets used across all experiments, Appendix B.2 presents precision-recall curve comparisons, Appendix B.3 reports $P@H \leq 2$ results, Appendix B.4 supplements parameter analysis experiments, Appendix B.5 offers ablation experiment of the covariance and sampling distribution, Appendix B.6 demonstrates the effectiveness of IDGH under long-tail scenario, and Appendix B.7 validates the model's robustness to noise.

A. Theoretical Basis of Properties

A.1. Reasonableness of Distributions-aware Adaptive Generation

The design of distribution-aware adaptive generation stems from a significant recent discovery in research: deep features within deep neural networks often exhibit linearisation properties [1, 5, 14]. Specifically, numerous semantic directions exist within deep feature spaces. Translating sample features along these directions yields new features corresponding to another sample within the same category but representing a different semantic aspect.

For instance, a particular direction may correspond to the semantic dimension of 'facial expression variation': translating the features of an expressionless individual along this direction may yield new features representing the same person smiling. Thus, identifying multiple such semantic directions effectively expands the training set, providing representative and informative training samples for similarity learning.

A.2. Selection of Sampling Distribution

Explicitly identifying specific semantic directions typically demands extensive manual annotation [14, 15], whereas randomly sampled directions, though computationally efficient, may result in semantically meaningless transformations (e.g., applying a 'wearing spectacles' direction to the 'car' category is nonsensical). To address this issue, we propose an intra-class distribution estimation approach that captures class-specific distribution patterns.

Based on the estimated distribution, we then sample directions from a multivariate normal distribution with zero mean and covariance proportional to the intra-class covariance matrix. These sample directions are applied to the features of the corresponding category to generate semantically meaningful and diverse synthetic samples adaptively. This process effectively approximates meaningful semantic transformation directions within the deep feature space. For example, features in the 'animals' category may shift along a 'posture variation' direction while showing minimal variation along directions that are semantically irrelevant to the category (e.g., 'vehicle structure' or 'building layout'). As a result, the likelihood of producing meaningless samples is substantially reduced.

B. Additional Experiments

Without loss of generality, this paper employs four commonly used image-text datasets to evaluate cross-modal performance. Detailed information regarding these experimental datasets is provided below, with statistical results summarised in Table 1.

B.1. Datasets

MIRFlickr-25K: consists of 24,581 image-text pairs from the Flickr website [6]. The dataset involves 24 categories, and each one is annotated with at least one annotation from all classes.

NUS-WIDE: consists of 269,648 pairs of image-text instances annotated with 81 categories [3]. In the experiment, we removed several categories according to [10]. The processed dataset involves 195,834 sample pairs corresponding to 21 categories.

IAPR TC-12: consists of 20,000 natural images collected worldwide, including photographs of people, landscapes, etc., which contain 255 different categories in total [4]. Each image has the corresponding English caption.

XMediaNet: consists of five media types, including images, texts, videos, audios, and 3D models [12]. The dataset covers 200 semantic categories and contains 40,000 images, 40,000 text descriptions, 10,000 videos, 10,000 audio clips, and 2,000 3D models. Samples from different modalities correspond to 200 non-overlapping categories.

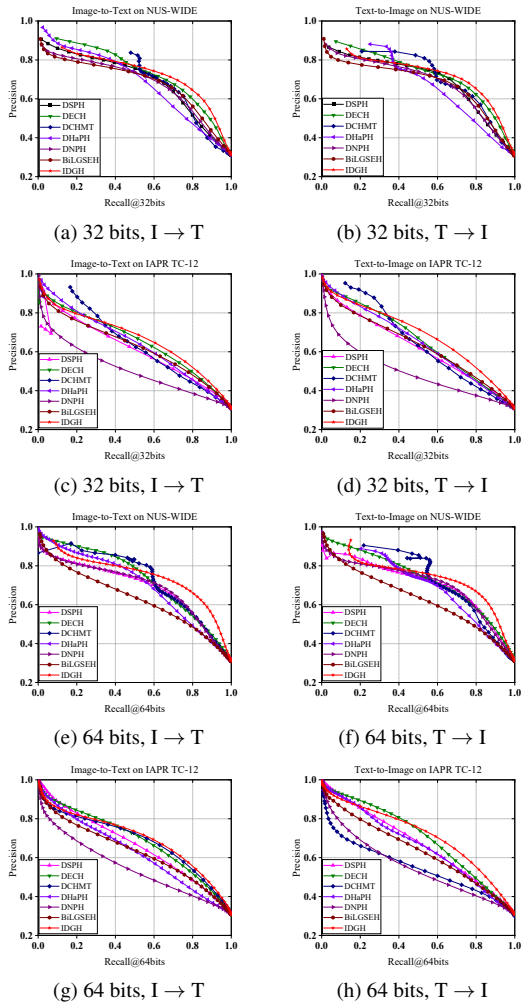
B.2. Precision-recall Curve Comparisons

To comprehensively compare the proposed IDGH method with other baseline approaches (i.e., DCHMT [13], DSPH [7], DNPH [8], DHaPH [9], BiLGSEH [16], and DECH [11]), precision-recall (P-R) curves were plotted, as shown in Figure 1. Precision and recall are two mutually constraining metrics; an improvement in one typically leads to a decrease

Table 1. Statistical information regarding the datasets employed in the experiment, where in '* / * / *' denote the respective counts of image-text pairs in the training, query, and retrieval sets. The symbol C represents the total number of categories, while $f^{\mathcal{I}}$ and $f^{\mathcal{T}}$ denote the dimensionality of the image and text features, respectively.

Dataset	train / query / retrieval	C	$f^{\mathcal{I}}$	$f^{\mathcal{T}}$
MIRFlickr-25K [6]	10,000 / 5,000 / 19,581	24	512	512
NUS-WIDE [3]	10,000 / 5,000 / 190,834	21	512	512
IAPR TC-12 [4]	10,000 / 5,000 / 14,626	291	512	512
XMediaNet [12]	10,000 / 5,000 / 35,000	200	4,096	300

in the other. Within the P-R curve, methods positioned higher generally exhibit superior performance. Results demonstrate that the IDGH method outperforms other baseline methods in both image-to-text (I $\rightarrow$ T) and text-to-image (T $\rightarrow$ I) retrieval tasks. This further validates the effectiveness of IDGH in cross-modal hashing retrieval tasks.



(a) 32 bits, I $\rightarrow$ T (b) 32 bits, T $\rightarrow$ I

(c) 32 bits, I $\rightarrow$ T (d) 32 bits, T $\rightarrow$ I

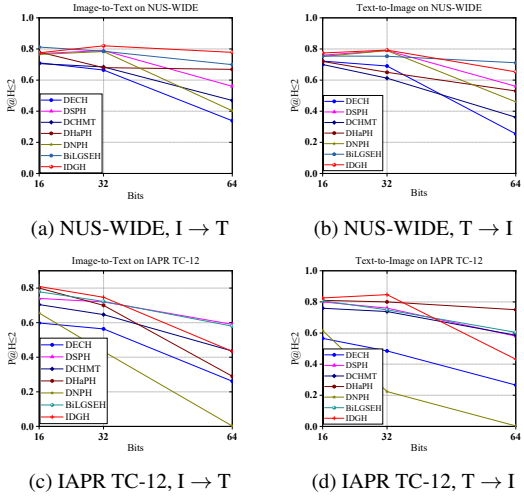
(e) 64 bits, I $\rightarrow$ T (f) 64 bits, T $\rightarrow$ I

(g) 64 bits, I $\rightarrow$ T (h) 64 bits, T $\rightarrow$ I

Figure 1. The Precision-recall curves with 32 and 64 bits on NUS-WIDE and IAPR TC-12. The red curve is the result of our method.

B.3. $P@H \leq 2$ Results

To intuitively assess the model's retrieval quality in Hamming space, we also compute the precision under a Hamming radius ≤ 2 ($H \leq 2$), as shown in Figure 2. This metric measures the discriminative power of hash codes within locally compact neighbourhoods: higher $P@H \leq 2$ indicates that retrieved neighbors are more semantically consistent with the query. The results demonstrate that our method effectively separates negative samples, thereby learning a discriminative embedding space.



(a) NUS-WIDE, I $\rightarrow$ T (b) NUS-WIDE, T $\rightarrow$ I

(c) IAPR TC-12, I $\rightarrow$ T (d) IAPR TC-12, T $\rightarrow$ I

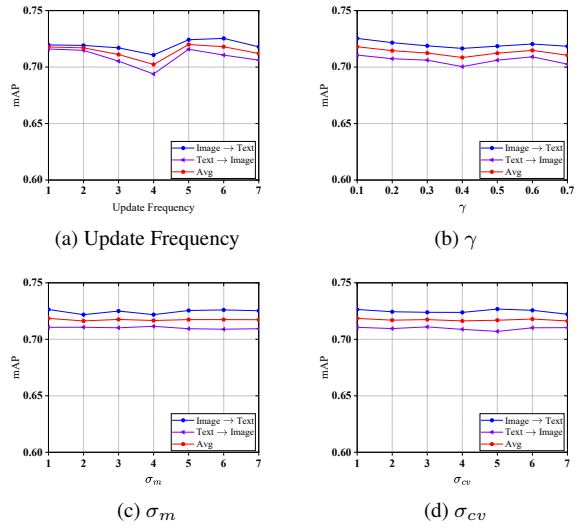
Figure 2. $P@H \leq 2$ results with different code lengths on NUS-WIDE and IAPR TC-12. The red curve is the result of our method.

B.4. Additional Parameter Analysis

Update Frequency: Frequent updates to the estimated covariance matrix ensure the distribution remains up-to-date during training, though this incurs additional computational overhead. As shown in Figure 3 (a), performance exhibits no significant degradation within an appropriate range of iteration intervals. We set the interval to 5 to strike a balance between performance and computational cost.

Refinement Weights: As shown in Figure 3 (b), γ controls the strength of neighborhood refinement weights, where a smaller γ yields stronger refinement effects. Experiments

demonstrate that sufficient neighborhood information is crucial for correcting unreliable estimates; thus, we set $\gamma = 0.1$. The parameters σ_m and σ_{cv} also influence the refinement process. Specifically, σ_m regulates the sensitivity to differences in mean distance, while σ_{cv} controls the sensitivity to covariance distance. As illustrated in Figure 3 (c) and (d), the model achieves optimal performance when $\sigma_m = \sigma_{cv} = 1$.



(a) Update Frequency (b) γ (c) σ_m (d) σ_{cv}

Figure 3. Additional Parameter analysis with 128 bits on IAPR TC-12.

B.5. Additional Ablation Study

Covariance Matrix: To further verify the effectiveness of the estimated covariance matrix, we evaluated the performance of the model under different configurations, as reported in Table 2. Using only the diagonal elements of the covariance matrix yields competitive results. Moreover, removing the class-level covariance matrix leads to unreliable estimations and degraded performance, whereas leveraging global information effectively alleviates overfitting. In conclusion, to balance efficiency and effectiveness, we adopt the diagonal covariance matrix in our implementation.

Sampling Distribution: To validate the effectiveness of IDGH's generation strategy, we compared two sampling schemes for generating synthetic samples. The IDGH synthetic features are obtained by sampling the dynamic distribution $\mathcal{N}(\widetilde{b}_i, \eta\boldsymbol{\Sigma}_{l_i})$, where each original sample serves as the mean. In contrast, w/fix involves sampling from a fixed distribution, $\mathcal{N}(\hat{\boldsymbol{\mu}}_{l_i}, \eta\boldsymbol{\Sigma}_{l_i})$, where the category mean is used as the center. As shown in Figure 4, direct sampling of the fixed distribution leads to a significant performance drop. This is because the generated synthetic samples lack a one-to-one correspondence with the original samples and often produce noisy samples that interfere with training. By contrast, IDGH generates semantically consistent samples from intra-class diversity, thereby facilitating similarity learning.

Table 2. Ablation experiment of the covariance matrix on MIRFLICKR-25K.

Task	Method	MIRFLICKR-25K			
		16bits	32bits	64bits	128bits
I → T	w/o Σ_c	0.8289	0.8520	0.8601	0.8649
	w/o Σ_{global}	0.8341	0.8506	0.8609	0.8647
	IDGH	**0.8392**	**0.8526**	**0.8621**	**0.8687**
T → I	w/o Σ_c	0.7949	0.8209	0.8284	0.8287
	w/o Σ_{global}	0.7975	0.8104	0.8264	0.8283
	IDGH	**0.8151**	**0.8255**	**0.8379**	**0.8411**

(a) I → T (b) T → I

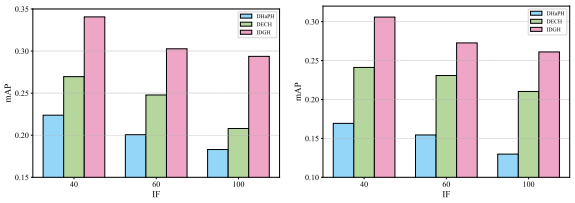
Figure 4. Ablation experiment of the sampling distribution with 128 bits on MIRFLICKR-25K.

B.6. Effectiveness under Long-tail Scenario

To further evaluate the effectiveness of the sampling distribution selection of the IDGH approach, we conducted experiments using a long-tail dataset with an extremely unbalanced training sample distribution. Following prior work [2], we first set the number of head classes to d_h, then adjusted the number of tail classes to d_t based on the IF value. Finally, we employed an exponential decay formula to predetermine the sample size for intermediate classes, ensuring a monotonically decreasing sample count from head to tail while satisfying IF = d_h / d_t. As shown in Figure 5, the performance of the Baseline methods drops sharply on long-tail datasets. This is because they rely on semantic labels to guide hash code learning, but tail labels lack sufficient training samples, thus weakening the discriminative capacity of the model. In contrast, our IDGH framework adaptively generates representative and informative synthetic samples directly from the estimated intra-class distributions, enabling the model to learn more discriminative hash codes even under severe data imbalance.

B.7. Noise Robustness

To validate the robustness of IDGH under noisy and anomalous conditions, we simulate varying degrees of label noise by applying 20%, 50%, and 80% random perturbations to



(a) I → T (b) T → I

Figure 5. Experiment results of long-tail scenario with 128 bits on XMediaNet.

the training labels. As shown in Table 3, IDGH achieves optimal retrieval performance at both noise-free and various noise levels, demonstrating that the method can generate reliable synthetic samples even when labels are unreliable, thus providing informative training signals for similarity learning.

Table 3. mAP results of the IDGH and baselines under different noise rates with 128 bit on MIRFLICKR-25K.

Task	Method	MIRFLICKR-25K			
		0%	20%	50%	80%
I → T	DNPH	0.8370	0.8353	0.8332	0.8108
	DHaPH	0.8549	0.8264	0.7977	0.7728
	BiLGSEH	0.8207	0.8000	0.7932	0.7746
	IDGH	**0.8687**	**0.8585**	**0.8456**	**0.8286**
T → I	DNPH	0.8232	0.8126	0.8076	0.7945
	DHaPH	0.8279	0.7931	0.7699	0.7438
	BiLGSEH	0.8347	0.7955	0.7736	0.7672
	IDGH	**0.8411**	**0.8338**	**0.8206**	**0.8082**

References

[1] Yoshua Bengio, Grégoire Mesnil, Yann Dauphin, and Salah Rifai. Better mixing via deep representations. In *International conference on machine learning*, pages 552–560. PMLR, 2013. 1

[2] Yong Chen, Yuqing Hou, Shu Leng, Qing Zhang, Zhouchen Lin, and Dell Zhang. Long-tail hashing. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1328–1338, 2021. 3

[3] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9, 2009. 1, 2

[4] Hugo Jair Escalante, Carlos A Hernández, Jesus A Gonzalez, Aurelio López-López, Manuel Montes, Eduardo F Morales, L Enrique Sucar, Luis Villasenor, and Michael Grubinger. The segmented and annotated iapr tc-12 benchmark. *Computer vision and image understanding*, 114(4):419–428, 2010. 1, 2

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[6] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 39–43, 2008. 1, 2

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

[8] Yadong Huo, Qin Qibing, Jiangyan Dai, Wenfeng Zhang, Lei Huang, and Chengduan Wang. Deep neighborhood-aware proxy hashing with uniform distribution constraint for cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(6):1–23, 2024. 1

[9] Yadong Huo, Qibing Qin, Wenfeng Zhang, Lei Huang, and Jie Nie. Deep hierarchy-aware proxy hashing with self-paced learning for cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 2024. 1

[10] Qing-Yuan Jiang and Wu-Jun Li. Deep cross-modal hashing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3232–3240, 2017. 1

[11] 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*, pages 18566–18574, 2025. 1

[12] Yuxin Peng, Jinwei Qi, and Yuxin Yuan. Modality-specific cross-modal similarity measurement with recurrent attention network. *IEEE Transactions on Image Processing*, 27(11): 5585–5599, 2018. 1, 2

[13] Junfeng Tu, Xueliang Liu, Zongxiang Lin, Richang Hong, and Meng Wang. Differentiable cross-modal hashing via multimodal transformers. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 453–461, 2022. 1

[14] Paul Upchurch, Jacob Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, and Kilian Weinberger. Deep feature interpolation for image content changes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7064–7073, 2017. 1

[15] Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Gao Huang, and Cheng Wu. Implicit semantic data augmentation for deep networks. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[16] Lei Zhu, Runbing Wu, Xinghui Zhu, Chengyuan Zhang, Lin Wu, Shichao Zhang, and Xuelong Li. Bi-direction label-guided semantic enhancement for cross-modal hashing. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 1