Absorbing Gradient Conflicts: Modeling Semantic Variance via Kent Distributions for Cross-Modal Hashing

Hengjie Zhu[1,2], **Dayan Wu**[1*], **Zihao Zhang**[1,2], **Xinze Liu**[1,2], **Jingxuan Yu**[3], **Peng Fu**[1], **Zheng Lin**[1], **Weiping Wang**[1]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
[3]School of Cyber Science and Engineering, Southeast University
{zhuhengjie, wudayan, zhangzihao, liuxinze, fupeng, linzheng, wangweiping}@iie.ac.cn,
yujingxuan24@seu.edu.cn

Abstract

Supervised proxy-based deep cross-modal hashing has become the dominant paradigm for large-scale retrieval. However, prevalent methods model class proxies as deterministic points in the embedding space. This rigid assumption causes severe gradient conflicts in multi-label scenarios, where gradient conflicts arising from label co-occurrence lead to severe gradient contention and optimization collapse. To resolve this, we propose Kent-based Distributional Proxy Hashing (KDPH), a novel framework that shifts proxy representation from static points to flexible anisotropic Kent distributions on the hypersphere. Unlike point proxies that must shift their positions to accommodate conflicting gradients, KDPH absorbs these conflicts by dynamically adjusting its directional variance. This allows the proxy to maintain a stable semantic mean direction while stretching to cover diverse label correlations. Furthermore, to ensure stable training of these geometric parameters, we derive a tailored loss function incorporating the Cayley transform to enforce strict orthogonality. To the best of our knowledge, KDPH is the first framework to successfully introduce the Kent distributions into cross-modal hashing. Experiments on three benchmark datasets demonstrate that KDPH mitigates proxy collapse and chaotic oscillation, significantly outperforms state-of-the-art methods. Code is available at https://github.com/Senmo996/KDPH-official-code.

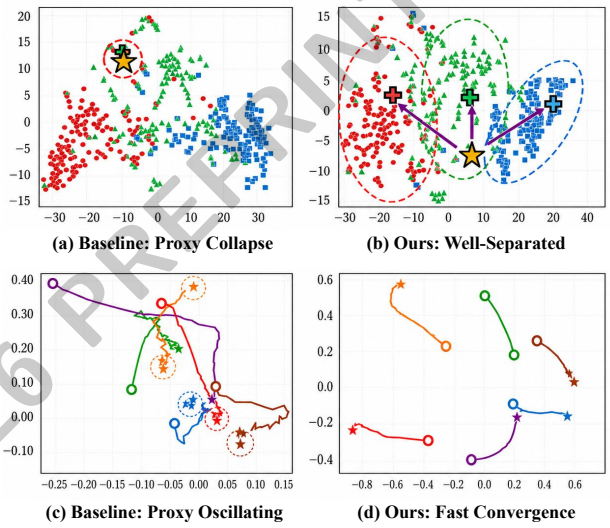


Figure 1: **Comparison of proxy optimization behaviors. (a)-(b)** Conventional proxy learning suffers from *proxy collapse*, where learned proxies move toward the anchor and become poorly separated, whereas KDPH preserves distinct proxy distributions. **(c)-(d)** Conventional methods exhibit slow and oscillatory optimization trajectories, while KDPH achieves smoother trajectories and faster convergence. Colors denote different categories.

1 Introduction

Driven by the surge in multimodal data [Xu *et al.*, 2024], Cross-Modal Hashing Retrieval has emerged as a standard method for large-scale information retrieval [Cao *et al.*, 2022; Singh and Gupta, 2022; Yan *et al.*, 2025; **?**]. By projecting high-dimensional features into a compact Hamming space, this approach circumvents the prohibitive computational and storage costs of conventional Euclidean methods, enabling efficient similarity approximation via bitwise XOR operations [Wang *et al.*, 2022; Lu *et al.*, 2019; Sun *et al.*, 2022]. Despite rapid progress towards Supervised Deep Cross-Modal Hashing Retrieval—leveraging deep neural networks and semantic supervision for end-to-end representation learning—fundamental challenges still persist. Real-world data are inherently multi-faceted, with single instances often associated with multiple overlapping labels. This multi-label characteristic poses severe difficulties for existing supervised frameworks, which struggle to capture complex, non-exclusive semantic correlations, necessitating more robust modeling strategies.

Despite the dominance of proxy-based strategies in deep cross-modal hashing [Tu *et al.*, 2023; Huo *et al.*, 2024b], fundamental limitations persist in multi-label scenarios. The prevailing paradigm typically abstracts each semantic category as a singular, deterministic point [Teh *et al.*, 2020;

*Corresponding Author

Kim *et al.*, 2020; Movshovitz-Attias *et al.*, 2017; Qian *et al.*, 2019] within the embedding space. While effective for single-label tasks, this rigid zero-variance assumption provokes severe gradient contention when confronted with the complex semantic conflicts inherent in label co-occurrence. Specifically, in a point-proxy framework, a multi-label sample simultaneously exerts pulling forces on divergent proxies. Modeled as rigid points, these proxies lack the spatial capacity to accommodate such conflicting semantic signals. Consequently, this unresolved contention drives the model towards optimization collapse. These two limitations manifesting in two degenerate states: (1) *Chaotic Proxy Oscillation*, where the proxy, unable to satisfy conflicting gradients, is forced to constantly shift its position, preventing the model from anchoring to a stable semantic center; and (2) *Proxy Collapse*, where the optimization process forcibly pulls proxies of distinct categories together to minimize aggregate loss—particularly causing low-frequency tail classes to collapse onto high-frequency head classes—thereby eliminating fine-grained decision boundaries and impairing discriminative power.

To address these challenges, we propose the Kent-based Distributed Proxy Hashing (KDPH) framework. We depart from the single-point paradigm by modeling each class proxy as a learnable Kent distribution—an anisotropic spherical probability model defined by a mean direction and an elliptical covariance structure. Unlike rigid point proxies that must drastically shift position to satisfy conflicting gradients, this distributed representation resolves conflicts by changing specific directional variances, thereby endowing each class with a semantic shape and volume to flexibly absorb the contextual variance introduced by label co-occurrence. Specifically, observing the anisotropic geometry of the Kent distribution, we design a tailored loss function to achieve gradient decoupling mechanism by stretching shape rather than moving centroid. This approach effectively resolves conflict, preventing chaotic oscillation and proxy collapse. Furthermore, by incorporating the Cayley transform to enforce strict orthogonality for the principal axes and utilizing this probabilistic module solely to guide the learning process, KDPH achieves state-of-the-art performance with zero additional computational overhead during inference, successfully resolving the persistent trade-off between retrieval accuracy and efficiency.

In summary, our contributions are as follows:

- We introduce KDPH, a novel framework that reformulates proxy-based hashing by substituting rigid point proxies with anisotropic Kent distributions. To the best of our knowledge, KDPH is the first work to demonstrate that modeling proxy uncertainty through directional distributions can eliminate the optimization collapse caused by gradient contention in multi-label cross-modal retrieval.
- We derive a manifold-constrained loss function leveraging the Cayley transform, which enables proxies to adaptively absorb gradient conflicts via directional variance rather than centroid displacement, thereby effectively eradicating **Proxy Collapse** and suppressing **Chaotic Proxy Oscillation** during the learning process.
- Extensive experiments demonstrate that KDPH achieves **state-of-the-art** performance on three benchmarks, empirically validating the efficacy of our gradient decoupling strategy and showing superior robustness particularly in complex multi-label scenarios.

2 Related Work

2.1 Supervised Deep Cross-Modal Hashing and Proxy Learning

Supervised deep cross-modal hashing has advanced significantly, positioned such as CNN-RNN and adversarial baselines [Jiang and Li, 2017; Li *et al.*, 2018; Bai *et al.*, 2020]. Recently, graph and transformer-based models [Xu *et al.*, 2019; Tu *et al.*, 2022; Liu *et al.*, 2023; Chen *et al.*, 2024; Liu *et al.*, 2024] have further pushed the boundaries of performance. In parallel, these models improving their algorithmic design from simple pairwise constraints [Qin *et al.*, 2024] to adaptive gradient-triplet losses [Zhu *et al.*, 2025a] or sophisticated geometric [Qin *et al.*, 2025b] strategies which ranging from spherical mutual information to boundary-sensitive mining [Qin *et al.*, 2025a]—to refine the Hamming space structure. To address the scalability bottlenecks of pairwise comparisons, proxy-based metric learning has been effectively incorporated. By representing categories as distinct point proxy, this paradigm allows the model to efficiently organize the embedding space [Tu *et al.*, 2023; Huo *et al.*, 2024b], and has recently explored hyperbolic geometries for hierarchical modeling [Huo *et al.*, 2024c]. However, a major challenge persists that the deterministic point formulation's inherent zero-variance assumption fundamentally limits its applicability to multi-label scenarios with high contextual variance.

2.2 Probabilistic Embeddings and Non-Isotropy

Probabilistic embedding learning has emerged as a robust alternative to deterministic point estimates [Lin *et al.*, 2018]. To respect the unique geometry of hyperspherical feature spaces, the field has evolved from early Euclidean Gaussian models [Shi and Jain, 2019] to spherical probability density functions, such as the von Mises-Fisher (vMF) distribution [Zhe *et al.*, 2018; Park *et al.*, 2019; Li *et al.*, 2021]. However, the efficacy of vMF-based approaches remains constrained by their inherent isotropy. One major challenge is the implicit assumption of uniform variance in all directions and this rigid simplification fails to accommodate the complex, anisotropic semantic manifolds inherent in multi-label data. Although recent metric learning approaches have explored non-isotropy through lightweight mechanisms, such as variance scaling or geometric regularization [Roth *et al.*, 2022; Kirchhof *et al.*, 2022], these low-parameter heuristics often lack the sufficient geometric flexibility required for large-scale multi-label cross-modal hashing, where semantic manifolds are densely entangled. To bridge this gap, we propose the KDPH, which introduces the parameter-rich Kent distribution to deliberately increase geometric capacity. We further facilitate stable training by resolving gradient contention and optimization collapse via the Cayley transform and a tailored loss function.

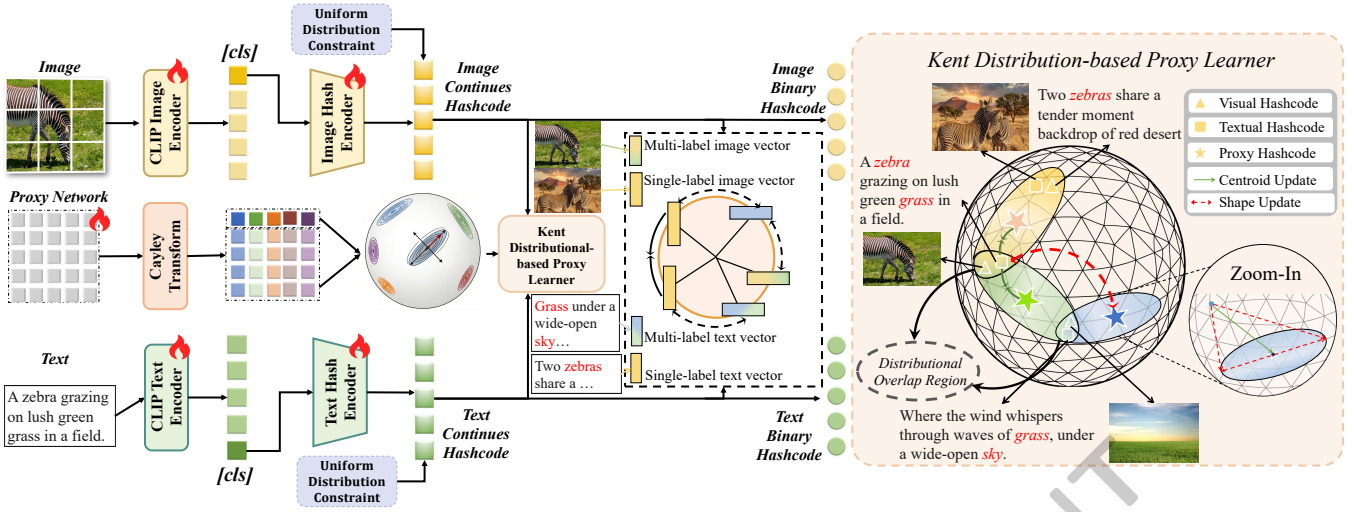


Figure 2: **Overview of the Kent-based Distributed Proxy Hashing (KDPH) framework.** It consists of a CLIP-based feature extraction module and a Kent-based Distributional Proxy Learner. As shown in the right panel, we model class proxies as learnable anisotropic distributions rather than static points. Through the gradient decoupling mechanism by separating centroid μ and variance B, Γ, the model dynamically stretches the proxy's semantic volume to accommodate multi-label variance, effectively mitigating proxy collapse and oscillation.

3 Method

3.1 Formulation and Architecture

Given a training dataset $O = \{(\boldsymbol{i}_n, \boldsymbol{t}_n, \boldsymbol{l}_n)\}_{n=1}^{N}$ with images $\boldsymbol{i}_n$, texts $\boldsymbol{t}_n$, and multi-label vectors $\boldsymbol{l}_n \in \{0,1\}^C$, we extract D-dimensional semantic features $\boldsymbol{f}_n^i$ and $\boldsymbol{f}_n^t$ via a dual-stream Transformer backbone initialized with pre-trained CLIP. Modality-specific hashing networks $H_I, H_T : \mathbb{R}^D \to \mathbb{R}^K$ then project these features into K-dimensional continuous representations $\boldsymbol{h}_n^i$ and $\boldsymbol{h}_n^t$. The final binary hash codes are obtained via the sign function:

$$\boldsymbol{b}_n^i := \text{sign}(\boldsymbol{h}_n^i) \in \{-1,1\}^K, \quad \boldsymbol{b}_n^t := \text{sign}(\boldsymbol{h}_n^t) \in \{-1,1\}^K \quad (1)$$

3.2 Kent-based Distributed Proxy Mechanism

Conventional hashing methods typically abstract each class as deterministic point whose rigid zero-variance assumption fundamentally unsuitable for multi-label scenarios, and conflicting semantic signals from co-occurring labels lead to gradient contention and optimization collapse. To resolve this, KDPH shifts the paradigm from static points to distributional proxies on the hypersphere [Wang and Isola, 2020]. By modeling each class as a flexible Kent distribution with learnable centroid and variance, our framework effectively absorbs the gradients variance arising from label co-occurrence.

Kent Distribution for Proxy Modeling

To instantiate this distributional paradigm, we employ the generalized Kent distribution. This strategic choice addresses the inherent limitations of simpler spherical models, such as the von Mises-Fisher (vMF) distribution. While vMF offers mathematical convenience, its isotropic nature rigidly imposes uniform variance across all dimensions, rendering it ill-suited for capturing the complex, directional manifolds typical of multi-label features. In contrast, the Kent distribution is intrinsically anisotropic. It provides the necessary geometric flexibility to model elliptical contours on the hypersphere $\mathbb{S}^{K-1}$, where K denotes the hash code dimension, distinct from the feature dimension D.

Formally, the probability density function (PDF) for a class c is defined as:

$$p(\boldsymbol{z}|\boldsymbol{\Theta}_c) = \frac{1}{c(\kappa, B)} \exp\left(\kappa_c \boldsymbol{\mu}_c^T \boldsymbol{z} + \sum_{i=1}^{N_{axes}} \beta_{c,i} (\boldsymbol{\gamma}_{c,i}^T \boldsymbol{z})^2\right) \quad (2)$$

where $c(\kappa, B)$ denotes the normalizing constant. The distribution is parameterized by $\boldsymbol{\Theta}_c = \{\boldsymbol{\mu}_c, \kappa_c, \Gamma_c, B_c\}$, in which each component serves a distinct semantic role: the mean direction $\boldsymbol{\mu}_c \in \mathbb{S}^{K-1}$ acts as the semantic centroid representing the stable core identity of the class, while κ_c governs the global concentration of samples around this centroid. Crucially, to model non-uniform variance, Γ_c defines the orthogonal principal axes, and $B_c = \{\beta_{c,i}\}$ controls the anisotropic shaping along these dimensions. By learning distinct $\beta_{c,i}$ values, the proxy can dynamically stretch or compress along specific semantic directions. This mechanism allows the distribution to accommodate the high semantic variance introduced by multi-label co-occurrences without drastically shifting the centroid $\boldsymbol{\mu}_c$.

3.3 Hash learning

To jointly optimize the deep networks and proxy parameters $\boldsymbol{\Theta}_c$ while respecting manifold geometry, we introduce a differentiable reparameterization technique. This mechanism is integrated with a Kent-based log-likelihood scoring function and a multi-task objective to strictly enforce both semantic discrimination and quantization constraints.

Orthogonal Frame Construction via Cayley Transform

Directly optimizing $\boldsymbol{\mu}_c$ and $\boldsymbol{\Gamma}_c$ is challenging due to the strict orthogonality required by the Kent distribution. To address this, we construct a unified orthonormal basis using the Cayley

Transform. Specifically, for each class c, we learn a skew-symmetric parameter matrix $\mathbf{A}_c \in \mathbb{R}^{K\times K}$. By exploiting the mapping between the Lie algebra $\mathfrak{so}(K)$ and the Lie group $SO(K)$, we generate a strict rotation matrix $\mathbf{Q}_c$:

$$\mathbf{Q}_c = (\mathbf{I} + \mathbf{A}_c)^{-1}(\mathbf{I} - \mathbf{A}_c) \in SO(K) \tag{3}$$

This transformation ensures that $\mathbf{Q}_c$ remains orthogonal regardless of the updates to $\mathbf{A}_c$. We then define the proxy components by slicing $\mathbf{Q}_c$: the first column serves as the centroid $\boldsymbol{\mu}_c$, while the following columns form the shape axes $\boldsymbol{\Gamma}_c$. This parameterization naturally enforces geometric constraints, guaranteeing orthogonality not only between the centroid and the axes but also pairwise among all principal axes within $\boldsymbol{\Gamma}_c$, effectively bypassing the need for expensive orthogonalization procedures.

Anisotropic Energy Scoring and Parameterization

To evaluate the semantic affinity between a query sample and the class proxies, we first project the continuous hash code $\boldsymbol{h}$ onto the unit hypersphere, $\boldsymbol{z} = \boldsymbol{h}/\|\boldsymbol{h}\|_2 \in \mathbb{S}^{K-1}$, ensuring geometric consistency with the directional statistics.

While the canonical Kent distribution defines a rigorous probability density function, its optimization is hindered by a computationally intractable normalization constant dependent on high-dimensional Bessel functions. Since our primary objective in hashing is discriminative ranking rather than exact density estimation, we circumvent this bottleneck by reinterpreting the log-density formulation as a geometric Compatibility Score. We define the anisotropic energy score $S(\boldsymbol{z}, c)$ for a class c as follows:

$$S(\boldsymbol{z}, c) = \kappa_c(\boldsymbol{\mu}_c^T \boldsymbol{z}) + \sum_{i=1}^{N_{\text{axes}}} \beta_{c,i}(\boldsymbol{\gamma}_{c,i}^T \boldsymbol{z})^2 - \mathcal{R}(\kappa, \boldsymbol{\beta}) \tag{4}$$

Here, we employ a logarithmic regularization term $\mathcal{R}(\kappa, \boldsymbol{\beta}) = \log(\kappa_c + \|\boldsymbol{\beta}_c\|_1)$. This term serves as a soft magnitude constraint, effectively transforming the probabilistic objective into a flexible quadratic energy metric on the hypersphere, preserving the topological benefits of Kent geometry while avoiding the gradient instability of Bessel approximations.

To strictly enforce the definition of concentration and magnitude, the parameters κ_c and $\beta_{c,i}$ are constrained to be positive via a Softplus activation on learnable scalars $\hat{\kappa}_c$ and $\hat{\beta}_{c,i}$. Crucially, regarding the shape constraints, while the canonical Kent distribution requires $2|\beta_{c,i}| < \kappa_c$ for uni-modality, we adopt a generalized Fisher-Bingham formulation by removing this upper bound. By allowing the network to learn $\boldsymbol{\beta}_c$ freely, the proxy gains the flexibility to model complex, potentially multi-modal distributions. This relaxation is particularly advantageous for capturing the high variance and distinct semantic manifolds inherent in multi-label data.

Distribution-Aware Proxy Triplet Loss

To strictly enforce discriminative margins within the hyperspherical manifold, we employ a structure-preserving Distribution-Aware Proxy Triplet Objective. Unlike standard cross-entropy which treats classes as independent points, this loss explicitly optimizes the anisotropic energy alignment between samples and learnable Kent distributions. By maximizing the energy score for relevant classes while minimizing it for irrelevant ones, the objective ensures that an instance resides firmly within the high-density acceptance region of its ground-truth semantic distributions, effectively pushing decision boundaries away from the ambiguous overlapping zones.

Formally, given a normalized query sample $\boldsymbol{z}$, a positive proxy distribution $c_p \in \mathcal{P}$ (where $\mathcal{P}$ is the set of ground-truth labels), and a negative proxy $c_n \in \mathcal{N}$, we enforce a strict separation constraint based on the energy score $S(\cdot)$:

$$S(\boldsymbol{z}, c_p) > S(\boldsymbol{z}, c_n) + m \tag{5}$$

where m is a fixed margin hyperparameter governing the minimum semantic gap.

To accommodate the multi-label setting where a sample is associated with multiple semantic centroids, the final loss aggregates these constraints over all valid triplets. This formulation allows the gradients to dynamically adjust the shape parameters $\boldsymbol{\beta}_c$ of the Kent distributions, absorbing the conflict from hard negatives:

$$\mathcal{L}_{Proxy} = \frac{1}{|\mathcal{P}||\mathcal{N}|} \sum_{y\in\mathcal{P}} \sum_{j\in\mathcal{N}} \max\big(0, S(\boldsymbol{z}, c_n) - S(\boldsymbol{z}, c_p) + m\big) \tag{6}$$

Multi-Modal Irrelevance Regularization

To strictly separate semantically disjoint samples, we incorporate the Multi-Modal Irrelevance Loss [Huo *et al.*, 2024b]. This objective penalizes high similarities between irrelevant pairs—defined as instances sharing no common labels (i.e., $l_i \cdot l_j = 0$). Taking the image intra-modal regularization as an example, the loss is defined as the average cosine similarity over all such pairs in a batch:

$$\mathcal{L}_{reg\_i} = \frac{\sum_{i=1}^{B}\sum_{j=1}^{B} \mathbb{I}(l_i \cdot l_j = 0)\cos(\boldsymbol{h}_i^x, \boldsymbol{h}_j^x)}{\sum_{i=1}^{B}\sum_{j=1}^{B} \mathbb{I}(l_i \cdot l_j = 0)} \tag{7}$$

where $\mathbb{I}(\cdot)$ is the indicator function and B is the batch size. The text intra-modal loss $\mathcal{L}_{reg\_t}$ and the cross-modal loss $\mathcal{L}_{reg\_it}$ are computed analogously by substituting the continues hash feature pairs with $(\boldsymbol{h}^y, \boldsymbol{h}^y)$ and $(\boldsymbol{h}^x, \boldsymbol{h}^y)$, respectively. The final regularization term aggregates these three components: $\mathcal{L}_{Reg} = \mathcal{L}_{reg\_i} + \mathcal{L}_{reg\_t} + \mathcal{L}_{reg\_it}$.

Cross-Modal Contrastive Alignment

To bridge the heterogeneity between modalities, we employ the symmetric InfoNCE [He *et al.*, 2020] objective to project representations into a shared, modality-invariant hypersphere. This contrastive loss maximizes the mutual information between matched image-text pairs. The image-to-text alignment loss is formulated as:

$$\mathcal{L}_{i\to t} = -\frac{1}{B}\sum_{k=1}^{B} \log \frac{\exp(\cos(\boldsymbol{h}_k^x, \boldsymbol{h}_k^y)/\tau)}{\sum_{j=1}^{B}\exp(\cos(\boldsymbol{h}_k^x, \boldsymbol{h}_j^y)/\tau)} \tag{8}$$

where τ is a learnable temperature parameter scaling the distribution sharpness. The text-to-image loss $\mathcal{L}_{t\to i}$ is computed symmetrically by swapping the anchor and positive samples. The final cross-modal alignment objective is the sum of both directions: $\mathcal{L}_{CM} = \mathcal{L}_{i\to t} + \mathcal{L}_{t\to i}$.

Task	Method	Reference	MIRFLICKR-25K			NUS-WIDE			MS COCO		
			16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits
I2T	DCPH	TKDE'22	0.7679	0.7733	0.7764	0.6161	0.6156	0.6180	0.5206	0.5806	0.5939
	DCMHT	MM'22	0.8263	0.8272	0.8441	0.6799	0.6992	0.7038	0.6447	0.6757	0.6915
	nivMF	ECCV'22	0.8241	0.8245	0.8352	0.7023	0.6902	0.7024	0.7075	0.7242	0.7523
	MIAN	TKDE'22	0.8123	0.8220	0.8355	0.6130	0.5894	0.6057	0.5350	0.5579	0.5404
	DSPH	TCSVT'23	0.8129	0.8482	0.8541	0.6830	0.6979	0.7162	0.7044	0.7510	0.7694
	DNPH	MC'24	0.8108	0.8269	0.8229	0.6689	0.6811	0.6939	0.6438	0.6910	0.7294
	DNpH	TMM'24	0.8423	0.8552	0.8558	0.6921	0.7022	0.7071	0.6727	0.6903	0.6860
	RDPH	TMM'24	0.7420	0.7749	0.7867	0.6330	0.6317	0.6523	0.5867	0.7151	0.7368
	DDBH	TCSVT'24	0.8450	0.8534	0.8610	0.7041	0.7145	0.7229	0.7165	0.7454	0.7681
	DNcH	ESWA'25	–	–	–	0.7150	0.7305	0.7387	0.7199	0.7341	0.7509
	DAGtH	ESWA'25	0.8268	0.8514	0.8599	0.6827	0.6882	0.6927	–	–	–
	KDPH	Ours	**0.8770**	**0.8639**	**0.8639**	**0.7430**	**0.7365**	**0.7392**	**0.7636**	**0.7761**	**0.7843**
T2I	DCPH	TKDE'22	0.7992	0.8046	0.8061	0.6187	0.6292	0.6394	0.5981	0.5879	0.5745
	DCMHT	MM'22	0.8116	0.8137	0.8281	0.6876	0.7104	0.7253	0.6513	0.6832	0.7052
	nivMF	ECCV'22	0.8125	0.8102	0.8213	0.6912	0.6952	0.6924	0.6421	0.6892	0.7042
	MIAN	TKDE'22	0.8058	0.8151	0.8182	0.6365	0.6312	0.6416	0.5489	0.5493	0.5365
	DSPH	TCSVT'23	0.8000	0.8238	0.8294	0.6997	0.7153	0.7304	0.7040	0.7556	0.7713
	DNPH	MC'24	0.8015	0.8176	0.8166	0.6871	0.6994	0.7182	0.6468	0.7012	0.7388
	DNpH	TMM'24	0.8147	0.8292	0.8361	0.6992	0.7137	0.7139	0.6562	0.6860	0.6928
	RDPH	TMM'24	0.7200	0.7504	0.7713	0.6732	0.7067	0.7190	0.5925	0.6910	0.7139
	DDBH	TCSVT'24	0.8245	0.8318	0.8390	0.7194	0.7211	0.7325	0.7167	0.7394	0.7595
	DNcH	ESWA'25	–	–	–	0.7196	0.7343	0.7430	0.7130	0.7306	0.7398
	DAGtH	ESWA'25	0.8072	0.8270	0.8382	0.6936	0.7042	0.7098	–	–	–
	KDPH	Ours	**0.8351**	**0.8439**	**0.8480**	**0.7473**	**0.7399**	**0.7458**	**0.7627**	**0.7785**	**0.7795**

Table 1: mAP comparisons on three benchmark datasets with references. The best results are highlighted in bold.

Uniform Distribution Constraint for Hashing

To maximize code entropy and prevent bit-collapse, we enforce a Uniform Distribution Constraint via Optimal Transport. Specifically, we minimize the Wasserstein-1 distance between the learned batch codes $\boldsymbol{B}^x$ and a sampled ideal uniform distribution $\boldsymbol{A} \in \{-1, 1\}^{B\times K}$. This is formulated as finding the optimal permutation matrix $\boldsymbol{O}^x$ that minimizes the transport cost:

$$\mathcal{L}_{U,x} = \min_{\boldsymbol{O}^x \in \mathcal{O}} -\text{tr}(\boldsymbol{O}^x \boldsymbol{A} (\boldsymbol{B}^x)^T) \tag{9}$$

where $\mathcal{O}$ is the set of all $B \times B$ permutation matrices. The constraint is applied symmetrically to both modalities, yielding the total loss $\mathcal{L}_U = \mathcal{L}_{U,x} + \mathcal{L}_{U,y}$.

Overall Objective Function

Our comprehensive training objective $\mathcal{L}_{Total}$ is a multi-task function, designed to integrate the four distinct components discussed previously. It simultaneously optimizes the model for (1) proxy learning, (2) cross-modal alignment, (3) semantic irrelevance separation, and (4) high-entropy hash code generation.The final objective is a weighted summation of these components:

$$\mathcal{L}_{Total} = \mathcal{L}_{Proxy} + \epsilon \mathcal{L}_{Reg} + \zeta \mathcal{L}_{CM} + \eta \mathcal{L}_U \tag{10}$$

where ϵ, ζ, and η are scalar hyperparameters that balance the contribution of each term.

4 Experiment

To demonstrate the validity of our proposed KDPH framework, we conducted comprehensive experiments on three publicly available cross-modal multi-label datasets MIRFLICKR-25K, NUS-WIDE and MS COCO. Besides, we introduce the datasets for the experiments and explain the details of the implementation of KDPH and evaluate metrics.

4.1 Experimental Settings

Datasets

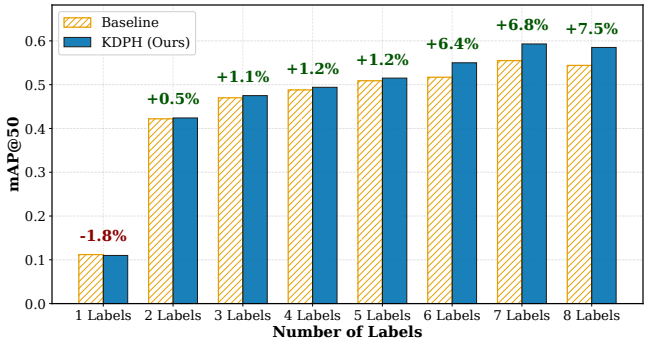
We evaluate KDPH on three standard multi-label datasets. MIRFLICKR-25K [Huiskes and Lew, 2008] contains 24,581 image-text pairs across 24 classes. NUS-WIDE [Chua *et al.*, 2009] comprises 269,648 pairs; following standard protocols, we utilize a subset of 195,834 pairs from 21 frequent categories. MS COCO [Lin *et al.*, 2014] consists of 123,289 images, each with 5 captions, spanning 80 categories. Across all datasets, we randomly sample 5,000 pairs as the query set and use the remainder as the database, from which 10,000 pairs are uniformly sampled to construct the training set.

Baseline Methods

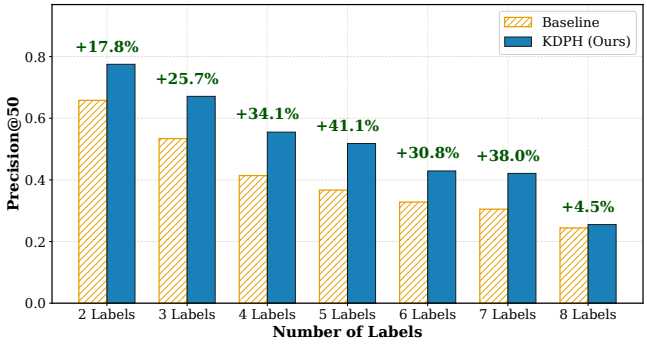
In our experiments, we selected 9 state-of-the-art deep cross-modal hashing methods for comparison, which contain DCPH[Tu *et al.*, 2023], nivMF[Kirchhof *et al.*, 2022], DCHMT[Tu *et al.*, 2022], MIAN[Zhang *et al.*, 2023], DSPH[Huo *et al.*, 2024b], MITH[Liu *et al.*, 2023], DNPH[Huo *et al.*, 2024a], DNpH[Qin *et al.*, 2024], DDBH[Qin *et al.*, 2025a], DNcH[Zhu *et al.*, 2025b] and DAGtH[Zhu *et al.*, 2025a].

Experimental Details

We implemented KDPH in PyTorch on an NVIDIA A800 GPU, optimizing via Adam (LR=0.001, batch size=128). The



(a) Retrieval Performance vs. Semantic Complexity



(b) Precision of 'Hard-to-Find' Labels

Figure 3: Impact of Semantic Complexity on Model Performance.

hyper-parameters are set as $\epsilon = 40$, $\zeta = 0.3$, $\eta = 0.1$. Regarding the anisotropic geometry, we set the number of principal axes as $N_{axes} = K/8$, where K denotes the target hash code length (e.g., $N_{axes} = 8$ for 64-bit codes). To ensure a fair comparison, we follow all baselines that use the identical CLIP backbone and data splits.

4.2 Main Results

To evaluate the effectiveness of our KDPH framework, we conduct a comprehensive comparison with state-of-the-art baselines on three benchmark datasets. The comparative results are summarized in Table 1. The proposed KDPH method outperforms all state-of-the-art baselines with significant performance improvements across all hash code lengths. This phenomenon demonstrates that the anisotropic Kent distributions possess stronger feature extraction and discriminative capabilities than the deterministic point-based baselines.In particular, our KDPH yields superiority over nivMF, a representative approach employing non-isotropic von Mises-Fisher distributions. It is observed that such methods suffer from the limitations of geometric inflexibility, particularly in scenarios enriched with complex semantic entanglements. This limitation is empirically evidenced by the substantial margin of over **7.92%** on the challenging MS COCO dataset ($I2T$ @ 16 bits).Despite the reliable performance achieved by previous methods, our KDPH benefits from the Kent-based distributional proxy and modeling elliptical semantic variance,

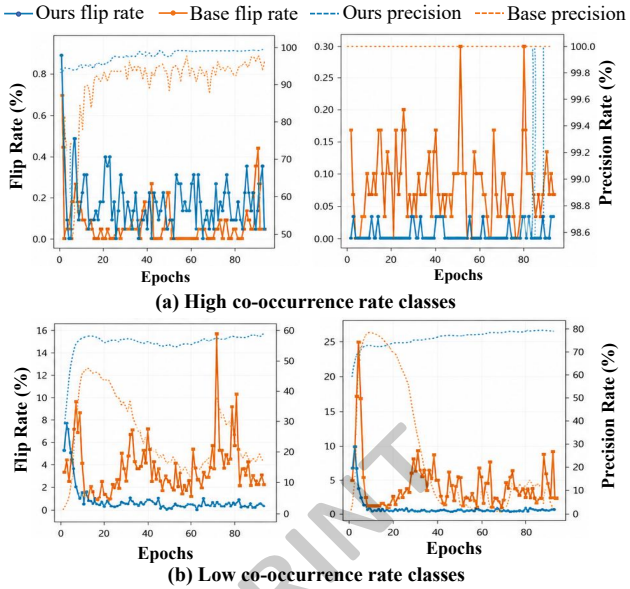


Figure 4: Training dynamics comparing DSPH and KDPH.

demonstrates superior performance.

Furthermore, to verify the performance of our KDPH under highly compressed spaces, we investigate the Bit-Efficiency. In comparison to the state-of-the-art methods (e.g., DCPH, DSPH) which suffer from sharp performance degradation under limited encoding capacity, KDPH maintains high discriminative power at 16 bits. Specifically, our KDPH yields an improvement of approximately 6.07% on MS COCO over the second-best method. These findings suggest that the meticulously designed Kent distributions offer superior information density, effectively absorbing gradient contentions even within highly compressed Hamming spaces. KDPH also achieves good balance between modalities, indicating that the optimization strategy successfully harmonizes visual and textual distributions within the shared latent space, thereby preventing the modality domination often observed in existing baselines.

4.3 Analysis

Mitigation of Chaos Proxy Oscillation via KDPH

We first investigate the robustness of our KDPH approach in preventing optimization collapse. So we monitor the track of Proxy Flip Rate $R_{flip}^{(t)}$ [Xie *et al.*, 2016], defined as the ratio of samples changing their nearest proxy assignment between consecutive epochs. As illustrated in Figure 4, it can be observed that in Low Co-occurrence Classes (Fig. 4b), both the baseline and KDPH exhibit stable convergence, confirming that rigid point proxies suffice for simple, unimodal manifolds.In contrast, regarding High Co-occurrence Classes (Fig. 4a), the baseline suffers from severe oscillation and jagged precision curves, leading to a deficiency in optimization stability. Benefiting from the advantages of anisotropic Kent distributions, KDPH maintains exceptional stability with a flip rate $< 1\%$. Specifically, our framework is designed to dynamically absorb gradient variance through shape deformation rather than

through straightforward centroid shifting. In this way, the proposed method effectively mitigates the chaotic oscillations that plague traditional point-based methods, ensuring better adaptability to complex semantic environments.

Prevention of Proxy Collapse via the KDPH

To comprehensively evaluate the effectiveness in prevention of proxy collapse, we track the *Subordinate Label* as the category with the least frequency in an sample. As illustrated in Figure 3, Baseline suffers from the limitations in high-complexity situations where rigid point proxies degrade rapidly under the pressure of dominant classes. In contrast, KDPH method consistently demonstrates the superiority and robustness over state-of-the-art competitors, yielding a significant improvement of 41.1% in scenarios with 5 labels. This promising performance confirms that the meticulously designed anisotropic distributions effectively mitigate the collapse issue by reserving sufficient *semantic variance* for tail concepts.

Despite the reliable performance achieved by the Baseline in single-label regimes due to the advantage of point proxies, it is noted that this advantage vanishes in multi-label scenarios. This slight redundancy introduced by KDPH serves as a necessary trade-off, providing the parametric flexibility required to bridge the semantic gaps between conflicting manifolds.

Components				mAP @ 64 bits	
Kent	$\mathcal{L}_{Reg}$	$\mathcal{L}_{CM}$	$\mathcal{L}_{U}$	I → T	T → I
	✓	✓	✓	0.7615	0.7542
$\beta = 0$	✓	✓	✓	0.7693	0.7628
✓		✓	✓	0.7326	0.7305
✓	✓		✓	0.7566	0.7585
✓	✓	✓		0.7537	0.7574
✓	✓	✓	✓	**0.7843**	**0.7795**

Table 2: Ablation study of different components on the MS COCO.

Ablation Study on Model Components

We conduct ablation studies by systematically evaluating the mAP of each component in our framework on the MS COCO dataset. The comparative results are summarized in Table 2. The performance metric exhibits a distinct hierarchy. Specifically, mAP decreases from 0.7693 with the isotropic vMF distribution ($\beta = 0$) to 0.7615 with rigid point proxies, suggesting only a marginal advantage for distributional proxies. Crucially, our anisotropic Kent model surpasses all state-of-the-art variants, owing to its capacity to capture *anisotropic* semantic conflicts, enabling effective recalibration of semantic interactions to resolve gradient contentions.Furthermore, the Multi-Modal Irrelevance Regularization $\mathcal{L}_{Reg}$ and Cross-Modal Alignment $\mathcal{L}_{CM}$ are designed to collaboratively enhance the contextual awareness and semantic capabilities, ensuring better adaptability to nuanced semantic variations while Uniform Constraint $\mathcal{L}_{U}$ provide essential complementary gains for modality invariance and code entropy.

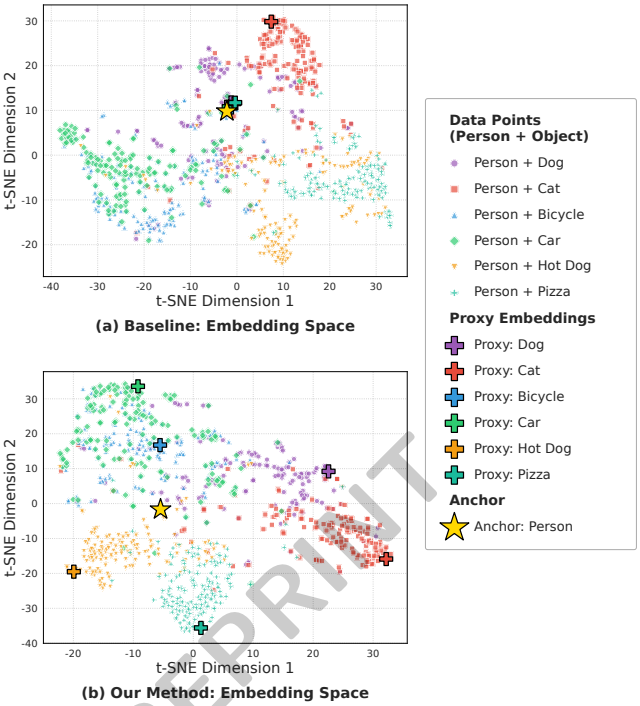


Figure 5: t-SNE visualization of the embedding spaces learned by DSPH and KDPH.

Visualizing Distributions Across Multi-label Scenarios

To empirically verify the topological superiority of KDPH, we visualized the embedding space of the complex Person category on MS COCO using t-SNE (Figure 5). The Baseline (Fig. 5a) exhibits severe proxy collapse, where proxies of distinct co-occurring classes (e.g., Car, Bicycle) physically collapse onto the central anchor. This visually confirms that rigid point proxies suffer from gradient contention, eliminating decision boundaries. In sharp contrast, KDPH (Fig. 5b) reveals a clear semantic-adaptive topology. The embedding space disentangles into coherent radial clusters based on high-level affinity—grouping animate entities, vehicles, and food along distinct directional axes. This structured layout demonstrates that our anisotropic Kent proxies successfully capture latent semantic hierarchies, accommodating contextual diversity without destabilizing the core identity.

5 Conclusion

In this paper, we address the critical bottleneck of gradient contention in multi-label cross-modal hashing with the Kent-based Distributional Proxy Hashing (KDPH) framework. By shifting the paradigm from point proxies to anisotropic Kent distributions, our method enables proxies to absorb contextual conflicts by dynamically adjusting directional variance. This geometric flexibility allows the model to maintain stable semantic centroids while stretching to accommodate diverse label correlations. Extensive experiments on three benchmarks demonstrate that KDPH outperforms state-of-the-art methods.

Acknowledgments

This work was supported by the Institute of Information Engineer ing, Chinese Academy of Sciences, under Grant E4101311F3, E4V06811F3. The reimbursement for this work was covered by this grant.

References

[Bai *et al.*, 2020] Cong Bai, Chao Zeng, Qing Ma, Jinglin Zhang, and Shengyong Chen. Deep adversarial discrete hashing for cross-modal retrieval. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, page 525–531, New York, NY, USA, 2020. Association for Computing Machinery.

[Cao *et al.*, 2022] Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713*, 2022.

[Chen *et al.*, 2024] Bingzhi Chen, Zhongqi Wu, Yishu Liu, Biqing Zeng, Guangming Lu, and Zheng Zhang. Enhancing cross-modal retrieval via visual-textual prompt hashing. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 623–631. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Main Track.

[Chua *et al.*, 2009] 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.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[Huiskes and Lew, 2008] Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, New York, NY, USA, 2008. ACM.

[Huo *et al.*, 2024a] 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 Trans. Multimedia Comput. Commun. Appl.*, 20(6), March 2024.

[Huo *et al.*, 2024b] 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, 2024.

[Huo *et al.*, 2024c] 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 Trans. on Knowl. and Data Eng.*, 36(11):5926–5939, November 2024.

[Jiang and Li, 2017] Qing-Yuan Jiang and Wu-Jun Li. Deep cross-modal hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[Kim *et al.*, 2020] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[Kirchhof *et al.*, 2022] Michael Kirchhof, Karsten Roth, Zeynep Akata, and Enkelejda Kasneci. A non-isotropic probabilistic take on proxy-based deep metric learning. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022.

[Li *et al.*, 2018] Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao. Self-supervised adversarial hashing networks for cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[Li *et al.*, 2021] Shen Li, Jianqing Xu, Xiaqing Xu, Pengcheng Shen, Shaoxin Li, and Bryan Hooi. Spherical confidence learning for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[Lin *et al.*, 2018] Xudong Lin, Yueqi Duan, Qiyuan Dong, Jiwen Lu, and Jie Zhou. Deep variational metric learning. In *Proceedings of the European Conference on Computer Vision*, September 2018.

[Liu *et al.*, 2023] Yishu Liu, Qingpeng Wu, Zheng Zhang, Jingyi Zhang, and Guangming Lu. Multi-granularity interactive transformer hashing for cross-modal retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 893–902, New York, NY, USA, 2023. Association for Computing Machinery.

[Liu *et al.*, 2024] Kaiming Liu, Yunhong Gong, Yu Cao, Zhenwen Ren, Dezhong Peng, and Yuan Sun. Dual semantic fusion hashing for multi-label cross-modal retrieval. In *International Joint Conferences on Artificial Intelligence Organization*, pages 4569–4577, 2024.

[Lu *et al.*, 2019] Xu Lu, Lei Zhu, Zhiyong Cheng, Liqiang Nie, and Huaxiang Zhang. Online multi-modal hashing with dynamic query-adaption. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 715–724, 2019.

[Movshovitz-Attias *et al.*, 2017] Yair Movshovitz-Attias, Alexander Toshev, Thomas K. Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *Proceedings of the IEEE International Conference on Computer Vision*, Oct 2017.

[Park *et al.*, 2019] Junyoung Park, Subin Yi, Yongseok Choi, Dong-Yeon Cho, and Jiwon Kim. Discriminative few-shot learning based on directional statistics. *arXiv preprint arXiv:1906.01819*, 2019.

[Qian *et al.*, 2019] Qi Qian, Lei Shang, Baigui Sun, Juhua Hu, Hao Li, and Rong Jin. Softtriple loss: Deep metric learning without triplet sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2019.

[Qin *et al.*, 2024] Qibing Qin, Yadong Huo, Lei Huang, Jiangyan Dai, Huihui Zhang, and Wenfeng Zhang. Deep neighborhood-preserving hashing with quadratic spherical mutual information for cross-modal retrieval. *IEEE Transactions on Multimedia*, 26:6361–6374, 2024.

[Qin *et al.*, 2025a] Qibing Qin, Yadong Huo, Wenfeng Zhang, Lei Huang, and Jie Nie. Deep discriminative boundary hashing for cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(10):10557–10570, 2025.

[Qin *et al.*, 2025b] Qibing Qin, Lei Wu, Wenfeng Zhang, Lei Huang, and Jie Nie. Deep semantic-consistent penalizing hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*, 27:4613–4626, 2025.

[Roth *et al.*, 2022] Karsten Roth, Oriol Vinyals, and Zeynep Akata. Non-isotropy regularization for proxy-based deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7420–7430, 2022.

[Shi and Jain, 2019] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6902–6911, 2019.

[Singh and Gupta, 2022] Avantika Singh and Shaifu Gupta. Learning to hash: a comprehensive survey of deep learning-based hashing methods. *Knowledge and Information Systems*, 64(10):2565–2597, 2022.

[Sun *et al.*, 2022] Changchang Sun, Hugo Latapie, Gaowen Liu, and Yan Yan. Deep normalized cross-modal hashing with bi-direction relation reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4941–4949, 2022.

[Teh *et al.*, 2020] Eu Wern Teh, Terrance DeVries, and Graham W Taylor. Proxynca++: Revisiting and revitalizing proxy neighborhood component analysis. In *European conference on computer vision*, pages 448–464. Springer, 2020.

[Tu *et al.*, 2022] 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*, MM '22, page 453–461, New York, NY, USA, 2022. Association for Computing Machinery.

[Tu *et al.*, 2023] Rong-Cheng Tu, Xian-Ling Mao, Rong-Xin Tu, Binbin Bian, Chengfei Cai, Hongfa Wang, Wei Wei, and Heyan Huang. Deep cross-modal proxy hashing. *IEEE Trans. on Knowl. and Data Eng.*, 35(7):6798–6810, July 2023.

[Wang and Isola, 2020] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020.

[Wang *et al.*, 2022] Yongxin Wang, Zhen-Duo Chen, Xin Luo, and Xin-Shun Xu. A high-dimensional sparse hashing framework for cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8822–8836, 2022.

[Xie *et al.*, 2016] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016.

[Xu *et al.*, 2019] Ruiqing Xu, Chao Li, Junchi Yan, Cheng Deng, and Xianglong Liu. Graph convolutional network hashing for cross-modal retrieval. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 982–988. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[Xu *et al.*, 2024] Cai Xu, Jiajun Si, Ziyu Guan, Wei Zhao, Yue Wu, and Xiyue Gao. Reliable conflictive multi-view learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 16129–16137, 2024.

[Yan *et al.*, 2025] Shuanglin Yan, Jun Liu, Neng Dong, Liyan Zhang, and Jinhui Tang. Cross-modal collaborative representation learning for text-to-image person retrieval. In *Proceedings of the thirty-fourth international joint conference on artificial intelligence*, 2025.

[Zhang *et al.*, 2023] 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, 2023.

[Zhe *et al.*, 2018] Xuefei Zhe, Shifeng Chen, and Hong Yan. Directional statistics-based deep metric learning for image classification and retrieval. *arXiv preprint arXiv:1802.09662*, 2018.

[Zhu *et al.*, 2025a] Congcong Zhu, Wei Hu, Jinkui Hou, Qibing Qin, Wenfeng Zhang, and Lei Huang. Deep adaptive gradient-triplet hashing for cross-modal retrieval. *Expert Systems with Applications*, 291:128566, 2025.

[Zhu *et al.*, 2025b] Congcong Zhu, Qibing Qin, Wenfeng Zhang, and Lei Huang. Deep neighbor-coherence hashing with discriminative sample mining for supervised cross-modal retrieval. *Expert Systems with Applications*, 279:127365, 2025.