Proxy Zero-Shot Hashing with Multimodal Fusion via Stable Diffusion

Hui Zhang[1], **Weikang Gao**[1], **Tao Yang**[1], **Yuan Cao**[*1]

[1]School of Computer Science and Technology, Ocean University of China, China
{zh9796, gaoweikang, yangtao5635}@stu.ouc.edu.cn, cy8661@ouc.edu.cn

Abstract

With the rapid growth of visual content in open-world environments, zero-shot hashing image retrieval (ZSHIR) has emerged to tackle the challenge of recognizing novel classes using attribute-level and semantic information. However, existing methods often rely on shallow fusion of multi-source cues (e.g., attributes, labels, and visual features) through external supervision or feature concatenation, failing to capture the underlying semantic structure in a generative way. Particularly, current bridging strategies between modalities suffer from information fragmentation and weak alignment, hindering the model's ability to fully understand complex attribute-visual relations. Moreover, subtle semantic gaps or "semantic drift" between seen and unseen classes further degrade inter-class separability and the scalability of hashing models. To address these issues, we propose a novel framework called Proxy Zero-Shot Hashing with Multimodal Fusion via Stable Diffusion (PZSH), which integrates generative modeling and contrastive learning. PZSH leverages a pre-trained Stable Diffusion model to synthesize multimodal content, and uses dual BLIP encoders to enhance semantic alignment across modalities. We further design a proxy hashing loss to enforce discriminative binary representations. Extensive experiments on benchmark datasets show that PZSH achieves state-of-the-art performance with stronger generalization to unseen classes.

Code — https://github.com/caoyuan618/PZSH

Introduction

The recent decade has witnessed the fast development of hashing for semantic image retrieval (Luo et al. 2023; Wang et al. 2016, 2018; Zhu et al. 2024). Hashing-based methods are particularly popular for encoding images into compact binary codes, enabling efficient similarity search. Supervised deep hashing approaches (Rongkai et al. 2014; Liu et al. 2016a; Su et al. 2018; Yang, Lin, and Chen 2018; Qiu et al. 2018; Chen et al. 2021; Zhang et al. 2022), integrating semantic labels and convolutional neural networks (CNNs) (Cong et al. 2020; Mi, Lei, and Gui 2013; Cheng et al. 2025; Huo et al. 2025; Sun et al. 2023; Pu et al. 2025a,b), have demonstrated strong performance by preserving semantic similarity. However, most existing methods depend on known, predefined categories, limiting their practical use in dynamic scenarios where new categories frequently emerge (Qiu et al. 2017). Thus, there is an urgent demand for hashing models capable of generalizing to unseen categories, supporting Zero-Shot Hashing (ZSH) retrieving images from novel classes without annotated examples.

ZSH techniques aim to enable retrieval for novel categories by transferring knowledge from seen to unseen classes using auxiliary semantic information such as attributes, word embeddings, or class descriptions. A range of methods (Wang et al. 2021; Yong et al. 2024) have explored embedding alignment, similarity transfer, and structural constraints to improve generalization, with notable efforts integrating semantic vectors into the hashing process or leveraging transductive strategies. These advances have made initial progress in bridging the domain gap between seen and unseen classes.

Despite recent advances, existing deep hashing methods for ZSH still face significant limitations. First, the integration of visual and semantic information is often superficial—typically via simple feature concatenation or auxiliary losses—failing to model complex cross-modal interactions and often treating attribute and visual cues as disjoint inputs without unified fusion. Some methods (Shen et al. 2018; Dong et al. 2024) explicitly train visual-semantic mapping networks, but they frequently come with high computational costs and limited scalability. Second, some approaches (Guo et al. 2017; Jiang et al. 2025) rely on coarse, label-level supervision without effectively aligning detailed structural relationships between visual and semantic spaces, limiting their ability to discriminate fine-grained semantics. Lastly, although semantic descriptions (e.g. attributes or class labels) are used to represent unseen classes, most methods lack mechanisms to synthesize corresponding visual representations, which is critical in cold-start scenarios where no visual data is available.

To overcome these challenges, we introduce a novel framework named Proxy Zero-Shot Hashing with Multimodal Fusion via Stable Diffuison (PZSH). This method unifies visual, semantic, and label information to enhance generalizability, particularly under zero-image scenarios. At its core, PZSH leverages a pre-trained Stable Diffusion (SD)

*Corresponding author

model to generate multimodal images that blend class-level semantics and visual features at the data level. Unlike existing strategies that fuse modalities in shallow feature space, our framework achieves deep semantic integration through generative visual representation, resulting in more expressive and transferable features. A key component of our approach is a dual-branch contrastive learning mechanism built on BLIP (Li et al. 2022) encoders. By independently encoding both real and SD-generated images and aligning them in feature space via contrastive loss, the model is guided to capture consistent, fine-grained visual semantics across variations. This facilitates a richer understanding of subtle attribute-level distinctions which are overlooked by conventional label-supervised methods. In scenarios where real visual data for unseen categories is entirely absent, PZSH further introduces a zero-image data augmentation strategy. Instead of relying on annotated samples, we synthesize images directly from semantic prompts—such as class names or attributes—using the SD model. These synthetic samples serve as informative proxies, enabling zero-image training for unseen categories, thus greatly alleviating the cold-start problem faced by conventional ZSH models.

Taken together, these innovations form a comprehensive and scalable solution to the problem of generalized Zero-Shot hashing, combining generative modeling, contrastive learning, and semantic synthesis in a unified framework. Our main contributions can be summarized as follows

- We propose a novel Zero-Shot hashing framework with strong multimodal fusion capability, which effectively aligns semantic and visual representations under unseen-category settings and significantly improves image retrieval performance.
- A generative augmentation strategy is introduced to enhance the model's generalization ability in zero-image scenarios. Furthermore, we present a novel proxy loss to enforce discriminative binary codes.
- We conduct extensive experiments on two common zero-shot retrieval benchmarks (AWA2 and CUB). The results demonstrate that our method achieves superior performance compared with state-of-the-art baselines.

Related Work

Hashing

Hashing is widely used for large-scale image retrieval due to its ability to compress high-dimensional visual data into compact binary codes while preserving semantic relationships. Traditional methods such as Locality Sensitive Hashing (LSH) (Gionis, Indyk, and Motwani 1999) and Iterative Quantization (ITQ) (Gong et al. 2013) introduce data-independent hash functions. However, their semantic modeling capacity is limited. To address this issue, supervised data-dependent hashing approaches emerge. Kernel-based Supervised Hashing (KSH) (Liu et al. 2012) and Supervised Discrete Hashing (SDH) (Shen et al. 2015) incorporate label supervision into the learning process, enhancing semantic consistency. The introduction of deep learning further revolutionize this field. End-to-end deep hashing frameworks such as DSH (Liu et al. 2016b), DPSH (Li, Wang, and Kang 2016), and DSDH (Li et al. 2017) enable simultaneous learning of image representations and binary codes in a unified network. Subsequent advances focus on tackling the discrete optimization challenge and improving quantization quality. HashNet (Cao et al. 2017), DHN (Zhu et al. 2016), and BNNH (Zhang et al. 2021) apply continuation methods and specialized quantization losses to reduce the performance gap between continuous and binary representations. Other work lines explore semantic-aware hashing by introducing class centers (CSQ) (Yuan et al. 2020a) or structural guidance through pairwise and graph-based constraints (DAPH, DFH) (Shen et al. 2017; Li et al. 2019) to enhance the discriminability of hash codes. However, these methods show poor performance when dealing with unseen classes.

Zero-Shot Hashing

Zero-Shot hashing (ZSH) extends supervised hashing to settings where test-time categories are unseen during training. This is achieved by leveraging auxiliary semantic information such as class attributes, word embeddings, or textual descriptions to enable knowledge transfer from seen to unseen classes. Early methods like TSK-ZSH (Yang et al. 2016) and SitNet (Guo et al. 2017) construct similarity-preserving frameworks that transfer supervision from seen to unseen domains. To improve semantic alignment, structural regularization strategies are proposed. ZSH-OP (Zhang, Long, and Shao 2019) and CHOP (Yuan et al. 2021) introduce orthogonal constraints to preserve inter-class relationships, while transductive methods like VSB2-Net (Li et al. 2021) and TZS-ML (Zou et al. 2022) utilize unlabeled test data during training to better capture unseen-class structure. More recently, generative approaches have gained traction in ZSH. DSH-GAN (Qiu et al. 2017), COMAE (Li et al. 2025), and CRAR (Wang et al. 2024) synthesize representative visual features from semantic embeddings using generative adversarial networks (GANs), which mitigates the zero-image challenge by enriching training data for unseen classes. Meanwhile, prompt-based and part-aware frameworks have emerged. AgNet (Ji et al. 2020) incorporates attribute-guided embedding learning for cross-modal transfer. PIXEL (Dong et al. 2024) introduces prompt-driven alignment between vision and language using transformer encoders. RAZH (Jiang et al. 2025) proposes part-aware reconstruction to enhance fine-grained alignment between textual semantics and visual regions.

Despite recent progress, existing ZSH methods still exhibit key limitations. Many rely on shallow fusion strategies—such as feature concatenation or auxiliary losses—that fail to capture the complex interplay between attributes, text, and visual content, thus limiting semantic transfer in fine-grained scenarios. Generative methods using GANs often produce low-fidelity features with limited diversity, and rarely explore image-level synthesis for richer training signals. Moreover, fine-grained alignment across modalities remains challenging, especially under subtle attribute shifts. These issues highlight the need for a more expressive framework capable of unified semantic-visual modeling and robust zero-image generalization.

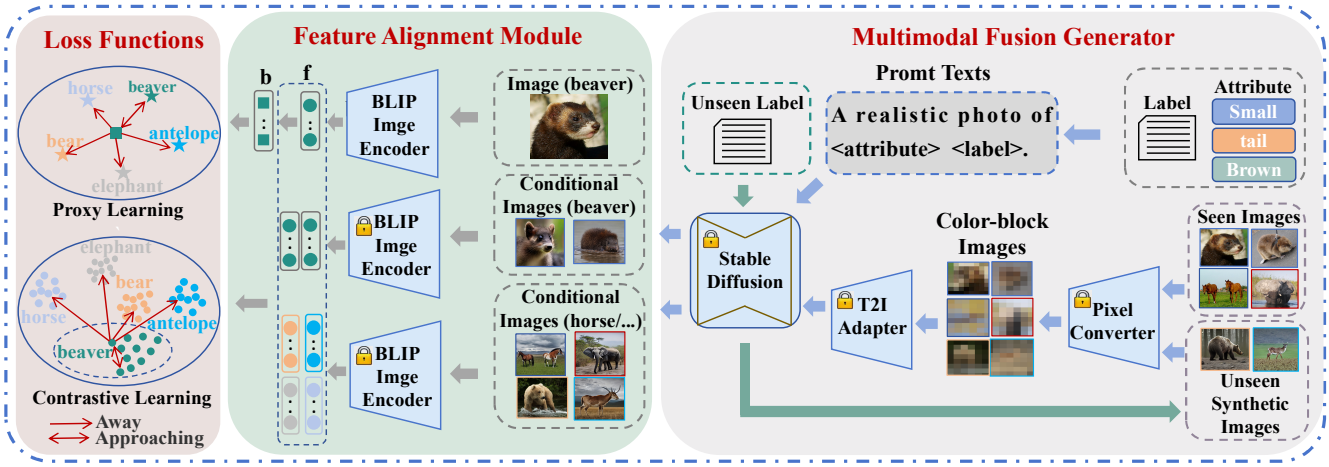

Figure 1: An overview of the proposed PZSH framework. The system integrates semantic labels and class attributes to synthesize pseudo-images via Stable Diffusion. A dual-branch contrastive mechanism aligns visual and synthetic features. A hashing network generates compact binary codes for retrieval.

Methodology

Framework Overview

As shown in Figure 1, our framework consists of a Multimodal Fusion Generator to fuse conditions of images, labels, and attributes and a Feature Alignment Module to supervise the encoding model aligning the fused conditional information. Furthermore, a proxy hashing loss is used for hash training.

Problem Definition

Zero-Shot hashing image retrieval (ZSHIR) aims to retrieve instances from unseen categories without using their visual samples during training. Let the training set be $\mathcal{X}_s = \{x_1^s, \ldots, x_{N_s}^s\}$, where each image x_i^s belongs to a seen class from the label space $\mathcal{C}_s = \{c_1^s, \ldots, c_{k_s}^s\}$. The training set is annotated by a one-hot label matrix $\mathcal{Y}_s \in \{0,1\}^{N_s \times k_s}$.

At test time, retrieval is conducted over instances from an unseen class set $\mathcal{C}_u = \{c_1^u, \ldots, c_{k_u}^u\}$, where $\mathcal{C}_s \cap \mathcal{C}_u = \emptyset$. The objective is to learn a hash function $h : \mathcal{X} \to \{-1, +1\}^K$ such that semantically similar images—regardless of being from seen or unseen categories—are mapped to nearby hash codes in Hamming space. To facilitate semantic knowledge transfer across disjoint categories, we leverage manually defined attribute vectors as semantic embeddings for each class $c \in \mathcal{C}_s \cup \mathcal{C}_u$, which serve as auxiliary guidance for training hash functions with generalization ability.

Multimodal Fusion Generator

In Zero-Shot hashing image retrieval, a key challenge lies in effectively integrating visual features with semantic information such as class attributes and labels. Most prior methods perform this fusion via shallow strategies—feature concatenation, auxiliary loss terms, or additive weighting—which fail to capture the nonlinear interactions between visual structures, category semantics, and fine-grained attributes. As a result, such methods suffer from limited transferability to unseen classes.

To address this issue, we propose a Multimodal Fusion Generator based on Stable Diffusion (SD), enabling deep-level semantic-visual fusion at the image generation stage. Each training sample is represented by

- A stylized image x_i^{style}, obtained by transforming the original image x_i into a color-block representation that preserves its layout and color distribution;
- A textual prompt T_i, which combines the class attribute and label:

$$T_i = \text{“A photo of a } \langle \text{attribute}_i \rangle \langle \text{label}_i \rangle \text{”}.$$

These two inputs are passed into the pre-trained SD model via a T2I-Adapter (Mou et al. 2024) to generate a multimodal pseudo-image:

$$G_i = \mathcal{G}(x_i^{\text{style}}, T_i), \tag{1}$$

where $\mathcal{G}(\cdot)$ denotes the diffusion-based generation function, and G_i is the resulting image that visually mimics x_i while semantically reflecting both attribute and label information.

This design enables deep cross-modal fusion at the data level, effectively capturing category-level semantics through the generation process. The resulting image G_i can be seen as a semantically enriched extension of the original sample, improving the model's ability to generalize to unseen categories.

Feature Alignment Module

To enhance the model's ability to capture fine-grained intra-class variation and inter-class distinctions, we introduce a feature alignment module based on a dual-branch BLIP encoder. For each sample, we extract semantic embeddings from the original image x_i and its synthesized multimodal

counterpart G_i as $f_i = \text{BLIP}(x_i)$ and $f_i' = \text{BLIP}(G_i)$, respectively. These representations are aligned in the feature space through a log-softmax contrastive objective.

We first normalize both f_i and f_i', then compute the pairwise similarity matrix as

$$\text{sim}(f_i, f_j') = \frac{f_i^\top f_j'}{\tau}, \tag{2}$$

where τ is a temperature parameter controlling the sharpness of similarity distributions. A row-wise log-softmax operation is then applied

$$\mathcal{L}_{\text{log-soft}}(i,j) = \log \frac{\exp(\text{sim}(f_i, f_j'))}{\sum_{k=1}^{N} \exp(\text{sim}(f_i, f_k'))}. \tag{3}$$

To identify positive pairs, we construct a binary mask $M \in \{0,1\}^{N\times N}$, where $M_{ij} = 1$ if and only if x_i and G_j share the same class label:

$$M_{ij} = \mathbb{I}[y_i = y_j]. \tag{4}$$

To avoid trivial identity alignment, the diagonal is removed $M_{ii} = 0$. Then, we normalize each row to ensure equal weight for each valid positive:

$$\tilde{M}_{ij} = \frac{M_{ij}}{\sum_j M_{ij} + \epsilon}. \tag{5}$$

The final contrastive loss encourages alignment of all same-class image pairs in the log-softmax space:

$$\mathcal{L}_{\text{CL}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \tilde{M}_{ij} \cdot \mathcal{L}_{\text{log-soft}}(i,j). \tag{6}$$

This formulation enables the model to leverage all available positive samples for alignment while avoiding overfitting to self-matching. By directly enforcing consistency between the original and generated image representations, the module promotes robust and generalizable semantic embeddings, crucial for effective Zero-Shot retrieval.

Hashing Module

The final component of our framework is designed to learn compact, discriminative, and generalizable binary hash codes that preserve semantic similarities in Hamming space. Given an input image x_i, the BLIP-encoded feature $\mathbf{f}_i \in \mathbb{R}^D$ is projected by a lightweight hash head into a latent representation $\mathbf{u}_i \in \mathbb{R}^K$, followed by a `tanh` activation to produce a continuous hash code $\mathbf{b}_i \in (-1,+1)^K$, where K is the hash length.

To explicitly enhance the class separability and intra-class compactness of hash codes, we employ a **center-based hash supervision strategy**. A set of semantic hash centers $\mathbf{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_k\} \subset \{-1,+1\}^K$ is generated via a Bernoulli-sampling strategy that ensures high inter-class dispersion in Hamming space (Yuan et al. 2020b). Specifically, each center $\mathbf{c}_j$ is sampled from a balanced binary distribution and selected from multiple trials to maximize average pairwise Hamming distances:

$$\mathbf{c}_j \sim \text{Bernoulli}(0.5) \quad \text{with} \quad \mathbb{E}[\text{Ham}(\mathbf{c}_j, \mathbf{c}_l)] \geq \frac{K}{2}, \quad \forall j \neq l. \tag{7}$$

We define a cosine-softmax classification loss to match each hash code with its corresponding center. Let $\mathbf{C} \in \mathbb{R}^{k\times K}$ be the matrix of class-wise centers, $\mathbf{y}_i \in \{0,1\}^k$ be the one-hot label vector of image i. The classification loss is computed as

$$\mathcal{L}_{\cos} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{k} y_{ij} \log \frac{\exp\left(\alpha \cdot \cos(\mathbf{b}_i, \mathbf{c}_j)\right)}{\sum_{l=1}^{k} \exp\left(\alpha \cdot \cos(\mathbf{b}_i, \mathbf{c}_l)\right)}, \tag{8}$$

where $\cos(\mathbf{b}_i, \mathbf{c}_j) = \frac{\mathbf{b}_i^\top \mathbf{c}_j}{\|\mathbf{b}_i\| \|\mathbf{c}_j\|}$, and $\alpha = \sqrt{K}$ is a scaling factor.

To encourage binarization, we introduce a quantization loss:

$$\mathcal{L}_{\text{quant}} = \frac{1}{N} \sum_{i=1}^{N} \| |\mathbf{b}_i| - \mathbf{1} \|_2^2, \tag{9}$$

which minimizes the deviation between the continuous code $\mathbf{b}_i$ and its binary target $\{-1,+1\}^K$.

In addition, to further refine inter-sample discriminability, we incorporate a proxy contrastive loss. With the historical codes as $\mathbf{t}_j$, the proxy loss is defined as

$$\mathcal{L}_{\text{proxy}} = \mathbb{E}_{(i,j)\sim\mathcal{P}} \left[\log \left(1 + \exp \left(\frac{1 - \cos(\mathbf{b}_i, \mathbf{b}_j)}{2} \right) \right) \right], \tag{10}$$

where $\mathcal{P}$ contains positive pairs (same class). The Proxy loss quantifies semantic consistency between continuous hash codes $\mathbf{b}_i$ and $\mathbf{b}_j$ via cosine similarity, enforcing intra-class hash code compactness by minimizing loss for high similarity and amplifying loss for deviations—ultimately enhancing the discriminative power of binary representations to advance zero-shot retrieval efficacy.

The final objective for the hashing module combines all terms:

$$\mathcal{L}_{\text{hash}} = \mathcal{L}_{\cos} + \mathcal{L}_{\text{quant}} + \beta \mathcal{L}_{\text{proxy}}, \tag{11}$$

where β is the balancing hyperparameter. This center-supervised cosine-softmax formulation enables our hash codes to align with well-separated semantic anchors, effectively improving the discriminability and generalization ability of the learned binary representations, especially under Zero-Shot settings.

Finally, we formulate the overall loss of PZSH as:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{hash}} + \mathcal{L}_{\text{CL}}, \tag{12}$$

where α is the balancing hyperparameter.

Model Optimization Strategy

For the purpose to further enhance support to unseen classes, we propose a data-level optimization strategy that synthesizes pseudo-images exclusively from the class names of unseen categories. Given a zero-shot setting where visual samples from unseen classes $\mathcal{C}_u$ are absent, we construct semantic prompts for each class label using the following template:

$$\text{Prompt}_i = \text{"A photo of a } \langle \text{label} \rangle \text{"}.$$

Method	AWA2				CUB			
	24 bits	48 bits	64 bits	128 bits	24 bits	48 bits	64 bits	128 bits
LSH (1998 STOC)	0.0106	0.0151	0.0204	0.0306	0.0055	0.0076	0.0095	0.0095
SH (2008 NeurIPS)	0.1833	0.2729	0.2955	0.3441	0.0568	0.0810	0.0886	0.1191
ITQ (2011 CVPR)	0.1999	0.2821	0.2964	0.3764	0.0533	0.0765	0.0892	0.1182
IMH (2011 CIKM)	0.1282	0.1536	0.1613	0.1681	0.0330	0.0361	0.0364	0.0386
PCA (2009 NeurIPS)	0.2165	0.2530	0.2701	0.2719	0.0547	0.0598	0.0632	0.0695
HashNet (2017 ICCV)	0.2086	0.2386	0.2516	0.2749	0.0528	0.0566	0.0595	0.0633
GreedyHash (2018 NeurIPS)	0.3420	0.4169	0.4240	0.4639	0.1132	0.1707	0.1841	0.2326
JMLH (2019 ICCV)	0.3607	0.4364	0.4408	0.4711	0.1078	0.1555	0.1987	0.2310
ADSH (2018 AAAI)	0.3360	0.4787	0.5105	0.5454	0.0858	0.1607	0.1827	0.2424
CSQ (2020 CVPR)	0.3194	0.3988	0.3773	0.4072	0.0996	0.1588	0.1712	0.2201
DPN (2020 IJCAI)	0.1783	0.2086	0.2378	0.2565	0.0445	0.0728	0.0772	0.1003
BiHalf (2021 AAAI)	0.3440	0.4036	0.4223	0.4577	0.0794	0.1280	0.1573	0.2142
OrthoCos (2021 NeurIPS)	0.1709	0.2302	0.2312	0.2566	0.0451	0.0660	0.0736	0.0984
CIBHash (2021 IJCAI)	0.2113	0.2304	0.2481	0.2618	0.0351	0.0404	0.0411	0.0453
TBH (2020 CVPR)	0.0941	0.1201	0.1073	0.1730	0.0157	0.0176	0.0226	0.0252
TSK (2016 MM)	0.2262	0.3109	0.3873	0.4151	0.0739	0.1200	0.1394	0.1112
SASH (2022 TIP)	0.2560	0.3421	0.3898	0.3947	0.0744	0.1278	0.1426	0.1535
ASZH (2023 TKDE)	0.2619	0.3787	0.4032	0.4158	0.0764	0.1192	0.1294	0.1727
SitNet (2017 IJCAI)	0.2344	0.2406	0.2549	0.2650	0.0880	0.1127	0.1141	0.1167
OPZSH (2023 ICASSP)	0.1056	0.1390	0.1618	0.1961	0.0632	0.0879	0.0962	0.1143
AH (2022 CVPR)	0.2275	0.1989	0.3154	0.3557	0.0480	0.0897	0.1089	0.1445
PIXEL (2024 CIKM)	0.3819	0.4792	0.5133	0.5465	0.1136	0.1777	0.1994	0.2519
PZSH(Ours)	**0.5129**	**0.5921**	**0.6297**	**0.6477**	**0.1640**	**0.2284**	**0.2651**	**0.2775**

Table 1: Comparison of mAP@all results for different methods on AWA2 and CUB with 24, 48, 64, and 128 bits. The best results are highlighted in bold, the second-best results are underlined.

These prompts are then fed into the pre-trained SD model to generate a set of auxiliary pseudo-images:

$$\mathcal{G}_u = \{\text{SD}(\text{Prompt}_i) \mid c_i^u \in \mathcal{C}_u\}. \tag{13}$$

This strategy effectively enriches the training distribution with diverse and class-consistent pseudo-visual information for unseen classes. Unlike prior methods that rely solely on attributes or embeddings for semantic transfer, our approach injects visual context into the learning process, enabling the model to acquire transferable semantics in a data-driven manner. Furthermore, this zero-image augmentation mechanism significantly alleviates the cold-start challenge in real-world retrieval systems where new categories may appear dynamically without labeled data.

Why using BLIP rather than CLIP or ViT? We adopt the BLIP encoder instead of CLIP (Radford et al. 2021) or ViT (Dosovitskiy et al. 2021) due to its superior capability in modeling fine-grained multimodal semantics. CLIP focuses on global image-text alignment via coarse caption supervision, which often fails to capture subtle intra-class variations. ViT, while strong in visual representation, lacks explicit multimodal alignment. In contrast, BLIP is pretrained on both generative and understanding tasks (e.g., VQA, image-grounded dialogue), enabling it to encode object-level attributes and contextual relationships more effectively. This fine-grained semantic reasoning is essential for aligning real and synthetic images in Zero-Shot settings, where robust cross-domain generalization is critical.

Experiments

We conducte extensive experiments on two widely used benchmark datasets for zero-shot learning and retrieval tasks, to evaluate our method in terms of both search accuracy and scalability. We also compare it with several state-of-the-art hashing and Zero-Shot hashing methods.

Datasets

AWA2 (Animals with Attributes 2) (Xian et al. 2018) includes 37,322 images of 50 animal species, each annotated with an 85-dimensional attribute vector for visual traits (color, shape, habitat). Using the standard Zero-Shot split (40 seen, 10 unseen classes), we train on 100 real images per seen class (4,000 total) and 40 Stable Diffusion-generated pseudo-images per unseen class (400 total), yielding 4,400 training images. For evaluation, 1,000 queries (100 per unseen class) are retrieved from a database of remaining unseen and all seen images.

CUB (Caltech-UCSD Birds-200-2011) (Wah et al. 2011) has 11,788 images of 200 bird species, each with a 312-dimensional binary attribute vector for fine-grained properties (beak shape, feather color). With the standard Zero-Shot split (150 seen, 50 unseen classes), we train on 30 real images per seen class (4,500 total) and 30 pseudo-images per unseen class (1,500 total), totaling 6,000 training images. The query set has 1,500 images (30 per unseen class), and the database includes all remaining unseen and seen images.

Variant	AWA2				CUB			
	24 bits	48 bits	64 bits	128 bits	24 bits	48 bits	64 bits	128 bits
w/o USI	0.4913	0.5430	0.5804	0.6183	0.1158	0.1506	0.1915	0.2280
w/o FAM	0.4858	0.4992	0.5561	0.6135	0.1032	0.1370	0.1599	0.1938
w/o L_{pair}	0.4884	0.5321	0.5618	0.6220	0.1396	0.2002	0.2375	0.2551
w/o L_{quant}	0.5009	0.5858	0.6206	0.6291	0.1545	0.2123	0.2502	0.2613
BLIP $\rightarrow$ CLIP	0.5044	0.5789	0.5942	0.6233	0.1290	0.1861	0.2085	0.2626
PZSH	**0.5129**	**0.5921**	**0.6297**	**0.6477**	**0.1640**	**0.2284**	**0.2651**	**0.2775**

Table 2: Ablation study of individual components in our framework on AWA2 and CUB. The full model PZSH includes BLIP encoder, Unseen Synthetic Images (USI) and Feature Alignment Module (FAM).

Evaluation Metrics

Following conventional ZSH evaluation, we report mean Average Precision (mAP) retrieved results to assess semantic consistency of retrieved instances. All methods are evaluated under identical training/test splits and hash lengths (24, 48, 64, 128 bits).

Implementation Details

Our model is optimized using RMSProp with a learning rate of 1×10^{-5} and weight decay of 1×10^{-5}. The batch size is set to 32. The hyperparameters in the loss functions α and β are set as (0.1,0.7) on CUB and (0.5,1.0) on AWA2, respectively. The temperature parameter τ is fixed at 0.07 throughout. We train the network for 50 epochs and reduce the learning rate by a factor of 0.5 every 15 epochs.

Results and Analyses

Due to the unavailability of official code for several baseline models, we directly adopt the mAP results reported in the PIXEL paper (Dong et al. 2024). This ensures consistent and fair comparison across all methods under the same evaluation settings. Compared to conventional and deep supervised hashing methods, Zero-Shot hashing (ZSH) methods demonstrate a clear advantage in handling unseen categories by leveraging semantic side information. As shown in Table 1, traditional unsupervised methods yield very limited performance due to their inability to incorporate semantic signals. Deep supervised hashing methods like JMLH and ADSH perform notably better on datasets like AWA2, which is a coarse-grained dataset where semantic differences between categories are large and easier to distinguish. This allows supervised models to perform relatively well given sufficient label supervision. However, their performance degrades on the fine-grained CUB dataset, where semantic distinctions between classes are much more subtle and demand stronger transfer learning capabilities. Since deep supervised hashing models typically lack mechanisms for cross-class generalization, they struggle to effectively retrieve instances from unseen fine-grained categories. In contrast, Zero-Shot hashing approaches like TSK, SASH, ASZH, and PIXEL offer improved retrieval accuracy on both AWA2 and CUB, with our proposed method PZSH achieving the highest mAP scores across all bit lengths. This is attributed to its generative semantic augmentation and fine-grained contrastive alignment, which better capture transferable visual semantics. Notably, even compared to strong ZSH baselines such as PIXEL and ASZH, PZSH shows a significant margin, especially at higher bit lengths, validating the effectiveness of our SD-based multimodal synthesis and BLIP-guided feature alignment in zero-image scenarios.

Ablation Study

To evaluate the contribution of each component in our framework, we perform ablation experiments on AWA2 and CUB across different code lengths, as shown in Table 2. Removing the unseen-class pseudo-image generation leads to consistent performance drops (from 0.6477 to 0.6183 on AWA2-128 bits), confirming its effectiveness in reducing the semantic gap. Further discarding the Multimodal Fusion Generator and Feature Alignment Module results in even lower performance (0.5561 on AWA2-64 bits), showing that modality-level integration and contrastive alignment are crucial for learning discriminative codes. Lastly, replacing BLIP with CLIP yields slight degradations (from 0.6297 to 0.5942 on AWA2-64 bits), highlighting the benefit of BLIP's fine-grained multimodal pretraining. These results jointly demonstrate the necessity of each proposed design choice in our framework.

Hyperparameter Sensitivity Analyses

We perform a grid-based sensitivity analysis to examine how the weights of contrastive loss (α) and proxy loss (β) affect the retrieval performance. Figure 3 displays the 3D surface plots of mAP@all across varying (α, β) pairs with 64 hash bits on the CUB and AWA2 datasets. On CUB, the performance surface reveals clear peaks around $(0.1, 0.7)$ and $(0.3, 1.0)$, reaching mAP values of 0.2651 and 0.2595, respectively. This indicates that relatively low contrastive loss combined with stronger center regularization yields more discriminative hash codes, especially in fine-grained scenarios. In contrast, AWA2 exhibits a flatter response surface with its highest mAP of 0.6297 appearing at $(0.5, 1.0)$. The wider plateau of strong results suggests that AWA2 is less sensitive to hyperparameter variation and benefits from moderate to strong regularization on both terms. All in all, the results demonstrate that our method remains robust across a broad range of hyperparameter choices, with optimal regions differing slightly between fine-grained and coarse-grained datasets.

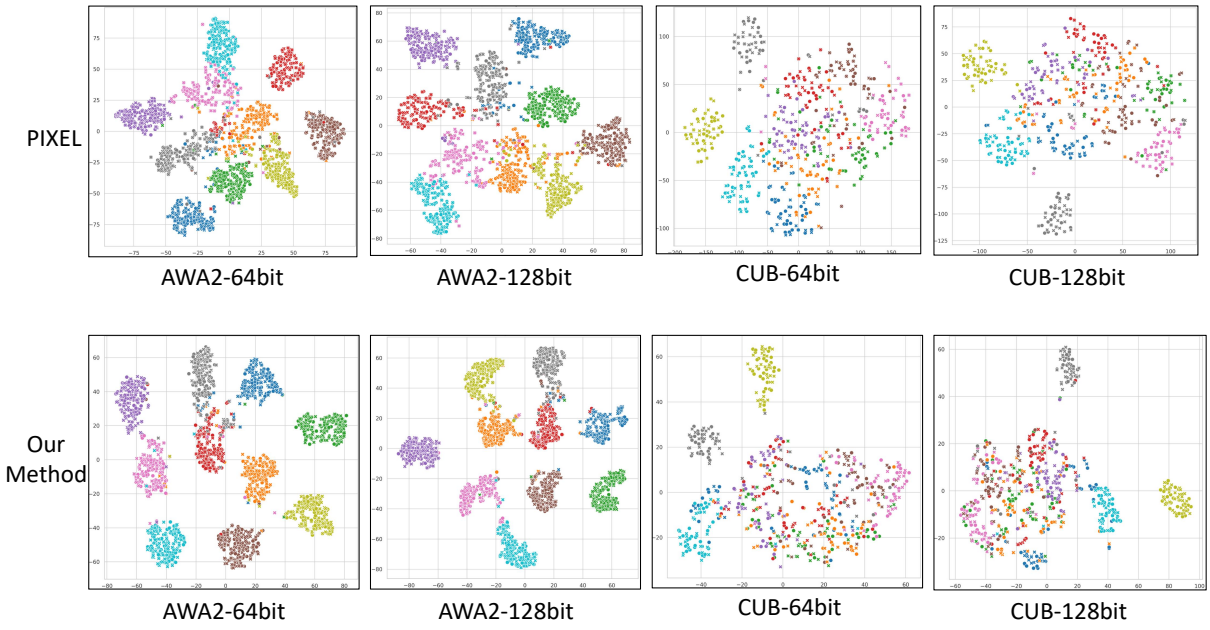

Figure 2: t-SNE visualizations of hash codes based on PIXEL and PZSH on AWA2 and CUB under 64-bit and 128-bit hash lengths. For AWA2, all 10 unseen classes are included. For CUB, 10 unseen classes are randomly selected for clarity. Circles represent query codes and crosses denote database codes.

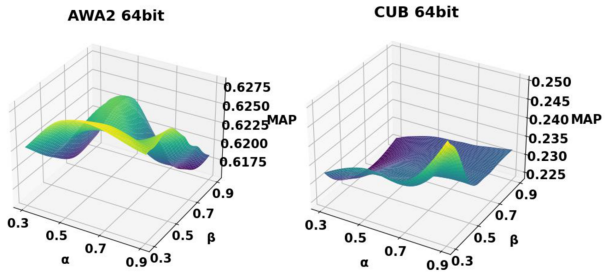

Figure 3: Hyperparameter sensitivity curves on AWA (left) and CUB (right) with shared triplets under 64-bit hash length.

T-SNE Visualizations

To qualitatively assess the semantic alignment and generalization ability of learned hash codes, we visualize the t-SNE projections of query and database embeddings from 10 unseen classes on both the AWA2 and CUB datasets under 64-bit and 128-bit settings (as shown in Figure 2). The circles denote queries and crosses represent corresponding database samples. Compared to PIXEL, PZSH exhibits noticeably more compact and semantically consistent clusters, where query points are tightly aligned with database counterparts. On the coarse-grained AWA2 dataset, semantic distinctions are preserved more effectively, while on the fine-grained CUB dataset, our method demonstrates stronger intra-class cohesion and inter-class separability. These results confirm that the proposed generative-contrastive design enables more robust Zero-Shot generalization in both coarse and fine-grained settings.

Conclusion

In this paper, we propose PZSH, a novel zero-shot image retrieval framework that integrates multimodal image synthesis via Stable Diffusion with dual-branch contrastive learning based on BLIP encoders. PZSH generates synthetic images conditioned on class-level semantic prompts and leverages BLIP-based contrastive alignment to enforce fine-grained semantic consistency between real and generated modalities. This synergy enables PZSH to construct a unified embedding space capturing both global category-level semantics and subtle attribute variations. By aligning heterogeneous modalities in a discriminative representation space, our method effectively bridges the domain gap between seen and unseen categories. Extensive experiments on AWA2 and CUB demonstrate consistent improvements over state-of-the-art methods across multiple hash lengths, especially in challenging fine-grained settings.

Acknowledgements

This work is partially supported by the National Science Foundation of China under grant No. 62202438; the Natural Science Foundation of Shandong Province Grant No. ZR2024MF128.

References

Cao, Z.; Long, M.; Wang, J.; and Yu, P. S. 2017. HashNet: Deep Learning to Hash by Continuation. In *Proceedings of the IEEE International Conference on Computer Vision*.

Chen, Y.; Wang, S.; Lu, J.; Chen, Z.; Zhang, Z.; and Huang, Z. 2021. Local Graph Convolutional Networks for Cross-Modal Hashing. In *Proceedings of the ACM International Conference on Multimedia*, 1921–1928.

Cheng, K.; Qin, Q.; Zhang, W.; Huang, L.; and Nie, J. 2025. Deep Probabilistic Binary Embedding via Learning Reliable Uncertainty for Cross-Modal Retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 6393–6402.

Cong, X.; Gui, J.; Miao, K.-C.; Zhang, J.; Wang, B.; and Chen, P. 2020. Discrete haze level dehazing network. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1828–1836.

Dong, Z.; Long, Q.; Zhou, Y.; Wang, P.; Zhu, Z.; Luo, X.; Wang, Y.; Wang, P.; and Zhou, Y. 2024. PIXEL: Prompt-based Zero-Shot Hashing via Visual and Textual Semantic Alignment. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, 487–496.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations*.

Gionis, A.; Indyk, P.; and Motwani, R. 1999. Similarity Search in High Dimensions via Hashing. In *Proceedings of the International Conference on Very Large Data Bases*, 518–529.

Gong, Y.; Lazebnik, S.; Gordo, A.; and Perronnin, F. 2013. Iterative Quantization: A Procrustean Approach to Learning Binary Codes for Large-Scale Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12): 2916–2929.

Guo, Y.; Ding, G.; Han, J.; and Gao, Y. 2017. SitNet: Discrete Similarity Transfer Network for Zero-Shot Hashing. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1767–1773.

Huo, Y.; Qin, Q.; Zhang, W.; Huang, L.; and Nie, J. 2025. Factorized Transformer Hashing with Adaptive Routing for Large-scale Image Retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 6066–6074.

Ji, Z.; Sun, Y.; Yu, Y.; Pang, Y.; and Han, J. 2020. Attribute-Guided Network for Cross-Modal Zero-Shot Hashing. *IEEE Transactions on Neural Networks and Learning Systems*, 31(1): 321–330.

Jiang, Y.; Qi, Z.; Li, J.; Qian, J.; Wang, C.; and Xin, Y. 2025. Zero-Shot Hashing based on Reconstruction with Part Alignment. arXiv:2503.07037.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the International Conference on Machine Learning*, 12888–12900.

Li, Q.; Sun, Z.; He, R.; and Tan, T. 2017. Deep Supervised Discrete Hashing. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30.

Li, W.-J.; Wang, S.; and Kang, W.-C. 2016. Feature Learning based Deep Supervised Hashing with Pairwise Labels. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1711–1717.

Li, X.; Wang, X.; Jin, B.; Zhang, W.; Wang, J.; and Zha, H. 2021. VSB2-Net: Visual-Semantic Bi-Branch Network for Zero-Shot Hashing. In *Proceedings of the International Conference on Pattern Recognition*, 1836–1843.

Li, Y.; Long, Q.; Zhou, Y.; Zhang, R.; Ning, Z.; Zhu, Z.; Zhou, Y.; Wang, X.; and Xiao, M. 2025. COMAE: COMprehensive Attribute Exploration for Zero-Shot Hashing. arXiv:2402.16424.

Li, Y.; Pei, W.; zha, Y.; and van Gemert, J. 2019. Push for Quantization: Deep Fisher Hashing. arXiv:1909.00206.

Liu, H.; Wang, R.; Shan, S.; and Chen, X. 2016a. Deep Supervised Hashing for Fast Image Retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Liu, H.; Wang, R.; Shan, S.; and Chen, X. 2016b. Deep Supervised Hashing for Fast Image Retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Liu, W.; Wang, J.; Ji, R.; Jiang, Y.-G.; and Chang, S.-F. 2012. Supervised Hashing with Kernels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2074–2081.

Luo, X.; Wang, H.; Wu, D.; Chen, C.; Deng, M.; Huang, J.; and Hua, X.-S. 2023. A Survey on Deep Hashing Methods. *ACM Trans. Knowl. Discov. Data*, 17(1).

Mi, J.-X.; Lei, D.; and Gui, J. 2013. A novel method for recognizing face with partial occlusion via sparse representation. *Optik*, 124(24): 6786–6789.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. T2I-Adapter: Learning Adapters to Dig Out More Controllable Ability for Text-to-Image Diffusion Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5): 4296–4304.

Pu, R.; Qin, Y.; Song, X.; Peng, D.; Ren, Z.; and Sun, Y. 2025a. SHE: Streaming-media Hashing Retrieval. In *Forty-second International Conference on Machine Learning*.

Pu, R.; Sun, Y.; Qin, Y.; Ren, Z.; Song, X.; Zheng, H.; and Peng, D. 2025b. Robust Self-Paced Hashing for Cross-Modal Retrieval with Noisy Labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 19969–19977.

Qiu, Q.; Lezama, J.; Bronstein, A.; and Sapiro, G. 2018. ForestHash: Semantic Hashing With Shallow Random Forests and Tiny Convolutional Networks. In *Proceedings of the European Conference on Computer Vision*.

Qiu, Z.; Pan, Y.; Yao, T.; and Mei, T. 2017. Deep Semantic Hashing with Generative Adversarial Networks. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 225–234.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning*, volume 139, 8748–8763.

Rongkai, X.; Yan, P.; Hanjiang, L.; Cong, L.; and Shuicheng, Y. 2014. Supervised Hashing for Image Retrieval via Image Representation Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1).

Shen, F.; Gao, X.; Liu, L.; Yang, Y.; and Shen, H. T. 2017. Deep Asymmetric Pairwise Hashing. In *Proceedings of the ACM International Conference on Multimedia*, 1522–1530.

Shen, F.; Shen, C.; Liu, W.; and Tao Shen, H. 2015. Supervised Discrete Hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Shen, Y.; Liu, L.; Shen, F.; and Shao, L. 2018. Zero-Shot Sketch-Image Hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Su, S.; Zhang, C.; Han, K.; and Tian, Y. 2018. Greedy Hash: Towards Fast Optimization for Accurate Hash Coding in CNN. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31.

Sun, Y.; Ren, Z.; Hu, P.; Peng, D.; and Wang, X. 2023. Hierarchical consensus hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*, 26: 824–836.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, California Institute of Technology.

Wang, J.; Liu, W.; Kumar, S.; and Chang, S.-F. 2016. Learning to Hash for Indexing Big Data—A Survey. *Proceedings of the IEEE*, 104(1): 34–57.

Wang, J.; Zhang, T.; song, j.; Sebe, N.; and Shen, H. T. 2018. A Survey on Learning to Hash. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4): 769–790.

Wang, R.; Yu, G.; Liu, L.; Cui, L.; Domeniconi, C.; and Zhang, X. 2021. Cross-modal Zero-shot Hashing by Label Attributes Embedding. arXiv:2111.04080.

Wang, S.; Chang, J.; Wang, Z.; Li, H.; Ouyang, W.; and Tian, Q. 2024. Content-Aware Rectified Activation for Zero-Shot Fine-Grained Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6): 4366–4380.

Xian, Y.; Lampert, C. H.; Schiele, B.; and Akata, Z. 2018. Zero-Shot Learning a Comprehensive Evaluation of the Good, the Bad and the Ugly. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Yang, H.-F.; Lin, K.; and Chen, C.-S. 2018. Supervised Learning of Semantics-Preserving Hash via Deep Convolutional Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(2): 437–451.

Yang, Y.; Luo, Y.; Chen, W.; Shen, F.; Shao, J.; and Shen, H. T. 2016. Zero-Shot Hashing via Transferring Supervised Knowledge. In *Proceedings of the ACM International Conference on Multimedia*, 1286–1295.

Yong, K.; Shu, Z.; Yu, J.; and Yu, Z. 2024. Zero-Shot Discrete Hashing with Adaptive Class Correlation for Cross-modal Retrieval. *Knowledge-Based Systems*, 295: 111820.

Yuan, L.; Wang, T.; Zhang, X.; Tay, F. E.; Jie, Z.; Liu, W.; and Feng, J. 2020a. Central Similarity Quantization for Efficient Image and Video Retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Yuan, L.; Wang, T.; Zhang, X.; Tay, F. E.; Jie, Z.; Liu, W.; and Feng, J. 2020b. Central similarity quantization for efficient image and video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3083–3092.

Yuan, X.; Wang, G.; Chen, Z.; and Zhong, F. 2021. CHOP: An Orthogonal Hashing Method for Zero-Shot Cross-Modal Retrieval. *Pattern Recognition Letters*, 145: 247–253.

Zhang, H.; Long, Y.; and Shao, L. 2019. Zero-Shot Hashing with Orthogonal Projection for Image Retrieval. *Pattern Recognition Letters*, 117: 201–209.

Zhang, P.-F.; Li, Y.; Huang, Z.; and Xu, X.-S. 2022. Aggregation-Based Graph Convolutional Hashing for Unsupervised Cross-Modal Retrieval. *IEEE Transactions on Multimedia*, 24: 466–479.

Zhang, W.; Wu, D.; Zhou, Y.; Li, B.; Wang, W.; and Meng, D. 2021. Binary Neural Network Hashing for Image Retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1318–1327.

Zhu, H.; Long, M.; Wang, J.; and Cao, Y. 2016. Deep Hashing Network for Efficient Similarity Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2415–2421.

Zhu, L.; Zheng, C.; Guan, W.; Li, J.; Yang, Y.; and Shen, H. T. 2024. Multi-Modal Hashing for Efficient Multimedia Retrieval: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(1): 239–260.

Zou, Q.; Cao, L.; Zhang, Z.; Chen, L.; and Wang, S. 2022. Transductive Zero-Shot Hashing for Multilabel Image Retrieval. *IEEE Transactions on Neural Networks and Learning Systems*, 33(4): 1673–1687.