Enhancing Cross-Modal Retrieval via Visual-Textual Prompt Hashing

Bingzhi Chen[1], **Zhongqi Wu**[2], **Yishu Liu**[3*], **Biqing Zeng**[2], **Guangming Lu**[3] and **Zheng Zhang**[3]
[1]Beijing Institute of Technology, Zhuhai, China
[2]South China Normal University, Guangzhou, China
[3]Harbin Institute of Technology, Shenzhen, China
chenbingzhi.smile@gmail.com, {wuzhongqi, zengbiqing}@m.scnu.edu.cn,
liuyishu.smile@gmail.com*, luguangm@hit.edu.cn, darrenzz219@gmail.com

Abstract

Cross-modal hashing has garnered considerable research interest due to its rapid retrieval and low storage costs. However, the majority of existing methods suffer from the limitations of *context loss* and *information redundancy*, particularly in simulated textual environments enriched with manually annotated tags or virtual descriptions. To mitigate these issues, we propose a novel Visual-Textual Prompt Hashing (VTPH) that aims to bridge the gap between simulated textual and visual modalities within a unified prompt optimization paradigm for cross-modal retrieval. By seamlessly integrating robust reasoning capabilities inherent in large-scale models, we design the visual and textual alignment prompt mechanisms to collaboratively enhance the contextual awareness and semantic capabilities embedded within simulated textual features. Furthermore, an affinity-adaptive contrastive learning strategy is dedicated to dynamically recalibrating the semantic interaction between visual and textual modalities by modeling the nuanced heterogeneity and semantic gaps between simulated and real-world textual environments. To the best of our knowledge, this is *the first attempt* to integrate both visual and textual prompt learning into cross-modal hashing, facilitating the efficacy of semantic coherence between diverse modalities. Extensive experiments on multiple benchmark datasets consistently demonstrate the superiority and robustness of our VTPH method over state-of-the-art competitors.

1 Introduction

With the explosive growth of multimedia data, cross-modal retrieval [Messina *et al.*, 2021] [Bogolin *et al.*, 2022] has become a hot issue and attracted continuous research attention from both academia and industry. Its primary objective is to retrieve relevant samples of one modality by the query from another modality. As a promising solution for similarity queries, cross-modal hashing (CMH) [Cao *et al.*, 2018][Zhang *et al.*, 2022] aims to map heterogeneous multimedia data into the Hamming space, ensuring that similar content possesses analogous representations within this hash space. Due to the computational efficiency and storage cost advantages, CMH has been extensively investigated in recent years for the retrieval and analysis of multimodal data.

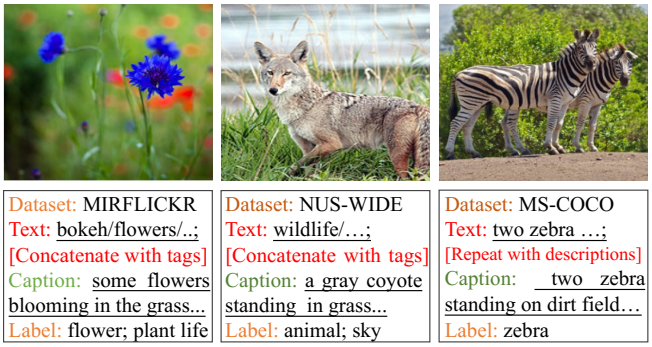


Figure 1: Illustration of image instances with the corresponding texts and captions produced by mPLUG. It is observed that texts extracted from MIRFLICKR-25K and NUS-WIDE lack contextual relationships, whereas texts utilized in MS-COCO often exhibit semantic redundancy. In contrast, captions produced by the large-scale model demonstrate a higher degree of realism in describing images.

In recent years, the rapid advancement of Deep Neural Networks (DNN) has stimulated the proposal of numerous deep cross-modal hashing (DCMH) approaches. According to the involvement of semantic label supervision signals, existing DCMH methods can be roughly categorized into two types: supervised DCMH [Zhu *et al.*, 2022] [Ou *et al.*, 2023a] and unsupervised DCMH [Luo *et al.*, 2021a][Mikriukov *et al.*, 2022][Hu *et al.*, 2023]. Technically, unsupervised DCMH methods leverage co-occurrence information to excavate consistency features from multimodal data and learn the hash function. In contrast, supervised DCMH methods generally attain superior retrieval performance than unsupervised DCMH methods by leveraging label-level information to learn more discriminative and general representations. To this end, the objective of our work is to generate discriminative unified binary codes across modalities for handling cross-modal retrieval tasks in the supervised learning paradigm.

Despite progress in learning with DCMH, it is noted that current methods are grounded in an implicit yet ideal

*Corresponding author.

assumption-namely, that **existing manually-annotated multimodal datasets can effectively simulate cross-modal retrieval scenarios akin to real-world environments**. Due to the challenges of real-world data collection from noisy websites and expensive annotation processes, the benchmark datasets used in the realm of cross-modal retrieval, including MIRFLICKR-25K [Huiskes and Lew, 2008], NUS-WIDE [Huiskes and Lew, 2008], and MS-COCO [Lin *et al.*, 2014], remains the inherent limitations posed by *context loss* and *information redundancy*, particularly in the representation of textual data. As illustrated in Figure 1, the textual content in both the MIRFLICKR-25K and NUS-WIDE datasets is generated by a straightforward concatenation of multiple tags, leading to a deficiency in contextual information. Furthermore, the textual content in the MS-COCO dataset is composed by merging multiple manually annotated image descriptions, which may introduce a certain level of semantic redundancy. The challenges mentioned above pose obstacles to the accuracy and generalization capabilities of existing DCMH methods when applied in real-world scenarios.

To address these challenges, this paper proposes a novel Visual-Textual Prompt Hashing (VTPH) paradigm that leverages the powerful semantic reasoning capabilities of large-scale vision-language models to reconstruct and enrich simulated textual environments, leading to a more effective and robust cross-modal retrieval process. In comparison to the state-of-the-art methods [Liu *et al.*, 2023b][Tu *et al.*, 2022][Ou *et al.*, 2023b], the proposed VTPH approach mainly benefits from the advantages of cross-modal prompt engineering. Specifically, two well-established prompt mechanisms, i.e., visual alignment prompt (VAP) and textual alignment prompt (TAP), are proposed to collaboratively enhance the contextual awareness and semantic capabilities embedded within simulated textual features. On the one hand, the VAP mechanism is meticulously designed to enhance salient features and suppress irrelevant information for text representations under the guidance of the global context from visual features. On the other hand, the TAP mechanism aims to capture more authentic and context-rich textual representations by bridging the semantic gap between simulated texts and captions generated by the large-scale model mPLUG [Li *et al.*, 2022]. By incorporating visual and textual prompts within a unified framework, our VTPH approach aims to optimize the interaction and alignment between different modalities based on the assumption of a scenario that closely resembles real-world environments. Furthermore, we propose an affinity-adaptive contrastive learning module to explicitly model the nuanced heterogeneity and semantic gaps between simulated and real-world textual environments. It can dynamically recalibrate the semantic interaction between visual and textual modalities, providing a more accurate representation of the semantic relationships for cross-modal retrieval.

To the best of our knowledge, our work is *the first attempt* that incorporates both visual-textual prompt learning to address the limitations of context loss and information redundancy within the domain of cross-modal retrieval. Our VTPH framework is comprehensively evaluated on multiple large-scale datasets, and extension experiments are also conducted to demonstrate the robustness of VTPH on benchmark datasets with noisy correspondences. The promising performance collectively demonstrates the effectiveness and superiority of our VTPH method over state-of-the-art algorithms.

2 Related Work

2.1 Deep Cross-Modal Hashing

Cross-modal hashing (CMH) [Cao *et al.*, 2018][Yao *et al.*, 2021][Zhang *et al.*, 2022] aims to learn hash functions that map raw data into a binary hash space, along with the hash metric intended to preserve semantic similarity between original multimedia data. Benefiting from the powerful representation abilities of deep neural networks, deep cross-modal hashing (DCMH) [Zhu *et al.*, 2022] [Liu *et al.*, 2023b] has gained significant attention in addressing large-scale cross-modal retrieval tasks. Due to the absence of semantic information, the retrieval performance of unsupervised DCMH remains unsatisfactory. By contrast, numerous supervised DCMH methods have been introduced to utilize manual annotation label information to facilitate the learning of hash functions. Specifically, one group of these studies is based on the CNN framework, such as DADH [Bai *et al.*, 2020], MSSPQ [Zhu *et al.*, 2022], and CMGCAH [Ou *et al.*, 2023b]. With the proposal of large-scale vision-language architectures (i.e., CLIP), transformer-based methods as another group have achieved more promising performances than traditional CNN-based methods. For instance, DCMHT [Tu *et al.*, 2022] firstly employed a visual transformer for encoding image content to improve the correlation modeling of CMH. In particular, MITH [Liu *et al.*, 2023b] hierarchically considered intra-modal interaction and inter-modal alignment with multi-granularity in one unified transformer-based framework. Additionally, CIMON [Luo *et al.*, 2021b] presents a novel approach to explore noisy data scenarios, potentially enhancing the robustness against data imperfections. However, the existing methods rely on simulated environments crafted from manually annotated datasets, which suffer from the challenges related to context loss and information redundancy.

2.2 Prompt Learning

Initially originating from the natural language processing (NLP) domain, prompt learning leverages pre-trained models to undertake downstream tasks by seamlessly incorporating handcrafted prompt templates into the input [Brown *et al.*, 2020]. Nevertheless, the meticulous design of crafted prompt templates requires extensive domain expertise and common knowledge, ultimately limiting the model's flexibility. In contrast, prompt tuning enables the model to adapt prompts as continuous vectors and optimize them directly during the fine-tuning process. Inspired by the significant success of prompt tuning in NLP, recent studies [Zhou *et al.*, 2022b] [Zhou *et al.*, 2022a] [Khattak *et al.*, 2023] have tried to incorporate the concept of prompt learning in vision-language models. Based on the CLIP pre-trained vision-language architecture, CoOp [Zhou *et al.*, 2022b] was the earliest work to apply trainable text prompt vectors for few-shot transferring. Co-CoOp [Zhou *et al.*, 2022a] introduced conditional context optimization that dynamically made a prompt conditioned on

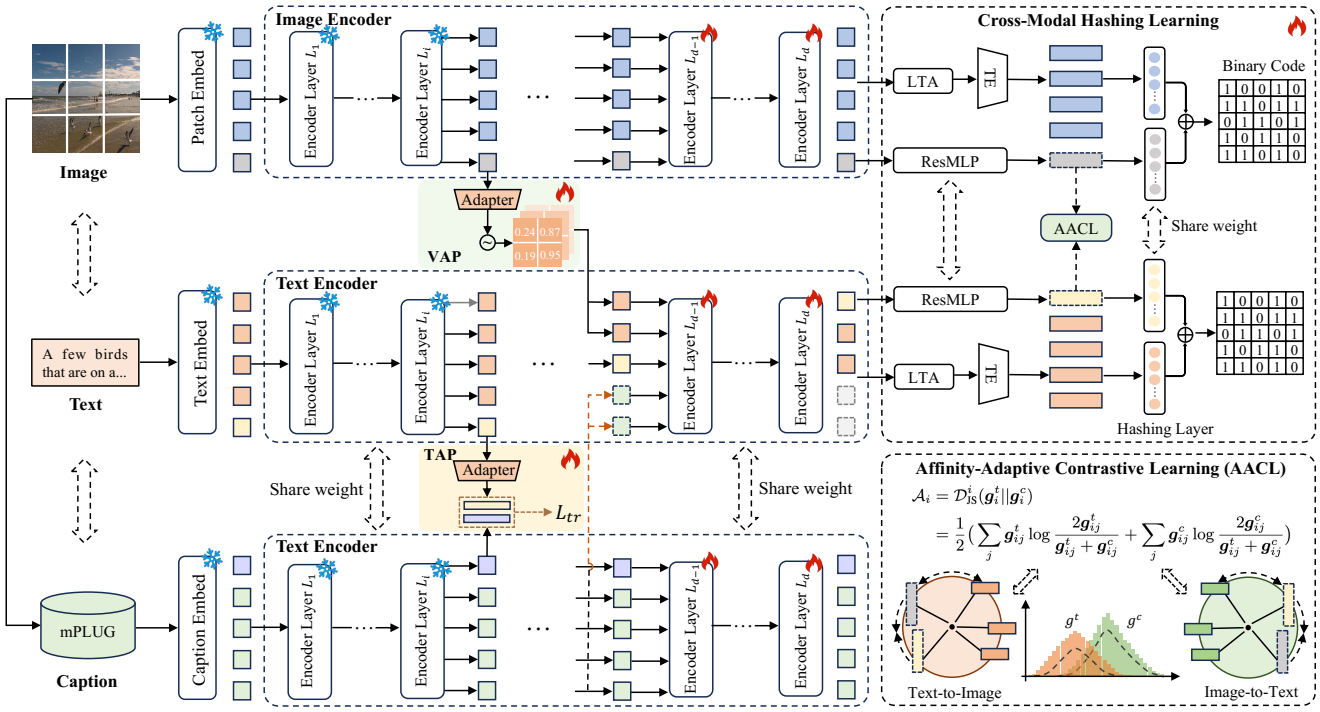


Figure 2: Illustration of the proposed Visual-Textual Prompt Hashing (VTPH) framework for cross-modal retrieval. Two prompt strategies, i.e., visual alignment prompt and textual alignment prompt, are dexterously established in the inter and intra-modal interaction phases. Meanwhile, an affinity-adaptive contrastive learning module is designed to model the heterogeneity and semantic gaps across modalities.

each input image instead of fixed ones. Recently, MaPLe [Khattak *et al.*, 2023] and DCP [Liu *et al.*, 2023a] facilitated a strong coupling between visual and language prompts to ensure their mutual collaboration. However, whether prompt learning is effective for cross-modal hashing remains under-explored and investigated. In this paper, we first attempt to design a multi-modal prompt learning paradigm to enhance the interaction between visual and textual representations for tackling the challenge of cross-modal hashing.

3 Methodology

3.1 Notations and Problem Formulation

In the context of N image-text pairs denoted as $\mathcal{P} = \{\mathcal{P}_i\}_i^N$, where $\mathcal{P}_i = \{(\boldsymbol{v}_i, \boldsymbol{t}_i); \boldsymbol{c}_i; \boldsymbol{l}_i\}$, $\boldsymbol{v}_i$ and $\boldsymbol{t}_i$ represents the image and text modalities of the i-th image-text pair, $\boldsymbol{c}_i$ denotes the image captions generated offline by mPLUG, and $\boldsymbol{l}_i$ indicates the associated label matrix with $\mathcal{Q}$ categories. In our work, the similarity matrix $\mathcal{S}$ for cross-modal retrieval is generated based on labels. With the supervision guidance of $\mathcal{S}$, the objective of this study is to acquire unified hash codes by projecting both image and text data from a high-dimensional space into a common K-bit discrete Hamming space, where K is the length of the hash codes. Based on the Hamming distance between similar instances, we aim to learn two hashing functions, i.e., $\boldsymbol{b}_i^v = \mathcal{H}^v(v_i; \theta^v) \in \{-1, +1\}^K$ and $\boldsymbol{b}_i^t = \mathcal{H}^t(t_i; \theta^t) \in \{-1, +1\}^K$, where $\boldsymbol{b}_i^v$ and $\boldsymbol{b}_i^t$ represent the learned hash codes to preserve the semantic similarities between visual and textual modalities, θ_v and θ_t denote the trainable parameters during the prompt-tuning stage.

3.2 Modality-Specific Feature Embedding

Figure 2 provides a detailed pipeline of the proposed VTPH framework. Following previous work [Liu *et al.*, 2023b], our VTPH framework adopts the pre-trained Vision Transformer (VIT) [Dosovitskiy *et al.*, 2021] and GPT-2 [Radford *et al.*, 2019] as image and text encoders to extract modality-specific feature representations, i.e., $\mathcal{F}_i^v = [\boldsymbol{g}_i^v, \boldsymbol{z}_i^v]$, $\mathcal{F}_i^t = [\boldsymbol{g}_i^t, \boldsymbol{z}_i^t]$, and $\mathcal{F}_i^c = [\boldsymbol{g}_i^c, \boldsymbol{z}_i^c]$, where $\boldsymbol{g}$ represents the global class embeddings, and $\boldsymbol{z}$ represents the sequence of local token embeddings. It is noted that the obtained image captions can be considered as additional text modalities to augment the original textual features. In this part, we implement a weight-sharing strategy between the text encoder and caption encoder to ensure the consistency of textual representations.

3.3 Visual-Textual Prompt Learning

By keeping the backbone network fixed during fine-tuning, we incorporate two well-designed prompt mechanisms, i.e., visual alignment prompt (VAP) and textual alignment prompt (TAP), to jointly enrich the contextual content and semantic representations of simulated textual features.

Visual Alignment Prompt. The goal of the VAP component is to enhance the salient features and suppress irrelevant information within textual features under the guidance of global embeddings from images. Specifically, the distribution of textual feature information is reshaped by applying the attention prompts to local text features from the image branch. The prompted feature $\overline{\boldsymbol{z}}_i^t$ is formulated as:

$$\overline{\boldsymbol{z}}_i^t = \boldsymbol{z}_i^t \cdot \mathrm{Softmax}(\boldsymbol{g}_i^v \cdot \mathcal{W}_{\mathrm{vap}}), \tag{1}$$

where $\mathcal{W}_{\text{vap}} \in \mathbb{R}^{d\times d}$ are learnable weights from the visual adapter, d represents the dimensionality of feature representations. During the adaptation process, the dimensions of both input and output remain constant. Importantly, the adapted features could seamlessly replace the original local text features as enhanced text features, that effectively filter out irrelevant information while retaining discriminative features to facilitate the subsequent multimodal fusion process.

Textual Alignment Prompt. To harness the robust reasoning capabilities inherent in large-scale models, we design the TAP component to augment the semantic capabilities within simulated textual features and align them more closely with real-world scenarios. Following previous work [Pei *et al.*, 2023], the prompted feature $\overline{\boldsymbol{g}}_i^t$ is adapted as follows:

$$\overline{\boldsymbol{g}}_i^t = \text{RELU}(\boldsymbol{g}_i^t \cdot \mathcal{W}_{\text{tap}}^1) \cdot \mathcal{W}_{\text{tap}}^2, \tag{2}$$

where $\mathcal{W}_{\text{tap}}^1 \in \mathbb{R}^{d\times h}$ and $\mathcal{W}_{\text{tap}}^2 \in \mathbb{R}^{h\times d}$ are learnable weights serving as textual adapters, $h = d/4$ is hidden dimensionality of feature representations. Subsequently, the output $\overline{\boldsymbol{g}}_i^t$ is directly supplied to the forward process as the global textual feature. Instead of employing a straightforward projection for attention weight generation, we introduce a cosine triplet contrastive learning objective [Khan *et al.*, 2023] to optimize the global text features, ensuring a more seamless alignment with authentic textual expressions,

$$\begin{aligned}\mathcal{L}_{\text{triplet}} = \sum_{i=1}^{M}\sum_{j=1}^{M}\big([1- <\overline{\boldsymbol{g}}_i^t, \boldsymbol{g}_i^c> + <\overline{\boldsymbol{g}}_i^t, \boldsymbol{g}_j^c>]_+ \\ +[1- <\overline{\boldsymbol{g}}_i^t, \boldsymbol{g}_i^c> + <\overline{\boldsymbol{g}}_j^t, \boldsymbol{g}_i^c>]_+\big),\end{aligned} \tag{3}$$

where M is the number of batch size, $i \neq j$, $<\cdot,\cdot>$ denotes the cosine simarility function, and $[\boldsymbol{x}]_+ = \max(0, \boldsymbol{x})$.

Visual-Textual Collaboration. To obtain more comprehensive and realistic feature representations, we insert $\boldsymbol{z}_i^c$ extracted from image captions into the text encoder as additional textual tokens. Notably, we selectively unfreeze the last two layers of the transformer-based encoder to fully utilize the informative cues provided by prompts and enhance the overall learning capability. Formally, the proposed visual and textual prompt mechanisms can be formulated as:

$$\begin{aligned}[\overline{\boldsymbol{g}}_{i,d-1}^t, \overline{\boldsymbol{z}}_{i,d-1}^t, \boldsymbol{z}_{i,d-1}^c] &= \text{SA}\big([\overline{\boldsymbol{g}}_{i,d-2}^t, \overline{\boldsymbol{z}}_{i,d-2}^t, \boldsymbol{z}_{i,d-2}^c]\big) \\ [\overline{\boldsymbol{g}}_{i,d}^t, \overline{\boldsymbol{z}}_{i,d}^t, -] &= \text{SA}\big([\overline{\boldsymbol{g}}_{i,d-1}^t, \overline{\boldsymbol{z}}_{i,d-1}^t, \boldsymbol{z}_{i,d-1}^c]\big),\end{aligned} \tag{4}$$

where d represents the depth of the self-attention (SA) layer in the text encoder, $[\cdot,\cdot]$ represents the concatenation function, and the symbol "$-$" indicates that the output tokens at the corresponding positions are discarded.

3.4 Affinity-Adaptive Contrastive Learning

Based on the aforementioned operations, both global embeddings $\boldsymbol{g}_i^*$ provided by visual and textual modalities are applied to the residual multi-layer perceptrons (ResMLP) to align them to the same dimensionalities,

$$\widetilde{\boldsymbol{g}}_i^{(*)} = \text{ResMLP}(\overline{\boldsymbol{g}}_i^{(*)}). \tag{5}$$

Simultaneously, the local embeddings $\boldsymbol{z}_i^{(*)}$ employ a localized token aggregation (LTA) strategy [Liu *et al.*, 2023b] to localize the preservation of the most crucial implicit semantic knowledge from the global embeddings by selecting the top-m features of high confidence to form embedding,

$$\widetilde{\boldsymbol{z}}_i^{(*)} = \text{TE}\big((\boldsymbol{W}_i^{(*)})^\top \overline{\boldsymbol{z}}_i^{(*)}\big), \tag{6}$$

where TE represents a two-layer transformer encoder.

To mitigate the heterogeneity and semantic gaps between simulated and real-world textual environments, we design the Affinity-Adaptive Contrastive Learning (AACL) module to dynamically recalibrate the semantic interaction between visual and textual modalities. Different from traditional contrastive learning [Chen *et al.*, 2020] [He *et al.*, 2020] [Grill *et al.*, 2020], the affinity $\mathcal{A}_i$ is specifically designed to capture the nuanced heterogeneity and semantic gaps between simulated and caption textual environments. Specifically, we employ Jensen-Shannon divergence [Lin, 1991] to compute the affinity between the global features $\boldsymbol{g}_i^t$ extracted from image captions and original global text features $\boldsymbol{g}_i^c$,

$$\begin{aligned}\mathcal{A}_i &= \mathcal{D}_{\text{JS}}^i(\boldsymbol{g}_i^t || \boldsymbol{g}_i^c) \\ &= \frac{1}{2}\Big(\sum_j \boldsymbol{g}_{ij}^t \log\frac{2\boldsymbol{g}_{ij}^t}{\boldsymbol{g}_{ij}^t + \boldsymbol{g}_{ij}^c} + \sum_j \boldsymbol{g}_{ij}^c \log\frac{2\boldsymbol{g}_{ij}^c}{\boldsymbol{g}_{ij}^t + \boldsymbol{g}_{ij}^c}\Big),\end{aligned} \tag{7}$$

where $\mathcal{D}_{\text{JS}} \in [0, 1]$ indicates the Jensen-Shannon divergence.

Given the i-th image-text pairs $(\widetilde{\boldsymbol{g}}_i^v, \widetilde{\boldsymbol{g}}_i^t)$ in a minibatch, we treat two modality data as queries and keys alternatively to learn the positive image-text pairs and the remaining pairs as considered negative samples. By automatically adjusting the temperature hyperparameter to fine-tune the strength of traditional contrastive learning, the objective of the AACL module can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{\text{aacl}} = \frac{1}{M}\sum_{i=1}^{M}\Big(&- log\frac{\exp((\widetilde{\boldsymbol{g}}_i^v)^\top \widetilde{\boldsymbol{g}}_i^t/\widehat{\tau}_i)}{\sum_{c=1}^{M}\exp((\widetilde{\boldsymbol{g}}_i^v)^\top \widetilde{\boldsymbol{g}}_c^t/\widehat{\tau}_i)} \\ &- log\frac{\exp((\widetilde{\boldsymbol{g}}_i^t)^\top \widetilde{\boldsymbol{g}}_i^v/\widehat{\tau}_i)}{\sum_{c=1}^{M}\exp((\widetilde{\boldsymbol{g}}_i^t)^\top \widetilde{\boldsymbol{g}}_c^v/\widehat{\tau}_i)}\Big),\end{aligned} \tag{8}$$

where $\widehat{\tau}_i$ denotes the temperature parameter, which is adaptively determined by the affinity $\mathcal{A}_i$,

$$\widehat{\tau}_i = \tau + \gamma \cdot \mathcal{A}_i. \tag{9}$$

By multiplying the affinity value $\mathcal{A}_i$ with a hyperparameter γ, the temperature can be automatically updated in a residual manner. In this way, the dynamic bridging of heterogeneity and semantic gaps across different modalities brings the representation closer to real-world scenarios.

3.5 Cross-Modal Hashing Learning

The primary purpose of the cross-modal hashing module is to map features into the Hamming space and ensure that the distance relationships between hash codes reflect the semantic similarity of different modalities [Jiang and Li, 2017] [Cao *et al.*, 2018] [Tu *et al.*, 2022]. To this end, a hashing linear projection layer (HashLayer) with the $tanh$ activation function is designed to decompose the projected features to global semantic features $\boldsymbol{h}_i^{(*)}$, i.e.,

$$\boldsymbol{h}_i^{(*)} = \text{HashLayer}(\widetilde{\boldsymbol{g}}_i^{(*)}). \tag{10}$$

Similarly, the linear projection hashing layer is designed to map $\widetilde{\boldsymbol{z}}_i^{(*)}$ to the K-bit Hamming space as local semantic features $\boldsymbol{f}_i^{(*)}$, which can be expressed as:

$$\boldsymbol{f}_i^{(*)} = \text{HashLayer}(\widetilde{\boldsymbol{z}}_i^{(*)}). \tag{11}$$

To learn the unified hash code, we employ the sign function to integrate global and local semantic features to collaboratively accomplish the encoding process,

$$\boldsymbol{b}_i = \text{sign}(\lambda(\boldsymbol{h}_i^v + \boldsymbol{h}_i^t) + (1-\lambda)(\boldsymbol{f}_i^v + \boldsymbol{f}_i^t)), \tag{12}$$

where $\lambda \in [0, 1]$ denotes a tunable weight parameter. Moreover, the quantization loss is used to learn a uniform semantic representation and generate compact hash codes, and the objective functions can be defined as follows,

$$\mathcal{L}_{\text{quan}} = \frac{1}{KM}\sum_{i=1}^{M}\left(\|\boldsymbol{b}_i - \frac{1}{2}(\boldsymbol{h}_i^v + \boldsymbol{f}_i^v)\|_2^2 + \|\boldsymbol{b}_i - \frac{1}{2}(\boldsymbol{h}_i^t + \boldsymbol{f}_i^t)\|_2^2\right). \tag{13}$$

Inspired by [Liu *et al.*, 2023b], both intra-modal similarity preservation loss and inter-modal similarity preservation loss are introduced in our cross-modal hashing learning. In particular, the intra-modal similarity preservation loss aims to preserve semantic similarities within modalities,

$$\mathcal{L}_{\text{intra}} = -\frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\mathcal{S}_{ij}\Omega_{ij}^{(*)} - log\left(1 + e^{\Omega_{ij}^{(*)}}\right)\right), \tag{14}$$

where $\Omega_{ij}^{(*)} = \frac{1}{2}(\boldsymbol{f}_i^{(*)})^T\boldsymbol{f}_j^{(*)}$ indicates the inner product among local semantic representations. Meanwhile, inter-modal similarity preservation is designed to preserve semantic similarities across modalities, that is,

$$\begin{aligned}\mathcal{L}_{\text{inter}} = -\frac{1}{MN}\Big(&\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\mathcal{S}_{ij}\Theta_{ij} - log\left(1 + e^{\Theta_{ij}}\right)\right) \\ &+\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\mathcal{S}_{ij}\Phi_{ij} - log\left(1 + e^{\Phi_{ij}}\right)\right)\Big),\end{aligned} \tag{15}$$

where $\Theta_{ij} = \frac{1}{2}(\boldsymbol{h}_i^t)^T\boldsymbol{h}_j^v$ and $\Phi_{ij} = \frac{1}{2}(\boldsymbol{h}_i^v)^T\boldsymbol{h}_j^t$ denotes the inner product among global semantic representations. Thus, the overall objective of cross-modal hashing learning is defined as follows:

$$\mathcal{L}_{\text{hash}} = \alpha \cdot \mathcal{L}_{\text{inter}} + \beta \cdot \mathcal{L}_{\text{quan}} + \mathcal{L}_{\text{intra}}, \tag{16}$$

where α and β are trade-off hyper-parameters.

3.6 Training and Optimization

Based on the above analyses, the comprehensive training objective of the proposed VPTH approach encompasses a combination of various loss functions, i.e.,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{triplet}} + \mathcal{L}_{\text{aacl}} + \mathcal{L}_{\text{hash}}. \tag{17}$$

By jointly optimizing these losses, our approach can further enhance the performance of cross-modal retrieval, ensuring better adaptability to noisy data in real-world scenarios.

4 Experiments

4.1 Experimental Settings

Datasets. Three commonly used multi-label image-text cross-modal datasets, i.e., MIRFLICKR-25K, NUS-WIDE, and MS-COCO, are selected for our experiments. In addition, our settings follow the data splitting protocol used in [Tu *et al.*, 2022][Liu *et al.*, 2023b], which are shown in the supplementary document for details. To demonstrate the robustness of our VTPH approach, we also conduct extension experiments on these benchmark datasets with randomly 30% noisy correspondence, which are referred to as MIRFLICKR-25K-N, NUS-WIDE-N, and MS-COCO-N [Huang *et al.*, 2021].

Metrics. To comprehensively evaluate the performance of our method, we perform two cross-modal retrieval tasks, i.e., image-to-text retrieval (I $\rightarrow$ T) and text-to-image retrieval (T $\rightarrow$ I). These tasks involve searching relevant texts by using images as queries and vice versa. As standard evaluation metrics, the mean Average Precision (mAP) and precision-recall curve (PR-curve) are also considered to measure the effectiveness of different methods under the hamming ranking protocol and hash lookup protocol, respectively.

Baselines. In our experiments, we conduct a comprehensive comparison with the state-of-the-art DCMH methods, including CNN-based methods and transformer-based methods. Specifically, the CNN-based methods consist of DCMH [Jiang and Li, 2017], SSAH [Li *et al.*, 2018], GCH [Xu *et al.*, 2019], DADH [Bai *et al.*, 2020], TEACH [Yao *et al.*, 2021], MSSPQ [Zhu *et al.*, 2022], and CMGCAH [Ou *et al.*, 2023b]. In addition, the transformer-based methods contain DCMHT [Tu *et al.*, 2022] and MITH [Liu *et al.*, 2023b].

4.2 Comparisons with State-of-The-Art

Hamming Ranking Protocol. To evaluate the effectiveness of our VTPH framework, we conduct a comprehensive comparison with a range of state-of-the-art baselines on three benchmark datasets by the image-to-text and text-to-image retrieval task. The comparative results are summarized in Table 1. We can obtain the following observations: 1) The proposed VTPH method outperforms all state-of-the-art baselines with significant performance improvements across all hash code lengths. 2) In particular, the latest Transformer-based framework, namely MITH, surpasses all classic previous works such as DADH, DCHMT, and CMGCAH. This phenomenon demonstrates that the multi-modal CLIP architecture has stronger feature extraction and discriminative hash code learning capabilities than the CNN-based baseline. 3) Despite the reliable performance achieved by MITH, our VTPH approach, with its visual-textual prompt tuning strategy and affinity-adaptive contrastive learning, demonstrates superior performance. Importantly, it can consistently achieve the best performance and surpass the second-best method by the mean mAP of 6.02%, 6.97%, 1.34% for I $\rightarrow$ T, and 8.13%, 5.55%, 1.36% for T $\rightarrow$ I, respectively. These findings underscore the efficacy of our approach in leveraging visual-textual prompts to enhance text feature representations, thereby optimizing the interaction and alignment across modalities for the learning of discriminative hash codes.

	Methods	Reference	MIRFLICKR-25K				NUS-WIDE				MS-COCO			
			16bits	32bits	64bits	mean	16bits	32bits	64bits	mean	16bits	32bits	64bits	mean
I → T	DCMH	CVPR'2017	0.7321	0.7464	0.7465	0.7416	0.5493	0.5925	0.6178	0.5865	0.4928	0.5007	0.5145	0.5026
	SSAH	CVPR'2018	0.7641	0.7790	0.7867	0.7766	0.6318	0.6542	0.6488	0.6449	0.5520	0.5884	0.5893	0.5765
	AGAH	ICMR'2019	0.7625	0.7805	0.7902	0.7777	0.5776	0.5706	0.6246	0.5909	0.5663	0.6011	0.6065	0.5913
	GCH	IJCAI'2019	0.7514	0.7620	0.7672	0.7601	0.6054	0.6343	0.6241	0.6212	0.5581	0.6010	0.5973	0.5854
	DADH	ICMR'2020	0.7876	0.8027	0.8128	0.8010	0.6447	0.6739	0.6736	0.6640	0.6233	0.6458	0.6498	0.6396
	DCHMT	MM'2022	0.8217	0.8262	0.8280	0.8252	0.6685	0.6709	0.6876	0.6757	0.6081	0.6156	0.6328	0.6188
	MSSPQ	ICMR'2022	0.7868	0.8011	0.8172	0.8017	0.6346	0.6478	0.6615	0.6480	0.5710	0.5881	0.5862	0.5818
	CMGCAH	TITS'2023	0.7901	0.8030	0.8123	0.8018	0.6213	0.6440	0.6462	0.6372	–	–	–	
	MITH	MM'2023	0.8464	0.8645	0.8718	0.8609	0.7004	0.7148	0.7297	0.7150	0.7039	0.7359	0.7664	0.7354
	VTPH	Ours	**0.9056**	**0.9249**	**0.9328**	**0.9211**	**0.7733**	**0.7870**	**0.7936**	**0.7846**	**0.7202**	**0.7477**	**0.7786**	**0.7488**
		Increased↑	**5.92%**	**6.04%**	**6.10%**	**6.02%**	**7.29%**	**7.22%**	**6.39%**	**6.97%**	**1.63%**	**1.18%**	**1.20%**	**1.34%**
T → I	DCMH	CVPR'2017	0.7640	0.7725	0.7757	0.7707	0.5675	0.5829	0.6200	0.5901	0.5268	0.5393	0.5327	0.5329
	SSAH	CVPR'2018	0.7790	0.7885	0.8041	0.7905	0.6484	0.6645	0.6677	0.6602	0.5589	0.5819	0.5844	0.5750
	AGAH	ICMR'2019	0.7590	0.7732	0.7890	0.7737	0.5441	0.5367	0.5894	0.5567	0.5492	0.5848	0.5948	0.5762
	GCH	IJCAI'2019	0.7644	0.7789	0.7863	0.7765	0.6063	0.6321	0.6281	0.6221	0.5543	0.6035	0.5908	0.5828
	DADH	ICMR'2020	0.7876	0.8053	0.8166	0.8031	0.6489	0.6752	0.6932	0.6724	0.6076	0.6270	0.6336	0.6227
	DCHMT	MM'2022	0.7896	0.7972	0.7974	0.7947	0.6863	0.6893	0.6993	0.6916	0.6088	0.6140	0.6308	0.6179
	MSSPQ	ICMR'2022	0.7946	0.7885	0.8022	0.7951	0.6312	0.6631	0.6882	0.6608	0.5472	0.5630	0.5985	0.5696
	CMGCAH	TITS'2023	0.7823	0.7932	0.8045	0.7933	0.6782	0.6801	0.6844	0.6809	–	–	–	
	MITH	MM'2023	0.8228	0.8389	0.8468	0.8362	0.7140	0.7276	0.7401	0.7272	0.7017	0.7358	0.7661	0.7345
	VTPH	Ours	**0.9020**	**0.9226**	**0.9278**	**0.9175**	**0.7708**	**0.7840**	**0.7933**	**0.7827**	**0.7185**	**0.7515**	**0.7743**	**0.7481**
		Increased↑	**7.92%**	**8.37%**	**8.10%**	**8.13%**	**5.68%**	**5.64%**	**5.32%**	**5.55%**	**1.68%**	**1.57%**	**0.82%**	**1.36%**

Table 1: Comparison of mAP scores on MIRFLICKR-25K, NUS-WIDE, and MS-COCO datasets, where the best and second-best results are highlighted in **bold** and underlined, respectively. Additionally, '–' denotes the unavailable results due to the unreleased codes.

Hash Lookup Protocol. To verify the performance of our VTPH under the lookup protocol, we calculate the PR-curve metric for the returned instances, and the comparison results with variations of different hash codes (i.e., 16bits, 32bits, 64bits) on three datasets are illustrated in Figure 3. From these figures, it can be observed that our proposed method consistently achieves the best retrieval results in comparison with all the state-of-the-art baselines over three datasets.

4.3 Extended Robustness Evaluation

We further investigate the robustness of our VTPH approach in noisy environments. Following the cross-modal matching studies [Huang *et al.*, 2021] [Qin *et al.*, 2022] [Yang *et al.*, 2023], we perform extension experiments on the MIRFLICKR-25K-N, NUS-WIDE-N, and MS-COCO-N datasets with 16 bits of 30% noisy correspondence. Particularly, we generate synthetic noisy correspondence by randomly shuffling the training images and captions to simulate real-world environments. The comparative results are summarized in Table 2. We can observe that all methods suffer from varying degrees of performance degradation under the influence of noisy data. Nonetheless, the proposed method consistently achieves competitive performance with all cases on three datasets. Specifically, our VTPH yields an improvement of 8.45%, 11.52%, 6.6% for I → T and 10.77%, 10.11%, 6.61% for T → I in average mAP than the second-best method, respectively. Moreover, our VTPH approach exhibits the least impact on the experimental results in most cases, indicating that our method can mitigate the negative impact of noisy correspondence to a certain extent.

	Methods	MIRFLICKR -25K-N	NUS-WIDE-N	MS-COCO-N
I → T	DCHM	0.7146 (-1.75%)	0.5139 (-3.45%)	0.4440 (-4.88%)
	SSAH	0.7060 (-5.81%)	0.5596 (-7.22%)	0.4988 (-5.32%)
	AGAH	0.6813 (-8.12%)	0.5351 (-4.52%)	0.5235 (-4.28%)
	DADH	0.7124 (-7.52%)	0.5742 (-7.05%)	0.5880 (-3.53%)
	DCHMT	0.8101 (-1.16%)	0.6498 (-1.87%)	0.5976 (**-1.05%**)
	MITH	0.7633 (-8.31%)	0.6508 (-4.96%)	0.6265 (-9.37%)
	VTPH	0.8946 (**-1.10%**)	0.7660 (**-0.73%**)	0.6925 (-2.77%)
T → I	DCHM	0.7303 (-3.37%)	0.5453 (-2.22%)	0.4862 (-4.06%)
	SSAH	0.7520 (-2.70%)	0.6058 (-4.26%)	0.5489 (-1.00%)
	AGAH	0.7259 (-3.31%)	0.5134 (-3.07%)	0.5270 (-2.22%)
	DADH	0.7509 (-3.67%)	0.5947 (-5.42%)	0.5983 (**-0.93%**)
	DCHMT	0.7490 (-4.06%)	0.6449 (-4.14%)	0.5659 (-4.29%)
	MITH	0.7870 (-3.58%)	0.6637 (-5.03%)	0.6260 (-9.25%)
	VTPH	0.8947 (**-0.73%**)	0.7648 (**-0.60%**)	0.6921 (-2.64%)

Table 2: Comparison of mAP scores using 16 bits on MIRFLICKR-25K-N, NUS-WIDE-N, and MS-COCO-N datasets containing 30% randomly assigned corresponding noise. In particular, the changes in mAP from clean data to noisy data are shown in parentheses.

4.4 Ablation Study

In this part, we conduct comprehensive ablation studies by systematically evaluating the impact of each component in VTPH on the MIRFLICKR-25K dataset. Five variations are involved, including a) **BASE** is regarded as the base model that only utilizes two transformer-based encoders of CLIP and hash layers for hashing learning; b) **BASE + VAP** adds the visual alignment prompt component based on the basic model; c) **BASE + TAP** adds the textual alignment prompt component to the basic model, along with the image caption branch; d) **BASE + VAP + TAP** integrate both the visual and textual prompt learning strategy into the base model; d) **VTPH** is considered as the "full" model, incorporating the affinity-adaptive contrastive learning.

The comparative results are presented in Table 3. It can be observed that both the visual alignment prompt and textual alignment prompt can work cooperatively with different

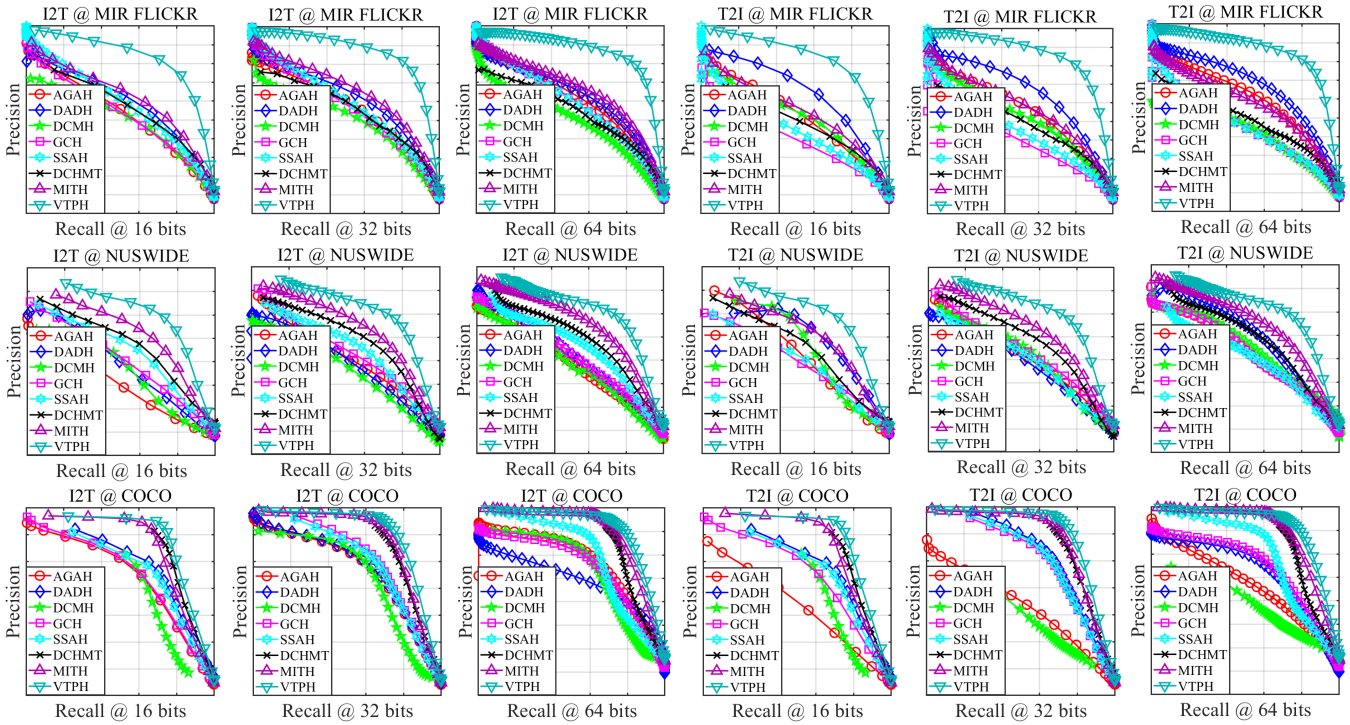


Figure 3: The PR curves w.r.t. different code lengths on the MIRFLICKR-25K, NUS-WIDE and MS-COCO datasets.

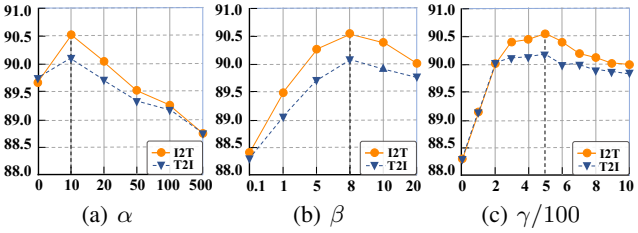
	Methods	mAP Scores			
		16bits	32bits	64bits	mean
I → T	BASE	0.8421	0.8606	0.8649	0.8559
	BASE+VAP	0.8655	0.8966	0.9045	0.8889
	BASE+TAP	0.8803	0.9035	0.9125	0.8988
	BASE+VAP+TAP	0.8830	0.9125	0.9249	0.9068
	VTPH	**0.9056**	**0.9249**	**0.9328**	**0.9211**
T → I	BASE	0.8224	0.8366	0.8441	0.8344
	BASE+VAP	0.8541	0.8857	0.9041	0.8813
	BASE+TAP	0.8767	0.8980	0.9030	0.8926
	BASE+VAP+TAP	0.8834	0.9132	0.9247	0.9071
	VTPH	**0.9020**	**0.9226**	**0.9278**	**0.9175**

Table 3: Comparison of mAP scores on the MIRFLICKR-25K dataset with different components.

hash code lengths, leading to more powerful semantic similarity learning capabilities. Furthermore, affinity-adaptive contrastive learning effectively mitigates heterogeneity and semantic gaps across modalities by introducing an augmented contrastive relationship, leading to an improvement in mAP scores from 0.8834 to 0.9020. The above results of the ablation studies demonstrate the importance of each component and their collective integration for cross-modal retrieval.

4.5 Parameter Sensitivity

To assess the sensitivity of parameters, we perform an exhaustive parameter analysis of the proposed VTPH method on MIRFLICKR-25K with 16 bits under different parameter configurations. Specifically, we focus on analyzing the effects of three hyper-parameters, i.e., α, β, and γ, as shown in Eqn. (9) and Eqn. (16). Through careful experimentation



(a) α (b) β (c) $\gamma/100$

Figure 4: Comparison of mAP scores on the MIRFLICKR-25K dataset with different parameter configurations.

and analysis in Figure 4, it can be observed that our VTPH method achieves the best performance when $\alpha = 10$, $\beta = 8$, and $\gamma = 500$, respectively. Hence, we can summarize that our VTPH model can obtain superior performance through an optimal combination of these hyperparameters.

5 Conclusion

In this paper, we identified the challenge of context loss and information redundancy in existing manually annotated cross-modal retrieval datasets. To overcome these challenges, we proposed a novel Visual-Textual Prompt Hashing framework that integrated both visual and textual prompt learning with a unified framework for cross-modal retrieval. Importantly, the proposed affinity-adaptive contrastive learning module modeled the affinity differences between simulated and real-world environments to augment the contrastive relationship across modalities. Benefiting from these powerful components, our VTPH approach can effectively mitigate the heterogeneity and semantic gaps among different modalities, even in real-world environments with noisy correspondences.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62302172, 62176077), in part by the Guangdong International Science and Technology Cooperation Project (Grant No. 2023A0505050108), in part by the Guangdong University Young Innovative Talents Program Project (Grant No. 2023KQNCX020), and in part by the Opening Project of GuangDong Province Key Laboratory of Information Security Technology (Grant No. 2023B1212060026).

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 ACM International Conference on Multimedia Retrieval (ICMR)*, pages 525–531, 2020.

[Bogolin *et al.*, 2022] Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5194–5205, 2022.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901, 2020.

[Cao *et al.*, 2018] Yue Cao, Bin Liu, Mingsheng Long, and Jianmin Wang. Cross-modal hamming hashing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 202–218, 2018.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1597–1607. PMLR, 2020.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[Grill *et al.*, 2020] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, volume 33, pages 21271–21284, 2020.

[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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020.

[Hu *et al.*, 2023] Peng Hu, Hongyuan Zhu, Jie Lin, Dezhong Peng, Yin-Ping Zhao, and Xi Peng. Unsupervised contrastive cross-modal hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(3):3877–3889, 2023.

[Huang *et al.*, 2021] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. In *Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, volume 34, pages 29406–29419, 2021.

[Huiskes and Lew, 2008] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 39–43, 2008.

[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 (CVPR)*, pages 3232–3240, 2017.

[Khan *et al.*, 2023] Muhammad Gul Zain Ali Khan, Muhammad Ferjad Naeem, Luc Van Gool, Didier Stricker, Federico Tombari, and Muhammad Zeshan Afzal. Introducing language guidance in prompt-based continual learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 11463–11473, 2023.

[Khattak *et al.*, 2023] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19113–19122, 2023.

[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 (CVPR)*, pages 4242–4251, 2018.

[Li *et al.*, 2022] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, He Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7241–7259, 2022.

[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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[Lin, 1991] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory (TIT)*, 37(1):145–151, 1991.

[Liu *et al.*, 2023a] Xuejing Liu, Wei Tang, Jinghui Lu, Rui Zhao, Zhaojun Guo, and Fei Tan. Deeply coupled cross-modal prompt learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7957–7970, 2023.

[Liu *et al.*, 2023b] Yishu Liu, Qingpeng Wu, Zheng Zhang, Jingyi Zhang, and Guangming Lu. Multi-granularity interactive transformer hashing for cross-modal retrieval. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 893–902, 2023.

[Luo *et al.*, 2021a] Xiao Luo, Daqing Wu, Zeyu Ma, Chong Chen, Minghua Deng, Jianqiang Huang, and Xian-Sheng Hua. A statistical approach to mining semantic similarity for deep unsupervised hashing. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 4306–4314, 2021.

[Luo *et al.*, 2021b] Xiao Luo, Daqing Wu, Zeyu Ma, Chong Chen, Minghua Deng, Jinwen Ma, Zhongming Jin, Jianqiang Huang, and Xian-Sheng Hua. Cimon: Towards high-quality hash codes. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

[Messina *et al.*, 2021] Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Transactions on Multimedia Computing, Communications and Applications (TOMM)*, 17(4):1–23, 2021.

[Mikriukov *et al.*, 2022] Georgii Mikriukov, Mahdyar Ravanbakhsh, and Begüm Demir. Unsupervised contrastive hashing for cross-modal retrieval in remote sensing. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4463–4467, 2022.

[Ou *et al.*, 2023a] Weihua Ou, Jiaxin Deng, Lei Zhang, Jianping Gou, and Quan Zhou. Cross-modal generation and pair correlation alignment hashing. *IEEE Transactions on Intelligent Transportation Systems (TITS)*, 24(3):3018–3026, 2023.

[Ou *et al.*, 2023b] Weihua Ou, Jiaxin Deng, Lei Zhang, Jianping Gou, and Quan Zhou. Cross-modal generation and pair correlation alignment hashing. *IEEE Transactions on Intelligent Transportation Systems (TITS)*, 24:3018–3026, 2023.

[Pei *et al.*, 2023] Wenjie Pei, Tongqi Xia, Fanglin Chen, Jinsong Li, Jiandong Tian, and Guangming Lu. SA^{2}VP: Spatially aligned-and-adapted visual prompt. *arXiv preprint arXiv:2312.10376*, 2023.

[Qin *et al.*, 2022] Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 4948–4956, 2022.

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[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 ACM International Conference on Multimedia (MM)*, pages 453–461, 2022.

[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 International Joint Conference on Artificial Intelligence (IJCAI)*, volume 2019, pages 982–988, 2019.

[Yang *et al.*, 2023] Shuo Yang, Zhaopan Xu, Kai Wang, Yang You, Hongxun Yao, Tongliang Liu, and Min Xu. Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19883–19892, 2023.

[Yao *et al.*, 2021] Hong-Lei Yao, Yu-Wei Zhan, Zhen-Duo Chen, Xin Luo, and Xin-Shun Xu. Teach: Attention-aware deep cross-modal hashing. In *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR)*, pages 376–384, 2021.

[Zhang *et al.*, 2022] 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 (TKDE)*, 35(5):5091–5104, 2022.

[Zhou *et al.*, 2022a] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16816–16825, 2022.

[Zhou *et al.*, 2022b] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 130(9):2337–2348, 2022.

[Zhu *et al.*, 2022] Lei Zhu, Liewu Cai, Jiayu Song, Xinghui Zhu, Chengyuan Zhang, and Shichao Zhang. Msspq: Multiple semantic structure-preserving quantization for cross-modal retrieval. In *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR)*, pages 631–638, 2022.