Adaptive Prompt-Based Semantic Embedding with Inspire Potential of Implicit Knowledge for Cross-Modal Retrieval

Xin Huang[1], Shilong Wang[1], Tong Jia[2], Zhihang Gou[1], Jingjing Li[1*]

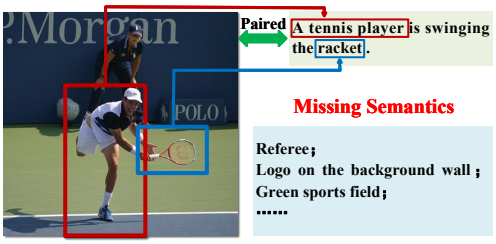
[1]School of Artificial Intelligence and Software Engineering, Nanyang Normal University, Henan, China
[2]Institute for Artificial Intelligence, Peking University, Beijing, China
huangxin@nynu.edu.cn, wangshilong@nynu.edu.cn, jia.tong@pku.edu.cn, gouzhihang@nynu.edu.cn, jingjl101@nynu.edu.cn

Abstract

In the era of big data, cross-modal retrieval is increasingly important in research and application. Given the latent complexity and non-intuitive nature of cross-modal relationships, leveraging external knowledge such as large models has become a popular approach to facilitate modality alignment. Existing methods typically address these challenges by fine-tuning model encoders or using a fixed number of prompts. However, these approaches struggle with the significant information asymmetry between image-text pairs and the high distribution diversity of image data. These limitations not only introduce noise during training but also reduce the accuracy and generalization capabilities in cross-modal retrieval tasks. To address the above issues, this paper proposes **A**daptive **P**rompt-Based **S**emantic **E**mbedding with **I**nspired **P**otential of **I**mplicit **K**nowledge (APSE-IPIK). On one hand, we propose an inspire potential strategy to extract fine-grained and multi-perspective text descriptions from large-scale pre-trained multimodal models, which can be seen as implicit knowledge injection. These descriptions are integrated into the visual-semantic embedding through cross-modal semantic alignment with images, balancing the information asymmetry between modalities and reducing the embedding of inaccurate mapping relationships. On the other hand, we construct an instance-level query-based prompt pool strategy to adaptively extract the most relevant prompts, addressing alignment biases caused by intra-modal (especially image) data diversity and improving alignment accuracy. Extensive experiments are conducted on two widely used datasets, Flickr30k and MSCOCO, which show the effectiveness of the proposed method.

Introduction

Under the background of today's digital society, the value and significance of cross-modal retrieval are increasingly prominent. With the rapid growth of primary modalities such as images and text, information retrieval and utilization are no longer confined to a single modality (Peng, Huang, and Zhao 2018). Cross-modal retrieval allows users to query in one modality (e.g., text) and retrieve results in another (e.g., images or videos), significantly enhancing the scope and efficiency of information retrieval. By effectively handling semantic differences across modalities, cross-modal retrieval offers more accurate and personalized services, driving the development of next-generation intelligent systems.



(a) Information asymmetry between modalities

(b) Distribution diversity of images

Figure 1: Subfigure (a) illustrates the asymmetry of information between modalities (some visual information in the image is not described in the text description) , while subfigure (b) demonstrates the distribution diversity of images, which are randomly sampled from the MSCOCO dataset.

The main challenge of cross-modal retrieval is the inconsistent representation space for different modalities, i.e., heterogeneity gap. The most direct and mainstream approach is to learn a unified representation that is structurally consistent, thereby eliminating inconsistencies in modalities while preserving semantic consistency. Traditional methods use linear mapping (i.e., mapping matrices) as the unified representation model. These types of methods is exemplified by Canonical Correlation Analysis (CCA) (Hotelling 1992) and Orthogonal Canonical Correlation Analysis (OCCA) (Wang et al. 2020a). However, due to the complexity of the relationships in cross-modal data, linear models often have significant limitations in their effectiveness. To overcome these limitations, researchers have turned to deep networks as the basic mapping model (Ngiam et al. 2011; Wei et al. 2017).

*Corresponding author.

These deep learning methods have demonstrated significant advantages in performance, becoming the mainstream approach.

In recent years, we have witnessed the rise of multimodal pre-trained large models (Radford et al. 2021; Brown et al. 2020; Li et al. 2022a). These large models can be seen as sources of implicit knowledge, which significantly benefit the cross-modal semantic understanding and alignment. For example, CLIP (Radford et al. 2021) has shown strong capabilities in integrating and understanding vision-language information, particularly excelling in zero-shot and few-shot learning (Li et al. 2022b). Given their potential for generalization, a new phase of cross-modal retrieval has emerged, where large models serve as foundational components (Zhuo et al. 2022; Yu et al. 2022; Zhu and Li 2023).

Following this idea, current methods primarily adjust encoder parameters through direct fine-tuning on downstream retrieval task data. For example, Bin et al. (Bin et al. 2023) utilized pre-trained encoders for images and texts, and further trained them on cross-modal retrieval datasets. During training, a triplet loss function is used to maximize the similarity between matched image-text pairs while minimizing the similarity between unmatched pairs, enhancing the encoder's performance in cross-modal retrieval tasks. Alternatively, fixed prompts can be used to guide and optimize encoder performance. Prompts, which are special tokens added to input data, help the model better learn cross-modal interaction. For instance, Zang et al. (Zang et al. 2022) proposed a method that learns a tiny neural network to jointly optimize prompts across different modalities, effectively capturing and integrating information from both text and visual modalities, thereby enabling the model to perform well in cross-modal retrieval tasks.

However, in cross-modal retrieval tasks, there is often significant information asymmetry between different modalities, such as images and text. Images capture a direct and detailed reflection of the natural world, while text offers a rationalized, abstract, and symbolic description of the image content, which typically fails to convey all the intricate elements contained within the image, as illustrated in **Fig. 1 (a)**. This asymmetry can lead to inaccurate mappings when fine-tuning pre-trained models, preventing the model from capturing detailed image features in the shared semantic space. Additionally, within modalities (especially the image modality), there is often a high degree of distribution diversity, as shown in **Fig. 1 (b)**. A fixed set of prompts, on the one hand, has limited expressive range, failing to capture subtle differences and rich contexts in images, and on the other hand, lacks the flexibility to adapt to rapidly changing content, making it difficult to encompass the multi-layered semantics present in the images. These challenges, to some extent, limit the accuracy and generalization ability of existing methods in cross-modal retrieval tasks.

Based on the above analysis, we propose the Adaptive Prompt-Based Semantic Embedding with Inspire Potential of Implicit Knowledge for Cross-Modal Retrieval(APSE-IPIK) method, which aims to address the challenges of information asymmetry in cross-modal data alignment and distribution diversity within modalities. Addressing information asymmetry between modalities, inspired by the powerful contextual understanding and content generation capabilities of generative large models, we propose the inspire potential strategy. This strategy leverages the rich implicit knowledge within generative multimodal large models to generate multi-perspective image descriptions, thereby supplementing the semantic information in the original text and effectively mitigating the noise introduced by information asymmetry. Furthermore, existing methods typically use a fixed number of prompts when handling the distribution diversity within modalities, especially the image modality. However, fixed prompts have limited expressive power, making it difficult to capture subtle differences and rich contexts within images. To solve this problem, we introduce the instance-level query-based prompt pool strategy, which dynamically selects prompts through instance-level queries to capture more detailed image information based on the input image's characteristics, thereby optimizing retrieval accuracy. Ultimately, by integrating the inspire potential strategy with the instance-level query-based prompt pool strategy, the APSE-IPIK framework effectively addresses the challenges posed by information asymmetry and distribution diversity within modalities. Main contributions of this paper are as follows:

- This paper proposes an **Inspire Potential Strategy** that leverages latent knowledge from large-scale pre-trained multimodal models to extract fine-grained, multi-perspective text descriptions, alleviating information asymmetry between different modalities and reducing the risk of inaccurate mappings.
- This paper designs an **Instance-Level Query-Based Prompt Pool Strategy**, which can flexibly handle the intra-modal data diversity of images, significantly enhancing cross-modal alignment accuracy and improving the model's generalization ability in complex scenarios.

Related Work

Cross-Modal Retrieval

Cross-modal retrieval is a retrieval paradigm that spans different modalities such as images, videos, and texts. Early, the traditional methods typically learn linear mapping matrices by optimizing specific statistical measures. Canonical Correlation Analysis (Hotelling 1992) is the first widely used cross-modal model, which optimizes the model by maximizing the correlation between paired data from different modalities.

With the powerful learning capabilities demonstrated by deep networks in fields such as image recognition and video classification, some efforts have been made to use deep networks to learn unified representations for cross-modal retrieval. Huang et al. (Huang, Peng, and Yuan 2017) proposed the CHTN method, which achieves effective cross-modal alignment and improves model generalization by transferring knowledge from a single media source domain to a cross-media target domain. Lee et al. (Lee et al. 2018) proposed a stacked cross-attention mechanism for fine-grained image-text alignment, improving cross-modal matching accuracy and efficiency. Chen et al. (Chen et al. 2020) proposed a multi-step alignment mechanism and a memory dis-

tillation unit, the method progressively captures fine-grained semantic correspondences between images and text, significantly enhancing the performance of bidirectional retrieval between vision and language. Cheng et al. (Cheng et al. 2021) designed a semantic alignment module within a cross-modal retrieval network to fully explore the potential correspondence between images and text. Zhang et al. (Zhang et al. 2022) developed a dual-branch mechanism exploiting mismatched fragments to improve image-text matching robustness and discriminability. Feng et al. (Feng, He, and Peng 2023) introduced a two-step reasoning mechanism leveraging intra- and inter-modal relations to enhance image-text embeddings. Li et al. (Li et al. 2023) proposed a cross-modal association learning method based on a deep residual shrinkage network, adding a deep residual shrinkage block to a dual-stream residual network to improve training efficiency and obtain more discriminative embedded features.

In recent years, large-scale pre-trained models have demonstrated significant capabilities in the field of cross-modal retrieval. With the success of these models, researchers have widely adopted them as foundational architectures to further advance vision-language retrieval. For example, Zhai et al. (Zhai et al. 2022) built on ALIGN by introducing the concept of locked image-text tuning, improving generalization by freezing the image encoder and only fine-tuning the text encoder. Wang et al. (Wang et al. 2021) integrated multiple pre-trained vision and language experts in a unified framework built on BLIP. Additionally, Bin et al. (Bin et al. 2023) proposed a framework combining dual-stream encoders with transformers for cross-modal retrieval, effectively leveraging the power of transformers to align and fuse information from different modalities for improved retrieval performance.

Prompt-Tuning of Large Models

Among existing strategies for fine-tuning large models, common approaches include traditional full fine-tuning, adapter-based fine-tuning, prompt-based fine-tuning, and Low-Rank Adaptation. Prompt Tuning has been widely applied in downstream tasks, offering the advantage of enhancing the model's expression and understanding by incorporating or adjusting prompts in the input. This method guides the model to capture key features more accurately, improving downstream task performance. For example, Zhou et al. (Zhou et al. 2022) proposed CoOp, which dynamically learns and optimizes prompts in image classification tasks, allowing the model to generate contextually relevant prompts based on different image categories, significantly enhancing the model's classification performance in new categories. Jia et al. (Jia et al. 2022) introduced VPT, which adds a small number of trainable parameters before image embedding while keeping the backbone model frozen, achieving performance improvements in downstream visual tasks.

In continual learning, Wang et al. (Wang et al. 2022) proposed L2P, which employs dynamic retrieval of task-relevant prompts to preserve the contextual knowledge of prior tasks and mitigate catastrophic forgetting by preventing the overwriting of earlier knowledge during new task training. In the cross-modal analysis area, Xing et al. (Xing et al. 2024) proposed a dual-modal prompt tuning method that simultaneously tunes visual and textual prompts, enhancing the adaptability and expressiveness of pre-trained vision-language models in downstream tasks. Liu et al. (Liu et al. 2023) proposed a deeply coupled cross-modal prompt learning method that strengthens the interaction between vision and language through multi-head attention modules, improving prompt learning effectiveness in cross-modal tasks.

The Proposed Method

The following is a detailed description of the overall architecture and workflow of the APSE-IPIK method (as shown in Fig. 2).

CLIP-Based Retrieval Baseline

In this subsection, we present the CLIP-based retrieval baseline of our APSE-IPIK. Considering that the vision-language retrieval task is mainly to match paired images and text, we employ the CLIP-based encoders to learn the discriminative deep embedding. Firstly, we define the input data as a set of image-text pairs $D = (v_i, t_i)_{i=1}^{N}$, where v_i represents the i-th image and t_i denotes the corresponding textual description. We use CLIP's image embedding layer EMB_i to obtain the image embeddings and the text embedding layer EMB_t to obtain the text embeddings. We let f_v and f_t be the encoders for images and text, respectively. Due to the requirements of the instance-level query-based prompt pool strategy, we take the feature z_p^i after the image passes through the final layer of the transformer in the image encoder as the query feature. The deep embedding z_v^i and z_t^i can be formulated as:

$$\begin{cases} i_{emb} = EMB_i(v_i) \\ t_{emb} = EMB_t(t_i) \\ z_v^i = f_v(v_i) \\ z_t^i = f_t(t_i) \end{cases} \tag{1}$$

To maximize the similarity of paired image-text pairs, a contrastive loss function L_{CLIP} is used to ensure that the similarity between paired image-text pairs is as high as possible, while the similarity between non-paired image-text pairs is minimized. The L_{CLIP} is defined as:

$$\begin{cases} L_v^i = -log(\frac{e^{(z_v^i, z_t^i)/\tau}}{\sum_{k=1}^{N} e^{(z_v^i, z_t^k)/\tau}}) \\ L_t^i = -log(\frac{e^{(z_t^i, z_v^i)/\tau}}{\sum_{k=1}^{N} e^{(z_t^i, z_v^k)/\tau}}) \\ L_{CLIP} = \sum_{i=1}^{N} (L_v^i + L_t^i)/2 \end{cases} \tag{2}$$

Inspire Potential Strategy

In cross-modal retrieval, there is often significant information asymmetry between images and text. Images, as direct representations of the natural world, contain rich visual information, while text, being a rationalized, abstract, and symbolic expression, often fails to fully capture the details of images. This asymmetry can lead to incorrect mappings

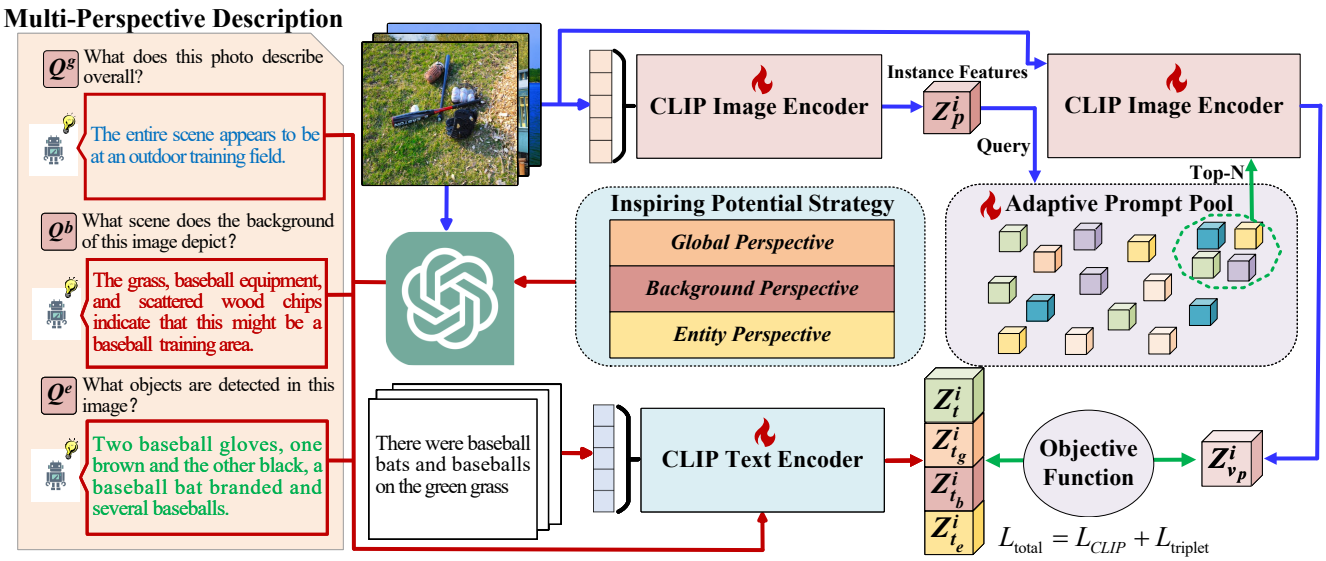


Figure 2: This is the overall framework diagram of our APSE-IPIK. On one hand, we extract detailed image information from the rich implicit knowledge embedded in generative multimodal large models, and inject it into the original text information to achieve data augmentation, thereby alleviating information asymmetry. On the other hand, we employ an instance-level query-based strategy to retrieve relevant prompts from an adaptive prompt pool, reducing the alignment biases caused by the distribution diversity of images.

when fine-tuning pre-trained models, causing the loss of image detail in the shared semantic space. However, generative multimodal large models can generate multi-perspective image descriptions, which, when injected into the original text, can effectively mitigate this asymmetry. Inspired by this, we propose the Inspire Potential Strategy. This strategy uses multi-perspective questioning to guide the generative multimodal large model (BLIP2 adopted in this paper) in creating rich, fine-grained descriptions, which are then injected into the text as external implicit knowledge to address the information asymmetry. We consider three main perspectives for questioning.

Global perspective: This aims to comprehensively capture the overall semantics and contextual details of the image. Specifically, given an image v_i, we utilize the generative multimodal large model to generate a corresponding global description t_i^g , ensuring that the textual information covers the key context of the image. The generated form is as follows:

$$t_i^g = BLIP2(Q^g, v_i) \tag{3}$$

where Q^g is the question: "What does this photo describe overall" ?

Background perspective: This aims to capture richer semantic context within the image. Specifically, given an image v_i , the generative multimodal large model is used to generate a corresponding background description t_i^b , ensuring that the model takes into account the surrounding environment and contextual factors when processing the image. The generated form is as follows:

$$t_i^b = BLIP2(Q^b, v_i) \tag{4}$$

where Q^b is the question: "What scene does the background of this image depict" ?

Entity perspective: This aims to capture the key content and core elements within the image, enabling the model to accurately identify and match corresponding relationships across different modalities. Similar to the previous processing, this perspective generates descriptions based on the entities within the image to capture the main objects and important details. The generated form is as follows:

$$t_i^e = BLIP2(Q^e, v_i) \tag{5}$$

where Q^e is the question: "What objects are detected in this image" ?

At this point, we have obtained a richer and more comprehensive set of text descriptions $T = \{t_i, t_i^g, t_i^b, t_i^e\}$, which focus on different aspects of the image, capturing comprehensive semantic information. These descriptions serve as supplements to the original text, alleviating the limitations of the original text in expressing image details. This enhances the model's adaptability when dealing with image-text information asymmetry and effectively reduces retrieval errors caused by semantic gaps or biases. The original text, along with the descriptions generated from different perspectives, is then used as input to the CLIP text encoder $f_t(x)$ to obtain the corresponding set of text embeddings $Z = \{z_t^i, z_{t_g}^i, z_{t_b}^i, z_{t_e}^i\}$.

Note that only in the prompt training phase, we follow the above steps to generate extended text descriptions T, where the texts are independently aligned with the images, thereby expanding each original image-text pair into 4 positive pairs to enhance the model's understanding of the relationship between images and text. In the testing phase, given that the model's modality alignment capability has been effectively improved, to ensure model consistency and simplify computation, we only use the embedded

representation of the original text t_i , while the additional text descriptions generated based on the inspire potential strategy are not performed.

Instance-Level Query-Based Prompt Pool Strategy

In cross-modal retrieval tasks, textual data, characterized by its abstract and symbolic nature, can be effectively processed by modern large-scale model encoders, which are sufficiently robust to handle most textual information. In contrast, image data exhibits significant distribution diversity, with varying content, styles, and contexts across different images. Given that the semantic search scope of users cannot be explicitly constrained, this diversity necessitates highly adaptive models capable of capturing the unique features of each image. Inspired by L2P (Wang et al. 2022) , we propose an instance-level query-based prompt pool strategy (as illustrated in Fig. 3). Building upon this foundation, we adapt and extend this concept to the cross-modal retrieval problem to address the unique challenges of information asymmetry between modalities and image distribution diversity. Specifically, our strategy employs an instance-level query mechanism to dynamically select the most relevant prompts for each image, capturing key visual features and guiding the model to focus on these features, thereby enhancing retrieval accuracy. The prompt pool is defined as:

$$P = [P_1, P_2, \cdots, P_M], M \neq 0 \tag{6}$$

where $P_i \in R^{L_p \times D}$ is a single prompt with token length L_p and the same embedding size D as i_{emb} . Denoting $\{s_i\}_{i=1}^{N}$ as a subset of N indices from $[1, M]$, we can then adapt the input embedding as follows:

$$x_p = [P_{s_1}; P_{s_2}; \cdots; P_{s_N}; i_{emb}], 1 \leq N \leq M \tag{7}$$

where ; represents concatenation along the token length dimension. Different images correspond to their respective sets of prompts, enabling the joint encoding of knowledge for the model to process. The diverse combinations of prompts effectively handle the variability in image distributions, reducing retrieval errors caused by the diversity of image distributions. To achieve this, we designed an instance-level query-based strategy. Specifically, for each image, the selection of prompts is dynamically determined through a key-value matching mechanism. Each prompt in the prompt pool is associated with a learnable key, allowing the model to efficiently match and select the most appropriate prompts based on the input:

$$\{(k_1, P_1), (k_2, P_2), \cdots, (k_M, P_M)\} \tag{8}$$

where $k_n \in R^{D_k}$. We denote the set of all keys as $K = \{k_i\}_{i=1}^{M}$. To enable the input instance to decide which prompt to select through query-key matching, we introduce a query function $q : R^{H \times W \times C} \Longrightarrow R^{D_k}$, which encodes the input v_i into the same dimensional space as the keys. We directly use the CLIP image encoder as the query function q to obtain the query features. Specifically, we take the intermediate image features z_p^i , which are the output of the final layer of the transformer in the image encoder as the query features.

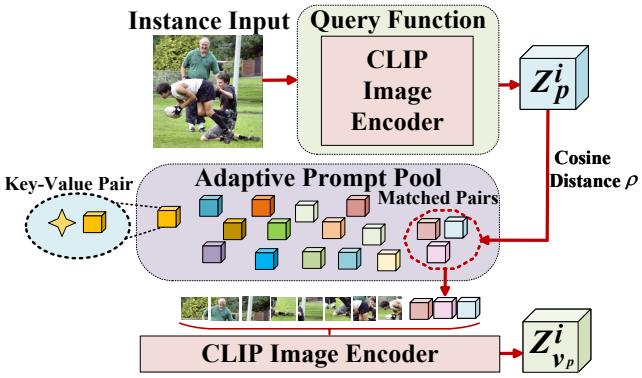


Figure 3: This is a diagram of our adaptive prompt pool method. Through the instance query strategy, the appropriate prompt is selected, then combined with the image embedding vector and input into the CLIP encoder for downstream image to text retrieval tasks.

We define $\rho : R^{D_k} \times R^{D_k} \Longrightarrow R^D$ as a function to score the match between the query and the prompt keys. In this paper, we use cosine distance as the scoring function. Given an input v_i, we find the top-N keys by simply solving the following objective function:

$$k_x = \underset{\{s_i\}_{i=1}^{N} \subseteq [1,M]}{\operatorname{argmin}} \sum_{i=1}^{N} \rho(z_p^i, k_{s_i}) \tag{9}$$

where k_x represents the subset of top-N keys selected specifically for v_i from K. We concatenate the N prompts obtained from the query based on the above strategy with the image embeddings i_{emb} as the input to the image encoder, and take the output corresponding to the [CLS] token as the final image feature representation $z_{v_p}^i$.

Optimization Objective

We take the image features $z_{v_p}^i$ obtained after prompt guidance and perform loss calculation with the original text features z_t^i as well as the multi-angle descriptive features Z obtained based on the inspire potential strategy. On the one hand, to extract the semantic information missing from the original text through multi-perspective image-based descriptions and avoid introducing noise. On the other hand, to enable the prompt pool to capture more detailed features of the image, we optimize the entire model using a triplet loss function, which is defined as follows:

$$\begin{aligned} L_{triplet} = &[m - sim(z_{v_p}^i, Z) + sim(z_{v_p}^i, \overline{Z})]_+ \\ &+[m - sim(Z, z_{v_p}^i) + sim(Z, \overline{z_{v_p}^i})]_+ \end{aligned} \tag{10}$$

where m is a margin set as 0.2, and $[\cdot]_+ = max(0, \cdot)$. $\overline{z_{v_p}^i}$ and $\overline{Z}$ denote the negative samples to push away. Because Z here is a collection of text features, our $sim(\cdot)$ here refers to the similarity calculation between image features and every text feature in the text feature set. Finally, the total objective function of our proposed APSE-IPIK is calculated as follows:

$$L_{total} = L_{CLIP} + L_{triplet} \tag{11}$$

where the L_{CLIP}, similar to $L_{triplet}$, is also computed using both the original data and the additional textual descriptions generated by the inspire potential strategy.

Experiments

In this section, we conduct extensive experiments on two datasets, MSCOCO and Flickr30K, to verify the effectiveness of our proposed APSE-IPIK method. The source code is available at https://github.com/nynu-BDAI/APSE-IPIK.

Datasets and Evaluations

We conduct the validation experiments on two real-world benchmark image-text retrieval datasets:

1) MSCOCO (Lin et al. 2014) contains 123,287 images, where each image is described by 5 different sentences. For fair comparison, we use Karpathy split (Karpathy and Fei-Fei 2015). We perform retrieval tasks on the MSCOCO dataset using the full set of 5K images, which includes retrieving target images from the 5K image set or retrieving relevant sentences from the corresponding corpus.

2) Flickr30K (Young et al. 2014) consists of 31,783 images collected from Flickr, each of which is annotated with 5 description sentences. Following previous works (Bin et al. 2023), we split this dataset into 29,783, 1,000, 1,000 images for training, validation, and testing respectively.

Evaluation Metrics: We follow previous works (Karpathy and Fei-Fei 2015) to evaluate the performance with the $Recall@K$ metric, short in $R@K$(K=1,5,10). It measures the percentage of ground-truth hits in the top-K ranking list. The higher $R@K$ indicates the better performance.

Methods	Flickr30K Dataset						RSUM
	Image-to-text			Text-to-image			
	R@1	R@5	R@10	R@1	R@5	R@10	
SCAN	67.4	90.3	95.8	48.6	77.7	85.2	465.0
BFAN	68.1	91.4	-	50.8	78.4	-	-
SGM	71.8	91.7	95.5	53.5	79.6	86.5	478.6
IMRAM	74.1	93.0	96.6	53.9	79.4	87.2	484.2
GSMN	76.4	94.3	97.3	57.4	82.3	89.0	496.8
SMFEA	73.7	92.5	96.1	54.7	82.1	88.4	487.5
SHAN	74.6	93.5	96.9	55.3	81.3	88.4	490.0
VSE∞	76.5	94.2	97.7	56.4	83.4	89.9	498.1
SGRAF	77.8	94.1	97.4	58.5	83.0	88.8	499.6
NAAF	81.9	96.1	98.3	61.0	85.3	90.6	513.2
MGCN	<u>82.9</u>	<u>96.5</u>	<u>98.9</u>	<u>63.2</u>	<u>87.1</u>	<u>92.5</u>	<u>521.1</u>
APSE-IPIK (ours)	**86.3**	**97.6**	**99.4**	**72.0**	**92.5**	**95.1**	**542.9**

Table 1: Comparison results on Flickr30K dataset.

Methods	MSCOCO (5K) Dataset						RSUM
	Image-to-text			Text-to-image			
	R@1	R@5	R@10	R@1	R@5	R@10	
SCAN	50.4	82.2	90.0	38.6	69.3	80.4	410.9
PVSE	45.2	74.3	84.5	32.4	63.0	75.0	374.3
SGM	50.0	79.3	87.9	35.3	64.9	76.5	393.9
IMRAM	53.7	83.2	91.0	39.7	69.1	79.8	416.5
SGRAF	57.8	-	91.6	41.9	-	81.3	-
DIME	**59.3**	<u>85.4</u>	91.9	<u>43.1</u>	<u>73.0</u>	<u>83.1</u>	435.8
NAAF	58.9	85.2	92.0	42.5	70.9	81.4	430.9
MGCN	**59.3**	84.9	<u>92.6</u>	42.8	**73.2**	**83.4**	<u>436.2</u>
APSE-IPIK (ours)	<u>59.1</u>	**85.7**	**94.6**	**45.1**	72.8	82.5	**439.8**

Table 2: Comparison results on the MSCOCO (5K) dataset.

Implementation Details

The proposed APSE-IPIK framework was implemented in PyTorch using NVIDIA A40 GPUs and an Intel® Xeon® Gold 6330 CPU, running on Ubuntu 22.10. We utilized CLIP (ViT-Based/32) as our base model. The input images were resized to 224×224 pixels. The model was trained for 10 epochs with a batch size of 128. We employed the Adam optimizer with $(\beta_1, \beta_2) = (0.9, 0.99)$ and a weight decay of 0.2 to update the entire CLIP model. The initial learning rate was set to $1e-6$, and a cosine annealing learning rate scheduler was used to update the entire framework. In addition, the parameters of the adaptive prompt pool are set as follows: the prompt pool size $M = 10$, the number of most relevant prompts selected per image instance $N = 5$, and the length of each prompt $L_p = 5$.

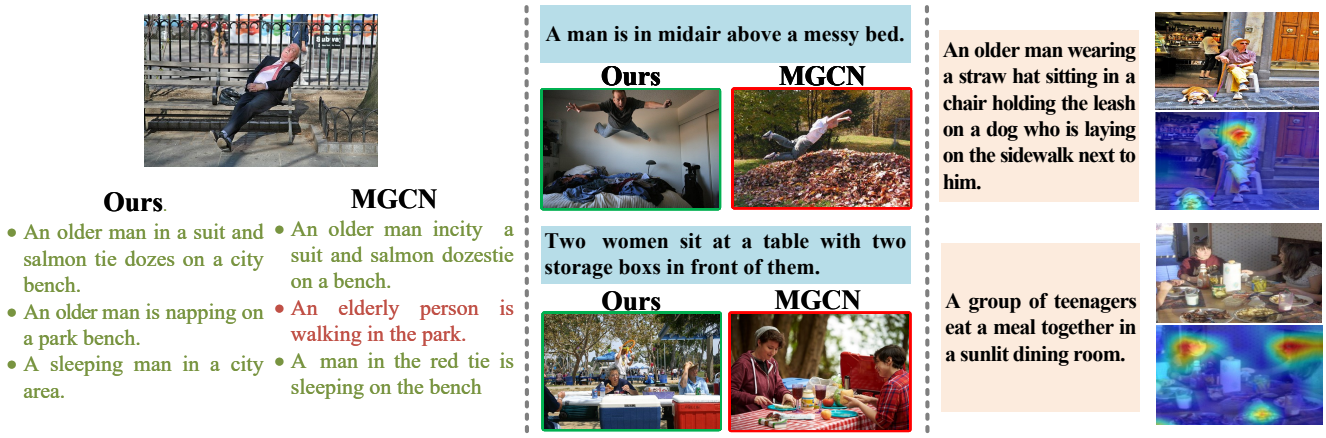
Comparison Results

We compared our proposed APSE-IPIK method with several recent models on two benchmark datasets. Table 1 presents the quantitative results of APSE-IPIK on the Flickr30K dataset, showing that our method significantly outperforms the existing state-of-the-art methods across most evaluation metrics. Specifically, for the image-to-text retrieval task, our method achieved an R@1 of 86.3, which is 3.4 higher than the previous best result by MGCN. Similarly, in the text-to-image retrieval task, APSE-IPIK reached an R@1 of 72.0, surpassing the second-best MGCN model by 8.8. Overall, APSE-IPIK achieved a substantial improvement in the RSUM score, showing a 21.8 increase compared to MGCN, further validating the effectiveness of our approach in enhancing retrieval accuracy.

Table 2 shows our experimental results on the larger and more complex MSCOCO dataset. As demonstrated by the results, our APSE-IPIK also exhibited strong performance. In the image-to-text retrieval task, our method achieved an R@1 of 59.1, slightly below the latest MGCN's 59.3. However, our method excelled in R@5 and R@10 scores, reaching 85.7 and 94.6, which are 0.8 and 2.0 higher than MGCN, respectively. In the text-to-image retrieval task, APSE-IPIK achieved competitive results. Overall, the RSUM score reached 439.8, representing a significant improvement compared to the latest MGCN and highlighting the model's efficiency in handling cross-modal retrieval tasks on complex datasets.

Bi-Directional Retrieval Results

Fig. 4 (a) and Fig. 4 (b) present examples between APSE-IPIK and MGCN models on bi-directional retrieval tasks using the Flickr30K dataset. In the image-to-text retrieval task, when we use an image as the query, APSE-IPIK effectively captures key details in the image, such as "a man in a suit dozes on a park bench, with a sign for the subway in the background", as shown in the first example. While MGCN also retrieves some relevant descriptions, the absence of additional knowledge leads to incorrect mappings during alignment, making it difficult to accurately capture important details. Consequently, the retrieved descriptions do not accurately reflect the actual scene in the image. In the text-to-image retrieval task, when we input the query "a man



(a) Image to Text retrieval **(b) Text to Image retrieval** **(c) Visualizations of critical regions**

Figure 4: Examples of bi-directional retrieval and visualizations of critical regions focused by the APSE-IPIK on the Flickr30K dataset.

in a black shirt and jeans is in midair above a messy bed", APSE-IPIK successfully retrieves the corresponding image, demonstrating its strong alignment capabilities in complex scenes. In contrast, MGCN tends to overlook subtle differences in complex scenes during retrieval, resulting in a failure to accurately match the correct image.

Furthermore, to better illustrate the effectiveness of our adaptive prompt selection strategy, we provide a visualization of the model's attention map (Fig. 4 (c)). This visualization showcases the critical regions in the image that APSE-IPIK focuses on.

Ablation Study

In this section, we conduct ablation studies to evaluate the effectiveness of each component in our APSE-IPIK approach (as shown in Table 3).

CLIP Baseline: Fine-tuning the CLIP model without additional strategies serves as the baseline, achieving an R@1 of 83.2 for image-to-text and 68.4 for text-to-image retrieval, with an RSUM of 526.7. This demonstrates the inherent limitations of using CLIP alone for cross-modal tasks.

CLIP-IP (Inspire Potential): Incorporating Inspire Potential into the fine-tuned CLIP model improves alignment performance, increasing R@1 to 83.9 (image-to-text) and 69.6 (text-to-image), with an RSUM of 532.1.

CLIP-AP (Adaptive Prompt Pool): Adding adaptive prompt pooling to the fine-tuned CLIP model further enhances performance, achieving an R@1 of 84.3 for image-to-text and 70.2 for text-to-image, with an RSUM of 534.8. This demonstrates the effectiveness of adaptive prompts in improving cross-modal alignment.

Methods	Flickr30K Dataset						RSUM
	Image-to-text			Text-to-image			
	R@1	R@5	R@10	R@1	R@5	R@10	
CLIP	83.2	94.3	97.8	68.4	89.4	93.6	526.7
CLIP-IP	83.9	95.2	98.1	69.6	91.0	94.3	532.1
CLIP-AP	84.3	96.5	98.6	70.2	90.4	94.8	534.8
APSE-IPIK	**86.3**	**97.6**	**99.4**	**72.0**	**92.5**	**95.1**	**542.9**

Table 3: Effectiveness of each component in our APSE-IPIK approach.

APSE-IPIK: Combining both Inspire Potential and adaptive prompt pooling yields the best performance, with R@1 scores of 86.3 (image-to-text) and 72.0 (text-to-image), and an RSUM of 542.9. These results highlight the synergy between the two strategies in enhancing cross-modal retrieval accuracy and generalization.

Parameter Analysis of the Prompt Pool

We conducted systematic parameter experiments on the key parameters of the adaptive prompt pool—prompt pool size M and the number of most relevant prompts selected per image instance N. The RSUM results obtained are as follows: 538.0(M=10, N=3), 538.6(M=10, N=5), 537.5(M=10, N=7), 539.1(M=20, N=3), 542.9(M=20, N=5), 539.7(M=20, N=7), 534.1(M=30, N=3), 537.8(M=30, N=5), and 535.9(M=30, N=7). Among these, the highest value 542.9 corresponds to M=20 and N=5, which we selected as the optimal result of this study. This finding suggests that a smaller prompt pool may limit the diversity of prompts, while a larger pool may introduce redundant information, ultimately degrading performance.

Conclusion

In this paper, we propose a novel cross-modal retrieval method called Adaptive Prompt-Based Semantic Embedding with Inspire Potential of Implicit Knowledge (APSE-IPIK). This method employs an inspire potential strategy to guide a generative multimodal model in creating multi-perspective image descriptions, enriching the text's semantic content and reducing noise from information asymmetry. Additionally, it integrates an adaptive prompt pool and an instance-level query-based strategy to dynamically select prompts, capturing image details and achieving more precise cross-modal alignment. Extensive experiments on various cross-modal retrieval models and datasets demonstrate significant performance improvements, achieving state-of-the-art results. In the future, we aim to extend the model's applicability to more scenarios and optimize the computational efficiency of the generative model to handle large-scale datasets effectively.

Acknowledgments

This work was supported by Humanities and Social Sciences Youth Foundation, Ministry of Education of the People's Republic of China (No. 24YJCZH135), and Science and Technology Projects of Henan Province (Nos. 242102211019 & 242102210184).

References

Bin, Y.; Li, H.; Xu, Y.; Xu, X.; Yang, Y.; and Shen, H. T. 2023. Unifying Two-Stream Encoders with Transformers for Cross-Modal Retrieval. In *ACM MM 2023*, 3041–3050.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *NeurIPS 2020*, volume 33, 1877–1901.

Chen, H.; Ding, G.; Liu, X.; Lin, Z.; Liu, J.; and Han, J. 2020. IMRAM: Iterative Matching With Recurrent Attention Memory for Cross-Modal Image-Text Retrieval. In *CVPR 2020*, 12652–12660.

Chen, J.; Hu, H.; Wu, H.; Jiang, Y.; and Wang, C. 2021. Learning the Best Pooling Strategy for Visual Semantic Embedding. In *CVPR 2021*, 15789–15798.

Cheng, Q.; Zhou, Y.; Fu, P.; Xu, Y.; and Zhang, L. 2021. A Deep Semantic Alignment Network for the Cross-Modal Image-Text Retrieval in Remote Sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14: 4284–4297.

Diao, H.; Zhang, Y.; Ma, L.; and Lu, H. 2021. Similarity Reasoning and Filtration for Image-Text Matching. In *AAAI 2021*, 1218–1226.

Feng, D.; He, X.; and Peng, Y. 2023. MKVSE: Multimodal Knowledge Enhanced Visual-semantic Embedding for Image-text Retrieval. *ACM TOMM*, 19(5): 1–21.

Ge, X.; Chen, F.; Jose, J. M.; Ji, Z.; Wu, Z.; and Liu, X. 2021. Structured Multi-modal Feature Embedding and Alignment for Image-Sentence Retrieval. In *ACM MM 2021*, 5185–5193.

Hotelling, H. 1992. Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution 1992*, 162–190.

Huang, X.; Peng, Y.; and Yuan, M. 2017. Cross-modal Common Representation Learning by Hybrid Transfer Network. In *IJCAI 2017*, 1893–1900.

Ji, Z.; Chen, K.; and Wang, H. 2021. Step-Wise Hierarchical Alignment Network for Image-Text Matching. In *IJCAI 2021*, 765–771.

Jia, M.; Tang, L.; Chen, B.; Cardie, C.; Belongie, S. J.; Hariharan, B.; and Lim, S. 2022. Visual Prompt Tuning. In *ECCV 2022*, volume 13693, 709–727.

Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR 2015*, 3128–3137.

Lee, K.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked Cross Attention for Image-Text Matching. In *ECCV 2018*, volume 11208, 212–228.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. C. H. 2022a. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML 2022*, volume 162, 12888–12900.

Li, J.; Tan, L.; Zhou, Y.; Mao, J.; Liu, Z.; and Bu, F. 2023. Voice-Face Cross-Modal Association Learning Based on Deep Residual Shrinkage Network. In *ICIPCA 2023*, 140–145.

Li, Y.; Zhao, J.; Lyu, M. R.; and Wang, L. 2022b. Eliciting Knowledge from Large Pre-Trained Models for Unsupervised Knowledge-Grounded Conversation. In *EMNLP 2022*, 10551–10564.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV 2014*, volume 8693, 740–755.

Liu, C.; Mao, Z.; Liu, A.; Zhang, T.; Wang, B.; and Zhang, Y. 2019. Focus Your Attention: A Bidirectional Focal Attention Network for Image-Text Matching. In *ACM MM 2019*, 3–11.

Liu, C.; Mao, Z.; Zhang, T.; Xie, H.; Wang, B.; and Zhang, Y. 2020. Graph Structured Network for Image-Text Matching. In *CVPR 2020*, 10918–10927.

Liu, X.; Tang, W.; Lu, J.; Zhao, R.; Guo, Z.; and Tan, F. 2023. Deeply Coupled Cross-Modal Prompt Learning. In *ACL 2023*, 7957–7970.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal Deep Learning. In *ICML 2011*, volume 11, 689–696.

Peng, Y.; Huang, X.; and Zhao, Y. 2018. An Overview of Cross-Media Retrieval: Concepts, Methodologies, Benchmarks, and Challenges. *IEEE TCSVT*, 28(9): 2372–2385.

Qu, L.; Liu, M.; Wu, J.; Gao, Z.; and Nie, L. 2021. Dynamic Modality Interaction Modeling for Image-Text Retrieval. In *ACM SIGIR 2021*, 1104–1113.

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 *ICML 2021*, volume 139, 8748–8763.

Song, Y.; and Soleymani, M. 2019. Polysemous Visual-Semantic Embedding for Cross-Modal Retrieval. In *CVPR 2019*, 1979–1988.

Wang, L.; Zhang, L.; Bai, Z.; and Li, R. 2020a. Orthogonal Canonical Correlation Analysis and Applications. *Optimization Methods and Software*, 35(4): 787–807.

Wang, S.; Wang, R.; Yao, Z.; Shan, S.; and Chen, X. 2020b. Cross-modal Scene Graph Matching for Relationship-aware Image-Text Retrieval. In *WACV 2020*, 1497–1506.

Wang, W.; Bao, H.; Dong, L.; and Wei, F. 2021. VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts. *ArXiv*, abs/2111.02358.

Wang, Z.; Zhang, Z.; Lee, C.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J. G.; and Pfister, T. 2022. Learning to Prompt for Continual Learning. In *CVPR 2022*, 139–149.

Wei, Y.; Zhao, Y.; Lu, C.; Wei, S.; Liu, L.; Zhu, Z.; and Yan, S. 2017. Cross-Modal Retrieval With CNN Visual Features: A New Baseline. *IEEE TCYB*, 47(2): 449–460.

Xing, Y.; Wu, Q.; Cheng, D.; Zhang, S.; Liang, G.; Wang, P.; and Zhang, Y. 2024. Dual Modality Prompt Tuning for Vision-Language Pre-Trained Model. *IEEE TMM*, 26: 2056–2068.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From Image Descriptions To Visual Denotations: New Similarity Metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78.

Yu, H.; Ding, S.; Li, L.; and Wu, J. 2022. Self-Attentive CLIP Hashing for Unsupervised Cross-Modal Retrieval. In *ACM MM 2022*, 1–7.

Zang, Y.; Li, W.; Zhou, K.; Huang, C.; and Loy, C. C. 2022. Unified Vision and Language Prompt Learning. *ArXiv*, abs/2210.07225.

Zhai, X.; Wang, X.; Mustafa, B.; Steiner, A.; Keysers, D.; Kolesnikov, A.; and Beyer, L. 2022. LiT: Zero-Shot Transfer with Locked-image text Tuning. In *CVPR 2022*, 18102–18112.

Zhang, K.; Mao, Z.; Wang, Q.; and Zhang, Y. 2022. Negative-Aware Attention Framework for Image-Text Matching. In *CVPR 2022*, 15640–15649.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhu, Y.; and Li, X. 2023. Iterative Uni-modal and Cross-modal Clustered Contrastive Learning for Image-text Retrieval. In *PRMVIA 2023*, 15–23.

Zhuo, Y.; Li, Y.; Hsiao, J.; Ho, C.; and Li, B. 2022. CLIP4Hashing: Unsupervised Deep Hashing for Cross-Modal Video-Text Retrieval. In *ICMR 2022*, 158–166.