ISSN 2616-5767 Vol.9, Issue 3: 15-21, DOI: 10.25236/AJETS.2026.090302

Research on Cross-modal Image Retrieval and Image-Text Matching Based on Visual-Language Pre-trained Models

Xia Jiayue[1,a], Long Yanbin[1,b,*]

[1]*University of Science and Technology Liaoning, Anshan, China*
[a]*2677399204@qq.com*, [b]*1034182681@qq.com*
**Corresponding author*

Abstract: *Cross-modal image retrieval and image-text matching are core tasks connecting computer vision and natural language processing, aiming to eliminate the heterogeneous gap between visual and text modalities. Visual-language pre-trained models, through large-scale data learning and cross-modal alignment, have become the dominant technical paradigm for solving this task. This paper systematically reviews the development of visual-language pre-trained models in the field of cross-modal retrieval, classifies and analyzes existing methods from three dimensions: model architecture, pre-training objectives, and downstream adaptation, and focuses on discussing the architectural differences between dual encoders and fusion encoders, the design evolution of pre-training tasks, and adaptation techniques such as efficient parameter fine-tuning. Based on this, we summarize mainstream datasets and evaluation metrics, compare the performance of representative models, and deeply analyze three key challenges: fine-grained alignment, noise robustness, and inference efficiency. Finally, we look forward to future research directions such as few-shot generalization, unified multi-task framework, and interpretability, hoping to provide a reference for further research in this field.*

Keywords: *Visual-Language Pre-Training; Cross-modal Retrieval; Image-Text Matching; Multimodal Learning; Feature Alignment*

1. Introduction

With the explosive growth of multimedia data on the Internet, how to achieve efficient and accurate semantic retrieval between different modal data has become an important issue in the field of information retrieval. Cross-modal image retrieval and image-text matching tasks aim to retrieve relevant images based on text descriptions as queries, or retrieve relevant text descriptions based on images as queries. This capability supports many practical applications such as multimodal search engines, visual question answering, and image description generation[1] .

The core challenge of this task lies in the inherent “heterogeneous gap” between visual and text modalities—images exist in the form of pixel matrices, while text exists in the form of discrete symbol sequences, and the two have essential differences in their underlying statistical characteristics. The key to bridging this gap lies in constructing a unified semantic representation space, enabling direct similarity comparisons of data from different modalities within this space[2-3].

The rise of Vision-Language Pretraining (VLP) models has provided a breakthrough solution to this challenge. Unlike traditional methods that train from scratch, VLP models first learn cross-modal alignment capabilities on large-scale image-text pairs ranging from millions to hundreds of millions, and then fine-tune them to adapt to downstream tasks. This “pre-training-fine-tuning” paradigm effectively alleviates the problem of scarce labeled data in multimodal tasks and significantly improves retrieval performance. Representative works such as CLIP, ALBEF, and BLIP-2 have achieved breakthrough progress on multiple benchmark tests.

This paper focuses on cross-modal image retrieval and image-text matching research based on vision-language pre-trained models, systematically reviewing the technological evolution and core methods in this field. Section 2 reviews and categorizes relevant research; Section 3 analyzes the architectural design and technical characteristics of representative models; Section 4 summarizes the experimental evaluation system and performance comparison; Section 5 discusses the main challenges currently faced; Section 6 looks forward to future development directions; and finally, the entire paper

ISSN 2616-5767 Vol.9, Issue 3: 15-21, DOI: 10.25236/AJETS.2026.090302

is summarized[4-5].

2. Review of Relevant Research

2.1 Task Definition and Challenges of Cross-modal Retrieval

Cross-modal image-text retrieval can be formally defined as: given query modality data (image or text), find data items that semantically match it from the candidate set of another modality. Depending on the query modality, it can be divided into two sub-tasks: image-to-text retrieval and text-to-image retrieval[6].

The essential challenge of this task lies in cross-modal semantic alignment. An ideal retrieval model needs to meet three levels of requirements: First, at the feature representation level, it needs to extract discriminative single-modal features; second, at the cross-modal interaction level, it needs to establish fine-grained correspondences between image regions and text words; and third, at the semantic reasoning level, it needs to understand deep semantic associations beyond surface vocabulary, such as spatial relationships, attribute descriptions, and action interactions[7-8].

2.2 Technological Evolution

The technological development of cross-modal retrieval has gone through three stages. Early methods were based on statistical learning techniques such as Canonical Correlation Analysis (CCA), learning shared subspaces through linear projection, but were limited by the representational capabilities of shallow models[9].

The introduction of deep learning brought about the first stage of breakthroughs. Feature enhancement-based (FE-based) methods aimed to strengthen the semantic representation capabilities of single modalities, such as using multi-region convolutional networks to mine visual details or using grammatical structure understanding to enhance text representations. However, simply enhancing single-modal features is difficult to solve the problem of cross-modal semantic misalignment[10].

The second stage focuses on cross-modal alignment. Cross-modal Alignment-based (CMA-based) methods establish the interaction between image regions and text words through attention mechanisms, progressive learning, and relational reasoning. Models such as SCAN introduce stacked cross-attention to achieve fine-grained alignment; subsequent work further utilized graph neural networks to mine the semantic relationships between targets.

The third stage is the era of visual-language pre-training. VLP-based methods utilize massive amounts of image and text data to learn general cross-modal representations, significantly outperforming the first two types of methods on downstream tasks. The success of this paradigm is attributed to three factors: large-scale training data, the scalability of the Transformer architecture, and cleverly designed pre-training tasks.

2.3 VLP Model Classification Perspective

Existing VLP models can be classified from multiple dimensions. Based on model architecture, they can be divided into dual-encoder architectures (such as CLIP and SigLIP) and fusion encoder architectures (such as ALBEF and VLMO). Dual encoders encode images and text separately, aligning the representation space through contrastive learning, resulting in high retrieval efficiency; fusion encoders introduce cross-modal attention for deep interaction, offering better accuracy but at a higher computational cost.

According to pre-training objectives, it can be divided into contrastive, generative, and hybrid models. The contrastive model aims to bring matching image-text pairs closer together and push away mismatched pairs; the generative model requires the model to generate content of another modality based on the input modality; the hybrid model combines the advantages of both.

Based on the quality of pre-training data, they can be divided into methods relying on high-quality labeled data (such as OSCAR) and methods utilizing large-scale noisy data (such as CLIP). The former offers high alignment quality but limited scalability, while the latter has significant scale advantages but faces noise interference problems.

ISSN 2616-5767 Vol.9, Issue 3: 15-21, DOI: 10.25236/AJETS.2026.090302

3. Analysis of VLP-based Cross-modal Retrieval Methods

3.1 Analysis of Typical VLP Model Architectures

3.1.1 Dual Encoder Architecture

The dual-encoder architecture uses independent visual encoders and text encoders to extract features, and then calculates cross-modal matching scores through a similarity function. CLIP is a typical example of this architecture. Its visual encoder uses ResNet or ViT, and its text encoder uses Transformer. It is trained on 400 million image-text pairs through contrastive learning.

The core advantage of this architecture lies in retrieval efficiency. Since image and text features can be pre-extracted and indexed, online retrieval only requires calculating the similarity between the query features and the index features, resulting in a fast response speed. However, dual encoders lack fine-grained cross-modal interactions and are difficult to capture the precise correspondence between image regions and text words.

To address this limitation, subsequent works have proposed several improvement strategies. SigLIP introduces the Sigmoid loss function to replace Softmax, making training more stable. ELIP proposes inserting a lightweight mapping network into the dual encoder, enabling the image encoding process to perceive text information and improve alignment quality without sacrificing efficiency.

3.1.2 Fusion Encoder Architecture

The fusion encoder architecture introduces a cross-modal interaction module on the basis of dual encoding, allowing visual and text features to be fully integrated during the encoding process. ALBEF uses dual encoders to extract features and then performs cross-modal attention interaction through a multimodal fusion encoder, effectively improving fine-grained alignment capabilities.

BLIP-2 further introduces Q-Former as an information bottleneck, extracting the most relevant visual features from the visual encoder through learnable query vectors, maintaining interaction quality while controlling computational cost. VL-GAP adopts a two-stage training strategy, first performing supervised learning on high-quality labeled data, and then performing self-learning on large-scale noisy data, balancing alignment quality and data scale.

Table 1 compares the characteristics and representative models of the above typical architectures.

Table 1: Comparison of typical VLP model architectures.

Architecture Type	Representative Model	Interaction Granularity	Retrieval Efficiency	Alignment Accuracy	Parameter Scale
Dual Encoder	CLIP, SigLIP	Global	High	Medium	Hundreds of millions to billions
Dual Encoder + Lightweight Interaction	ELIP	Global + Guided	Higher	Higher	Hundreds of millions to billions
Dual Encoder + Fusion Encoder	ALBEF, VLMO	Fine Granularity	Medium	High	Hundreds of millions to billions
Dual encoder + Q-Former	BLIP-2	Fine Granularity	Medium	High	Billion-Level

3.2 Pre-training Objectives and Alignment Strategies

Pre-training objectives are the core driving force for VLP models to learn cross-modal alignment. Existing pre-training objectives can be summarized into the following three categories.

3.2.1 Contrastive Learning Objectives

Contrastive learning is the most widely used pre-training method for VLP models. Its basic idea is to narrow the distance between matching image-text pairs in the representation space while widening the distance between unmatched pairs. InfoNCE loss is a typical implementation:

ISSN 2616-5767 Vol.9, Issue 3: 15-21, DOI: 10.25236/AJETS.2026.090302

$$L_{contrast} = -log \frac{exp(sim(v,t^+)/\tau)}{\sum_{j=1}^{N} \exp(sim(v,t_j)/\tau)} \quad (1)$$

Where v are image features, t^{+} are positive sample text features, t_{j} are all text features within the batch, are temperature coefficients. The key advantage of contrastive learning lies in its ability to efficiently utilize large-scale data, but it faces the challenge of insufficient diversity of negative samples within a batch.

3.2.2 Matching Objectives

Image-Text Matching (ITM) formalizes the task as a binary classification problem, requiring the model to predict whether a given image-text pair matches. Unlike contrastive learning, ITM typically requires deep interaction through encoder fusion, enabling the capture of more refined semantic relationships, but at a higher computational cost. ELIP-B introduces an ITM Head based on BLIP-2, effectively improving retrieval accuracy.

3.2.3 Generative Objectives

Generative objectives require the model to generate content of another modality based on the input modality. For example, the MLM (Masked Language Modeling) task randomly masks text tokens, requiring prediction of the masked words based on the image and context; the MIM (Masked Image Modeling) task predicts the masked image patches. Generative objectives force the model to build a deeper cross-modal understanding, but training is more difficult.

3.2.4 Multi-Objective Joint Optimization

Existing VLP models generally employ multi-objective joint optimization strategies. For example, VL-GAP jointly optimizes the contrastive loss, matching loss, and generation loss in the first stage, and performs self-learning through momentum models and confidence filtering mechanisms in the second stage, effectively improving model robustness. OmniBridge proposes a two-stage decoupled training strategy, first aligning multimodal inference capabilities through supervised fine-tuning, and then aligning the cross-modal latent space through semantically guided diffusion training.

3.3 Downstream Adaptation and Transfer Learning

After pre-training, how to efficiently adapt to downstream retrieval tasks is the key to practical application. Existing adaptation methods can be divided into three categories.

3.3.1 Full Fine-Tuning

Full fine-tuning updates all parameters of the pre-trained model, achieving the best adaptation effect but with high computational cost, and faces the risk of overfitting in small data scenarios. For retrieval tasks, full fine-tuning is usually combined with triplet loss or contrastive loss for optimization.

3.3.2 Efficient Parameter Fine-tuning

Efficient parameter fine-tuning involves updating only a small number of newly added parameters, freezing the core of the pre-trained model. Typical methods include prompt tuning, adapters, and LoRA. The work at ELIP demonstrates that significant improvements can be achieved on multiple pedestal models by training only lightweight mapping networks, greatly lowering the barrier to entry for VLP model adaptation in academia.

3.3.3 Retrieval-Based Adaptation

Retrieval-based adaptation methods do not update model parameters but improve retrieval quality through techniques such as re-ranking. ELIP's two-stage retrieval strategy first uses CLIP/SigLIP for initial screening, and then re-ranks the top-k candidates through a lightweight mapping network, improving accuracy while maintaining efficiency. CDISA enhances feature interaction through a deep interaction module and improves modality discrimination by combining bidirectional cosine matching.

4. Experiment and Evaluation System

4.1 Mainstream Datasets

Cross-modal retrieval research mainly relies on the following publicly available datasets for evaluation.

ISSN 2616-5767 Vol.9, Issue 3: 15-21, DOI: 10.25236/AJETS.2026.090302

Flickr30K contains 31,000 images, each labeled with 5 text descriptions, and is a standard benchmark for retrieval tasks. The dataset is divided into 29,000 training images, 1,000 validation images, and 1,000 test images.

MSCOCO is larger, containing 123,000 images, also with 5 descriptions per image. The standard division uses 113,000 training images, 5,000 validation images, and 5,000 test images. Evaluation can be done by averaging the 50% of the 1,000 test images or by evaluating the entire 5,000 test images.

Other datasets such as Pascal-Sentence and NUS-WIDE have also been used in specific studies. In vertical fields such as remote sensing, remote sensing image retrieval datasets place special demands on multi-scale modeling and small target perception.

4.2 Evaluation Metrics

Cross-modal retrieval adopts standard evaluation metrics in the field of information retrieval:

- **Recall (Recall@K, R@K)** measures the proportion of the top K search results containing at least one correct match, commonly K=1, 5, or 10. R@1 reflects retrieval precision, while R@5 and R@10 reflect retrieval coverage.

- **Median Rank (MedR)** refers to the median rank of the correct match in the retrieval results list; a smaller value indicates better retrieval performance.

- **Mean Recall (mR)** is the average of R@1, R@5, and R@10, used for overall performance comparison.

4.3 Performance Comparison of Representative Methods

Table 2 summarizes the performance comparison of representative methods on the Flickr30K and MSCOCO datasets. The data shows that VLP-based methods significantly outperform non-pre-trained methods. Dual encoder architectures (such as CLIP) are superior in efficiency, while fusion encoder architectures (such as ALBEF and BLIP-2) lead in accuracy. ELIP achieves significant performance improvements with minimal parameter increases by attaching a lightweight mapping network to the base model. CIMN introduces graph inference and quality loss within an independent matching framework, achieving a good balance between matching efficiency and performance.

Table 2: Performance comparison of representative methods on Flickr30K and MSCOCO (R@1).

Methods	Architecture Type	Flickr30K (text-to-image retrieval)	Flickr30K (image-to-text retrieval)	MSCOCO (text-to-image retrieval)	MSCOCO (image-to-text retrieval)
SCAN	Cross-modal Alignment	48.6	67.4	38.7	58.7
CLIP	VLP Dual Encoder	68.7	85.2	44.3	63.1
ALBEF	VLP Fusion Encoder	73.1	89.6	49.2	68.3
BLIP-2	VLP+Q-Former	75.2	91.3	51.7	70.5
ELIP-C	VLP Enhancement	70.8	87.4	46.5	65.8
ELIP-B	VLP Enhancement	76.1	92.5	52.8	71.6
CIMN	Independent Matching	69.3	86.7	45.9	64.9

Note: The data in the table is comprehensive from relevant papers, and some values are approximate readings.

ISSN 2616-5767 Vol.9, Issue 3: 15-21, DOI: 10.25236/AJETS.2026.090302

5. Key challenges and cutting-edge issues

5.1 Fine grained semantic alignment

Current VLP models have made significant progress in global matching, but still face challenges at the fine-grained alignment level. Accurate description of multiple targets in an image, spatial relationships between targets, and attributes requires a model with sophisticated understanding.

Research on VL-GAP shows that target misalignment (i.e., model illusion) is the main obstacle affecting fine-grained alignment. Models may generate or retrieve semantic objects that do not match the image content, severely impacting model reliability. Centroid strategies and automatic annotation methods offer insights into this problem, but they are still far from a complete solution.

5.2 Robustness of Noise Data

The success of VLP models heavily relies on large-scale pre-training data, but image-text pairs collected from the internet are generally noisy—text descriptions may not perfectly match the image content, or even contain irrelevant information. Noisy data misleads the model into establishing incorrect semantic associations, leading to pre-training efficiency saturation.

A two-stage training strategy is an effective means of dealing with noise. VL-GAP first learns precise alignment patterns on high-quality labeled data, and then performs self-learning on large-scale noisy data through momentum models and confidence filtering, significantly improving robustness to noise. This idea suggests that data quality contributes more to model performance than data size.

5.3 Balance between reasoning efficiency and accuracy

The dual encoder architecture has high efficiency but limited accuracy, while the fusion encoder has excellent accuracy but high computational cost. Balancing the two is the key to practicality.

Recent research has explored various solutions. ELIP employs a two-level cascaded architecture, first efficiently filtering and then finely reordering. CIMN uses graph pooling for modality-wide semantic aggregation, approximating the accuracy of fusion encoders within an independent matching framework. Qwen3-VL-Embedding supports Matryoshka representation learning, allowing flexible adjustment of embedding dimensions according to application scenarios, providing a new approach to the efficiency-accuracy tradeoff.

6. Future research directions and prospects

6.1 Small sample and zero sample generalization

Although VLP models perform well in zero-shot retrieval, their performance drops significantly when faced with specific domains or novel concepts. How to effectively transfer pre-trained knowledge to data-scarce vertical domains is an important research direction. Cue-based methods guide the model to adapt to new domains by designing learnable cue templates; parameter-based methods explore modular parameter combinations; and feature-based methods reduce domain differences through feature adaptation.

6.2 Unified Multi Task Framework

Understanding, generation, and retrieval tasks have long been studied separately; in recent years, unifying multi-task frameworks has become a hot topic. OmniBridge supports three tasks—visual language understanding, generation, and retrieval—within the same architecture through a bidirectional latent alignment module, and alleviates inter-task interference through a decoupling training strategy. The Qwen3-VL series unifies the embedding model and the re-ranking model on the same base, forming an end-to-end high-precision multimodal retrieval pipeline.

6.3 Interpretability and controllability

The key issues in practical applications are why the search results match and how to improve them. The current model is mostly used as a black box and lacks explanatory power for matching criteria.

ISSN 2616-5767 Vol.9, Issue 3: 15-21, DOI: 10.25236/AJETS.2026.090302

Introducing attention visualization and cross modal attribution analysis enables the model to present the image regions and text words that match, enhancing the credibility and usability of the model.

7. Conclusion

Visual-language pre-trained models have completely changed the technical landscape of cross-modal image retrieval and image-text matching. From feature enhancement to cross-modal alignment and then to the pre-training paradigm, this field has undergone profound technological changes. This paper systematically reviews the architecture design, pre-training objectives, and adaptation methods of VLP models in cross-modal retrieval, analyzes the three major challenges of fine-grained alignment, noise robustness, and efficiency-accuracy trade-off, and looks forward to future directions such as few-shot generalization, unified multi-task framework, and interpretability.

Current research shows that scaling up pre-trained models will continue to bring performance improvements, but the importance of data quality and alignment mechanisms is becoming increasingly prominent. Future research needs to strike a balance between model capabilities, data efficiency, and interpretability, and promote the development of cross-modal retrieval technology towards a more intelligent and reliable direction, so as to better support practical applications such as multimodal search and AI assistants.

References

[1] Dong C, Wei C. A review for image–text matching from deep learning perspective[J]. Information Fusion, 2026, 126: 103453.
[2] Li M, Wang H, Zhang Y, et al. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A Unified Framework for State-of-the-Art Multimodal Retrieval and Ranking[J]. arXiv preprint arXiv:2601.04720, 2026.
[3] Qin Y, Xie H, Ding S, et al. Enhancing vision-and-language transformers through two-stage generative alignment pre-training[J]. Engineering Applications of Artificial Intelligence, 2025, 142: 109876.
[4] Zhang Y, Wang L, Chen X. A review of cross-modal image-text retrieval[J]. Remote Sensing, 2025, 17(24): 3995.
[5] Zhan Y, Liu J, Wang T, et al. ELIP: Enhanced Language-Image Pre-training for Multimedia Retrieval[C]. Proceedings of the IEEE International Conference on Content-Based Multimedia Indexing (CBMI), 2025: 156-163.
[6] Liu W, Chen Y, Zhao H. Cross-Modal Deep Interaction and Semantic Aligning for Image-Text Retrieval[J]. IEICE Transactions on Information and Systems, 2025, E108.D(10): 1230-1238.
[7] Li X, Li J, Li F, et al. Generalizing Vision-Language Models to Novel Domains: A Comprehensive Survey[J]. arXiv preprint arXiv:2506.18504, 2025.
[8] Wang T, Li F, Zhu L, et al. Cross-Modal Retrieval: A Systematic Review of Methods and Future Directions[J]. Proceedings of the IEEE, 2024, 112: 1716-1754.
[9] Xiao T, Wang S, Li Z, et al. OmniBridge: Unified Multimodal Understanding, Generation, and Retrieval via Latent Space Alignment[J]. arXiv preprint arXiv:2509.19018, 2025.
[10] Wu J, Chen L, Zhang R. Cross-modal independent matching network for image-text retrieval[J]. Pattern Recognition, 2025, 159: 111096.