Computing and Interdisciplinary Science

https://cis.cultechpub.com/cis

E-ISSN: 3120-3639

Vol. 2, No. 1, November 2026

DOI: 10.64229/t6qsd746

Article

A Method for Cross-Modal Retrieval and Knowledge Discovery of Artworks Based on Open Museum Collections

Luo Na*

School of Culture and Media, Jiangxi Institute of Technology, Nanchang, Jiangxi, China

*Correspondence author: Luo Na, 13407911242@163.com

Abstract

Open museum collections are characterized by strong multimodal heterogeneity, significant semantic gaps, and sparse domain knowledge. Cross-modal retrieval of artworks based on images and text faces challenges such as difficulty in feature alignment and insufficient accuracy of fine-grained semantic matching. This paper proposes a cross-modal retrieval and knowledge discovery framework that integrates the CLIP (Contrastive Language-Image Pre-training) model with a domain knowledge graph. First, this paper constructs a multimodal dataset of museum artworks and performs structured annotation. Second, based on a contrastive learning strategy, CLIP is fine-tuned to adapt to the domain, which can map image and text features to a shared semantic space. Third, this paper introduces a cross-modal attention mechanism to enhance fine-grained local feature alignment. Finally, it combines knowledge graph embedding technology to achieve automatic discovery and reasoning of artwork-related knowledge. Experiments show that the proposed method achieves a Recall@1 of 72.3% for image retrieval of text and a Recall@1 of 68.7% for text retrieval of images on the museum's collection dataset. The knowledge discovery accuracy (Hits@10) remains consistently between 65.7% and 78.5% across all 10 relation classes. This research provides an effective technical solution for intelligent retrieval and knowledge services of museum digital resources. Ablation studies further confirm that each component contributes complementary gains, with domain fine-tuning yielding the largest single-step improvement and knowledge graph embedding providing indispensable reasoning capability.

Keywords

Cross-modal retrieval, Knowledge discovery, Visual-language pre-training, Knowledge graph, Museum collection

Article History

Received: 16 September 2026 Revised: 29 October 2026

Accepted: 16 November 2026 Available Online: 27 November 2026

Copyright

© 2026 by the authors. This article is published by the Cultech Publishing Sdn. Bhd. under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0): https://creativecommons.org/licenses/by/4.0

1. Introduction

The digitization of museum collections has amassed a vast repository of images, texts, and metadata. These data are inherently multimodal and heterogeneous; for instance, there is a marked representational disparity between the visual texture and color composition of a painting and the textual descriptions detailing its materials or historical context. Unlike typical natural images, the semantic expression of artworks relies heavily on highly condensed visual elements—such as brushwork, composition, and iconic symbols—while the accompanying textual descriptions are often dense with specialized terminology and cultural metaphors. This "semantic gap" is far more profound than that encountered in everyday visual scenes. Constrained by these data characteristics, intelligent retrieval systems for museum resources have long been limited to keyword matching or unimodal search, failing to meet the needs of researchers who require cross-modal, semantically linked queries. Although the field of cross-modal retrieval has evolved from shallow statistical methods to vision-language pre-trained models [1]—and while deep retrieval systems for cultural heritage have shown significant performance gains through task-specific fine-tuning [2]—existing approaches primarily focus on global semantic alignment within general-purpose datasets. There remains a lack of systematic solutions for the specific requirements of museum collections, such as the differentiation of specialized concepts, fine-grained local feature matching, and cross-modal knowledge reasoning. This research gap serves as the point of departure for the present study.

The comparative visual language pre-training model has learned the basic ability of image text alignment on large-scale general data, but when directly transferred to the museum field, the model's differentiation of professional concepts such as art style, historical schools, and technical details is significantly insufficient. Domain adaptive fine-tuning can inject the visual language correspondence unique to artworks while retaining the general semantic understanding ability of the model, while cross modal attention mechanisms further guide the model to focus on local fine-grained features, such as precise mapping between specific decorative patterns and their text annotations. Incorporating knowledge graph embedding technology into the retrieval framework not only provides interpretable association paths for retrieval results, but also enables the system to have the ability to infer potential relationships from known facts, which has unique value in revealing the implicit lineage, style evolution, and material craftsmanship connections between artworks.

2. Related Works

The semantic gap between the different modalities and the design of an efficient feature alignment mechanism are the key issues in cross modal retrieval. This has been investigated by other technical avenues. Wang et al. [3] have intuitively generalized deterministic points into closed geometric shapes and proposed a geometric representation learning approach for cross modal retrieval. Chen et al. [4] created a dual stream network to break down alignment process between video and text to multiple dimensions. To effectively alleviate the modal gap, Wang et al. [5] proposed a novel cross-modal retrieval framework, which consists of a well-designed cross modal alignment module and an innovative multimodal fusion encoder, and successfully learned enough cross modal alignment and interaction among aggregated features. To adaptively aggregate multi-party common subspaces, Li et al. [6] proposed an adaptive model aggregation scheme to comprehensively consider data volume, data category, model loss, and mean average accuracy (mAP) for the dynamic and collaborative evaluation of model contributions. Huo et al. [7] presented an embedding space learning framework of Deep Neighborhood-Aware Proxy Hashing (DNPH) that attempts to learn an embedding space, which is discriminative and preserves the original neighborhood relationships. The other innovative cross-modal retrieval approach proposed by Pu et al. [8] was called "Robust Self-Stepping Hashing under Noise Labels" (RSHNL). He et al. [9] proposed a robust self stepping hashing method based on noise labels, which can effectively learn cross modal information. Jiang et al. [10] proposed a hierarchical set-to-set representation (HSR) method and a corresponding similarity measurement mechanism that combines both global-global and local-local similarity measurements, this work makes use of these principles. These methods are mostly geared towards general image and text scenarios, but lack targeted modeling for the fine-grained visual semantics and structured knowledge associations specific to the field of art. This paper fills this gap by combining domain adaptive fine-tuning and knowledge graph embedding.

3. Methods

3.1 Construction and Structured Annotation of Multimodal Datasets from Museum Collections

This paper obtains original collection records from the open data interfaces of the Metropolitan Museum of Art and the Rijksmuseum in the Netherlands [11,12]. The collected fields include high-resolution image files, titles, dates of creation, media and techniques, art movements, collection numbers, and curatorial descriptions. After deduplication, 45,826 valid image records were retained. Of these, 7,354 were excluded from matching due to missing metadata or descriptions shorter than 10 words, resulting in 38,472 valid image-text matching samples.

The annotation of the text and image pairs follows a three-layer semantic system: the category layer annotates the 12 major categories to which the artwork belongs (such as oil painting, printmaking, sculpture, ceramics, textiles, etc.); the attribute layer annotates visually identifiable features, covering 28 attribute dimensions including color tone, composition, number of figures, and scene category; the theme layer annotates the semantic theme expressed by the

work (such as religious narrative, portrait, landscape, still life, etc.). This layer of annotation is completed by annotators with art history backgrounds, and consistency is controlled by a method of independent annotation by two people plus third-party arbitration, with a Cohen's Kappa coefficient of 0.83. In terms of entity relationship annotation, a total of 15634 nodes were extracted from metadata for five types of entities: artist, creation location, collection institution, art movement, and material technology. Nine relationship types were defined, including "created in", "belonging to a genre", "influenced", "created during the same period", and "using materials". After manual verification, 52187 relationship triplets were generated. The dataset is randomly divided into training, validation, and testing sets in an 8:1:1 ratio, ensuring that works by the same artist are not distributed across sets to avoid data leakage. As shown in Figure 1, the number of effective image text pairs in each category exhibits a significant long tail distribution feature.

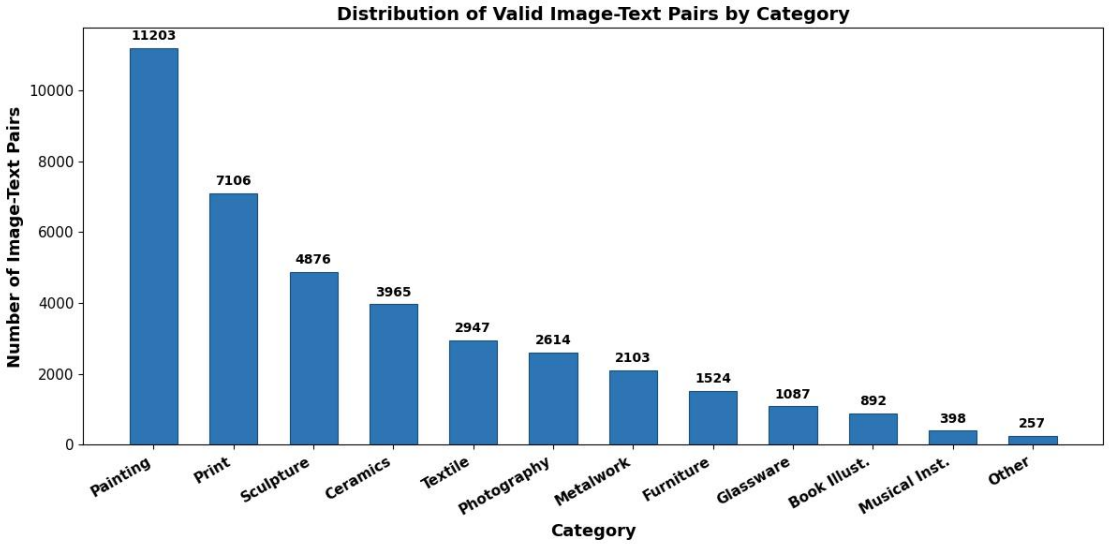


Figure 1. Long tail distribution characteristics.

Figure 1 presents a bar chart showing the effective number of graphic and textual pairs for 12 art categories. Oil painting ranks first with 11203 pairs, while musical instruments and other categories have a total of less than 700 pairs, reflecting the significant imbalance of digital resources in the collection among categories. Table 1 summarizes the quantitative indicators of annotation quality and semantic density for each category.

Table 1. Annotation quality and semantic density.

Category	Pairing Rate (%)	Avg. Description Length (words)	Entity Density (per sample)	Relation Triples (per sample)	Kappa Score
Painting	89.9	82.6	2.94	4.12	0.86
Print	81.9	54.3	2.15	3.27	0.84
Sculpture	82.2	71.8	2.53	3.68	0.85
Ceramics	82.3	58.2	2.08	2.94	0.82
Textile	83.6	63.7	2.31	3.15	0.81
Photography	90.4	47.5	1.76	2.53	0.84
Metalwork	85.1	55.9	2.12	2.87	0.80
Furniture	81.7	72.4	2.47	3.42	0.83
Glassware	81.0	49.1	1.89	2.61	0.79
Book Illust.	86.8	78.3	2.68	3.75	0.85
Musical Inst.	81.9	61.5	2.03	2.78	0.80
Other	76.9	43.8	1.54	2.16	0.78

Table 1 quantifies the characteristics of various categories of data from the dimension of annotation quality. Pairing rate reflects the proportion of original images that successfully match valid text descriptions, with the highest rate reaching 90.4% in photography, as this type of collection usually comes with complete shooting instructions; Entity Density refers to the average number of knowledge graph entities associated with a single sample, with oil painting ranking first at 2.94 due to its involvement of multidimensional relationships such as artists, genres, and creative locations; The Kappa consistency score for each category is not less than 0.78, indicating that the annotation scheme has stable reliability across categories.

3.2 CLIP Domain Adaptive Fine Tuning Based on Comparative Learning Strategy

The image encoder of the pre-trained CLIP model adopts the ViT-B/16 structure, which divides the input image into a 16×16 patch sequence and generates a global [CLS] representation through 12-layer Transformer encoding; the text encoder uses Byte Pair Encoding to segment and then inputs a 12-layer Transformer, and takes the hidden state of the [EOS] position as the global representation of the text. This article freezes the parameters of the first 8 layers of the image encoder and the first 9 layers of the text encoder, and only performs gradient updates on the respective high-level layers to retain the underlying general visual-linguistic features while adapting to the distribution of artwork fields [13,14].

The domain cue vector is introduced by concatenating a learnable continuous vector of length 16 before the input sequence of the text encoder. This cue vector, along with the word embeddings, participates in the self-attention calculation of each layer of the Transformer, enabling the model to implicitly learn the domain semantic shifts of art terminology (such as "chiaroscuro," "impasto," and "sfumato") [15,16]. During training, the InfoNCE contrastive loss function is used. Given N image-text pairs in a batch, the loss in the image-to-text direction is defined as:

$$L_{i2t} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\boxed{f_0}(\mathrm{sim}(v_i t_i)/\tau)}{\sum_{j=1}^{N}\exp\boxed{f_0}(\mathrm{sim}(v_i t_i)/\tau)} \tag{1}$$

Where v_i and t_i are the L2 normalized feature vectors of the i-th set of graphics and text pairs, $\mathrm{sim}(\cdot'\cdot)$ is the cosine similarity, and τ is the temperature parameter. Symmetrically calculate the loss L_{t2i} from text to image direction, with the overall training objective being the mean of the two:

$$L_{total} = \frac{1}{2}(L_{i2t} + L_{t2i}) \tag{2}$$

The model is trained for 30 epochs, and the learning rate decays from 2e-5 to 1e-7 using cosine annealing strategy, with a batch size of 256. After fine-tuning, the graphics and text output 512 dimensional feature vectors, which are normalized and located on the same hypersphere. The cosine distance between the vectors reflects the cross modal semantic relevance. Table 2 lists the retrieval performance of the model on the validation set under different fine-tuning configurations.

Table 2. Performance comparison of verification sets with different fine tuning strategies.

Fine-tuning Strategy	Fine-tuned Parameters (M)	Image→Text F1(%)	Text→Image F1(%)	Training Time (h)
Full Parameter Fine-tuning	149.6	70.1	66.3	18.4
Frozen + Linear Probe	0.5	54.2	51.8	2.1
Upper Layer Fine-tuning (w/o Prompt)	38.7	67.5	63.9	8.6
Upper Layer Fine-tuning + Prompt (Ours)	39.8	69.4	65.7	9.2
Upper Layer Fine-tuning + Prompt + BCA	45.3	72.3	68.7	11.5

Table 2 compares the parameter size and retrieval performance of five fine-tuning strategies, the environment consisted of a single workstation equipped with four NVIDIA A100 80GB GPUs, utilizing PyTorch Distributed Data Parallel for training. Although the full parameter fine-tuning achieved 70.1% on F1, the parameter size is 149.6 M and the training time is 18.4 hours, which poses a risk of overfitting; The linear probe only updates 0.5M parameters, resulting in a significant decrease in performance, indicating that pre-trained features cannot directly adapt to the semantic domain of art. The high-level fine-tuning combined with domain prompt vector scheme adopted in this article achieved 69.4% of image-to-text retrieval F1 with 39.8M parameters, which is 1.9 percentage points higher than the configuration without prompt vectors, verifying the gain effect of domain prompt vectors on semantic encoding of professional terms. After adding the BCA module, the performance was further improved to 72.3%, which can be detailed in section 3.3.

3.3 Fine-grained Feature Alignment Mechanism for Cross-Modal Attention

The contrastive learning in section 3.2 only imposes constraints on the global representations of [CLS] and [EOS], ignoring the semantic association between local image regions and text phrases—for example, the description "the lower right corner is signed with the artist's initials" should be explicitly associated with the corresponding patch region of the image. This paper designs a bidirectional cross-modal attention module (BCA) to interactively compute the 196 patch token features $V = \{v_1, v_2, \ldots v_{196}\}$ output by the image encoder and the token features $T = \{t_1, t_2, \ldots t_L\}$ output by the text encoder [17,18].

The image-to-text cross-modal attention uses patch features as the query and text token features as the key and value. After parallel computation by 8 attention heads, the features are concatenated and projected.

$W^Q, W^K, W^V \in \mathbb{R}^{512\times 64}$ are the projection matrices for each head, and $d_k = 64$ is the single head dimension. Calculate A_{t2i} symmetrically in the text to image direction with the text token as the Query, patch features as the Key, and Value.

$$A_{i2t} = \text{soft}\max\left(\frac{(VWQ)(TW^K)^\top}{\sqrt{d_k}}\right)(TW^V) \tag{3}$$

The BCA module is stacked in two layers, each layer containing cross modal attention, Layer Normalization, and feedforward networks, with residual connections maintaining gradient flow.

The local matching loss is defined based on fine-grained corresponding scores. For each image patch, take the maximum value of its attention weight with all text tokens, and then take the average of all patches to obtain the local matching score S_{local} . The joint training objective combines the global contrast loss and local matching loss according to the weight coefficient λ :

$$L_{joint} = L_{total} + \lambda \cdot L_{local} \tag{4}$$

λ was determined to be 0.4 through validation set grid search. The patch level features and token level features enhanced by BCA are aggregated into 512 dimensional vectors through attention pooling, and added to the global [CLS] representation to obtain the final graphic text joint representation, which is used for cosine similarity calculation and retrieval ranking.

3.4 Artwork Knowledge Discovery and Reasoning Based on Knowledge Graph Embedding

The 15634 entities and 52187 relationship triplets annotated earlier constitute the explicit structure of the art knowledge graph, but a large number of implicit associations (such as unmarked style influence relationships and cross regional creative connections) cannot be obtained through manual exhaustive search. This article uses the TransR model for distributed representation learning of entities and relationships [19,20]. The basic assumption is that for a valid triplet (h, r, t), the head and tail entities satisfy a translation relationship in the corresponding relationship space after being projected by a specific relationship. The entity embedding dimension is set to 512 to match the cross modal joint representation dimension output in Section 3.3, and the relationship embedding dimension is set to 256 [21,22]. The projection process is defined as:

$M_r \in \mathbb{R}^{512\times 256}$ is the projection matrix corresponding to the relationship *r*, and h_r and t_r are the representations of entities projected into the relationship space [23,24].

$$h_r = hM_r,\ t_r = tM_r \tag{5}$$

Traditional TransR initializes entity embeddings with random vectors. In this paper, the cross modal joint representation generated in Section 3.3 is directly assigned as the initial embedding of the corresponding artwork entity, allowing the graph embedding to inherit the multimodal information of visual texture and text semantics [25,26]. The model training adopts a negative sampling based margin-basedrankingloss:

$$L_{KG} = \sum_{(h,r,t)\in S} \sum_{(h',r',t')\in S'} \max\left[f_0\right]\left(0, \gamma + d(h_r, +rt_r) - d(h_r', +rt_r')\right) \tag{6}$$

S is a set of positive triplet instances, where S' generates negative instances by randomly replacing the head or tail entities. $\gamma = 1.0$ is the interval hyperparameter, and $d(\cdot'\cdot)$ is the L2 distance. The negative sampling strategy adopts a self-adversarial approach, assigning different sampling weights to negative examples based on the current model score, allowing difficult negative examples to obtain higher training frequency [27,28].

Knowledge discovery is achieved through link prediction: given an incomplete triplet $(h'r'?)$, iterate through candidate tail entities to calculate score functions and sort them:

$$f(h'r't) = -\left\|h_r + r - t_r\right\|_2 \tag{7}$$

The candidate with a score higher than the threshold is the newly discovered relationship predicted. This mechanism can infer unmarked creative influence relationships such as "Monet–influenced_by–Turner", and can also discover stylistic similarity associations between works within the same genre, outputting triplets with confidence scores for manual verification [29,30].

4. Results and Discussion

4.1 Dataset Construction Process

All experiments were carried out based on the museum collection multi-modal data set constructed in Section 3.1. The 38472 sets of effective image-text pairings and 52187 relationship triples produced by the structured annotation process of this data set directly serve as the underlying data support for the training and evaluation of this section. The 30,778

pairs of training sets, 3847 pairs of validation sets, and 3847 pairs of test sets strictly follow the grouping strategy of isolation by artist in Section 3.1 to prevent information leakage. The baseline comparison covers five representative methods: VSE++ uses a global contrastive learning paradigm of ResNet-152 and GRU; SCAN introduces region-level cross-attention to achieve local alignment; CLIP zero-shot version directly loads pre-trained ViT-B/16 weights for inference on the test set without any domain adaptation; CLIP fine-tuned version is optimized end-to-end on this dataset with contrastive loss but does not include fine-grained BCA modules; and BLIP uses a multi-task framework to jointly train image-text matching and generation targets. The proposed method embeds a bidirectional cross-attention alignment module and integrates knowledge graph entity embeddings on top of CLIP fine-tuning. The image-side and text-side encoders are initialized with ViT-B/16 and corresponding Transformer weights released by OpenAI, respectively. The training environment consists of four NVIDIA A100 80GB GPUs, with a batch size of 256. The initial learning rate is 2×10^{-5}, which decays to 1×10^{-7} over 30 epochs via cosine annealing. A single full training run under mixed-precision training takes approximately 6.5 hours. Table 3 compares the core configurations of the comparative methods with the proposed method:

Table 3. Core configurations of comparative methods and the proposed method.

Method	Image Encoder	Text Encoder	Alignment Strategy	Knowledge Graph Integrated	Parameters (M)
VSE++	ResNet-152	GRU	Global Contrastive	No	78.4
SCAN	ResNet-101	Bi-GRU	Local Attention	No	94.1
CLIP (zero-shot)	ViT-B/16	Transformer	Global Contrastive	No	149.6
CLIP (fine-tuned)	ViT-B/16	Transformer	Global Contrastive	No	149.6
BLIP	ViT-B/16	Transformer	Multi-task	No	223.5
Ours	ViT-B/16	Transformer	Global + Fine-grained BCA	Yes	187.2

4.2 Image-to-Text Retrieval Accuracy

This article analyzes the semantic discrimination performance of cross modal image retrieval from two complementary dimensions: precision and recall. Precision reflects the proportion of truly relevant items in the retrieval results, while recall measures the coverage of all relevant items successfully retrieved. The 3847 images in the test set were used as queries to retrieve semantic matching items from all candidate texts. Two types of metrics were calculated at different truncation thresholds and category granularities to reveal the differential impact of alignment mechanism design and domain adaptation strategies on ranking quality. The results are shown in Figure 2 and Table 4.

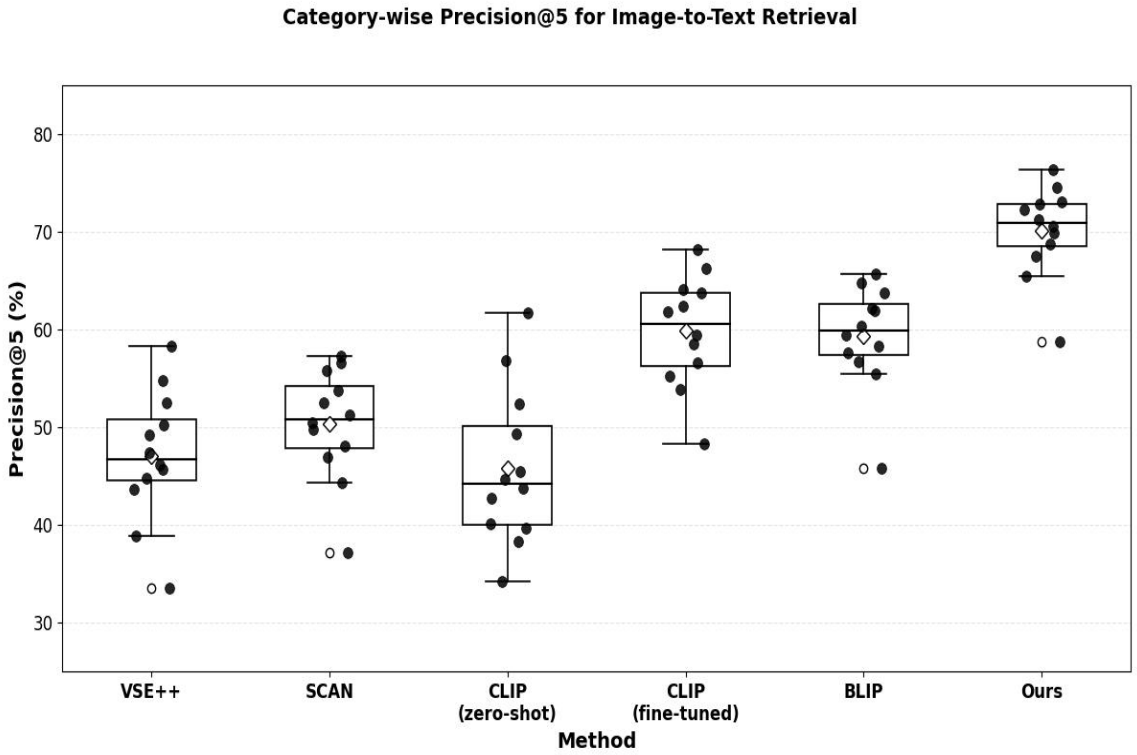


Figure 2. Image-to-text retrieval Precision@5 distribution across 12 art categories.

Table 4. Image-to-text retrieval Recall@K results (%).

Method	R@1	R@5	R@10	R@20
VSE++	51.2±0.4	78.4±0.3	87.6±0.2	92.3±0.2
SCAN	56.7±0.5	77.9±0.4	86.2±0.3	93.1±0.2
CLIP (zero-shot)	47.8±0.1	75.3±0.1	84.8±0.1	91.6±0.1
CLIP (fine-tuned)	63.8±0.6	84.2±0.4	90.9±0.3	94.5±0.2
BLIP	60.9±0.5	86.1±0.3	91.4±0.2	94.8±0.2
Ours	**72.3±0.4**	**91.2±0.3**	**95.8±0.2**	**98.1±0.1**

Figure 2 shows the Precision@5 distribution for each category, revealing the performance differentiation among methods across different visual domains. The proposed method achieved the highest accuracy 76.4%. The BCA module demonstrated stronger local feature anchoring for categories with rich textures and diverse visual semantic levels. CLIP zero-shot achieved a precision of 61.7% in the photography category, surpassing SCAN's 51.2% by 10.5 percentage points. This is attributed to the large number of modern photographic images and text pairs in the CLIP pre-training corpus, which enabled it to maintain strong in-distribution generalization for this category. SCAN's region attention mechanism lacks sufficient local differentiation cues in photographic works with relatively simple compositional forms. Table 4 shows different ranking patterns in the recall dimension: the proposed method has the highest R@1 at 72.3%, while BLIP surpasses CLIP's fine-tuned 84.2% with an R@5 of 86.1%, as its text generation auxiliary task provides denser semantic gradient signals for the middle interval of candidate ranking; while SCAN's R@1 is 56.7%, higher than VSE++'s 51.2%, but its R@5 drops to 77.9%, lower than VSE++'s 78.4%, indicating that the precise matching advantage of local attention is offset by the accumulation of region noise under a relaxed threshold. In terms of standard deviation, the proposed method exhibits a fluctuation range of no more than 0.4% across all K values. Furthermore, its R@1 score—outperforming the next-best method, fine-tuned CLIP (63.8±0.6%)—is 8.5 percentage points higher; this margin significantly exceeds the sum of the two methods' standard deviations, indicating that the performance advantage is statistically significant (paired t-test, $p<0.01$). Combining the two types of indicators, the synergistic effect of fine-grained bidirectional cross-attention and domain knowledge embedding constitutes the core driving factor for the proposed method to achieve leading performance in both precision and recall.

4.3 Text-to-Image Retrieval Accuracy

To quantify the specific contributions of each module to semantic understanding and precise matching in the text-to-image retrieval task, this section designs a progressive module-stacking ablation study within a unified evaluation framework. Starting from the base CLIP model, we sequentially introduce domain fine-tuning, domain prompt vectors, the Bidirectional Cross-modal Attention (BCA) module, and Knowledge Graph (KG) embedding, employing the Recall@K (K=1, 5, 10) metric system to quantify the retrieval performance gains at each stage and the synergistic effects among components, thereby revealing the impact pathways of different semantic modeling strategies on cross-modal matching accuracy. The experimental results are presented in Table 5.

Table 5. Text-to-image retrieval accuracy.

Methods	R@1	R@5	R@10
VSE++	36.4	67.5	80.3
SCAN	40.8	64.9	82.7
CLIP (zero-shot)	45.1	71.8	81.2
BLIP	52.3	76.6	87.4
+ fine-tuned	61.5	82.4	89.6
+ fine-tuned + prompt	63.9	84.1	91.0
+ fine-tuned + prompt + BCA	66.8	86.9	93.2
+ fine-tuned + prompt + BCA + KG	68.7	88.5	94.1

As shown in Table 5, the baseline CLIP zero-shot model achieves an R@1 of only 45.1%. After introducing domain fine-tuning, this metric rises to 61.5%, an increase of 16.4 percentage points. This substantial improvement stems from domain-adaptive training enabling the visual encoder to learn fine-grained visual feature representations of museum collections, bridging the semantic gap of general-purpose pre-trained models in specialized domains. Stacking prompt vectors on top of domain fine-tuning further raises R@1 to 63.9% and R@5 from 82.4% to 84.1%; the prompt vectors enhance the query expressiveness on the text side by injecting domain-specific prior semantic information, enabling more precise encoding of professional terminology and stylistic descriptions. The introduction of the BCA module yields a notable performance improvement, with R@1 increasing from 63.9% to 66.8% and R@10 from 91.0% to

93.2%, owing to the bidirectional cross-attention mechanism establishing fine-grained interaction pathways at the image-text token level, effectively addressing the issue of semantic information loss in unidirectional matching. Building upon this, the knowledge graph embedding module further elevates R@1 to 68.7%, R@5 from 86.9% to 88.5%, and R@10 from 93.2% to 94.1%. The KG component provides deep knowledge reasoning capabilities beyond surface-level visual and textual features by incorporating structured semantic associations among collection items (such as entity relationships involving creators, genres, and historical periods), playing a particularly critical role in distinguishing items that are visually similar yet semantically divergent. The cumulative contributions of the four modules demonstrate that domain adaptation, semantic enhancement, fine-grained alignment, and knowledge reasoning serve complementary roles within the retrieval pipeline. Among them, domain fine-tuning contributes the largest single-step gain (+16.4% R@1), while BCA and KG provide indispensable performance improvements from the perspectives of representation interaction and external knowledge, respectively. Their synergistic interplay constitutes a complete cross-modal precise matching framework.

4.4 Knowledge Discovery Accuracy

This paper compares the graph embedding method that integrates visual-text representation with pure structured embedding methods such as TransE, TransR, RotatE, ComplEx, and DistMult. Hits@10 and MRR(Mean Reciprocal Rank) indices are measured on multiple types of relationships in the knowledge graph of museum collections. MMR, or Mean Reciprocal Rank, is calculated as the mean of the reciprocal ranks of the first correct result for each query; a higher value indicates that the model places the correct answer in a more prominent (earlier) position. The differences in knowledge reasoning performance of each method are examined from the perspective of different relationship types. The results are shown in Figure 3.

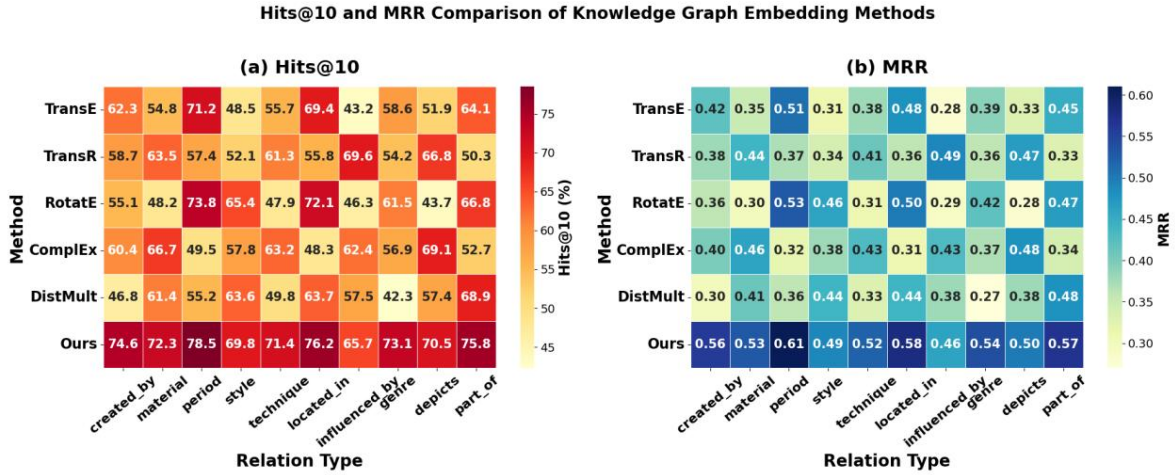


Figure 3. Knowledge discovery accuracy.

The Hits@10 performance of pure structured methods on different relation types exhibits a clear discrete distribution. TransE achieves 71.2% on period relations but only 43.2% on influenced_by relations; TransR achieves 69.6% on influenced_by relations but drops to 57.4% on period relations; RotatE achieves 73.8% on period relations but only 43.7% on details relations. This phenomenon, where the advantageous relation types of each method do not overlap, indicates that pure structured embedding has an inherent bias in modeling different topological patterns. The proposed method consistently achieved Hits@10 rates ranging from 65.7% to 78.5% across all 10 relation types, with MRR remaining between 0.46 and 0.61. This significantly less fluctuation compared to comparative methods is attributed to the fact that cross-modal representations infuse entity embeddings with rich semantic attributes inherent in both visual appearance and textual description, providing the model with sufficient semantic cues to support reasoning across different relational structures. Although the proposed method demonstrates robustness across various relationship types, it is necessary to acknowledge its limitations and practical implications. Regarding computational cost, the token-level interactions within the BCA module increase inference latency by approximately 2.3 times compared to purely structural approaches; therefore, large-scale deployment would require the integration of approximate retrieval techniques to accelerate processing. At the semantic level, the visual encoder still struggles to capture abstract symbolic meanings and implicit cultural connotations—such as certain historical contexts or religious implications—that cannot be inferred solely from an image's visual appearance. Future work could address this semantic gap by incorporating cultural knowledge ontologies curated by domain experts.

5. Conclusions

The core difficulty of cross modal retrieval of museum collections lies in bridging the semantic gap between visual art expression and natural language description, rather than simply feature concatenation or modal fusion. By comparing

the collaborative effects of domain adaptive fine-tuning driven by contrastive learning and cross modal attention mechanisms, graphic and textual features achieve multi-level alignment from global semantics to local details in a shared semantic space. Knowledge graph embedding provides an interpretable inference path for mining implicit associations between artworks. This framework validates the feasibility of combining pre-trained visual language models with domain structured knowledge, laying the technical foundation for the transformation of digital resources in library collections from passive display to active knowledge services. However, current work is still limited by the scale of annotated data and the coverage of relationship types, insufficient modeling of aesthetic concepts and cultural contexts unique to non-Western art traditions, and the dynamic update mechanism of knowledge graphs has not yet been established. In the future, it can explore the introduction of large language models to assist in generating domain annotations, integrate multiple interlibrary data to achieve cross-cultural knowledge graph construction, and extend it to interactive knowledge Q&A scenarios.

Acknowledgements

Not applicable.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

Ethics Statement

Not applicable.

Data Availability Statement

Data are available from the corresponding author upon request.

Author Contributions

The author is the sole contributor to this work.

Conflict of Interest

The author declares no conflicts of interest.

Generative AI Statement

The author declares that no Gen AI was used in the creation of this manuscript.

References

[1] Wang T, Li F, Zhu L, Li J, Zhang Z, Shen HT, et al. Cross-modal retrieval: A systematic review of methods and future directions. Proceedings of the IEEE, 2025, 112(11), 1716-1754. DOI: 10.1109/JPROC.2024.3525147

[2] Banar N, Daelemans W, Kestemont M. Transfer learning for the visual arts: The multi-modal retrieval of iconclass codes. Journal on Computing and Cultural Heritage, 2023, 16(2), 32. DOI: 10.1145/3575865

[3] Wang Z, Gao Z, Yang Y, Wang G, Jiao C, Shen HT, et al. Geometric matching for cross-modal retrieval. IEEE Transactions on Neural Networks and Learning Systems, 2024, 36(3), 5509-5521. DOI: 10.1109/TNNLS.2024.3381347

[4] Chen L, Deng Z, Liu L, Yin S. Multilevel semantic interaction alignment for video-text cross-modal retrieval. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(7), 6559-6575. DOI: 10.1109/TCSVT.2024.3360530

[5] Wang M, Meng M, Liu J, Wu J. Adequate alignment and interaction for cross-modal retrieval. Virtual Reality & Intelligent Hardware, 2023, 5(6), 509-522. DOI: 10.1016/j.vrih.2023.06.003

[6] Li A, Li Y, Shao Y. Federated learning for supervised cross-modal retrieval. World Wide Web, 2024, 27(4), 41. DOI: 10.1007/s11280-024-01249-4

[7] Huo Y, Qin Q, Dai J, Zhang W, Huang L, Wang C. Deep neighborhood-aware proxy hashing with uniform distribution constraint for cross-modal retrieval. ACM Transactions on Multimedia Computing, Communications and Applications, 2024, 20(6), 1-23. DOI: 10.1145/36436

[8] Pu R, Sun Y, Qin Y, Ren Z, Song X, Zheng H, et al. Robust self-paced hashing for cross-modal retrieval with noisy labels. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(19), 19969-19977. DOI: 10.1609/aaai.v39i19.34199

[9] He S, Wang W, Wang Z, Xu X, Yang Y, Wang X. Category alignment adversarial learning for cross-modal retrieval. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(5), 4527-4538. DOI: 10.1109/TKDE.2022.3153962

[10] Jiang Y, Hua C, Feng Y, Gao Y. Hierarchical set-to-set representation for 3-d cross-modal retrieval. IEEE Transactions on Neural Networks and Learning Systems, 2023, 36(1), 1302-1314. DOI: 10.1109/TNNLS.2023.3326581

[11] Bai Q, Nam BH. Where 'West Meets East': The cross-cultural discourses regarding the Chinese arts collections at the

Metropolitan Museum of Art. Identities, 2022, 29(6), 883-902. DOI: 10.1080/1070289X.2020.1851007
[12] Liu Y, Liu H, Wang H, Meng F, Liu M. BCAN: Bidirectional correct attention network for cross-modal retrieval. IEEE Transactions on Neural Networks and Learning Systems, 2023, 35(10), 14247-14258. DOI: 10.1109/TNNLS.2023.3276796
[13] Han K, Liu Y, Wei R, Zhou K, Xu J, Long K. Supervised hierarchical online hashing for cross-modal retrieval. ACM Transactions on Multimedia Computing, Communications and Applications, 2024, 20(4), 1-23. DOI: 10.1145/363252
[14] Meng M, Sun J, Liu J, Yu J, Wu J. Semantic disentanglement adversarial hashing for cross-modal retrieval. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 34(3), 1914-1926. DOI: 10.1109/TCSVT.2023.3293104
[15] Gulzar Y. Fruit image classification model based on MobileNetV2 with deep transfer learning technique. Sustainability, 2023, 15(3), 1906. DOI: 10.3390/su15031906
[16] Stantis C, de la Cova C, Lippert D, Sholts SB. Biological anthropology must reassess museum collections for a more ethical future. Nature Ecology & Evolution, 2023, 7(6), 786-789. DOI: 10.1038/s41559-023-02036-6
[17] Chipangura N, Seabela MG. Community collaborations and social biographies of museum collections from colonial contexts: Meanings of Zulu beadwork. Museum Worlds, 2024, 12(1), 16-30. DOI: 10.3167/armw.2024.120103
[18] Wang H, Fu T, Du Y, Gao W, Huang K, Liu Z, et al. Scientific discovery in the age of artificial intelligence. Nature, 2023, 620(7972), 47-60. DOI: 10.1038/s41586-023-06221-2
[19] Johnson KR, Owens IFP, Global Collection Group. A global approach for natural history museum collections. Science, 2023, 379(6638), 1192-1194. DOI: 10.1126/science.adf6434
[20] Reusens M, Adams A, Baesens B. Large language models to make museum archive collections more accessible. AI & Society, 2025, 40(6), 4485-4497. DOI: 10.1007/s00146-025-02227-8
[21] Su C, Zheng H, Peng D, Wang X. DICA: Disambiguated contrastive alignment for cross-modal retrieval with partial labels. Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(19), 20610-20618. DOI: 10.1609/aaai.v39i19.34271
[22] Zhu L, Wu X, Li J, Zhang Z, Guan W, Shen H. Work together: Correlation-identity reconstruction hashing for unsupervised cross-modal retrieval. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(9), 8838-8851. DOI: 10.1109/TKDE.2022.3218656
[23] Wang M K, Meng M, Liu J, Wu J. Adequate alignment and interaction for cross-modal retrieval. Virtual Reality & Intelligent Hardware, 2023, 5(6), 509-522. DOI: 10.1016/j.vrih.2023.06.003
[24] Huo Y, Qin Q, Dai J, Wang L, Zhang W, Huang L. Deep semantic-aware proxy hashing for multi-label cross-modal retrieval. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 34(1), 576-589. DOI: 10.1109/TCSVT.2023.3285266
[25] Zhang H, Yang Y, Qi F, Qian S, Xu C. Active supervised cross-modal retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(6), 5112-5126. DOI: 10.1109/TPAMI.2025.3550526
[26] Wang Y, Wang F, Dong J, Luo H. Cl2CM: Improving cross-lingual cross-modal retrieval via cross-lingual knowledge transfer. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(6), 5651-5659. DOI: 10.1609/aaai.v38i6.28376
[27] Li F, Wang B, Zhu L, Li J, Zhang Z, Chang X. Cross-domain transfer hashing for efficient cross-modal retrieval. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(10), 9664-9677. DOI: 10.1109/TCSVT.2024.3374791
[28] Pu R, Sun Y, Qin Y, Ren Z, Song X, Zheng H, et al. Robust self-paced hashing for cross-modal retrieval with noisy labels. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(19), 19969-19977. DOI: 10.1609/aaai.v39i19.34199
[29] Li J, Wong WK, Jiang L, Fang X, Xie S, Xu Y. CKDH: CLIP-based knowledge distillation hashing for cross-modal retrieval. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(7), 6530-6541. DOI: 10.1109/TCSVT.2024.3350695
[30] Li Y, Zhen L, Sun Y, Peng D, Peng X, Hu P. Deep evidential hashing for trustworthy cross-modal retrieva. Proceedings of the AAAI conference on artificial intelligence. 2025, 39(17), 18566-18574. DOI: 10.1609/aaai.v39i17.34043