

Cross-Modal Retrieval in Art Design via Attribute Fusion and Multi-Task Collaborative Learning

Ya'nan Guo

Lecturer, Arts Department, Zhengzhou Railway Vocational and Technical University, Zhengzhou, 450018, China, E-mail: guoyanan10@outlook.com

Project Management

Received January 7, 2026; revised March 9, 2026; September 14, 2026; accepted September 24, 2026
Available online October 26, 2026

Abstract: Existing cross-modal retrieval ignores key attribute features such as genre and author in art design, and independent encoding of images and texts leads to poor semantic alignment. Therefore, this study proposes an art design cross-modal retrieval model that combines attribute deep fusion and multi-task collaborative learning. This method first employs a prompt-template-based strategy to convert structured attribute information into natural-language sentences. From a causal inference perspective, separate encoding paths are designed to integrate attributes, text, and visual features in depth. In addition, the model introduces a multi-task collaborative learning framework to learn more robust shared representations by jointly optimizing cross-modal retrieval and artistic attribute classification. Experimental results on the Rijksmuseum data set showed that, in the text-retrieval image task, the average precision of the cross-modal retrieval model reached 62.7%, and the total recall rate reached 397.8%. In the image-retrieval text task, the average accuracy was 59.4%. The recall of the cross-modal retrieval model reached a maximum of 95.63%. Multi-task analysis showed that when using Genre as an auxiliary task, the Recall@50 (Number of relevant items in the top 50 results) of the image retrieval text task was 95.06%. The attribute deep fusion strategy based on prompt templates and the multi-task collaborative learning framework are the key to improving model performance. The combination of the two can effectively capture the specific semantics in the art field and achieve more accurate cross-modal alignment. The research results provide efficient and innovative solutions for solving cross-modal retrieval problems in art and design.

Keywords: Cross-modal retrieval; attribute features; collaborative learning; art design; prompt template.

Copyright © Journal of Engineering, Project, and Production Management (EPPM-Journal).
DOI 10.32738/JEPPM-2026-19

1. Introduction

Driven by global informatization, the digitization of art has become an inevitable trend in cultural inheritance and dissemination. Massive works of art have been digitized, forming a vast digital art archive. However, efficiently retrieving target works from massive amounts of data remains a major problem (Huo et al., 2024a). In this context, cross-modal retrieval technology has great potential in art and design (Li et al., 2024a; Li et al., 2024b). The main purpose of this technology is to break the semantic gap between the high-dimensional visual features of artworks and unstructured text descriptions, and provide users with a more natural way to explore and interact with art. Traditional retrieval methods primarily rely on statistical approaches to identify linear correlations between modalities but struggle to capture complex nonlinear semantic relationships (Huo et al., 2024b). Currently, deep learning-based retrieval methods have become mainstream. This method usually uses a convolutional neural network to extract image features, a recurrent neural network or a Transformer to extract text features, and then maps the two into a shared space to perform retrieval by computing similarity (Zhang et al., 2023; Bai et al., 2024). However, most of these methods encode images and text independently, ignoring the close semantic connection between them, thereby limiting the model's ability to fully leverage cross-modal complementary information during feature learning. Art criticism and description often contain complex semantic structures and logical connections. Existing methods fail to explore these high-level semantic expressions in depth when processing text, resulting in a one-sided understanding. In addition, attribute characteristics such as genre, author, technique, and era of artworks contain rich prior knowledge, but most existing research focuses on binary matching of images and texts, ignoring structured attribute information (Han et al., 2024; Meng et al., 2024). Therefore, more advanced cross-modal retrieval needs to be developed.

Existing research has proposed a series of innovative solutions to deal with specific challenges, such as noisy data and matching uncertainty in cross-modal retrieval. Aiming to address performance degradation caused by mismatched pairs in

cross-modal retrieval, Hu et al. (2025) developed a robust cross-modal learning framework that used a complementary contrastive learning paradigm to mitigate the effects of noisy data. The method demonstrated superior effectiveness and robustness compared with existing methods across five benchmark datasets. Wang et al. (2025) found that existing cross-modal retrieval methods struggled to handle one-to-many matching problems with deterministic point vectors. Therefore, they proposed point-to-rectangle matching and rectangle-to-rectangle matching methods based on geometric representation learning. The results showed that this strategy achieved significant retrieval performance across multiple image and video datasets. Wang et al. (2024a) proposed a cross-language-to-cross-modality general framework in light of existing cross-language and cross-modal retrieval methods that are trapped in the modal heterogeneity gap and translation noise. This method employed cross-language knowledge transfer and leveraged multilingual pre-training models to ensure reliable semantic correspondence for cross-modal networks. Experimental results on benchmark datasets showed that the model achieved excellent performance on large-scale retrieval tasks.

The development of the prompt learning paradigm has opened up new ways to solve complex retrieval and semantic alignment problems. General large-scale language models are difficult to answer highly contextualized course questions on student forums. Mitra et al. (2024) proposed an improved large-scale language model. This method built context by retrieving historical Q&A and course documents and incorporating specific prompt words to guide the model. The quality of the generated answers was significantly better than that of baseline models that did not rely on context or relied only on retrieval context. Zhai et al. (2024) developed a multi-prompt pedestrian re-identification framework to address the limited use of fine-grained attributes in existing pedestrian re-identification algorithms. This method used a generative model to construct explicit attribute hints, which were combined with implicit, learnable hints to assist the model. The results showed that experiments on datasets such as Market 1501 proved the effectiveness of this multi-prompt fusion strategy. Liu et al. (2023) developed a new prompt learning approach to address the problem that traditional supervised learning relies heavily on annotated data for new tasks. This method recasts the prediction task as a template-based text-filling task for language models. This framework achieved few-shot or even zero-shot learning with little or no labeled data.

To sum up, existing cross-modal retrieval research has made significant progress in improving model robustness, handling complex matching relationships, and coping with data incompleteness. However, most of these methods are limited to general image-text binary matching and often ignore structural attribute features that contain rich prior knowledge in art and design, making it hard for the model to capture the deep semantics of artistic works. The prompt learning paradigm offers new approaches to integrating contextual information and processing structured data. However, there is still a lack of a cross-modal retrieval framework specifically designed for the art and design field that can deeply integrate high-dimensional visual features, unstructured text descriptions, and structured attribute features. Therefore, it is necessary to develop more advanced cross-modal retrieval methods for art design. Therefore, a cross-modal retrieval model combining attribute features and multi-task collaborative learning (ADF-MTL) is proposed to effectively capture complex artistic semantics and improve retrieval performance through causal interventions via attribute separation encoding and multi-task attribute classification. The innovation of the research lies in proposing an attribute deep fusion using a prompt template strategy. This method converts discrete structured attribute features into natural-language sentences and designs separate encoding paths for causal intervention, thereby deeply fusing the three modalities of visual, textual description, and attributes. Meanwhile, a multi-task collaborative learning framework is constructed, taking cross-modal retrieval as the main task and artistic attribute classification as the auxiliary task. Through collaborative optimization, the model is forced to learn shared representations that are more discriminative in the semantics of the art field, which in turn enhances the performance of the retrieval task.

2. Methodology

2.1. Attribute Deep Fusion Based on Prompt Template Strategy

Due to the significant spatiotemporal differences and semantic diversity of data in the field of art and design, traditional retrieval models that rely solely on text and image pairs often ignore key attribute information and are difficult to capture complex nonlinear relationships between multiple modalities (Wang et al., 2024b; Pu et al., 2025; Zhang et al., 2024). To address this problem, the study proposes an improved attribute feature deep fusion method that incorporates a hint-template strategy. In the causal diagram, the attribute feature A is the common cause of text description T and image V . For example, an artwork belonging to Impressionism results in its image featuring specific brushstrokes and light and shadow, and in its text description appearing words such as light and shadow changes and outdoor sketching (Jin et al., 2024; Shu et al., 2023). To learn the true mapping relationship $P_r(V|T)$ from text to image, this biased path $P_r(A|T)$ requires caution. When the model is trained, if T and A are simply spliced together, it will lead to retrieval bias due to over-reliance on $P_r(A|T)$ learned in general domain knowledge instead of $P_r(A|T)$ specific to the art field. Therefore, the main goal is to cut off the dependence between T and A , implement intervention $do(T)$, and make the model learn $P_r(V|do(T))$. According to the backdoor adjustment theory of causal inference, it is mainly realized by layering and summing A . The specific expression is as shown in Eq. (1).

$$P_r(V|do(T)) = \sum_{a \in A} Pr(V|T, a)P_r(a) \quad (1)$$

In contrast, the expression of non-intervention traditional model learning is shown in Eq. (2).

$$P_r(V|T) = \sum_{a \in A} P_r(V|T, a)P_r(a|T) \quad (2)$$

The main difference between Eq. (1) and Eq. (2) is $P_r(a)$ and $P_r(a|T)$. $P_r(a)$ represents the prior probability of the attribute, and $P_r(a|T)$ introduces confounding bias. The designed model framework approximates Eq. (1) through structural separation and fusion. The research transforms the traditional image-and-text tuples into triples comprising an image, a text description and attribute features. By treating attribute features as an independent input modality alongside text descriptions, the model can effectively fuse ternary information in subsequent stages. To effectively process the three modalities of image, text description and attribute features and convert discrete, word-form attribute features into understandable input for the model, the study introduces a prompt template strategy, which mainly converts structured attribute features into natural language sentences so that they are consistent with the text description in form to activate relevant knowledge in the pre-trained language model. The process consists of three steps, starting with the design of specific prompt templates for each attribute in the SemArt dataset. In the data preprocessing stage, the specific attribute values of each sample are used to fill the slots in the template. Finally, after filling in a complete sentence t_{attr} is generated, which is the attribute text of the attribute, as shown in Eq. (3).

$$t_{attr} = \Phi_{prompt}(a_{raw}) \quad (3)$$

In Eq. (3), Φ_{prompt} represents the prompt template function. a_{raw} represents the original attribute value. In the semantic fusion stage, the research does not simply splice the generated attribute text t_{attr} and the original text description D , but designs separate encoding paths. In the attribute feature encoding process, the generated attribute text is sent to an independent attribute feature encoder, thereby implementing a separate encoding path. This encoder learns the semantics of specific attributes in the art domain, and its parameters are not shared with the encoder that processes the original text description. The encoder finally outputs the attribute embedding e_a , which is a dense vector representation that captures all attribute information. The attribute feature encoder employs an improved Transformer architecture, as shown in Fig. 1.

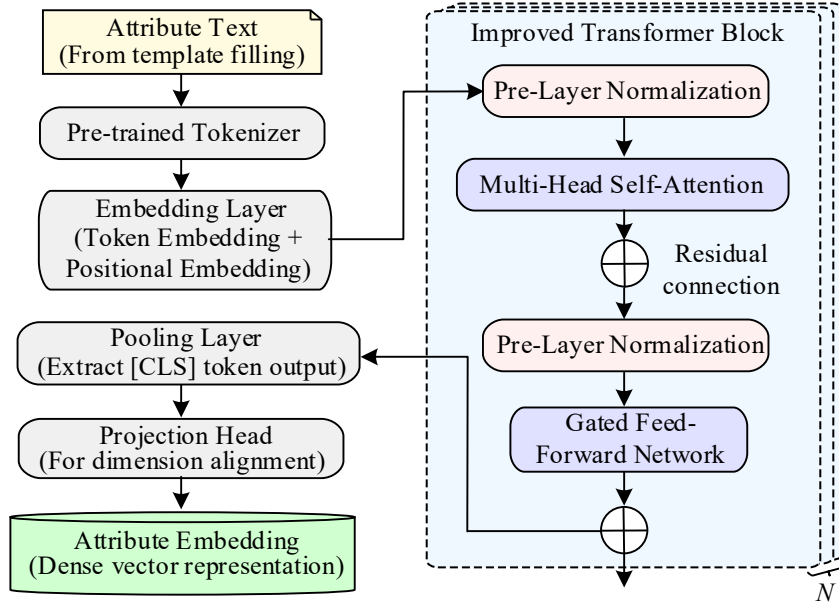


Fig. 1. Attribute feature encoder architecture

From Fig. 1, in the attribute feature encoder, the input attribute text is first converted into token identifiers using a pre-trained tokenizer, then passed to the embedding layer, which combines the token identifiers with their positional information to generate an initial vector representation. This is then fed into the encoder's improved Transformer block stack. Compared with the standard Transformer, the improved Transformer performs layer normalization before self-attention and feed-forward networks, which can make the training process more stable (Weng et al., 2025; Fan et al., 2024). The study uses a gated feedforward network to replace the standard ReLU activation. This gating structure introduces a multiplicative gating mechanism, allowing the network to more dynamically control the flow of information through the feedforward layer, thereby achieving stronger model expressiveness. Inside the improved Transformer block, the data are first normalized through the pre-layer, and then enter the multi-head self-attention layer. The output and the input of this sub-layer are added through a residual connection. The added result is passed through a pre-layer normalization layer again, and then into the gated feedforward network, and is residually connected again. After the data passes through N blocks, the final output sequence is fed into the pooling layer, which extracts the [CLS] token vector as an aggregated semantic representation of the entire attribute text. Finally, the vector is dimensionally transformed by a projection head to generate a dense vector of attribute embeddings. To achieve causal inference, the study designs separate coding paths for attribute text and text description and merges them in the final stage. The deep fusion process of attribute features is shown in Fig. 2.

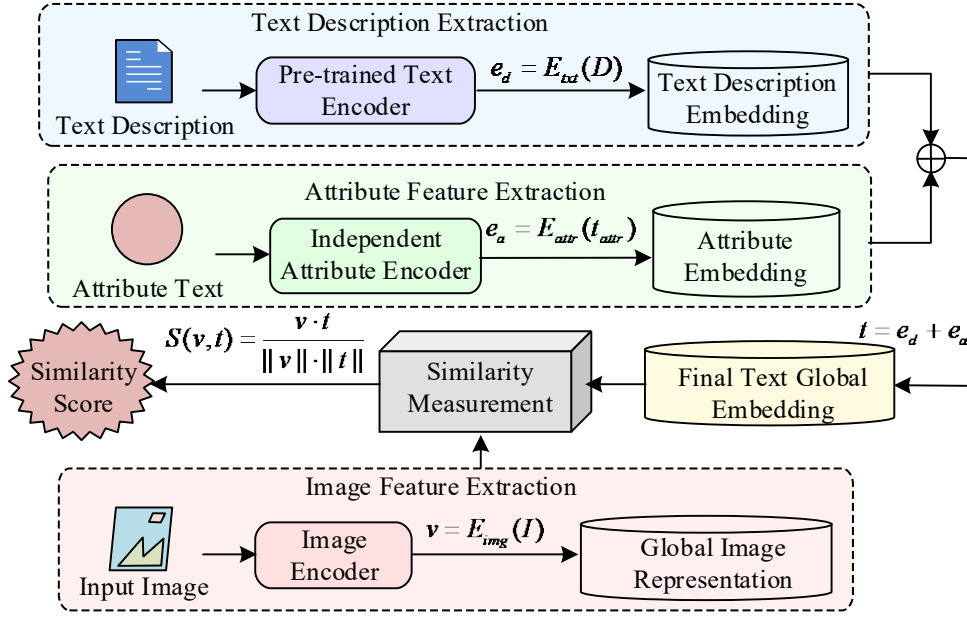


Fig. 2. Flow chart of deep fusion of attribute features (image source: original)

In Fig. 2, the fusion process consists of four steps. First, image features are extracted. The input image I passes through the image encoder E_{img} to obtain the global image representation v , as presented in Eq. (4).

$$v = E_{img}(I) \quad (4)$$

Next, text description feature extraction is performed. The text description D is input and the pre-trained text encoder E_{txt} obtains the text description embedding e_d . The expression is as shown in Eq. (5).

$$e_d = E_{txt}(D) \quad (5)$$

Then, the attribute text t_{attr} generated by the prompt template strategy is input into an independent attribute feature encoder E_{attr} to obtain the attribute embedding e_a , and the expression is shown in Eq. (6).

$$e_a = E_{attr}(t_{attr}) \quad (6)$$

E_{attr} does not share parameters with E_{txt} . The encoder can be a combination of an embedding layer trained from scratch and a multi-layer perceptron. This design method structurally separates the D and A , forcing the model to learn their representations separately, thereby avoiding the bias caused by $P_r(a|T)$ and instead learning the prior distribution of $P_r(a)$. Finally, the text description embedding e_d and the attribute embedding e_a are fused to obtain the final text global embedding representation t , whose expression is shown in Eq. (7).

$$t = e_d + e_a \quad (7)$$

This fusion strategy not only utilizes the rich semantic information provided by attribute features e_a to enhance text representation, but also achieves causal intervention in confounding factors through a separate encoder design, reducing the model's overreliance on common domain knowledge. The fused features v and t are used in subsequent retrieval tasks. The similarity measure $S(v, t)$ is shown in Eq. (8).

$$S(v, t) = \frac{v \cdot t}{\|v\| \cdot \|t\|} \quad (8)$$

By maximizing this similarity function, the model can project the heterogeneous visual space and the attribute-enhanced text space onto a unified metric manifold.

2.2. Cross-Modal Retrieval Based on Multi-Task Collaborative Learning

The research constructs a three-element input model that can deeply integrate image, text description and attribute features and obtains image features v and text features t that are richer in information and more semantically aligned. However, this model is currently limited to a single cross-modal retrieval task (Wang et al., 2023). In the field of art, attributing characteristics are not only inputs that assist understanding but also labels in their own right. The genre, type, and other attributes of artistic works contain highly condensed semantic information and serve as a bridge between visual images and language descriptions (Xu et al., 2023). Therefore, the study introduces a multi-task collaborative learning framework that relies on a deep fusion model. The schematic diagram of multi-task collaborative learning is shown in Fig. 3.

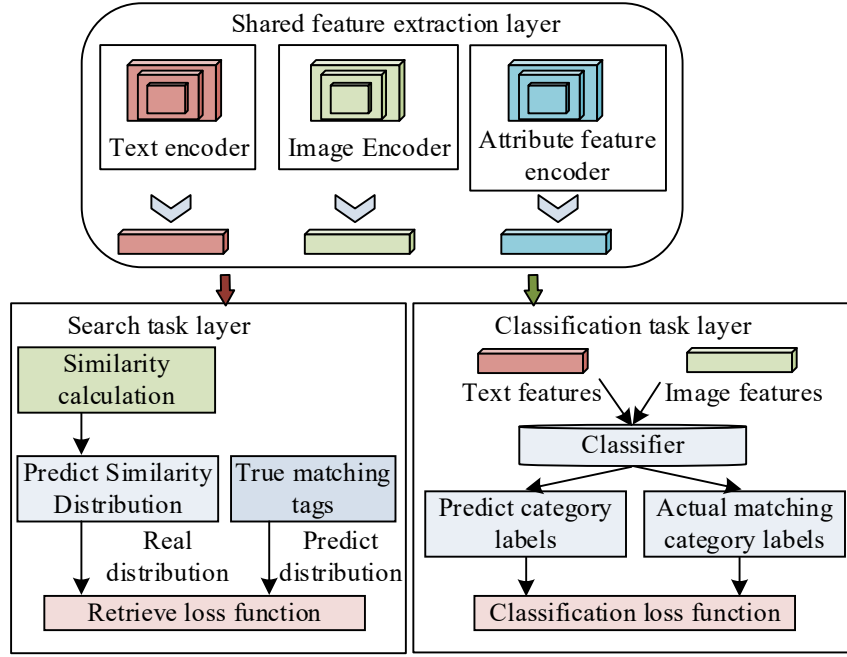


Fig. 3. Multi-task collaborative learning framework diagram

From Fig. 3, the multi-task collaborative learning framework includes a shared feature extraction layer and multiple task-specific layers. When a model is asked to predict the type of a painting simultaneously, it must find explicit evidence from both the image and the text. This will, in turn, enhance attention to these key semantics, thereby improving retrieval performance. The multi-task collaborative learning framework combines retrieval and classification tasks. The cross-modal retrieval task is the main task, and the goal is to learn a shared embedding space so that the matching image-text pair (v, t) has a high similarity score in this space, and the mismatched one has a low score. The artistic attribute classification task is an auxiliary task that mainly uses the fused image feature v or text feature t to accurately predict the attribute category corresponding to the sample. The research combines attribute features with multi-task collaborative learning to construct an ADF-MTL cross-modal retrieval model. The overall architecture diagram is shown in Fig. 4.

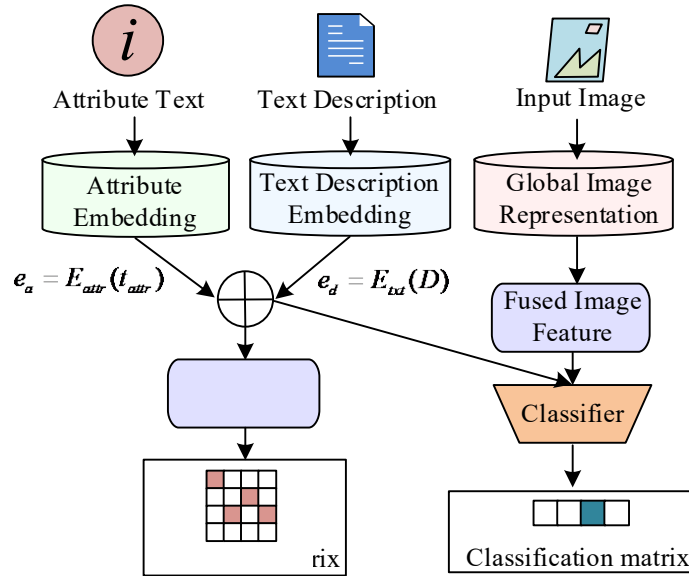


Fig. 4. ADF-MTL cross-modal retrieval model architecture (image source: original)

In Fig. 4 of the ADF-MTL, the input images, text and attributes pass through their respective encoders, and then feature fusion is performed. The fused features are used not only for similarity calculation (i.e., the retrieval task) but also for classification by a classification head. The detection head is a parameter-free computational layer that computes the similarity matrix between v and t and applies retrieval loss. The classification head is one or more fully connected layers that project v , t , or a fusion of the two into a K -dimensional logit space, and then calculate the probability distribution through Softmax. To collaboratively optimize these two tasks, a comprehensive multi-task loss function L_{MTL} is designed, whose expression is shown in Eq. (9).

$$L_{MTL} = \lambda_1 \cdot L_{Ret} + \lambda_2 \cdot L_{Cls} \quad (9)$$

In Eq. (9), L_{Ret} represents the retrieval loss. L_{Cls} represents classification loss. λ_1 and λ_2 respectively represent the hyperparameters of the two task learning scales, and are initially set to 1:1. In the retrieval task, the research mainly draws on the similarity distribution matching to construct the loss function. The specific calculation process of this loss function is shown in Fig. 5.

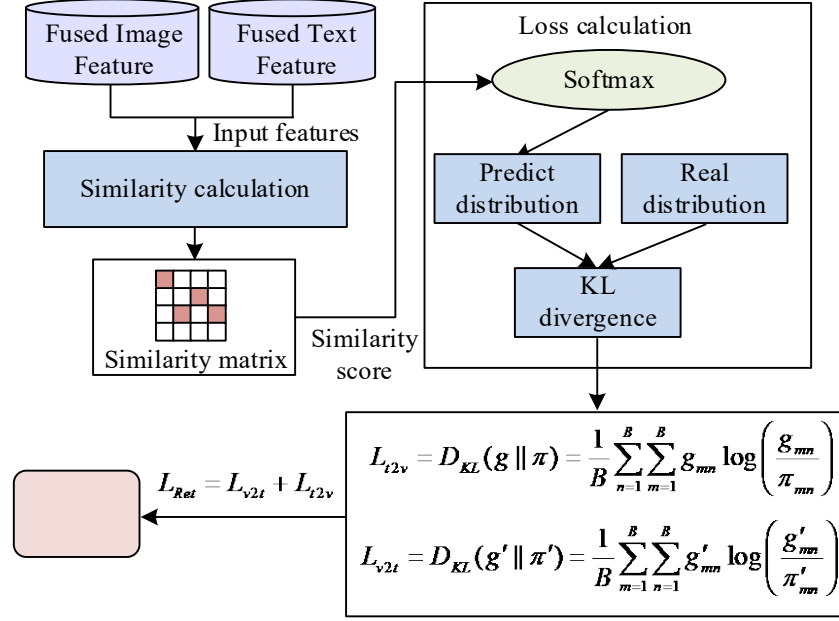


Fig. 5. Retrieval task loss calculation flowchart

As shown in Fig. 5, for a batch of image-text pairs, the cross-modal similarity matrix is first computed. In text-retrieval image tasks, the probability of retrieving images is computed using temperature hyperparameters, and the Kullback-Leibler (KL) divergence between this probability distribution and the true matching distribution is used as the loss function. The image retrieval text task is similar, and the final retrieval loss is composed of the sum of the losses in these two directions. For classification tasks, the uneven distribution of art attributes can cause standard cross-entropy loss to focus too heavily on numerous simple samples while overlooking complex ones. This study adopts the Focal Loss function. This method inputs the fused features into a classifier to obtain category prediction probabilities, and it significantly reduces weight loss for high-confidence, simple samples by introducing a modulation mechanism with focus parameters and balance factors, ensuring that model training can focus more on low-confidence, difficult samples and balance the importance of different categories.

3. Results and Analysis

3.1. Experimental Environment and Ablation Experiment Analysis

To verify the effectiveness and generalization ability of the ADF-MTL for intelligent art design and material mining tasks, experiments were conducted on the publicly available Rijksmuseum dataset. This dataset provides high-resolution images and rich metadata for more than 700,000 art collections. The study constructed a subset of 250,000 samples, each containing high-resolution images, corresponding English descriptions, and structured artistic attribute features. Five key attributes were selected for analysis, including Artist, Genre, Style, Century and Medium. The data set was divided into 80% for training, 10% for validation, and 10% for testing. The experiment is built on the PyTorch 2.0 deep learning framework, and the hardware environment uses an NVIDIA A100 GPU with 40GB of HBM memory for computational acceleration. During model training, the batch size is set to 64, and the AdamW optimizer is used for parameter updates. The initial learning rate is set to 1×10^{-5} . The National Museum Subset Dataset consists of 250,000 items created by 8500 different artists. The data is organized into a training set containing 200,000 objects, while the validation and testing sets each contain 25,000 objects. On average, each object in the dataset is associated with five attributes. To better understand the distribution of attributes, the study conducted statistics, as presented in Fig. 6. In Fig. 6(a), in the feature distribution of Medium, Oil Painting is the category with the highest proportion in the data set, with more than 100,000, followed by Print/Etching and Drawing. The Genre distribution in Fig. 6(b) showed that Portrait and genre paintings accounted for the largest proportions, with each exceeding 50,000 features. The distribution of the Century feature in Fig. 6(c) was extremely uneven, with the vast majority of data concentrated in the 18th century, followed by the 19th. The distribution of Style features in Fig. 6(d) showed that the number of Baroque-style works was 117,564, far exceeding that of Realism (37,698). This illustrates the common differences in historical retention and aesthetic preferences in digital art archives. For intelligent art design systems, models can easily over-fit to high-frequency categories such as the mainstream Baroque style, while ignoring low-frequency design elements.

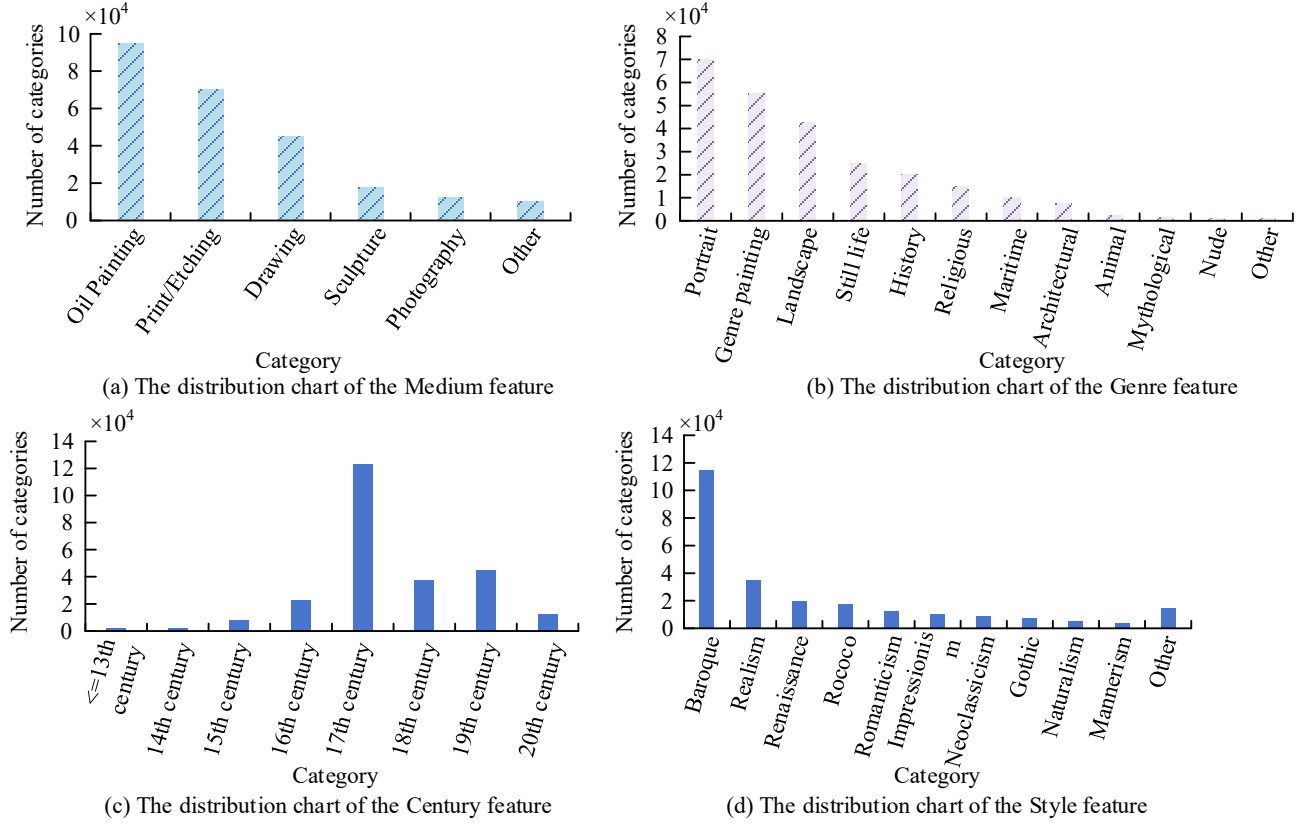


Fig. 6. Distribution of different attributes

To verify the effectiveness of each architecture in the ADF-MTL model, ablation experiments were conducted. The ablation contrast models include the fine-tuned Contrastive Language-Image Pre-training (CLIP) model, the Base-Concat model that simply concatenates the attribute text and the description text generated by the prompt template at the input end, and the ADF model that only deeply fuses the attributes. CLIP does not use any attribute information. Base-Concat simply splices the attribute text and description text generated by the prompt template at the input end. The study first used the common recall indices (Recall@K and R@K) for cross-modal retrieval as evaluation metrics, evaluating them at different K values. The R@K test results of each ablation model are shown in Fig. 7. From Fig. 7(a), ADF-MTL achieved the best performance at all K values, with R@1 reaching 52.17% and R@50 as high as 95.63%. Following ADF, the R@1 was 47.26%, and the R@50 increased to 91.63%. The CLIP model performed the worst, achieving only 78.11% R@50. From Fig. 7(b), the R@K of ADF-MTL remained significantly higher than those of other models, with R@30 reaching 93.84% and R@50 94.76%. The R@30 of the Base-Concat model using simple splicing was only 77.02%, and the R@50 of the CLIP model was only 74.89%, which was much lower than that of the ADF and ADF-MTL models. Compared with CLIP, the simply spliced Base-Concat model, and the ADF model, the ADF-MTL model has higher retrieval performance in the two-way retrieval task and successfully establishes a deep semantic association between visual features and art and design professional terms.

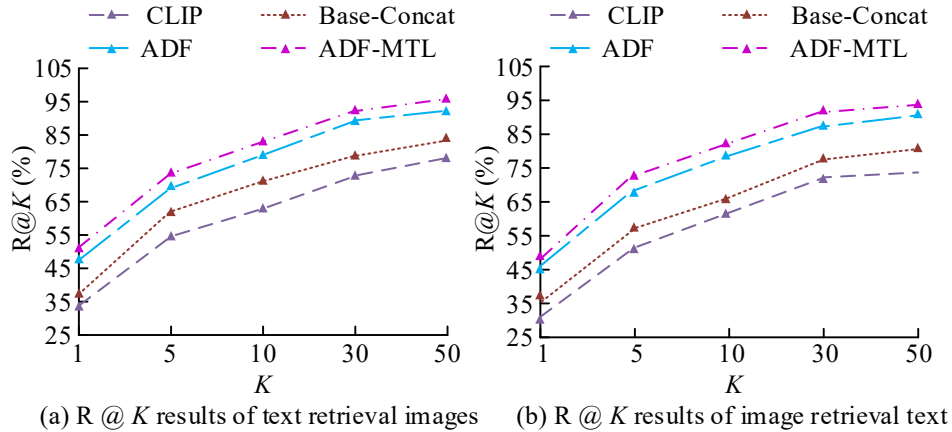


Fig. 7. R@K test results of various ablation models

The study used the mean Average Precision for the first 10 results (mAP@10) and the Sum of Recalls (SumR) for R@1, R@5, R@10, R@30, and R@50 as evaluation metrics. SumR is a comprehensive indicator that includes a precise

summation of K values, defined as $SumR = R@1 + R@5 + R@10 + R@30 + R@50$, reflecting the overall recall capability. $mAP@10$ measures ranking quality in the top 10 search results, giving greater weight to correct results ranked

higher. It is defined as $mAP@10 = \frac{1}{Q} \sum_{q=1}^Q \frac{\sum_{k=1}^{10} P_q(k) \times rel_q(k)}{N_{rel,q}}$, where Q represents the total number of queries, $P_q(k)$

represents the accuracy of the q -th query at position k , $rel_q(k)$ is an indicator function, and $N_{rel,q}$ is the total number of relevant documents in the top 10 results. The $mAP@10$ and SumR test results of different ablation models are shown in Fig. 8. From Fig. 8(a), the $mAP@10$ of the ADF-MTL model was 62.4%, and the SumR was 395.8%, both of which were better than those of the other models. Followed by the ADF, the $mAP@10$ and SumR were 58.9% and 375.9%, respectively. The benchmark model CLIP performed the worst, with $mAP@10$ and SumR only 41.5% and 317.2%, respectively. From Fig. 8(b), the performance ranking of the four models was exactly the same as the text retrieval image task. The ADF-MTL model still performed best, with $mAP@10$ and SumR as high as 59.3% and 392.6%, respectively, though slightly lower than in the text-retrieval image task. The $mAP@10$ and SumR of the Base-Concat model were only 39.2% and 307.5%, which were significantly lower than those of the ADF-MTL model. This once again confirms the superiority of attribute deep fusion and multi-task collaborative learning.

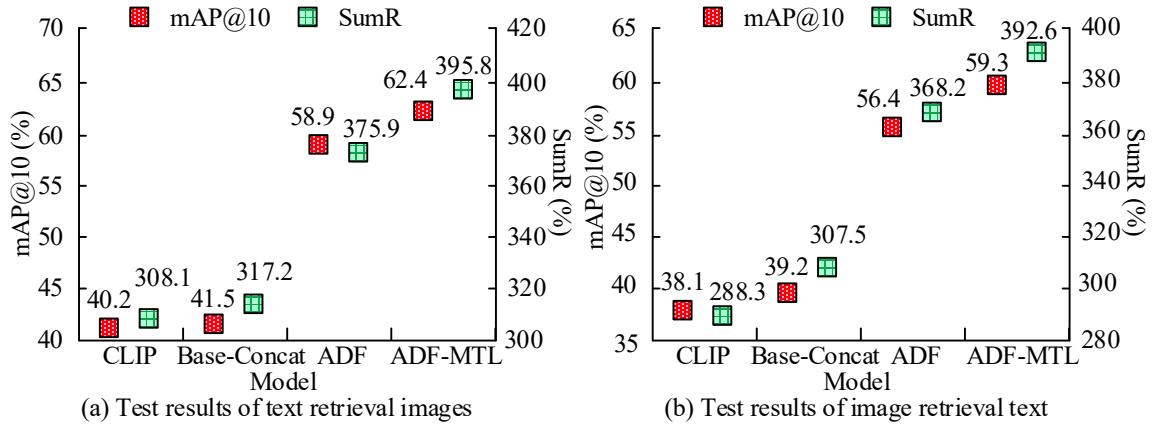


Fig. 8. The $mAP@10$ and SumR of different ablation models

3.2. Auxiliary Task Analysis for Multi-Task Collaborative Learning

To verify the rationality of auxiliary tasks, the study fixed the ADF method and conducted comparative experiments on different attribute classification tasks. The ADF model had no auxiliary tasks, and the auxiliary task for ADF-MTL-Medium was set to Medium and assigned to model 1. The auxiliary task of the ADF-MTL-Century was Century, which was set to model 2. The auxiliary task of the ADF-MTL-Style was Style, which was set to model 3. The auxiliary task of ADF-MTL-Genre was Genre, which was set to model 4. The $R@K$ test results for different models are presented in Fig. 9. In Fig. 9(a), for the text retrieval image task, the performance of all models with auxiliary classification tasks was consistently better than that of the ADF model without auxiliary tasks, demonstrating the effectiveness of the multi-task collaborative learning framework. The $R@1$ of the ADF-MTL-Genre model with Genre as the auxiliary task reached 52.21%, and the $R@50$ was as high as 95.72%. In the image retrieval text task (Fig. 9(b)), the $R@1$ of the ADF-MTL-Genre model was 48.17%, and the $R@50$ was 95.06%, which were significantly better than those of the other models. The $R@50$ of the ADF-MTL-Medium model was only 92.14%, which was poor performance.

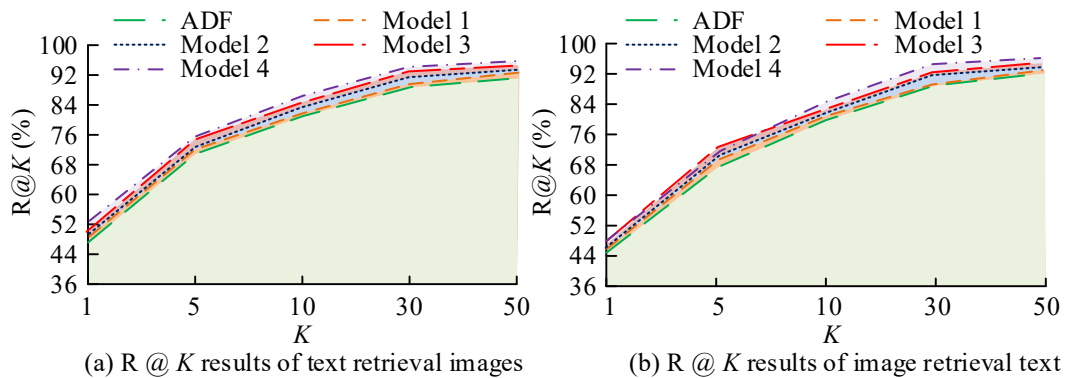


Fig. 9. $R@K$ test results for different auxiliary tasks

The study then compared $mAP@10$ and SumR under different auxiliary tasks. The test results are presented in Fig. 10. In Fig. 10(a), the $mAP@10$ of the ADF-MTL-Genre model was 62.4%, and the SumR reached 397.8%, which were

significantly higher than those of all other models. Following the ADF-MTL-Style model, the $mAP@10$ was 61.5%. From Fig. 10(b), the $mAP@10$ of the ADF-MTL-Genre model was 59.3%, and the SumR was 387.4%. Compared with the ADF model, they were increased by 2.9% and 19.2%, respectively. The $mAP@10$ and SumR of the ADF-MTL-Style model were 58.8% and 381.5%, respectively. Genre and Style are robust auxiliary tasks that can steadily improve the sorting quality and comprehensive recall in bidirectional retrieval tasks. In the field of art and design, accurately using these two high-level semantic features can effectively narrow the semantic gap between the underlying visual pixels and the top-level design image.

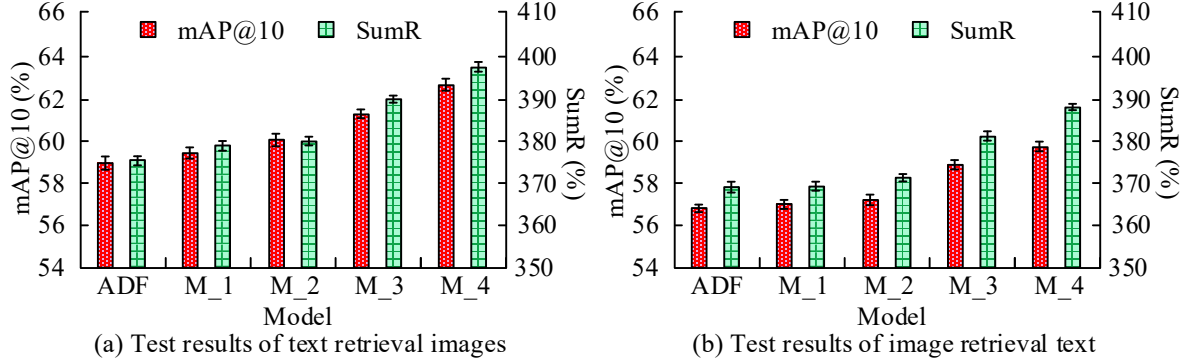


Fig. 10. The $mAP@10$ and SumR under different auxiliary tasks

3.3. Experimental Analysis of Prompt Strategy Engineering

The study further explored the impact of the prompt template's wording on model performance. Since the ADF-MTL-Genre model was the best, all model algorithms were fixedly integrated with auxiliary classification tasks based on Genre attributes. The study developed prompt templates for the two attributes: genre and style. To analyze the impact of prompt wording on model performance, this study evaluated language templates across several genres and style attributes. For the genre attribute, four templates were tested: G-1, "genre is [X]"; G-2, This is an [X] painting; G-3, Artwork depicting the scene of [X]; G-4, Theme: [X]. Similarly, the "Style" attribute of four templates was evaluated: S-1, "Style [X]"; S-2, "A work of art from [X] school"; S-3, "A [X] style artwork"; S-4, Art Movement: [X]. The study applied the template in Table 2 to the ADF-MTL-Genre model. The retrieval test results are presented in Fig. 11. In Fig. 11(a), for the Genre attribute, the $R@K$ value of the G-2 template was higher than that of the other templates, with $R@1$ at 52.5% and $R@50$ at 95.9%. The G-4 template had the lowest $R@K$, with an $R@50$ of only 92.3%. From Fig. 11(b), the Style attribute shows that the G-2 template's $R@K$ performance remained the best, with $R@50$ reaching 95.2%. The $R@50$ of the G-4 template was only 93.9%. It shows that the wording of the prompt template has a reproducible impact on the model performance. The G-2 template is closer to the natural language description and performs best in both retrieval directions. The G-4 template is too simple, and its performance is relatively weak. In the field of art and design, natural language descriptions rich in contextual information can better activate the model's in-depth understanding of complex aesthetic features than individual tags, thereby achieving precise alignment between visual imagery and design semantics.

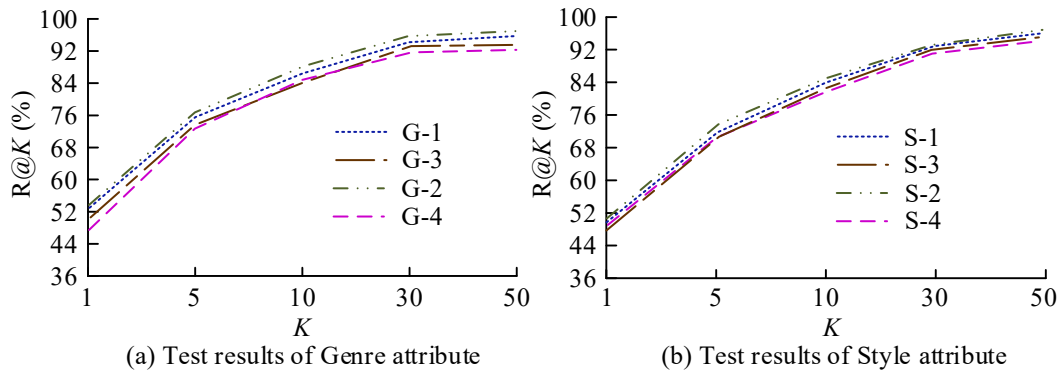


Fig. 11. The $R@K$ value under different templates

3.4. Comparative Experimental Analysis of Cross-Modal Retrieval

The study finally used several advanced retrieval models to compare with the ADF-MTL model, including Flexible Language and Vision Alignment (FLAVA), Knowledge Graph Model (KGM), and Bootstrapping Language-Image Pre-training (BLIP). To ensure the fairness of the input text in the comparative experiment, all baseline models only receive the original English text description as the text modality input during the training and testing stages, without including the attribute text generated by the prompt template. The experiment evaluated the $R@50$ and SumR of each model, as shown in Fig. 12. In the text retrieval image task (Fig. 12(a)), the ADF-MTL model achieved the best performance in both $mAP@10$ and SumR, reaching 62.7% and 397.8%, respectively. Followed by the KGM model, the $mAP@10$ and SumR were 58.0% and 379.5%, respectively. The various indicators of the FLAVA model were only 52.5% and 362.3%, which

were significantly lower than those of the ADF-MTL model. From Fig. 12(b), the $mAP@10$ and SumR of the ADF-MTL model in the image retrieval text task were as high as 59.4% and 387.4%, which were 7.5% and 27.8% higher than those of the BLIP. The $mAP@10$ and SumR of the FLAVA model were only 50.7% and 354.1%, respectively. It shows that the ADF-MTL model clearly leverages structured attribute information and achieves superior overall retrieval performance to existing models across both bidirectional tasks.

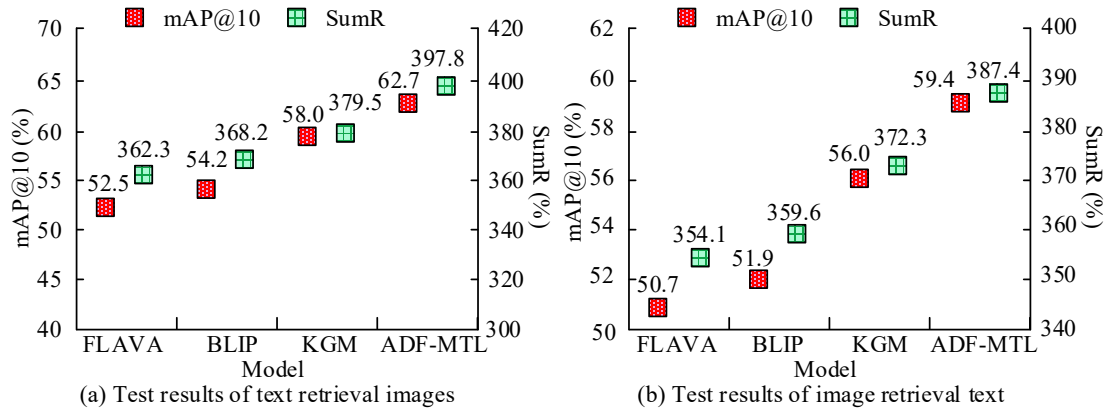


Fig. 12. The $R@50$ and SumR of different comparison models

To verify the practical application effect of the proposed cross-modal retrieval model for art and design, a text description was selected to generate a corresponding image. The selected text for the experiment is described as an opulent Baroque still life of precious metals, Venetian glass, and exotic fruits, showcasing rich tactile textures and themes of earthly wealth under dramatic chiaroscuro lighting. Structured attribute features: [Genre: Portrait] [Style: Baroque]. The generated result is shown in Fig. 13. The ADF-MTL model accurately retrieves artistic images with strong contrast effects by integrating Baroque still-life text and attribute features. The model successfully captured the rich textures of metal, glass, and fruit, achieving similarity scores up to 0.95.

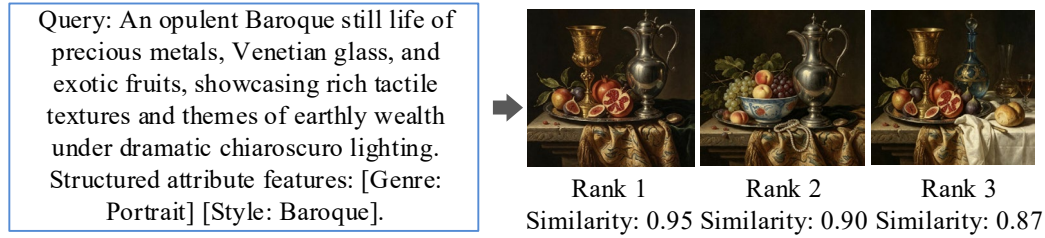


Fig. 13. Generate image effects

4. Discussion and Interpretation

To efficiently complete the art design cross-modal retrieval task, the ADF-MTL model is proposed. The ablation experimental results verify the effectiveness of deep fusion of theoretical attributes and multi-task collaborative learning. The $mAP@10$ and SumR of the CLIP in the text retrieval image task were only 41.5% and 317.2%, respectively. The $mAP@10$ of the ADF-MTL model was 62.4%, and the SumR was 397.8%. Followed by the ADF model, the $mAP@10$ and SumR were as high as 58.9% and 375.9%, respectively. The experiment confirmed that the separation encoding path based on causal reasoning is the key to effective fusion. Multi-task collaborative learning can enhance explicit feature differentiation by introducing high-semantic, universal auxiliary tasks, thereby improving the ability to align semantically. The experiments evaluated the ADF method and compared it across different classification tasks. The study found that when using Genre as an auxiliary task, the ADF-MTL-Genre performed best, with a SumR of 397.8% for text retrieval and an $R@50$ of 95.06% for image retrieval. When Style was used as an auxiliary task, the effect was second, with $R@50$ reaching 95.05%. The reason is that Genre can provide the most effective cross-modal semantic bridge, allowing the model to understand visual patterns and text concepts simultaneously. However, Medium or Century are highly unevenly distributed in the dataset, resulting in a weak regularization effect when used as an auxiliary task (Cai et al., 2024).

The experimental results of the prompt strategy project show that the G-2 template, that is, "This is a [X] painting", is better than the G-4, that is, "Subject: [X]", because the sentence structure of G-2 is closer to the natural language description that the large language model is exposed to in the pre-training stage. Cao et al. (2025) pointed out that reconstructing the prediction task as a language-model text-filling task using templates can enable efficient learning with little or no labeled data. The G-2 template used in the study uses this idea to convert structured attribute inputs into natural-language sentences that are familiar during model pre-training. This alignment activates prior knowledge within the model, allowing it to better understand the semantic connotations of the attributes. This is consistent with the findings drawn by Zhu et al. (2024) on the importance of optimizing prompt templates. In comparative experiments with advanced models, ADF-MTL outperforms general models such as FLAVA, KGM, and BLIP on bidirectional retrieval tasks. For example, the $mAP@10$

of the ADF-MTL model on the text-retrieval image task reached 62.7%, higher than that of the KGM (58.0%). Although KGM and BLIP models are powerful, they primarily rely on binary image and text information. The advantage of the research lies in the deep integration of attribute features in the art field through the ADF-MTL architecture, thereby achieving more accurate semantic understanding for art retrieval tasks.

In summary, in professional fields such as art and design, the ADF-MTL model can effectively improve cross-modal retrieval performance. However, this model still has certain limitations. First of all, the research is evaluated only on a single Rijksmuseum dataset, and its generalizability needs to be tested on additional art datasets from different styles. Secondly, the prompt templates currently rely on manual experience. Although the G-2 template performs best, automated prompt workers are a better solution. Follow-up research will consider extending the ADF-MTL framework to other domains rich in structured attributes. Meanwhile, research will explore learnable soft prompts to replace current manual templates and investigate more complex causal intervention models to account for possible interrelationships among multiple attributes.

5. Conclusion

An ADF-MTL model that combined attribute deep fusion and multi-task collaborative learning was designed, and the effectiveness of this method on the Rijksmuseum data set was verified. The ablation experiment results showed that in the text retrieval image task, the R@1 of the ADF-MTL model reached 52.17%, the R@50 was as high as 95.63%, the mAP@10 was 62.4%, and the SumR reached 395.8%, which were significantly better than those of all baseline models. The text retrieval image mAP@10 of the CLIP baseline model, which did not use any attribute information, was only 41.5%, and the SumR was only 317.2%. In the art field, missing attribute information can lead to a serious semantic gap. In the image retrieval text task, the Base-Concat model with simple splicing attributes achieved only 39.2% mAP@10. In contrast, the ADF model that uses only separate encoding paths has its text-retrieval image mAP@10 increase to 58.9%. When using Genre as an auxiliary task, the ADF-MTL-Genre performed best, achieving a SumR of 397.8% for text retrieval and an R@50 of 95.06% for text-image retrieval. When using the G-2 template, which was closer to a natural-language description, the R@50 reached 95.9%, which was higher than that of the other templates. The ADF-MTL successfully integrates structured attribute knowledge from the art field into image and text representations, achieving accurate and efficient cross-modal retrieval.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

Declaration of Artificial Intelligence (AI) Tools

The author only uses ChatGPT for language editing and readability improvement. The author has reviewed and verified all content and takes full responsibility for the accuracy and completeness of the manuscript.

References

- Bai, Y., Shu, Z., Yu, J., Yu, Z., and Wu, X. J. (2024). Proxy-based graph convolutional hashing for cross-modal retrieval. *Journal of Big Data*, 10(4), 371-385. doi: 10.1109/TBDATA.2023.3338951
- Cai, R., Dong, J., Liang, T., Liang, Y., Wang, Y., Yang, X., Wang, X., and Wang, M. (2024). Cross-lingual cross-modal retrieval with noise-robust fine-tuning. *IEEE Transactions on Knowledge and Data Engineering*, 36(11), 5860-5873. doi: 10.1109/TKDE.2024.3400060
- Cao, C., Sang, J., Arora, R., Chen, D., Kloosterman, R., Cecere, M., Gorla, J., Saleh, R., Drennan, I., Teja, B., Fehlings, M., Ronksley, P., Leung, A. A., Weisz, D. E., Ware, H., Whelan, M., Emerson, D. B., Arora, R. K., and Bobrovitz, N. (2025). Development of prompt templates for large language model-driven screening in systematic reviews. *Annals of Internal Medicine*, 178(3), 389-401. doi: 10.7326/ANNALS-24-02189
- Fan, X., Zhang, Y., Lu, Y., and Wang, H. (2024). PARFormer: Transformer-based multi-task network for pedestrian attribute recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(1), 411-423. doi: 10.1109/TCSVT.2023.3285411
- Han, K., Liu, Y., Wei, R., Zhou, K., Xu, J., and Long, K. (2024). Supervised hierarchical online hashing for cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 20(4), 1-23. doi: 10.1145/3632527
- Hu, P., Huang, Z., Peng, D., Wang, X., and Peng, X. (2023). Cross-modal retrieval with partially mismatched pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8), 9595-9610. doi: 10.1109/TPAMI.2023.3247939
- Huo, Y., Qin, Q., Dai, J., Wang, L., Zhang, W., Huang, L., and Wang, C. (2024a). Deep semantic-aware proxy hashing for multi-label cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(1), 576-589. doi: 10.1109/TCSVT.2023.3285266
- Huo, Y., Qin, Q., Zhang, W., Huang, L., and Nie, J. (2024b). Deep hierarchy-aware proxy hashing with self-paced learning for cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 36(11), 5926-5939. doi: 10.1109/TKDE.2024.3401050
- Jin, M., Hu, W., Zhu, L., Wang, X., and Hong, R. (2024). Based on spatial and temporal implicit semantic relational inference for cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11), 11286-11298. doi: 10.1109/TCSVT.2024.3411298
- Li, F., Wang, B., Zhu, L., Li, J., Zhang, Z., and Chang, X. (2024a). Cross-domain transfer hashing for efficient cross-modal

- retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10), 9664-9677. doi: 10.1109/TCSVT.2024.3374791
- Li, J., Wong, W. K., Jiang, L., Fang, X., Xie, S., and Xu, Y. (2024b). CKDH: CLIP-based knowledge distillation hashing for cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7), 6530-6541. doi: 10.1109/TCSVT.2024.3350695
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), 1-35. doi: 10.1145/3560815
- Meng, M., Sun, J., Liu, J., Yu, J., and Wu, J. (2024). Semantic disentanglement adversarial hashing for cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(3), 1914-1926. doi: 10.1109/TCSVT.2023.3293104
- Mitra, C., Miroyan, M., Jain, R., Kumud, V., Ranade, G., and Norouzi, N. (2024). RetLLM-E: retrieval-prompt strategy for question-answering on student discussion forums. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21), 23215-23223. doi: 10.1609/aaai.v38i21.30368
- Pu, R., Sun, Y., Qin, Y., Ren, Z., Song, X., Zheng, H., and Peng, D. (2025). Robust self-paced hashing for cross-modal retrieval with noisy labels. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(19), 19969-19977. doi: 10.1609/aaai.v39i19.34199
- Shu, Z., Li, L., Yu, J., Zhang, D., Yu, Z., and Wu, X. J. (2023). Online supervised collective matrix factorization hashing for cross-modal retrieval. *Applied Intelligence*, 53(11), 14201-14218. doi: 10.1007/s10489-022-04189-6
- Wang, J., Yue, K., and Duan, L. (2023). Models and techniques for domain relation extraction: A survey. *Journal of Data Science and Intelligent Systems*, 3(1), 16-25. doi: 10.47852/bonviewJDSIS3202973
- Wang, Y., Wang, F., Dong, J., and Luo, H. (2024a). Cl2cm: Improving cross-lingual cross-modal retrieval via cross-lingual knowledge transfer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6), 5651-5659. doi: 10.1609/aaai.v38i6.28376
- Wang, Y., Zhan, Y. W., Chen, Z. D., Luo, X., and Xu, X. S. (2024b). Multiple information embedded hashing for large-scale cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6), 5118-5131. doi: 10.1109/TCSVT.2023.3340102
- Wang, Z., Gao, Z., Yang, Y., Wang, G., Jiao, C., and Shen, H. T. (2025). Geometric matching for cross-modal retrieval. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3), 5509-5521. doi: 10.1109/TNNLS.2024.3381347
- Weng, W., Hou, F., Gong, S., Chen, F., and Lin, D. (2025). Attribute graph clustering via transformer and graph attention autoencoder. *Intelligent Data Analysis*, 29(2), 306-319. doi: 10.3233/IDA-230647
- Xu, S., Pang, L., Shen, H., and Cheng, X. (2023). Nir-prompt: A multi-task generalized neural information retrieval training framework. *ACM Transactions on Information Systems*, 42(2), 1-32. doi: 10.1145/3626092
- Zhai, Y., Zeng, Z., Huang, Z., Qin, Z., Jin, X., and Cao, D. (2024). Multi-prompts learning with cross-modal alignment for attribute-based person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7), 6979-6987. doi: 10.1609/aaai.v38i7.28524
- Zhang, F., Zhou, H., Hua, X. S., Chen, C., and Luo, X. (2024). HOPE: A hierarchical perspective for semi-supervised 2D-3D cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12), 8976-8993. doi: 10.1109/TPAMI.2024.3412760
- Zhang, P. F., Bai, G., Yin, H., and Huang, Z. (2023). Proactive privacy-preserving learning for cross-modal retrieval. *ACM Transactions on Information Systems*, 41(2), 1-23. doi: 10.1145/3545799
- Zhu, Y., Wang, Y., Qiang, J., and Wu, X. (2024). Prompt-learning for short text classification. *IEEE Transactions on Knowledge and Data Engineering*, 36(10), 5328-5339. doi: 10.1109/TKDE.2023.3332787.



Ya'nan Guo was born in December 1990 in Zhengzhou City, Henan Province and is of Han nationality. She graduated from Luoyang Normal University with a bachelor's degree in fine arts in 2015 and from Henan Normal University with a master's degree in fine arts in 2017. Since 2017, she has served as a design teacher in the Art Department at Zhengzhou Railway Vocational and Technical College. She published six papers, two academic books, twelve scientific research projects, and has filed four patents.