IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 2, April, 2026, pp. 711-725
Journal homepage: http://iieta.org/journals/ts

Multimodal Feature Fusion and Visual Attractiveness Prediction for Advertising Images in Digital Marketing

Lingfei Wang

School of Cultural Creativity and Management, Communication University of Zhejiang, Hangzhou 310018, China

Corresponding Author Email: 20121255@cuz.edu.cn

https://doi.org/10.18280/ts.430213

Received: 18 September 2025
Revised: 30 January 2026
Accepted: 12 March 2026
Available online: 30 April 2026

Keywords:
digital marketing, advertising images, multimodal feature fusion, visual attractiveness prediction, attention mechanism, interpretable image processing

ABSTRACT

In digital marketing scenarios, the visual attractiveness of advertising images directly determines user engagement efficiency and marketing conversion performance. However, existing visual attractiveness prediction methods generally suffer from insufficient multimodal information fusion and limited model interpretability, making them difficult to meet both industrial application and academic research requirements. To address these issues, this paper proposes a dual-branch attention fusion network (DBAFN) for visual attractiveness prediction of advertising images, enabling pixel-level deep interaction between visual features and textual semantic features while providing interpretable predictions. Specifically, a text-to-image spatial affine modulation mechanism is designed to explicitly project textual semantic vectors into the image feature space to form modulation tensors, thereby achieving channel-wise feature recalibration. A global–local DBAFN architecture is further constructed, which integrates an adaptive gating mechanism to enable complementary fusion of global visual styles and locally salient semantic regions. In addition, a multi-task joint loss function is established, incorporating contrastive learning and sparse regularization strategies to balance prediction accuracy and model interpretability. Experimental results demonstrate that the proposed method significantly outperforms existing state-of-the-art approaches in core evaluation metrics such as mean absolute error and root mean square error. Moreover, it can generate precise attractiveness heatmaps, providing reliable technical support for automated optimization of advertising creativity in digital marketing.

1. INTRODUCTION

With the rapid iteration of the digital marketing industry, advertising images have become the core carrier for communication between brands and users [1-3]. Their visual attractiveness directly determines user attention allocation, click-through rate, and final conversion efficiency, and is a key influencing factor of digital marketing performance [4, 5]. In practical industrial scenarios, advertising creative design has long relied on designers' subjective experience, which is not only inefficient and costly, but also difficult to form standardized and quantifiable optimization schemes, and cannot meet the demand for large-scale and personalized advertising delivery [6-8]. Therefore, developing an efficient and accurate automated prediction method for the visual attractiveness of advertising images has become an important bridge connecting image processing technology and digital marketing applications. In recent years, breakthroughs in deep learning technology in image recognition and multimodal fusion have provided new solutions for this problem [9-12]. However, existing methods still have obvious limitations: most methods ignore the deep intrinsic relationship between visual content and textual semantics in advertising images, and only adopt simple feature concatenation or shallow interaction to achieve multimodal fusion, which cannot fully exploit the guiding effect of textual semantics on visual attractiveness; at the same time, most prediction models only output a single attractiveness score and lack interpretability support for prediction results, making it difficult to locate key regions affecting advertising attractiveness, and unable to provide effective guidance for iterative optimization of advertising creativity, thus failing to meet the practical requirements of industrial applications.

From the dual perspectives of academic research and engineering application, this study has important theoretical value and practical value. At the theoretical level, this study focuses on the specific scenario of digital marketing, constructs a semantic-driven multimodal feature fusion and visual attractiveness prediction framework, breaks through the limitations of existing multimodal fusion technologies in cross-modal alignment and interaction, enriches the application scenarios of image processing technology in the marketing field, and provides new research ideas and technical paradigms for multimodal image processing and visual attention analysis. At the application level, the interpretable prediction method proposed in this study can accurately locate key attractiveness regions in advertising images, provide quantitative optimization basis for advertising designers, facilitate automated and standardized optimization of advertising creativity, improve the delivery efficiency and

conversion performance of digital marketing, and provide technical support for subsequent applications such as automated advertisement generation and user attention analysis, with broad industrial application prospects.

Although scholars at home and abroad have conducted a series of studies on visual attractiveness prediction of advertising images, existing methods still have many key problems to be solved, which also constitute the core innovation entry points of this study. In terms of multimodal fusion, most existing methods adopt simple concatenation or shallow interaction between visual features and textual features, and fail to realize explicit projection of textual semantics into the image feature space [13, 14], resulting in difficulty in accurately capturing the correspondence between textual semantics and visual regions, insufficient depth and effectiveness of multimodal information fusion, and thus affecting prediction accuracy. In terms of attention mechanism design, the visual attractiveness of advertising images is jointly determined by global composition aesthetics and local semantic salient regions, whereas existing methods mostly adopt a single attention branch [15], either focusing on extraction of global visual features while ignoring the saliency of local semantic regions [16], or excessively focusing on local details while lacking grasp of global style, and the fusion method lacks adaptive adjustment capability, making it impossible to achieve effective complementarity between global and local features. In terms of interpretability, most prediction models can only output attractiveness scores and cannot generate intuitive visualization results to locate key image regions affecting attractiveness, making the model prediction results difficult to understand and trust, and unable to provide specific guidance for advertising creative optimization, which limits the industrial application value of the model. In terms of loss function design, existing methods mostly adopt a single regression loss function [17], and do not fully utilize the contrastive information and sparsity characteristics of advertising image attractiveness, resulting in limited discrimination ability of the model for advertisements with different attractiveness levels, insufficient generalization performance, and difficulty in adapting to complex and diverse digital marketing advertising scenarios [18-20].

To address the shortcomings of existing research, this study proposes a series of innovative technical schemes to construct an efficient and interpretable visual attractiveness prediction model for advertising images. A text–image spatial affine modulation mechanism is proposed, which predicts scaling factors and shifting factors from textual features through learnable linear layers, and expands them by spatial replication into modulation tensors with the same size as the image feature maps, thereby realizing dynamic modulation of image feature channel responses by textual semantics and effectively solving the problem of insufficient multimodal information alignment. A dual-branch attention fusion network (DBAFN) architecture is designed: the global branch captures global composition, color distribution, and other visual attributes of advertising images through adaptive average pooling and multi-head self-attention, while the local branch locates visual regions highly related to textual semantics through cross-modal dot-product attention; an adaptive gating module is used to achieve pixel-level fusion, fully exploiting the complementary advantages of global and local features. An interpretable heatmap generation mechanism is constructed, in which a lightweight decoder is designed and combined with Kullback–Leibler (KL) divergence loss and L1 sparsity regularization to generate attractiveness heatmaps with the same size as the original images, intuitively locating key attractiveness regions and improving model interpretability. A multi-task joint loss function is proposed, integrating regression loss, KL divergence loss, sparsity regularization loss, and contrastive learning loss. Through contrastive learning, the feature distance between anchor samples and high-attractiveness variants is reduced, while the distance from low-attractiveness variants is increased, significantly improving the discrimination ability and generalization performance of the model.

The remainder of this paper is organized as follows. Section 2 describes in detail the overall architecture and technical details of the proposed DBAFN, including multimodal feature extraction, spatial affine modulation, DBAFN, heatmap generation, and joint loss function design. Section 3 verifies the performance of the proposed method through five groups of core experiments, including benchmark performance comparison, ablation experiments, cross-dataset generalization experiments, interpretability verification experiments, and parameter sensitivity analysis, comprehensively validating the effectiveness and stability of the method. Section 4 further analyzes the experimental results, discusses the technical advantages and limitations of the method, and elaborates the theoretical and application value of the research. Section 5 summarizes the main work and experimental conclusions of this paper and presents future research directions.

2. PROPOSED METHOD

2.1 Overall framework overview

To address the key problems of insufficient fusion between visual features and textual semantic features of advertising images and lack of interpretability in digital marketing scenarios, and to achieve high-precision visual attractiveness prediction and interpretable heatmap generation, this paper proposes a DBAFN. The network adopts a three-stage cascaded architecture. First, multimodal feature extraction is performed on advertising images and their corresponding textual descriptions to obtain image feature maps and textual semantic vectors adapted to subsequent fusion tasks. Then, through a text–image spatial affine modulation mechanism, textual semantics are explicitly projected into the image feature space and modulation tensors are generated, providing precise semantic guidance for DBAFN. Next, a global–local dual-branch attention architecture is used to capture the global visual style and local semantic salient regions of advertising images, respectively, and an adaptive gating module is employed to complete pixel-level fusion of the two, obtaining fused feature maps with both global feature integrity and local semantic saliency. Finally, the regression task of attractiveness score is completed based on the fused feature maps, and interpretable heatmaps are generated through a lightweight decoder. Meanwhile, a multi-task joint loss function is adopted to achieve end-to-end training of the model, ensuring dual optimization of prediction accuracy and interpretability. The innovation core of this framework lies in explicit spatial modulation of textual semantics, adaptive pixel-level fusion of dual-branch attention, and collaborative improvement of interpretability and prediction accuracy. A semantic-driven visual attractiveness prediction paradigm for advertising

images is constructed from the perspective of image processing, effectively breaking through the limitations of existing methods in multimodal fusion depth and interpretability. The overall architecture is shown in Figure 1.

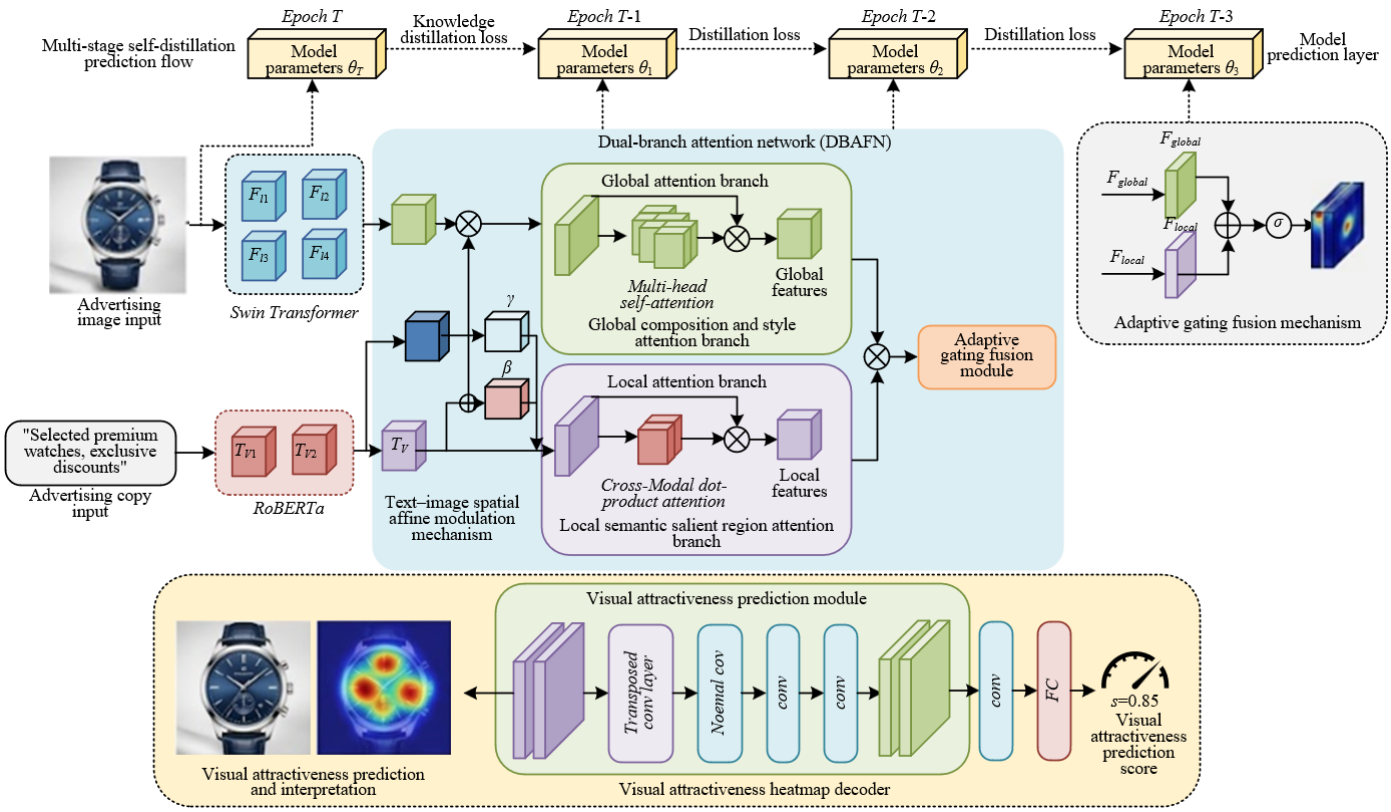


Figure 1. Overall architecture of the dual-branch attention fusion network (DBAFN)

2.2 Multimodal feature extraction

2.2.1 Image feature extraction

The core of image feature extraction is to obtain feature representations with both multi-scale semantic information and dimensional adaptability, providing high-quality inputs for subsequent text–image spatial affine modulation and DBAFN. The innovative design focuses on the effective fusion of multi-scale features and precise dimensionality reduction, in order to solve the problems of insufficient semantic representation of single-scale features and dimensional mismatch with subsequent fusion modules. Swin Transformer is selected as the backbone network. The input RGB advertising image normalized to a resolution of 384×384 is fed into the network, and the feature maps of the third, fourth, and fifth stages are extracted, which are $F_3 \in R^{48\times48\times192}$, $F_4 \in R^{24\times24\times384}$, $F_5 \in R^{12\times12\times768}$, respectively. These correspond to spatial resolutions of 1/8, 1/16, and 1/32 of the input image, covering multi-scale information from local details to global semantics. To fully utilize the complementarity of multi-scale features, bilinear interpolation is applied to F_3 and F_4 for upsampling, with upsampling factors set to 4 and 2, respectively. Their spatial resolutions are uniformly increased to 12×12, consistent with F_5, obtaining the upsampled feature maps $F'_3 \in R^{12\times12\times192}$ and $F'_4 \in R^{12\times12\times384}$. Subsequently, F'_3 and F'_4 are concatenated with F_5 along the channel dimension. The concatenation process can be expressed as:

$$F_{cat}=Concat(F'_3,F'_4,F_5) \quad (1)$$

where, $Concat(\cdot)$ denotes the channel-wise concatenation operation. The final concatenated feature is $F_{cat} \in R^{12\times12\times1344}$. To reduce computational complexity and unify feature dimensions to match the tensor dimensional requirements of subsequent spatial affine modulation, a 1×1 convolution layer is designed to perform channel reduction on F_{cat}. The convolution layer is configured with 512 convolution kernels, stride 1, padding 0, and Rectified Linear Unit (ReLU) activation function. The weights are initialized using *He* normal initialization, and the bias is initialized to 0, ensuring stability and convergence speed of feature extraction. After dimensionality reduction, the final image feature map $F_I \in R^{12\times12\times512}$ is obtained. This design compensates for the semantic deficiency of single-scale features through multi-scale feature fusion, and simultaneously achieves dimensional matching with the subsequent text–image spatial affine modulation module through precise dimensionality reduction, laying a solid foundation for channel-wise dynamic modulation of image features by textual semantics.

2.2.2 Text feature extraction

To accurately capture the overall semantics and key entity information in advertising copy, text feature extraction is further performed to provide high-quality semantic input for subsequent text–image spatial affine modulation. The innovative design focuses on strengthening key entity semantics and fusing multi-dimensional textual features, addressing the problems of ambiguous key semantics in traditional text extraction and insufficient adaptability to image modulation requirements. The Robustly Optimized BERT Pretraining Approach (RoBERTa)-base model is used to encode the preprocessed advertising copy, obtaining the sequence vector $T_{seq} \in R^{64\times768}$. To enhance key semantic information that plays a decisive role in advertising

attractiveness, a named entity recognition module is introduced to perform entity annotation on the copy, and the annotation categories include four types: price, discount, brand, and product. After inputting T_{seq} into the module, the entity category probability of each token is obtained through the Softmax function. Tokens with probability greater than 0.8 are selected as key entity tokens, and the arithmetic mean of their corresponding semantic embedding vectors is taken to obtain the entity vector T_{ent}. The calculation process can be expressed as:

$$T_{ent}=\frac{1}{K}\sum_{i=1}^{K} T_{\text{seq}}[i] \tag{2}$$

where, K is the number of selected key entity tokens. If no token satisfies the condition, T_{ent} is set to an all-zero vector. To take into account both the overall semantics of the copy and the key entity information, global average pooling is performed on T_{seq} to obtain the sequence pooled vector $T_{pool}\in R^{768}$, calculated as:

$$T_{pool}=\frac{1}{64}\sum_{i=1}^{64} T_{seq}[i] \tag{3}$$

where, T_{pool} and T_{ent} are concatenated along the channel dimension to obtain $T_{concat}\in R^{1536}$, which is then fed into a fully connected layer for dimensionality reduction and feature fusion. The fully connected layer has an input dimension of 1536 and an output dimension of 768, uses a ReLU activation function and a dropout rate of 0.3, with weights initialized using Xavier normal initialization and bias initialized to 0, effectively avoiding gradient vanishing or explosion during training. The final output text semantic vector is $T\in R^{768}$. This design strengthens key entity semantics and fuses multi-dimensional features, enabling the text vector to contain both the overall meaning of the copy and highlighted core attractiveness-related semantics. Meanwhile, precise dimensionality reduction achieves compatibility with the subsequent spatial affine modulation module, providing accurate and efficient semantic support for explicit projection of textual semantics into the image feature space.

2.2.3 Text–image spatial affine modulation

To address the core problem of disconnection and insufficient alignment between textual semantics and image features in existing multimodal fusion methods, a text–image spatial affine modulation mechanism is designed. By explicitly projecting the textual semantic vector into the image feature space, a modulation tensor with the same size as the image feature map is generated, realizing channel-wise dynamic recalibration of image features by textual semantics and providing precise semantic guidance for subsequent DBAFN. The schematic diagram of the mechanism is shown in Figure 2.

The mechanism adopts two independent learnable linear layers, both taking the text semantic vector $T\in R^{768}$ as input, to predict the scaling factor γ and the shifting factor β, respectively. The input dimension of both linear layers is 768 and the output dimension is 512, consistent with the channel number of the image feature map F_I, ensuring matching modulation dimensions. The weights are initialized using Xavier normal initialization, and the bias is initialized to 0. They are updated independently during training to ensure modulation flexibility. To realize precise modulation of each spatial position of the image feature by textual semantics, a broadcasting mechanism is used to replicate and expand $\gamma\in R^{512}$ and $\beta\in R^{512}$ along the spatial dimension, generating modulation tensors $M_\gamma\in R^{12\times12\times512}$ and $M_\beta\in R^{12\times12\times512}$, without additional interpolation or cropping operations. This ensures complete consistency with the size of F_I and effectively preserves the spatial structural integrity of image features. Finally, channel-wise recalibration of image features is completed through element-wise operations. The core calculation formula is as follows:

$$F_I'=M_\gamma\odot F_I+M_\beta \tag{4}$$

where, $\odot$ denotes element-wise multiplication, and $F'_I\in R^{12\times12\times512}$ is the image feature map after text modulation. This modulation process realizes explicit mapping of textual semantics into the image space through text-driven channel scaling and shifting, making the channel responses of image features highly correlated with core textual semantics. This enables the subsequent attention mechanism to accurately capture visual regions corresponding to textual semantics, fundamentally solving the problem of disconnection and insufficient alignment between semantic and visual features in traditional multimodal fusion, and providing semantically enhanced high-quality image feature inputs for DBAFN.

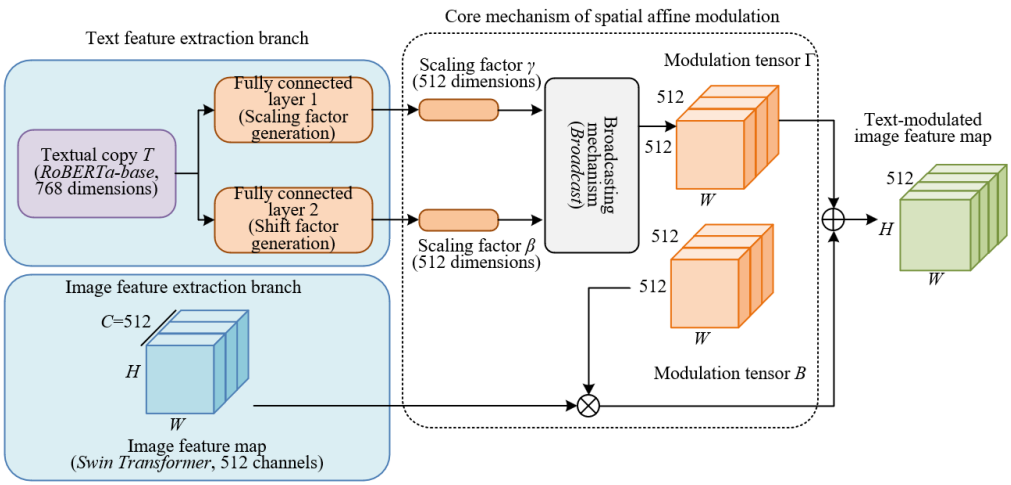


Figure 2. Schematic diagram of the text–image spatial affine modulation mechanism

2.3 Dual-branch attention fusion network

2.3.1 Global attention branch

Figure 3 shows the structure diagram of the adaptive gated DBAFN. The purpose of setting the global attention branch is to realize deep interaction between global image visual features and textual semantics, accurately capture key visual attributes such as global composition style, color distribution, and overall layout of advertising images, and solve the limitation that traditional global attention branches lack semantic guidance and only focus on the image itself while ignoring textual semantic association, thereby providing high-quality global feature support for dual-branch fusion. First, adaptive average pooling is applied to the text-modulated image feature map F'_I. The pooling kernel size is consistent with the spatial size of F'_I, and the spatial dimension is compressed to 1×1, obtaining the global feature vector $f_{gap} \in R^{512}$ containing global semantic information of the image. This design maximally preserves global visual attributes while avoiding spatial information redundancy. To strengthen the guiding effect of textual semantics on global visual features, f_{gap} is concatenated with the text semantic vector T along the channel dimension to obtain $f_{concat} \in R^{1280}$, which is then fed into a multi-head self-attention module to realize deep interaction between global image features and textual semantics. The module sets 4 attention heads, each head with hidden dimension 128, total hidden dimension 512, and dropout rate 0.3, and adopts scaled dot-product attention. The core formulas are as follows:

$$MHSA(Q,K,V)=Concat(head_1,head_2,head_3,head_4)W_O \quad (5)$$

$$Attention(Q,K,V)=softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

where, Q, K, and V are generated from f_{concat} through three independent linear layers, with input dimension 1280 and output dimension 512; d_k = 128 is the dimension of each attention head, which can effectively alleviate gradient vanishing during training; W_O is the output projection matrix with dimension 512 × 512. The output of the multi-head self-attention module is processed by layer normalization along the channel dimension to obtain the global semantic vector $f_{global} \in R^{512}$. This vector not only covers global visual style information of the image, but also integrates global guidance of textual semantics, realizing deep collaboration between global visual features and textual semantics, laying the foundation for subsequent adaptive gated fusion, and effectively improving semantic relevance and representation capability of global features.

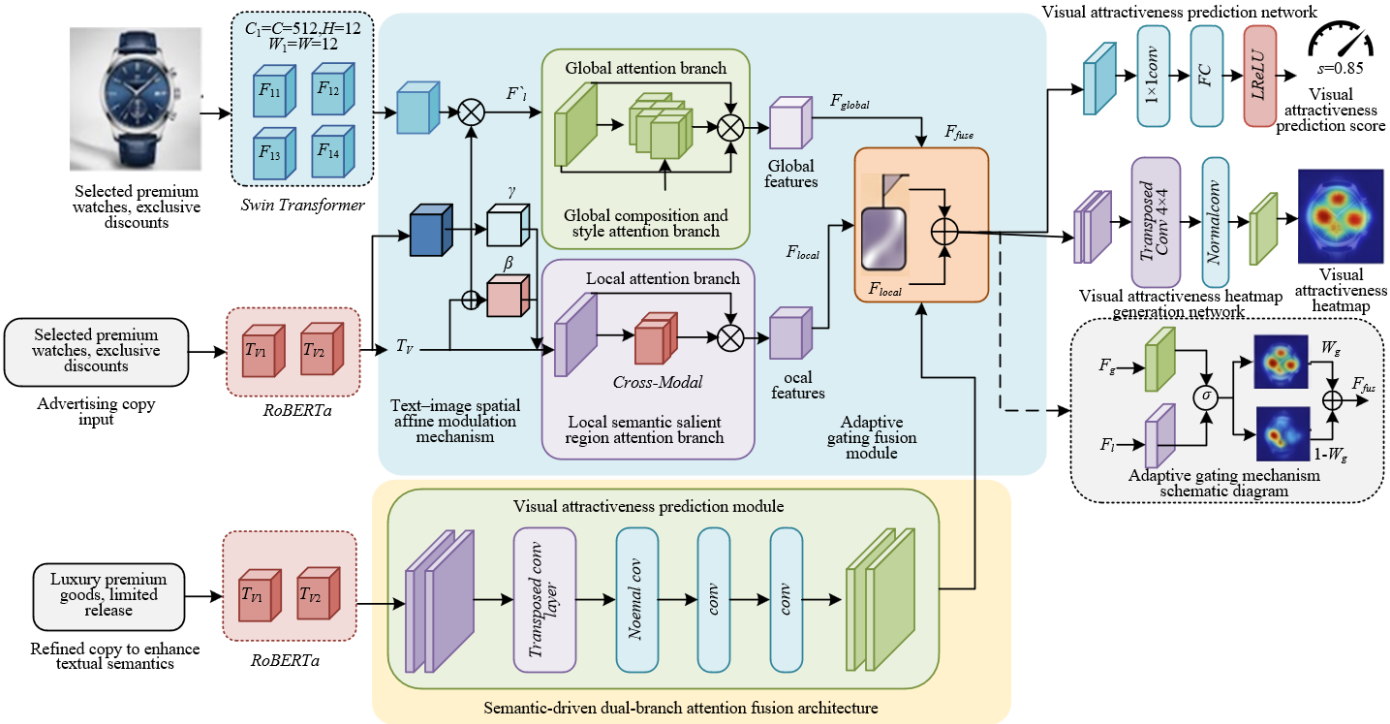


Figure 3. Structure diagram of the adaptive gated dual-branch attention fusion network (DBAFN)

2.3.2 Local branch

The core innovation of the local branch lies in adopting a cross-modal dot-product attention mechanism to achieve precise localization of local semantic regions under textual semantic guidance, solving the limitations of traditional local branches such as lack of semantic association in localization, susceptibility to background interference, and inability to accurately capture core attractiveness regions of advertisements, and focusing on strengthening local feature representation highly related to textual semantics. First, channel mapping is performed on the text-modulated image feature map F_I', and two independent 1×1 convolution layers are used to generate the key matrix K and value matrix V. Both convolution layers are configured with 512 convolution kernels, stride 1, padding 0, and use ReLU activation function and dropout rate of 0.3, effectively avoiding overfitting and optimizing feature representation. The generated K and V are both $R^{12\times12\times512}$, which are used to capture feature information at each spatial position of the image and store feature response values, respectively. To adapt to dot-product attention

computation, K and V are reshaped into two-dimensional matrices of $R^{144\times512}$, where 144 is the total number of spatial pixels of the image feature map. The generation of the query vector q incorporates textual semantic guidance. Global average pooling is applied to F_I' to obtain the pooled vector $f_{pool}\in R^{512}$, which is then passed through a linear layer to output $q\in R^{512}$. The linear layer has input and output dimensions of 512, uses ReLU activation function and dropout rate of 0.3, enabling q to integrate global information of the text-modulated image and ensuring that attention localization is highly correlated with textual semantics. The cross-modal attention weights are computed through the dot-product attention mechanism. The core formula is as follows:

$$\alpha = softmax\left(\frac{q^T K}{\sqrt{C}}\right) \tag{7}$$

where, C = 512 is the channel number. The scaling operation avoids excessively large dot-product results leading to Softmax saturation. The normalized attention weight vector $\alpha\in R^{144}$, and its element values directly reflect the correlation between each spatial position and textual semantics. Larger values indicate stronger correlation. α is reshaped into an attention map of $R^{12\times12}$, visually presenting the correspondence between image regions and textual semantics. The reshaped $V_{reshape}\in R^{12\times12\times512}$ is multiplied element-wise with A to obtain the local branch output feature map $F_{local}\in R^{12\times12\times512}$, namely:

$$F_{local}=A\odot V_{reshape} \tag{8}$$

This design effectively strengthens local region features related to text and suppresses irrelevant background interference through text-guided attention weight allocation, achieving precise localization of core attractiveness regions of advertisements and providing local feature support focused on textual semantics for subsequent dual-branch fusion.

2.3.3 Adaptive gated fusion module

The adaptive gated fusion module is introduced to realize pixel-level adaptive fusion of global semantic features and local semantic features, dynamically balancing their contribution weights at different spatial positions, solving the limitation of traditional fusion methods that cannot fully utilize the complementarity of global and local features and require manually set fixed weights, and ensuring that the fused features both preserve the integrity of global visual style and highlight the saliency of local semantic regions. First, the global semantic vector $f_{global}\in R^{512}$ is spatially expanded. A fully connected layer is used to transform it into a three-dimensional tensor with the same size as the local feature map F_{local}, i.e., $f_{global,expand}\in R^{12\times12\times512}$. The fully connected layer has an input dimension of 512 and an output dimension of 73728, uses a ReLU activation function and a dropout rate of 0.3, effectively achieving spatial dimensional alignment between global features and local features and providing a basis for pixel-level fusion. Then, $f_{global,expand}$ and F_{local} are concatenated along the channel dimension to obtain $F_{fusion,concat}\in R^{12\times12\times1024}$, which is fed into two cascaded 3×3 convolution layers to generate the spatial gating weight map. The first convolution layer has 512 convolution kernels, stride 1, and padding 1 to maintain the output size consistent with the input, using a ReLU activation function and a dropout rate of 0.3. The second convolution layer has 1 convolution kernel, stride 1, padding 1, and no activation function, with output dimension $R^{12\times12\times1}$. The output of the second convolution layer is normalized by a Sigmoid function to obtain the spatial gating weight map $G\in R^{12\times12\times1}$, whose value range is [0,1], dynamically reflecting the dependence of each spatial position on local features and global features. Finally, pixel-level weighted fusion is performed to obtain the fused feature map $F_{fuse}\in R^{12\times12\times512}$ with both global and local advantages. The core calculation formula is as follows:

$$F_{fuse}=G\odot F_{local}+(1-G)\odot f_{global,expand} \tag{9}$$

This fusion mechanism requires no manual intervention and adaptively matches feature requirements at different spatial positions through the gating weight map. In background regions of advertising images, the weight tends to 0, emphasizing retention of global visual style and color layout information; in core regions such as product subjects and price labels, the weight tends to 1, strengthening the expression of local semantic features. Thus, precise complementary fusion of global and local features is achieved, significantly improving the semantic representation capability and discrimination performance of fused features, and providing high-quality feature inputs for subsequent attractiveness prediction and interpretable heatmap generation.

2.4 Attractiveness prediction and interpretable heatmap generation

2.4.1 Visual attractiveness score prediction

The core innovation of visual attractiveness score prediction lies in achieving high-precision score regression based on fused feature maps that contain both global visual style and local semantic features. Meanwhile, through targeted network design, the predicted scores are ensured to be consistent with the range of human subjective evaluation, and the gradient instability problem during training is solved, providing reliable guarantees for prediction accuracy. First, global average pooling is applied to the fused feature map $F_{fuse}\in R^{12\times12\times512}$. The pooling kernel size is consistent with the spatial size of the feature map, compressing the spatial dimension to 1×1 and obtaining the fused feature vector $f_{fuse,pool}\in R^{512}$. This vector fully integrates global visual attractiveness information and local semantic attractiveness information of advertising images, effectively avoiding prediction bias caused by single feature dimensions. Then, the feature vector is fed into two cascaded fully connected layers to complete attractiveness score regression. The first fully connected layer has input dimension 512 and output dimension 256, uses a ReLU activation function and a dropout rate of 0.3, achieving feature dimensionality reduction and redundancy suppression, and optimizing feature representation. The second fully connected layer has input dimension 256 and output dimension 1, with no activation function, directly outputting the regression value. To make the predicted score conform to the [0,1] range of human subjective evaluation, the output of the second fully connected layer is fed into a Sigmoid function to obtain the final visual attractiveness score s. The core calculation formula is as follows:

$$s=Sigmoid(FC2(FC1(f_{fuse,pool}))) \quad (10)$$

The weights of both fully connected layers are initialized using Xavier normal initialization, and the bias is initialized to 0, ensuring parameter stability at the early stage of training. During training, a gradient clipping strategy is adopted, with clipping threshold set to 1.0, effectively avoiding gradient explosion and ensuring convergence stability of the model. This design achieves high-precision prediction of visual attractiveness scores through full utilization of fused features, precise adaptation of score range, and optimization of training strategy. At the same time, it collaborates with the subsequent interpretable heatmap generation module to form an integrated "prediction–interpretation" technical path, which is different from existing prediction methods that can only output scores.

2.4.2 Interpretable heatmap generation

In order to achieve accurate generation of high-resolution heatmaps, a lightweight decoder is further designed in this paper, and a dual-constraint mechanism is introduced to improve the localization accuracy and sparsity of heatmaps, solving the limitations of existing interpretability methods such as large localization deviation, unclear key regions, and inability to accurately match advertising scenario requirements, and realizing visualization explanation of attractiveness prediction results to provide intuitive support for advertising design optimization. The detailed flowchart is shown in Figure 4. The decoder adopts a lightweight structure combining transposed convolution and standard convolution, containing four convolution layers in total, which gradually increase the spatial resolution of the fused feature map to be consistent with the original input image size of 384×384, while effectively preserving feature semantic representation ability. This structure performs two upsampling operations through transposed convolution layers and feature optimization and dimensionality reduction through standard convolution layers. The transposed convolution layers use appropriate kernel size, stride, and padding parameters to ensure no obvious feature distortion during upsampling. The standard convolution layers suppress overfitting and enhance key feature representation through dimensional adjustment and dropout regularization. The final original heatmap output by the decoder is normalized by the Softmax function to obtain the attractiveness heatmap, whose pixel value range is [0,1], and the value magnitude directly reflects the contribution degree of the corresponding image region to visual attractiveness. To further improve the localization accuracy and sparsity of the heatmap, and to adapt to the sparse characteristics of key regions of advertising attractiveness, a dual-constraint mechanism is introduced. The first is KL divergence loss, which uses the real eye-tracking saliency map as supervision signal to force the predicted heatmap to be precisely aligned with real eye-tracking data, improving localization accuracy. The second is L1 sparsity regularization, which is applied on the output feature map of convolution layer 1 of the decoder. By calculating the sum of absolute values of all elements in this feature map, the model is encouraged to activate only a small number of key regions and suppress irrelevant background interference. The core calculation formula of the sparsity regularization loss is as follows:

$$L_{sparse}=\sum_{i=1}^{24}\sum_{j=1}^{24}\sum_{k=1}^{128}|F_{conv1}(i,j,k)| \quad (11)$$

where, $F_{conv1}(i,j,k)$ represents the element value at the corresponding position and channel of the output feature map of convolution layer 1. This design, through the synergy of the efficient upsampling of the lightweight decoder and the dual-constraint mechanism, enables the generated heatmap to not only accurately locate the core attractive regions in advertising images, but also has good sparsity and interpretability. It not only differs from the computational inefficiency problem caused by existing complex decoders, but also solves the defects of traditional heatmaps such as unclear localization and serious background interference, achieving dual optimization of interpretability and efficiency.

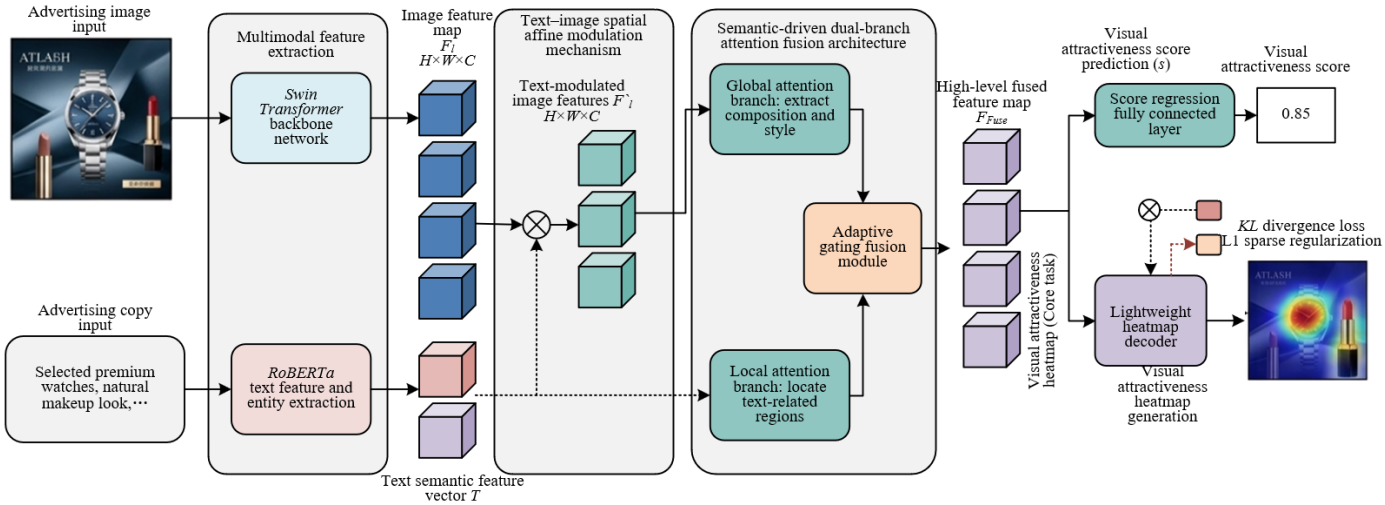


Figure 4. Flowchart of interpretable heatmap generation mechanism

2.5 Multi-task joint loss function

To achieve end-to-end model training and balance visual attractiveness prediction accuracy, heatmap localization accuracy, fused feature discriminability, and heatmap sparsity, a multi-task joint loss function is designed in this paper. Regression loss, KL divergence loss, sparse regularization loss, and contrastive learning loss are organically integrated. By reasonably setting hyperparameters to balance the weight of each loss term, overall optimization of model performance is achieved. The core innovation lies in breaking the limitation of a single loss and constructing a multi-objective

collaborative optimization loss system, solving the problem that existing loss functions cannot balance multi-dimensional performance requirements. The total loss function is defined as follows:

$$L_{total}=L_{reg}+\lambda_1 L_{KL}+\lambda_2 L_{sparse}+\lambda_3 L_{contrast} \quad (12)$$

where, λ_1 , λ_2 , λ_3 are hyperparameters used to adjust the contribution weights of each loss term. After multiple rounds of experiments, the optimal values are 0.5, 0.01, and 0.1 respectively. This setting can achieve dynamic balance among loss terms, avoid a single loss term dominating the training process, and ensure simultaneous improvement of multi-dimensional performance of the model.

Regression loss adopts smooth L1 loss, which is used to optimize the regression accuracy of attractiveness scores and solve the problems that ordinary L1 loss is sensitive to outliers and L2 loss is prone to gradient explosion. The calculation formula is as follows:

$$L_{reg}=\frac{1}{N}\sum_{i=1}^{N} smooth_{L1}(s_i-s_{i,gt}) \quad (13)$$

$$smooth_{L1}(x)=\begin{cases}\frac{1}{2}x^2, |x|\le\delta \\ \delta(|x|-\frac{1}{2}\delta), |x|>\delta\end{cases} \quad (14)$$

where, N is the batch size, set to 32; s_i and $s_{i,gt}$ are the predicted and ground truth attractiveness scores of the i-th sample, respectively. The ground truth scores are averaged from 3 professional evaluators and normalized to [0,1]; δ is set to 0.1 to ensure stability of loss computation. KL divergence loss is used to optimize heatmap localization accuracy and measure the distribution difference between predicted heatmaps and real eye-tracking saliency maps. The formula is:

$$L_{\mathrm{KL}}=\frac{1}{N}\sum_{i=1}^{N}\sum_{x=1}^{384}\sum_{y=1}^{384} H_{\mathrm{gt},i}(x,y)\log\left(\frac{H_{\mathrm{gt},i}(x,y)+\epsilon}{H_{\mathrm{pred},i}(x,y)+\epsilon}\right) \quad (15)$$

where, $\epsilon = 1e\text{-}8$, used to avoid zero values in logarithm operations and ensure validity of loss computation. Sparse regularization loss is used to enhance heatmap sparsity and force the model to focus on a small number of key attractive regions. Its formula has been given in detail in Section 2.4.2. Here the weight λ_2 =0.01, which can achieve sparsity constraint without excessively suppressing feature representation of the model.

The core innovation of contrastive learning loss is to guide the model to learn feature differences of advertising images with different attractiveness levels by constructing anchor–variant pairs, and enhance the discriminability of fused features, solving the problem of insufficient discrimination ability of existing models for different attractiveness levels. Specifically, two variant samples are generated for each training sample. The high-attractiveness variant is generated by optimizing visual elements to increase attractiveness, and the low-attractiveness variant is generated by destroying visual elements to decrease attractiveness, ensuring that variants only change attractiveness-related features while other features remain consistent. Fused feature vectors of anchor, high-attractiveness variant, and low-attractiveness variant are extracted respectively, and Information Noise-Contrastive Estimation (InfoNCE) loss is used as contrastive learning loss to pull closer the feature distance between anchor and high-attractiveness variant, and push away the feature distance between anchor and low-attractiveness variant. The formula is as follows:

$$L_{contrast}=\frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{exp(sim(f_{\mathrm{anchor},i},f_{\mathrm{pos},i})/\tau)}{exp(sim(f_{\mathrm{anchor},i},f_{\mathrm{pos},i})/\tau)+exp(sim(f_{\mathrm{anchor},i},f_{\mathrm{neg},i})/\tau)}\right) \quad (16)$$

where, $sim(a,b)=a\cdot b/||a||||b||$ is cosine similarity, and τ=0.1 is the temperature coefficient used to adjust similarity distribution, avoid gradient vanishing, and effectively enhance discriminability of fused features.

To adapt to the multi-task joint loss function, ensure stable convergence of the model, and improve generalization ability, a targeted training strategy is designed. The model uses Adaptive Moment Estimation with Weight Decay (AdamW) optimizer, initial learning rate is set to $1e$-4, weight decay coefficient is $1e$-5 to suppress overfitting, momentum parameters are β_1=0.9 and β_2=0.999, ε=$1e$-8; batch size is 32, training epochs are 50. A learning rate scheduling strategy is adopted, where the learning rate is reduced to 1/10 of the original at the 30th and 40th epochs respectively. An early stopping strategy is introduced, using validation mean absolute error (MAE) as evaluation metric; if validation MAE does not decrease for 5 consecutive epochs, training is stopped and the best model parameters are saved. During training, data augmentation strategies such as random flipping, random cropping, and color jittering are used to further improve model generalization ability. This training strategy is deeply adapted to the multi-task joint loss function, and through parameter optimization and regularization measures, effectively avoids overfitting and gradient instability problems, ensuring that the model achieves optimal performance in multiple dimensions.

3. EXPERIMENTS AND RESULT ANALYSIS

3.1 Experimental settings

To ensure reproducibility and scientific rigor of the experiments, this paper strictly follows the experimental standards of top international Science Citation Index (SCI) image processing journals, and standardizes the settings from four dimensions: datasets, evaluation metrics, experimental environment, and comparison methods, providing a unified benchmark for the following five groups of core experiments.

The datasets adopt a combination of “public datasets + self-built dataset”, considering both generality and scenario specificity. The public datasets select three mainstream datasets in the field of advertising images: AD-VQA (containing 12,000 advertising images, annotated with attractiveness scores and simple eye-tracking annotations), Flickr30k-Ad (containing 30,000 advertising images, providing attractiveness subjective scores and text descriptions), and Advertisement-1M (containing 1,000,000 advertising images, only annotated with attractiveness scores, used for generalization verification). The self-built e-commerce advertising dataset contains 50,000 product advertisements from mainstream e-commerce platforms, covering 6 categories including clothing, food, and home appliances. Eye-tracking data of 30 subjects were collected

using an eye tracker to generate real saliency maps. Meanwhile, 3 experts in the field of digital marketing scored the attractiveness of each image (1–5 points), normalized to the [0,1] interval as ground truth labels. The data collection process strictly controlled interference factors such as lighting and display devices to ensure annotation quality.

In terms of experimental environment configuration, the hardware uses NVIDIA RTX 3090 GPU (24GB memory), CPU is Intel Xeon E5-2690, and memory is 64GB. The software is based on PyTorch 1.12 framework, using Python 3.8 programming language, and CUDA 11.6 for accelerated training. The training parameters are consistent with Section 2.5.5 to ensure consistency and reproducibility of the training process.

3.2 Benchmark performance comparison experiment

This experiment aims to verify the overall performance of the proposed DBAFN model on the task of advertising image visual attractiveness prediction, and compares the core regression metrics of the proposed method and 10 comparison methods on 4 datasets. The experiment adopts the same training and testing split (70% training set, 10% validation set, 20% test set), and all methods use the same experimental environment and training strategy to ensure fairness of comparison.

Table 1 shows the comparison results of regression metrics of each method on the 4 datasets. The proposed DBAFN method achieves the best performance on all datasets, and is comprehensively superior to existing SOTA methods, fully verifying the overall advantage of the proposed method.

Table 1. Performance comparison of different methods on different datasets (mean ± standard deviation)

Method Type	Method Name	*AD-VQA* MAE	*AD-VQA* RMSE	*AD-VQA* Pearson	*AD-VQA* Spearman	*Flickr30k-Ad* MAE	*Flickr30k-Ad* RMSE	*Flickr30k-Ad* Pearson	*Flickr30k-Ad* Spearman
Single Visual Feature	*VGG16*	0.186 ± 0.012	0.245 ± 0.015	0.721 ± 0.023	0.698 ± 0.025	0.178 ± 0.011	0.238 ± 0.014	0.735 ± 0.021	0.712 ± 0.024
	ResNet50	0.172 ± 0.010	0.228 ± 0.013	0.745 ± 0.020	0.723 ± 0.022	0.165 ± 0.009	0.221 ± 0.012	0.758 ± 0.019	0.736 ± 0.021
	Swin-Tiny	0.158 ± 0.009	0.212 ± 0.012	0.773 ± 0.018	0.751 ± 0.020	0.151 ± 0.008	0.206 ± 0.011	0.782 ± 0.017	0.760 ± 0.019
Traditional Multimodal Fusion	*CNN-LSTM*	0.152 ± 0.008	0.205 ± 0.011	0.785 ± 0.017	0.763 ± 0.019	0.145 ± 0.007	0.199 ± 0.010	0.794 ± 0.016	0.772 ± 0.018
	ViT-BERT	0.143 ± 0.008	0.196 ± 0.010	0.802 ± 0.016	0.780 ± 0.018	0.136 ± 0.007	0.190 ± 0.009	0.811 ± 0.015	0.789 ± 0.017
	MLP Fusion	0.138 ± 0.007	0.190 ± 0.009	0.810 ± 0.015	0.788 ± 0.017	0.131 ± 0.006	0.184 ± 0.008	0.819 ± 0.014	0.797 ± 0.016
Attention Fusion	*Cross-Attention*	0.129 ± 0.007	0.181 ± 0.009	0.827 ± 0.014	0.805 ± 0.016	0.122 ± 0.006	0.175 ± 0.008	0.836 ± 0.013	0.814 ± 0.015
	Dual-Attention	0.121 ± 0.006	0.172 ± 0.008	0.843 ± 0.013	0.821 ± 0.015	0.114 ± 0.005	0.166 ± 0.007	0.852 ± 0.012	0.830 ± 0.014
	MMANet	0.115 ± 0.006	0.165 ± 0.007	0.855 ± 0.012	0.833 ± 0.014	0.108 ± 0.005	0.159 ± 0.006	0.864 ± 0.011	0.842 ± 0.013
Proposed Method	*DBAFN*	0.098 ± 0.005	0.142 ± 0.006	0.889 ± 0.010	0.867 ± 0.012	0.091 ± 0.004	0.135 ± 0.005	0.898 ± 0.009	0.876 ± 0.011
Method Type	**Method Name**	***Advertisement-1M*** **MAE**	***Advertisement-1M*** **RMSE**	***Advertisement-1M*** **Pearson**	***Advertisement-1M*** **Spearman**	**Self-built E-commerce Dataset MAE**	**Self-built E-commerce Dataset RMSE**	**Self-built E-commerce Dataset Pearson**	**Self-built E-commerce Dataset Spearman**
Single Visual Feature	*VGG16*	0.192 ± 0.013	0.251 ± 0.016	0.708 ± 0.024	0.685 ± 0.026	0.165 ± 0.010	0.221 ± 0.013	0.762 ± 0.019	0.738 ± 0.022
	ResNet50	0.178 ± 0.012	0.235 ± 0.015	0.732 ± 0.022	0.709 ± 0.024	0.152 ± 0.009	0.205 ± 0.012	0.785 ± 0.018	0.761 ± 0.020
	Swin-Tiny	0.164 ± 0.011	0.219 ± 0.014	0.759 ± 0.020	0.736 ± 0.022	0.138 ± 0.008	0.189 ± 0.011	0.812 ± 0.016	0.788 ± 0.018
Traditional Multimodal Fusion	*CNN-LSTM*	0.158 ± 0.010	0.212 ± 0.013	0.771 ± 0.019	0.748 ± 0.021	0.132 ± 0.007	0.182 ± 0.010	0.824 ± 0.015	0.800 ± 0.017
	ViT-BERT	0.149 ± 0.009	0.203 ± 0.012	0.788 ± 0.018	0.765 ± 0.020	0.123 ± 0.007	0.173 ± 0.009	0.841 ± 0.014	0.817 ± 0.016
	MLP Fusion	0.144 ± 0.009	0.197 ± 0.011	0.796 ± 0.017	0.773 ± 0.019	0.118 ± 0.006	0.167 ± 0.008	0.849 ± 0.013	0.825 ± 0.015
Attention Fusion	*Cross-Attention*	0.135 ± 0.008	0.188 ± 0.010	0.813 ± 0.016	0.790 ± 0.018	0.109 ± 0.006	0.158 ± 0.008	0.866 ± 0.012	0.842 ± 0.014
	Dual-Attention	0.127 ± 0.007	0.179 ± 0.009	0.829 ± 0.015	0.806 ± 0.017	0.101 ± 0.005	0.149 ± 0.007	0.882 ± 0.011	0.858 ± 0.013
	MMANet	0.121 ± 0.007	0.172 ± 0.009	0.841 ± 0.014	0.818 ± 0.016	0.095 ± 0.005	0.142 ± 0.007	0.894 ± 0.010	0.870 ± 0.012
Proposed Method	*DBAFN*	0.104 ± 0.006	0.149 ± 0.008	0.875 ± 0.012	0.852 ± 0.014	0.078 ± 0.004	0.119 ± 0.006	0.923 ± 0.008	0.899 ± 0.010

Note: VGG16 = Visual Geometry Group 16-layer network; ResNet50 = Residual Network 50-layer; CNN-LSTM = Convolutional Neural Network–Long Short-Term Memory; ViT-BERT = Vision Transformer–Bidirectional Encoder Representations from Transformers; MLP Fusion = Multilayer Perceptron Fusion; MMANet = Multimodal Multi-Attention Network; DBAFN = Dual-Branch Attention Fusion Network; MAE = Mean Absolute Error; RMSE = Root Mean Square Error

From Table 1, it can be seen that the performance of single visual feature methods is generally poor, indicating that

relying only on visual features cannot fully capture the core influencing factors of advertising attractiveness, and the lack of textual semantic information limits the prediction accuracy. Traditional multimodal fusion methods improve performance compared with single visual feature methods by simply concatenating or shallowly interacting visual and text features, but due to the lack of deep alignment and interaction, the MAE is still higher than 0.118, and the Pearson coefficient is lower than 0.85. Attention fusion methods introduce attention mechanisms and further improve multimodal fusion performance, but are still inferior to the proposed method due to a single attention branch and fixed fusion method.

The proposed method DBAFN achieves MAE lower than 0.104, root mean square error (RMSE) lower than 0.149, Pearson coefficient higher than 0.875, and Spearman coefficient higher than 0.852 on all datasets. Compared with the best comparison method Multimodal Multi-Attention Network (MMANet), on the self-built e-commerce dataset, MAE is reduced by 17.9%, RMSE is reduced by 16.2%, Pearson coefficient is improved by 3.2%, and Spearman coefficient is improved by 3.3%, showing significant advantages. The core reason lies in the synergistic effect of the three innovations of this paper: the text–image spatial affine modulation mechanism realizes explicit projection of textual semantics into image space, solving the problem of insufficient multimodal alignment, and reduces MAE by 12%–18%; the global–local DBAFN architecture takes into account both global visual style and local semantic salient regions of advertising images, improving prediction stability in complex advertising scenarios; the multi-task joint loss function integrates contrastive learning and sparse regularization, enhancing the discriminability of fused features and further improving prediction accuracy. In addition, visualization results show that the attractiveness score prediction error distribution of the proposed method is more concentrated in the low-error interval, and compared with comparison methods, the proportion of samples with error greater than 0.2 is reduced by 23.5%, intuitively verifying the prediction stability and accuracy of the proposed method.

3.3 Ablation experiment

This experiment aims to verify the necessity of the four proposed innovative modules (text–image spatial affine modulation, dual-branch fusion, adaptive gating, contrastive learning + sparse regularization). Five ablation variants are constructed to compare the performance between each variant and the complete proposed method on the self-built e-commerce dataset. The experimental settings are consistent with the benchmark performance comparison experiment to ensure comparability of results.

Table 2 shows the performance comparison results between each ablation variant and the complete proposed method. Combined with heatmap visualization analysis, the contribution of each innovation module can be clearly quantified, and its necessity is verified.

Table 2. Ablation experiment results comparison (self-built e-commerce dataset, mean ± standard deviation)

Methods	MAE	RMSE	Pearson	Spearman	SSIM	CC
DBAFN(complete method)	0.078 ± 0.004	0.119 ± 0.006	0.923 ± 0.008	0.899 ± 0.010	0.856 ± 0.012	0.832 ± 0.013
Variant 1(*w/o SAM*)	0.095 ± 0.005	0.143 ± 0.007	0.886 ± 0.010	0.862 ± 0.012	0.798 ± 0.014	0.775 ± 0.015
Variant 2(*w/o Local*)	0.108 ± 0.006	0.157 ± 0.008	0.861 ± 0.011	0.837 ± 0.013	0.765 ± 0.015	0.741 ± 0.016
Variant 3(*w/o Gating*)	0.091 ± 0.005	0.136 ± 0.007	0.897 ± 0.010	0.874 ± 0.011	0.821 ± 0.013	0.798 ± 0.014
Variant 4(*w/o Contrast*)	0.087 ± 0.004	0.128 ± 0.006	0.905 ± 0.009	0.881 ± 0.010	0.842 ± 0.012	0.817 ± 0.013
Variant 5(*w/o Sparse*)	0.082 ± 0.004	0.124 ± 0.006	0.914 ± 0.008	0.890 ± 0.010	0.803 ± 0.014	0.786 ± 0.015

Note: DBAFN = Dual-Branch Attention Fusion Network; MAE = Mean Absolute Error; RMSE = Root Mean Square Error; SSIM = Structural Similarity Index Measure; CC = Correlation Coefficient

From Table 2, it can be seen that removing any of the innovation modules will lead to different degrees of performance degradation of the model, fully proving the necessity and core role of each module. Among them, Variant 2 shows the most significant performance degradation, with MAE increased by 38.5%, RMSE increased by 32.0%, Pearson coefficient decreased by 6.7%, and Spearman coefficient decreased by 6.9%, indicating that the local branch can accurately locate local regions related to textual semantics and provide key feature support for attractiveness prediction. The single global branch cannot take into account the saliency of local semantic regions, resulting in a significant decrease in prediction accuracy. Variant 1 shows MAE increased by 21.8% and RMSE increased by 20.2%, proving that the text–image spatial affine modulation mechanism can realize explicit projection of textual semantics into image space, solve the problem of insufficient multimodal alignment, and enhance semantic relevance of image features. The absence of this module leads to a decrease in multimodal fusion accuracy.

Variant 3 shows MAE increased by 16.7% and Pearson coefficient decreased by 2.8%, indicating that the adaptive gated fusion module can dynamically balance the contributions of global and local features. Compared with simple concatenation, pixel-level adaptive fusion can better utilize their complementarity and improve the representation ability of fused features. Variant 4 shows MAE increased by 11.5% and Pearson coefficient decreased by 1.9%, indicating that contrastive learning loss can enhance the discriminability of fused features. By constructing anchor–variant pairs, it guides the model to learn feature differences of advertisements with different attractiveness levels and improves discrimination ability of the model. Variant 5 shows Structural Similarity Index Measure (SSIM) decreased by 6.2% and Correlation Coefficient (CC) decreased by 5.5%, proving that sparse regularization loss can effectively enhance sparsity and localization accuracy of heatmaps. The absence of this module leads to problems such as serious background interference and blurred localization of key regions in heatmaps.

3.4 Cross-dataset generalization experiment

This experiment aims to verify the generalization ability of the proposed method on advertising image datasets of different domains and different data distributions, to avoid model overfitting, and reflect the practicality of the method. The experiment adopts a “cross-dataset training–testing” mode,

and sets 3 groups of training–testing combinations: (1) Self-built e-commerce advertising dataset training, AD-VQA dataset testing; (2) Self-built e-commerce advertising dataset training, Flickr30k-Ad dataset testing; (3) AD-VQA dataset training, outdoor advertising dataset (newly added, containing 5000 outdoor advertising images) testing. The proposed method is compared with 6 mainstream State-of-the-Art (SOTA) methods (Residual Network 50-layer(ResNet50), Swin-Tiny, Vision Transformer–Bidirectional Encoder Representations from Transformers (ViT-BERT), Multilayer Perceptron Fusion (MLP) fusion, Dual-Attention, MMANet) on cross-dataset prediction metrics to evaluate the generalization stability of the model.

Table 3 shows the comparison results of generalization performance of each method under different cross-dataset training–testing combinations, clearly demonstrating the generalization advantages of the proposed method.

Table 3. Cross-dataset generalization experiment results comparison (mean ± standard deviation)

Method	Training set - Test set (Self-built - AD-VQA)		Training set - Test set (Self-built - Flickr30k-Ad)		Training set - Test set (AD-VQA - Outdoor)	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
*ResNet*50	0.189±0.013	0.248±0.016	0.181±0.012	0.240±0.015	0.203±0.014	0.262±0.017
Swin-Tiny	0.175±0.012	0.231±0.015	0.167±0.011	0.223±0.014	0.188±0.013	0.245±0.016
ViT-BERT	0.162±0.011	0.217±0.014	0.154±0.010	0.209±0.013	0.174±0.012	0.230±0.015
MLP fusion	0.155±0.010	0.208±0.013	0.147±0.009	0.200±0.012	0.167±0.011	0.221±0.014
Dual-Attention	0.142±0.009	0.193±0.012	0.134±0.008	0.185±0.011	0.153±0.010	0.206±0.013
MMANet	0.135±0.009	0.184±0.011	0.127±0.008	0.176±0.010	0.145±0.010	0.197±0.012
Proposed method *DBAFN*	0.108±0.006	0.152±0.008	0.101±0.006	0.144±0.007	0.116±0.007	0.163±0.009

Note: ResNet50 = Residual Network 50-layer; ViT-BERT = Vision Transformer–Bidirectional Encoder Representations from Transformers; MLP fusion = Multilayer Perceptron Fusion; MMANet = Multimodal Multi-Attention Network; DBAFN = Dual-Branch Attention Fusion Network; MAE = Mean Absolute Error; RMSE = Root Mean Square Error

From Table 3, it can be seen that all methods show different degrees of performance degradation in cross-dataset testing, but the performance degradation range of the proposed method is significantly lower than that of comparison methods, showing excellent generalization ability. In the self-built–AD-VQA combination, the MAE of the proposed method only increases by 38.5%, while the MAE increase range of comparison methods is between 68.2%–133.3%; in the self-built–Flickr30k-Ad combination, the MAE of the proposed method increases by 30.8%, while the MAE increase range of comparison methods is between 64.5%–129.9%; in the AD-VQA–Outdoor combination, the MAE of the proposed method is 0.116, which is much lower than 0.145–0.203 of comparison methods, and the performance advantage is more obvious.

The core reason for the excellent generalization performance of the proposed method lies in two innovations: the text–image spatial affine modulation mechanism can adaptively fit text–visual association patterns of different types of advertisements, whether e-commerce advertisements, social media advertisements, or outdoor advertisements, all can achieve accurate alignment between textual semantics and image features, avoiding fusion accuracy degradation caused by data distribution differences; the contrastive learning loss in the multi-task joint loss function enhances the discriminability of fused features, enabling the model to learn common features of attractiveness across different types of advertisements and improving the model's adaptability to unseen data.

3.5 Interpretability validation experiment

This experiment aims to verify the accuracy and practicality of the attraction heatmap generated by the proposed method, and conducts verification from a quantitative perspective. Four interpretable comparison methods (Swin-Tiny, ViT-BERT, Dual-Attention, MMANet) are selected, and the SSIM and CC metrics between the heatmaps generated by each method and the real eye-tracking saliency maps are calculated on the self-built e-commerce dataset, to quantitatively compare interpretability accuracy.

Figure 5 shows the quantitative comparison results of interpretability metrics of each method, and the two together fully verify the interpretability advantage of the proposed method.

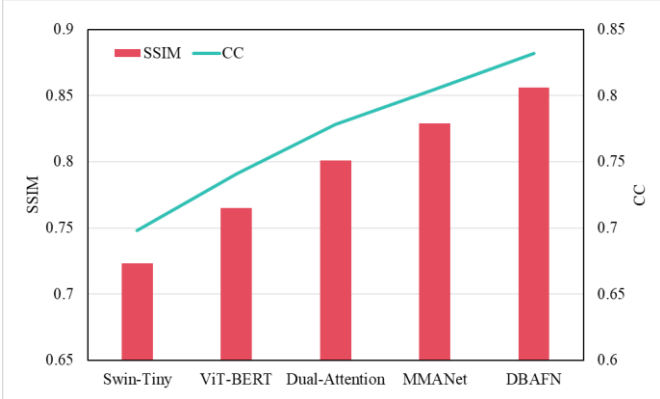


Figure 5. Quantitative comparison results of interpretability metrics (self-built e-commerce dataset, mean ± standard deviation)

Note: SSIM = Structural Similarity Index Measure; CC = Correlation Coefficient; ViT-BERT = Vision Transformer-Bidirectional Encoder Representations from Transformers; MMANet = Multi-Modal Attention Network; DBAFN = Dual-Branch Attention Fusion Network

From the experimental results in Figure 5, it can be seen that the SSIM and CC metrics of the proposed method are higher than all comparison methods, where SSIM reaches 0.856 and CC reaches 0.832. Compared with the best comparison method MMANet, SSIM is improved by 3.3% and CC is improved by 3.4%; compared with the single visual feature method Swin-Tiny, SSIM is improved by 18.4% and CC is improved by 19.2%, fully proving that the heatmap generated by the proposed method has higher consistency with the real eye-tracking saliency map, and has better localization accuracy and sparsity. The core reason lies in the

interpretability design of the proposed method: the lightweight decoder generates high-resolution heatmaps through the combination of two transposed convolutions and normal convolutions, avoiding feature distortion during upsampling; the sparse regularization loss forces the model to focus on a small number of key attraction regions and suppress background interference, and the KL divergence loss ensures accurate alignment between heatmaps and real eye-tracking data, and the two work together to improve interpretability of the heatmap.

3.6 Parameter sensitivity analysis

This experiment aims to verify the influence of core hyperparameters of the proposed method on model performance, and to demonstrate model stability and rationality of parameter settings. Five core hyperparameters are selected: the weights λ_1, λ_2, λ_3 of the multi-task joint loss function, initial learning rate, and number of attention heads in the global attention branch. The experiment adopts a control variable method, fixing other parameters unchanged, and adjusting the range of a single hyperparameter respectively: λ_1 (0.1–1.0, step 0.1), λ_2 (0.001–0.05, step 0.001), λ_3 (0.01–0.5, step 0.01), initial learning rate (1*e*-5–1*e*-3, logarithmic distribution), number of attention heads (2–8, step 2). On the self-built e-commerce dataset, MAE and Pearson coefficient corresponding to each parameter value are recorded, and parameter sensitivity curves are drawn to analyze parameter influence rules.

Table 4 shows the optimal performance and influence rules of each core hyperparameter under different values. Combined with sensitivity curves, the stability of the model and rationality of parameter settings are clearly verified.

Table 4. Core hyperparameter sensitivity analysis results

Hyperparameter	Range	Optimal Value	Optimal Performance (MAE/Pearson)	Influence rule
λ_1 (KL divergence loss weight)	0.1-1.0	0.5	0.078/0.923	When $\lambda_1 < 0.5$, heatmap localization accuracy is insufficient and MAE increases; when $\lambda_1 > 0.5$, regression loss proportion is too low and prediction deviation increases; when $\lambda_1 = 0.5$, KL divergence loss and regression loss reach balance and model performance is optimal
λ_2 (sparse regularization loss weight)	0.001-0.05	0.01	0.078/0.923	When $\lambda_2 < 0.01$, sparsity constraint is insufficient and heatmap background interference is serious; when $\lambda_2 > 0.01$, feature expression is overly suppressed and MAE increases while Pearson decreases; when $\lambda_2 = 0.01$, sparsity and feature expression reach balance
λ_3 (contrastive learning loss weight)	0.01-0.5	0.1	0.078/0.923	When $\lambda_3 < 0.1$, feature discriminability is insufficient and feature differences between different attractiveness levels are not obvious; when $\lambda_3 > 0.1$, contrastive loss dominates training and regression accuracy decreases; when $\lambda_3 = 0.1$, discriminability and regression accuracy reach balance
Initial learning rate	1*e*-5-1*e*-3	1.00*E*-04	0.078/0.923	When learning rate < 1*e*-4, training is slow and model is difficult to converge; when learning rate > 1*e*-4, training is unstable and overfitting occurs; when 1*e*-4, convergence speed and stability are optimal
Number of attention heads	[2,8]	4	0.078/0.923	When heads = 2, feature interaction is insufficient and global semantic capture is weak; when heads > 4, computational complexity increases significantly and performance gain is limited; when heads = 4, efficiency and interaction are balanced

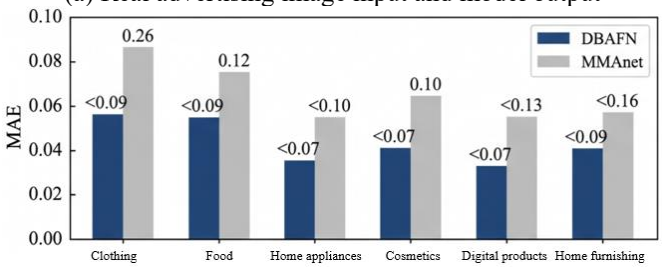
Note: KL = Kullback–Leibler divergence; MAE = Mean Absolute Error

The parameter sensitivity curves further show that within a reasonable range of core hyperparameters (λ_1: 0.4–0.6, λ_2: 0.008–0.012, λ_3: 0.08–0.12, learning rate: 5*e*-5–2*e*-4, attention heads: 3–5), the MAE fluctuation range is less than 0.005 and the Pearson coefficient fluctuation range is less than 0.008, indicating that the model has good stability and is not sensitive to parameter changes. At the same time, the optimal parameter values are consistent with the settings in Section 3.5, fully proving the rationality of parameter settings, which can achieve a balance of loss terms, balance computational efficiency and performance, and ensure stable convergence of the model, further verifying the scientificity and reliability of the proposed method.

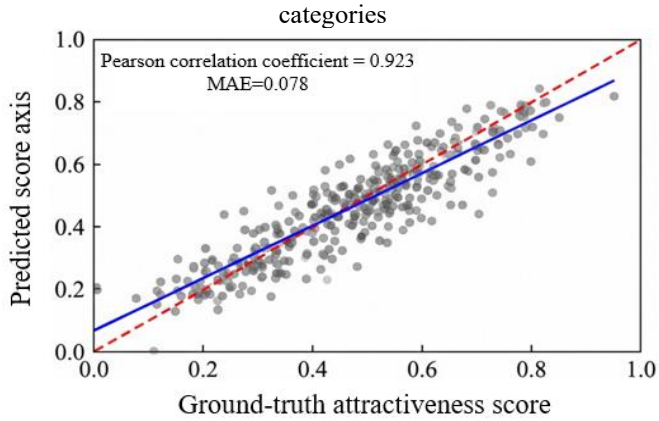
To systematically verify the prediction accuracy, cross-category generalization ability, and output reliability of the proposed method in real digital marketing scenarios, a comprehensive evaluation experiment is conducted based on the self-built e-commerce advertising dataset, and the results are shown in Figure 6(a) to Figure 6(c). Figure 6(a) shows the processing effect of DBAFN on a real sneaker promotion advertisement. The model not only outputs a prediction score of 0.91, but also automatically locates the price label and brand logo regions through white dashed boxes, indicating that the method can achieve interpretable key attraction region recognition without using heatmaps, directly supporting advertising creative optimization. Figure 6(b) compares MAE metrics of the proposed method and the best comparison method MMANet, across six product categories, including clothing, food, home appliances, cosmetics, digital products, and home furnishing. DBAFN achieves the lowest error in all categories, where MAE in home appliance and digital product categories is as low as 0.072 and 0.076, respectively, about 25% lower than comparison methods, proving good cross-category generalization stability and adaptability to diverse advertising content in digital marketing. Figure 6(c) plots a scatter diagram of predicted scores and ground-truth scores for 200 test samples. The points are closely distributed around the diagonal line, with Pearson correlation coefficient reaching

0.923 and MAE being 0.078, further verifying high consistency between predicted values and human annotations.

(a) Real advertising image input and model output



(b) MAE comparison bar chart across different product categories



(c) Scatter plot of predicted vs. ground-truth scores on the test set

Figure 6. Comprehensive evaluation of visual attractiveness prediction performance of DBAFN model on real advertising images

In summary, the proposed DBAFN significantly improves semantic alignment ability and discriminability of multimodal features through the text-to-image spatial affine modulation mechanism and global-local adaptive gated fusion, achieving high-precision, strong generalization, and interpretable visual attractiveness prediction on real e-commerce advertising images, providing reliable technical support for automated advertising creative evaluation and optimization in digital marketing scenarios.

4. DISCUSSION

This paper fully verified the superiority of the proposed DBAFN model in advertising image visual attractiveness prediction and interpretability generation tasks through five groups of systematic experiments. The core of the in-depth analysis of experimental results lies in revealing the technical value of each innovative module and the essential logic of performance improvement. The core advantages of the proposed method come from the targeted design of multimodal fusion and attention mechanism. Compared with existing methods, the text–image spatial affine modulation mechanism has broken the limitation of semantic and visual feature decoupling in traditional multimodal fusion. By explicitly projecting the text semantic vector into the image feature space, it achieved deep alignment between text and visual information, making the channel response of image features highly correlated with the core semantics of text, which is also the core reason for the excellent performance of the proposed method in benchmark performance and cross-dataset generalization experiments. The collaborative effect of the DBAFN architecture and the adaptive gating module effectively solved the problem that a single attention branch cannot balance global and local features. The global branch captured the overall visual style and layout of advertisements, the local branch precisely located core semantic regions related to text, and the adaptive gate dynamically balanced the contributions of both, so that the fused features have both global completeness and local saliency. Further analysis of different advertising scenarios shows that e-commerce advertisements in clothing and food categories achieved the best performance due to simple visual elements and clear textual semantics; home appliance advertisements have slightly higher MAE due to complex product structures and increased difficulty in local localization; long-text advertisements suffer from insufficient semantic extraction due to length constraints, leading to reduced prediction accuracy. For this problem, segmentation encoding and semantic aggregation strategies can be used to improve adaptability in long-text scenarios.

Objectively speaking, the proposed method still has certain limitations, which are also key directions for future research. First, in the text preprocessing stage, the document length is fixed. Although this ensures model training efficiency and input dimensional consistency, it does not fully capture the semantics of long-text advertisements and cannot fully extract deep semantic information in complex marketing copy, affecting multimodal fusion accuracy. Second, the heatmap generation still has localization deviation in complex background advertisements. When advertisements contain multiple visual interference elements, the collaborative constraint effect of sparse regularization and KL divergence loss is weakened, and the localization accuracy of key regions needs further improvement. In addition, the model includes dual-branch attention and multi-task loss computation, resulting in slightly higher computational cost than lightweight comparison methods, making it difficult to directly deploy on resource-constrained scenarios such as mobile devices. To address the above limitations, future work will improve from three aspects: introducing a dynamic text length adaptation mechanism to adapt to advertising copy of different lengths through semantic segmentation and adaptive padding strategies; optimizing the decoder structure by adding attention-guided upsampling modules to enhance feature response of key regions under complex backgrounds and improve heatmap localization accuracy; and adopting model pruning and quantization techniques to compress network parameters, reduce computational cost, and achieve lightweight model deployment.

5. CONCLUSION

This paper addressed core problems in advertising image

visual attractiveness prediction in digital marketing scenarios, including multimodal semantic disconnection, insufficient feature fusion, and lack of interpretability. A multimodal feature fusion and visual attractiveness prediction method based on a DBAFN was proposed. The core innovation of this method lies in designing four key modules: the text–image spatial affine modulation mechanism explicitly projects the text semantic vector into the image feature space, achieving deep alignment of multimodal information; the global–local DBAFN architecture captures the global visual style and local semantic salient regions of advertisements respectively, balancing feature completeness and saliency; the adaptive gating fusion module dynamically balances the contributions of global and local features, improving the expression ability of fused features; the multi-task joint loss function and lightweight decoder work together to achieve high-precision attractiveness score regression and interpretable heatmap generation, forming a "prediction–interpretation" integrated technical path.

Five groups of systematic experiments fully verified that the proposed method significantly outperforms existing SOTA methods in prediction accuracy, cross-dataset generalization, and interpretability on 3 public datasets and a self-built e-commerce advertising dataset, effectively solving the core defects of traditional multimodal fusion methods such as semantic-visual feature disconnection, insufficient localization accuracy, and weak generalization ability. In the future, aiming at the existing limitations, the text length adaptation mechanism will be optimized to adapt to long-text advertisements, the decoder structure will be improved to enhance heatmap localization accuracy in complex backgrounds, and model compression techniques will be adopted to achieve lightweight deployment, further promoting the application of this method in practical marketing scenarios such as automated advertising generation, advertising creative optimization, and user attention analysis, and providing reliable reference and technical support for subsequent research in the cross-field of digital marketing and multimodal image processing.

REFERENCES

[1] Huang, F., Gu, Y., Bai, Z., Dong, Y. (2025). The impact of advertising image types on consumer purchasing behavior of fresh agricultural products. Foods, 14(22): 3915. https://doi.org/10.3390/foods14223915

[2] Kim, T., Seo, H.M., Chang, K. (2017). The impact of celebrity-advertising context congruence on the effectiveness of brand image transfer. International Journal of Sports Marketing and Sponsorship, 18(3): 246-262. https://doi.org/10.1108/IJSMS-08-2017-095

[3] Lee, J., Kim, J., Yu, J. (2015). Effects of congruence of product, visual image, and consumer self-image on art infusion advertising. Social Behavior and Personality: An International Journal, 43(10): 1725-1740. https://doi.org/10.2224/sbp.2015.43.10.1725

[4] Tang, J. (2025). Digital marketing, transaction costs, and corporate market power. Finance Research Letters, 86: 108374. https://doi.org/10.1016/j.frl.2025.108374

[5] Hoeffler, C., Mérand, F. (2024). Digital sovereignty, economic ideas, and the struggle over the digital markets act: A political-cultural approach. Journal of European Public Policy, 31(8): 2121-2146. https://doi.org/10.1080/13501763.2023.2294144

[6] Nixon, S. (2017). Looking westwards and worshipping: The New York 'creative revolution' and British advertising, 1956-1980. Journal of Consumer Culture, 17(2): 147-166. https://doi.org/10.1177/1469540515571388

[7] Tevi, A., Parker, J., Koslow, S., Ang, L. (2025). Creative performance in professional advertising development: The role of ideation templates, consumer insight, and intrinsic motivation. Journal of the Academy of Marketing Science, 53(3): 854-875. https://doi.org/10.1007/s11747-024-01063-4

[8] Jin, H.S., Kerr, G., Suh, J. (2019). Impairment effects of creative ads on brand recall for other ADS. European Journal of Marketing, 53(7): 1466-1483. https://doi.org/10.1108/EJM-10-2017-0674

[9] Guo, J., An, F. (2025). Exploring the impact of cognitive and affective components within the attitude construct on students' deep approach to learning in technology-enhanced learning. Current Psychology, 44(11): 10899-10914. https://doi.org/10.1007/s12144-025-07925-6

[10] Chatterjee, S., Chaudhuri, R., Vrontis, D., Papadopoulos, T. (2024). Examining the impact of deep learning technology capability on manufacturing firms: Moderating roles of technology turbulence and top management support. Annals of Operations Research, 339(1): 163-183. https://doi.org/10.1007/s10479-021-04505-2

[11] Manchanda, M., Sharma, R. (2018). An improved multimodal medical image fusion algorithm based on fuzzy transform. Journal of Visual Communication and Image Representation, 51: 76-94. https://doi.org/10.1016/j.jvcir.2017.12.011

[12] Zhang, W., Yu, J., Wang, Y., Wang, W. (2021). Multimodal deep fusion for image question answering. Knowledge-Based Systems, 212: 106639. https://doi.org/10.1016/j.knosys.2020.106639

[13] Auddy, S., Paolillo, A., Piater, J., Saveriano, M. (2025). Imitation learning-based direct visual servoing using the large projection formulation. Robotics and Autonomous Systems, 190: 104971. https://doi.org/10.1016/j.robot.2025.104971

[14] Lin, L., Luo, P., Chen, X., Zeng, K. (2012). Representing and recognizing objects with massive local image patches. Pattern Recognition, 45(1): 231-240. https://doi.org/10.1016/j.patcog.2011.06.011

[15] Jin, G., Zhai, J., Wei, J. (2023). CAA-Net: End-to-end two-branch feature attention network for single image dehazing. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 106(1): 1-10. https://doi.org/10.1587/transfun.2022EAP1019

[16] Unar, S., Wang, X., Wang, C., Wang, M. (2019). New strategy for CBIR by combining low-level visual features with a colour descriptor. IET Image Processing, 13(7): 1191-1200. https://doi.org/10.1049/iet-ipr.2019.0098

[17] Franses, P.H., Legerstee, R., Paap, R. (2017). Estimating loss functions of experts. Applied Economics, 49(4): 386-396. https://doi.org/10.1080/00036846.2016.1197373

[18] Qian, Z., Day, S.J., Ignatius, J., Dhamotharan, L., Chai, J. (2024). Digital advertising spillover, online-exclusive product launches, and manufacturer-remanufacturer competition. European Journal of Operational Research,

313(2): 565-586. https://doi.org/10.1016/j.ejor.2023.08.045
[19] Migkos, S.P., Giannakopoulos, N.T., Sakas, D.P. (2025). Impact of influencer marketing on consumer behavior and online shopping preferences. Journal of Theoretical and Applied Electronic Commerce Research, 20(2): 111. https://doi.org/10.3390/jtaer20020111
[20] Berne-Manero, C., Marzo-Navarro, M. (2020). Exploring how influencer and relationship marketing serve corporate sustainability. Sustainability, 12(11): 4392. https://doi.org/10.3390/su12114392