JEI Journal of Electronic Imaging

RESEARCH PAPER

Beyond pixels: text-guided deep insights into graphic design image aesthetics

Guangyu Shi,[a] Luming Li,[b] and Mengke Song[b,*]

[a]Zhengzhou College of Finance and Economics, School of Art and Design, Zhengzhou, China

[b]China University of Petroleum (East China), School of Computer Science and Technology, Qingdao, China

ABSTRACT. The rapid development of computer vision and deep learning has significantly advanced image aesthetic assessment, yet traditional methods, which primarily rely on low-level visual features such as color and texture, often struggle with the complexity of graphic design images. These images are characterized by diverse design elements, including color, typography, and layout, as well as various styles such as minimalism, retro, and modernism, presenting substantial challenges to conventional assessment techniques. To overcome these limitations, we propose an innovative multimodal learning approach that integrates image content with textual descriptions to comprehensively analyze the aesthetic qualities of graphic design images. The core innovation of our method lies in the utilization of two distinct textual description methodologies: holistic descriptions, which capture the main theme of the design, and detailed descriptions, which focus on specific aspects such as composition, color, detail, and atmosphere. This dual approach allows for a more nuanced and complete assessment of aesthetic value. To effectively merge these descriptions with visual content, we introduce a feature similarity blending mechanism that aligns and integrates features from both modalities, enhancing the representation of aesthetic attributes. In addition, we employ a score bagging technique to aggregate scores from multiple fused features, ensuring robustness and reliability in the assessments. Our method is implemented within a multi-task learning framework, enabling simultaneous prediction across multiple rating dimensions. Experimental results demonstrate that, compared with the state-of-the-art TAHF method, our approach achieves notable improvements in Spearman's rank correlation coefficient—by 1.7%, 3.4%, and 2.6% on the HDDI, BAID, and TAD66K datasets, respectively—along with consistent gains in Pearson's linear correlation coefficient and accuracy. Moreover, our method achieves these performance improvements with fewer parameters and lower computational complexity, highlighting its efficiency and effectiveness in graphic design image aesthetic assessment.

 [DOI: 10.1117/1.JEI.33.5.053059]

Keywords: graphic design images; aesthetic assessment; multimodal learning; deep learning

Paper 240617G received Jun. 10, 2024; revised Aug. 22, 2024; accepted Oct. 7, 2024; published Oct. 30, 2024.

1 Introduction

The image aesthetic assessment (IAA) has long been a focal point in computer vision and deep learning research, driven by the need to understand and evaluate the visual appeal of images across various domains. For IAA, it can be classified into three types: natural IAA, artistic image aesthetic assessment (AIAA) and graphic design IAA (GDIAA).

*Address all correspondence to Mengke Song, bz23070005@s.upc.edu.cn

1017-9909/2024/$28.00

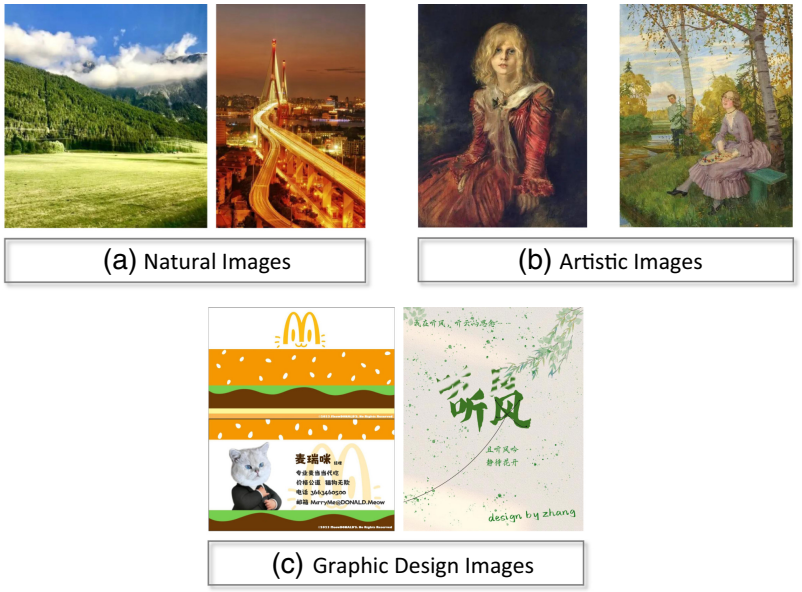

Fig. 1 Comparison demonstration of natural images (a), artistic images (b), and graphic design images (c).

Natural images typically capture scenes from the real world, focusing on aspects such as composition, lighting, and realism. Traditional natural IAA (NIAA) methods, such as those proposed by Refs. 1 and 2, have predominantly relied on low-level visual features such as color, texture, and edge information. Although these methods have proven effective for evaluating natural images [Fig. 1(a)], they often fall short when applied to more complex image types, such as artistic and graphic design images [Figs. 1(b) and 1(c)]. Artistic images [Figs. 1(b)] are created with an emphasis on creativity, emotion, and expression, often following certain artistic styles and principles. Existing methods for artistic IAA (AIAA)[3,4] have achieved reasonable success by effectively capturing essential aspects of artistic images, such as balance, contrast, and harmony. These methods have been specifically designed to understand and evaluate the creative and expressive elements that define artistic works, reflecting their unique aesthetic principles. Graphic design images [Fig. 1(c)], however, present a different set of challenges. These images incorporate a variety of design elements, including color schemes, typography, and layout, and are characterized by distinct stylistic approaches such as minimalism, retro, and modernism. Unlike artistic images, graphic design images are not only visually complex but also semantically rich, often created with specific functional purposes in mind, such as communication, branding, and user engagement.

The primary limitation of existing aesthetic assessment methods is their inability to capture the nuanced interplay of these diverse elements in graphic design images. In addition, graphic design images often follow strict guidelines and standards that are not typically present in artistic images, making the evaluation criteria different and more complex. To effectively assess the aesthetics of graphic design images, it is essential to develop methods that integrate both visual and semantic features. By combining these sources of information, we can form a more comprehensive understanding of what makes graphic design appealing, accommodating its inherent complexity and diversity.

In the realm of GDIAA, the dataset HDDI[5] is frequently employed as the standard. Despite being annotated by a diverse group of individuals with different social backgrounds and cognitive levels, the dataset's limitation lies in its approach to assigning a singular aesthetic score to the images based on their themes. This constraint hinders the dataset's capacity for nuanced analysis of graphic design images, potentially impeding the model's ability to determine an accurate aesthetic rating. Intuitively, a direct method to address this issue would involve re-annotating the dataset with a more granular focus, such as visual expressiveness, user engagement, and emotional resonance. However, this process is not only laborious but also costly, making it an impractical solution. The current challenge is how to obtain detailed content analysis of graphic design

images without engaging in the time-consuming and labor-intensive process of fine-grained data annotation.

With the rise of Transformers in the field of vision,[6] the gap between natural language and images has gradually been bridged. Compared with images, the advantages of natural language lie in its ability to capture abstract concepts and complex semantics more effectively. Natural language can convey emotions, intentions, and contextual information, which are often challenging to communicate through visual features alone. Moreover, textual descriptions can detail the choices and interactions of design elements, providing a comprehensive understanding and explanation of the visual effects.

Thus, inspired by this, we propose leveraging textual descriptions to represent the detailed content analysis of graphic design images. This approach not only mitigates the need for extensive re-annotation but also enriches the feature space, allowing for a more comprehensive evaluation of the aesthetic qualities. Textual descriptions offer several unique advantages for this purpose: (1) Textual descriptions can convey complex and abstract concepts that are difficult to quantify visually. For instance, they can describe the emotional tone, thematic elements, and stylistic nuances that contribute to the overall aesthetic appeal of a design. (2) Although visual features can capture color, texture, and basic shapes, textual descriptions can provide a more granular analysis by detailing specific design elements such as typography choices, spatial relationships, and compositional techniques. This level of detail allows for a more comprehensive evaluation of the design's aesthetics.

The core innovation of our method lies in the utilization of two distinct textual description methodologies (Sec. 3.3). Holistic descriptions capture the main theme and overall aesthetic impression of the design, providing a broad context for evaluation. In contrast, detailed descriptions focus on specific aspects such as composition, color, detail, and atmosphere, enabling a granular analysis of the design elements. By incorporating both holistic and detailed descriptions, our method achieves a more comprehensive assessment of the image aesthetics. To effectively merge these descriptions with image content, we introduce the feature similarity blending (FSB) mechanism (Sec. 3.4). This mechanism aligns features from both modalities, enhancing the representation of aesthetic qualities. The blended features are then used to predict aesthetic scores across multiple dimensions, reflecting various aspects of visual appeal. To ensure robust and reliable assessments, we employ a score bagging (SB) technique (Sec. 3.5) that aggregates the scores derived from multiple fused features, mitigating the impact of potential biases and inconsistencies. Our method is implemented within a multi-task learning framework, enabling the simultaneous prediction of multiple rating dimensions. This approach enhances the accuracy of individual assessments, offers a comprehensive view of the aesthetic qualities of graphic design images, and also excels in traditional natural IAAs. In summary, our contributions can be summarized as follows:

- We introduce a multimodal learning approach that integrates image content with textual descriptions to comprehensively analyze the aesthetic qualities of graphic design images.
- We employ two distinct textual description methodologies—holistic and detailed—to capture both the overall theme and specific design elements, achieving a more thorough assessment.
- We develop an FSB mechanism that aligns features from both visual and textual modalities, enhancing the representation of aesthetic attributes.
- Experimental results suggest that our method achieves state-of-the-art performance on both graphic design and natural image benchmark datasets, which demonstrates its effectiveness.

2 Related Work

2.1 Natural Image Aesthetic Assessment

IAA is a multifaceted tool with broad applications across various domains such as search engines, content ranking, and recommendation systems. It stands in contrast to technical quality assessment, which is concerned with aspects such as image distortion, improper cropping, or the presence of noise.[7,8] Instead, IAA is dedicated to gauging the intrinsic aesthetic appeal of images,

capturing the subjective and often intangible qualities that resonate with viewers and influence their perception of beauty and artistic value. Early approaches[1,2] primarily relied on low-level visual features such as color, texture, and edge information to predict aesthetic quality. In the era of deep learning, studies such as Refs. 9–14 have concentrated on data-driven approaches and amassed extensive datasets that comprise images alongside human-assigned ratings. Leveraging these datasets, Ref. 15 developed a ranking-based model, and Refs. 16 and 17 aimed to estimate the true distributions of aesthetic scores.

Several other methods, Refs. 18–20 leveraged a combination of local and global image features to enhance the predictive accuracy of aesthetic judgments. For instance, the RAPID[18] model integrated diverse inputs, including both global and local perspectives, to classify images based on their aesthetic level. Further advancements have been made with models such as A-Lamp,[19] which introduced an adaptive selection mechanism for multi-patches and incorporates layout-aware attribute graphs to refine the aesthetic assessment process. MPada,[20] on the other hand, employed an attention-based mechanism that dynamically adjusts the importance of each patch during training, thereby enhancing the efficiency of the learning process. Furthermore, VILA[21] proposed to learn image aesthetics from user comments, exploring vision-language pretraining methods to learn multimodal aesthetic representations. Li et al.[22] proposed theme-aware visual attribute reasoning by simulating the process of human perception in image aesthetics by performing bilevel reasoning.

Despite these advancements, traditional methods for natural IAA often fall short when applied to more complex and diverse graphic design images, necessitating the exploration of more sophisticated approaches.

2.2 Artistic Image Aesthetic Assessment

AIAA poses unique challenges due to the subjective nature of art and the diverse range of styles and mediums. However, in the realm of AIAA, there is a limited body of work. Earlier studies, such as Refs. 23–26, relied on manually crafted features and utilized support vector machines for classification purposes. More recently, researchers tend to utilize deep learning–based methods.[3,4,27–29] Among them, Yi et al.[30] proposed to extract and utilize style-specific and generic aesthetic information to evaluate artistic images. Shi et al.[31] presented a novel approach called semantic and style-based multiple reference learning for artistic and general IAA, which leverages semantic and stylistic features of images through multiple reference learning and graph reasoning to improve the prediction accuracy of artistic and general image aesthetics.

Although significant progress has been made in both natural and artistic IAAs, the complexity and diversity of graphic design images require innovative methods that integrate multiple sources of information to provide a comprehensive evaluation. Our proposed method leverages these advancements by combining visual content with detailed textual descriptions, offering a robust framework for assessing the aesthetic qualities of graphic design images.

3 Proposed Method

3.1 Preliminary

In the domain of aesthetic assessment for graphic design images, the integration of textual descriptions is essential for capturing the multifaceted nature of visual appeal. Traditional methods[25,32] that rely solely on visual features often fail to encapsulate the abstract concepts and complex semantics inherent in graphic design. To address this limitation, our approach leverages textual descriptions that provide rich semantic information complementary to visual features. We utilize textual descriptions from the perspectives of "main theme, composition, color, detail, and atmosphere." These aspects were chosen because they collectively provide a comprehensive and nuanced understanding of visual aesthetics, addressing the holistic and granular elements of design.

3.1.1 *Main theme*

The main theme of a graphic design image encapsulates the overall aesthetic impression and the primary message intended by the designer. It provides a broad context and a general feel of the

design, which is crucial for understanding its aesthetic direction. By capturing the main theme, we can comprehend the overarching concept that ties all design elements together, ensuring a holistic evaluation of the image's aesthetic quality.

3.1.2 *Composition*

Composition refers to the arrangement of visual elements within the image, dictating their spatial relationships and balance. According to the Gestalt principles of perception,[33] the composition significantly influences how viewers perceive and interpret the image. A well-composed image guides the viewer's eye smoothly across the design, enhancing its aesthetic appeal. Therefore, assessing the composition is vital for understanding the structural integrity and visual flow of the design.

3.1.3 *Color*

Color schemes play a pivotal role in evoking emotions and setting the mood of an image. Color theory[34] highlights the psychological impacts of different colors and their combinations. Colors can influence viewers' emotional responses and are fundamental to the image's overall harmony and appeal. By analyzing the color aspects, we can evaluate how effectively the color palette contributes to the aesthetic quality of the design.

3.1.4 *Detail*

Details such as textures, patterns, and intricate elements add richness and depth to a design. Theories such as the processing fluency theory[35] suggest that an optimal level of detail and complexity enhances the aesthetic experience by making the image more engaging and interesting. Evaluating the details helps in understanding the intricacy and craftsmanship involved in the design, which are critical components of its aesthetic value.

3.1.5 *Atmosphere*

The atmosphere or mood conveyed by an image plays a significant role in its aesthetic perception. Cognitive psychology research, including the affect-infusion model,[36] indicates that the emotional tone of an image can influence viewers' judgments and overall experience. By assessing the atmosphere, we can evaluate how effectively the design communicates its intended mood and emotional impact, which is essential for a comprehensive aesthetic assessment.

The integration of textual descriptions from the perspectives of the main theme, composition, color, detail, and atmosphere can provide a holistic and detailed analysis of graphic design images.

3.2 Method Overview

The key insight of our method is to leverage the power of multimodal learning by integrating visual features with textual descriptions to achieve a more nuanced and comprehensive aesthetic assessment of graphic design images. Figure 2 illustrates the method pipeline of our proposed multimodal learning approach. Our approach comprises three main components: (1) textual description generation (TDG, Sec. 3.3), (2) feature similarity blending (FSB, Sec. 3.4), and (3) score bagging (SB, Sec. 3.5).

First, TDG generates holistic and detailed textual descriptions to provide rich semantic information complementing visual features. Then, FSB aims to align and combine features from both visual and textual modalities to enhance the representation of aesthetic attributes. Finally, SB aggregates aesthetic scores from multiple fused features to ensure robust and reliable assessments across multiple dimensions.

We employ two distinct methodologies to generate textual descriptions: holistic and detailed descriptions. Holistic descriptions capture the main theme and overall aesthetic impression of the design, providing a broad context for evaluation. In contrast, detailed descriptions focus on specific aspects such as composition, color, detail, and atmosphere, enabling a granular analysis

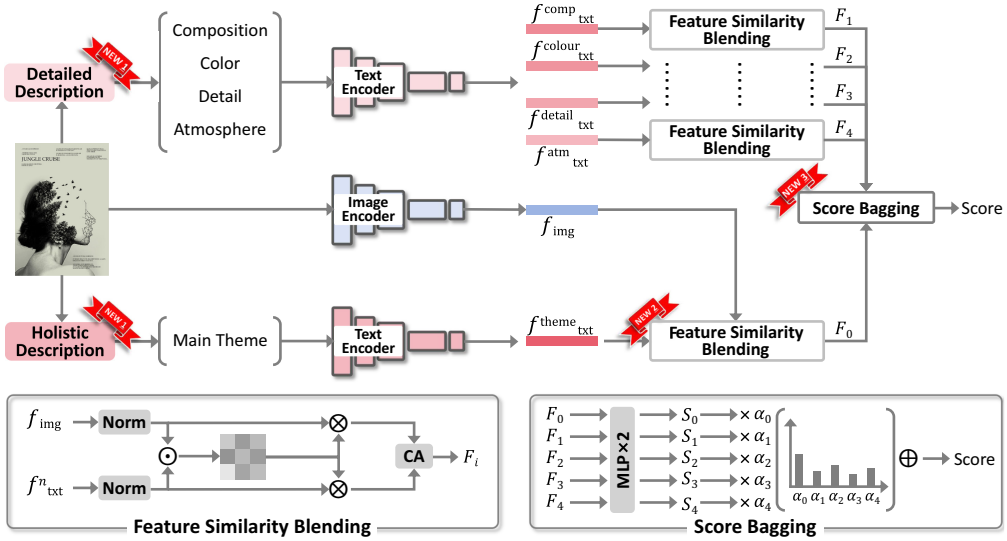

Fig. 2 Framework of the proposed method, which includes three main components. First, textural description generation (Sec. 3.3) aims to generate holistic and detailed textual descriptions to provide rich semantic information complementing visual features. Second, FSB (Sec. 3.4) aims to align and combine features from both visual and textual modalities to enhance the representation of aesthetic attributes. Third, SB (Sec. 3.5) aggregates aesthetic scores from multiple fused features to ensure robust and reliable assessments across multiple dimensions.

of the design elements. These textual descriptions offer a rich semantic layer that complements the visual features of the images.

To effectively combine visual and textual information, we introduce the FSB mechanism. This mechanism aligns features extracted from both modalities, enhancing the representation of aesthetic attributes. By blending features from the image content with those from the textual descriptions, we create a unified feature space that encapsulates both visual and semantic characteristics of the graphic design images.

For robust and reliable aesthetic assessments, we employ an SG technique. This technique aggregates aesthetic scores derived from multiple fused features, mitigating potential biases and inconsistencies. Our method is implemented within a multi-task learning framework, enabling the simultaneous prediction of multiple rating dimensions. This comprehensive approach enhances the accuracy of individual assessments and provides a holistic evaluation of the aesthetic qualities.

3.3 Textural Description Generation

Traditional methods for aesthetic assessment of graphic design images primarily rely on visual features such as color, texture, and layout. Although these methods have proven effective for natural and artistic images, they often fail to capture the complexity and diversity of graphic design images. These images not only incorporate various design elements but also embody distinct stylistic approaches that convey different themes and emotional tones. To address this limitation, our approach leverages textual descriptions, which can capture abstract concepts, complex semantics, and detailed design elements that are challenging to quantify visually. By integrating both holistic and detailed textual descriptions, we aim to provide a more comprehensive analysis of graphic design images' aesthetic qualities.

To implement this, we propose the TDG methodology, which is designed to create rich semantic representations of graphic design images by generating both holistic and detailed descriptions. These descriptions are generated by existing pretrained language-image models (e.g., BLIP,[37] ViLT,[38] and GPT-3).[39] The process is illustrated in Fig. 2-NEW1 and consists of two steps, taking the graphic design image shown in Fig. 2 as an example.

First, holistic descriptions capture the main theme and overall aesthetic impression of the design. They provide a broad context that encapsulates the general feel and intention behind the

design. To achieve this, we input general prompts into existing pretrained language-image models such as GPT-3. For instance, by inputting the prompt "Describe the main theme and aesthetic impression of the image," we can obtain a textual description such as

> "This image is a poster for a movie or literary work, which combines natural landscapes with artistic rendering of a human portrait, creating a mysterious, dark, and slightly eerie atmosphere."

Second, detailed descriptions focus on specific design elements such as composition, color, detail, and atmosphere. They offer a granular analysis of the visual components that contribute to the aesthetic quality of the image. To achieve this, we input targeted prompts into existing pretrained language-image models such as ViLT. For example, by inputting prompts such as "Describe the composition of the image," "Describe the color scheme of the image," "Detail the visual elements present in the image," and "Describe the atmosphere conveyed by the image," we can obtain textual descriptions such as:

1. "The composition of the image features a silhouette of a human profile overlapped with trees and birds. The title is centered, with text distributed on both sides, achieving a balanced composition;"
2. "The color of the image features a silhouette of a human profile overlapped with trees and birds. The title is centered, with text distributed on both sides, achieving a balanced composition;"
3. "The detail of the image features that the textures of the trees and birds are clear and geometric lines outline the human face, enhancing the sense of technology;"
4. "The atmosphere of the image features that the black-and-white tone and the double exposure effect create a mysterious and tense atmosphere, with text adding a narrative element."

By leveraging these pretrained models with specific prompts, we generate comprehensive textual descriptions that encompass both the holistic and detailed aspects of graphic design images, enriching the semantic representation.

3.4 Feature Similarity Blending

In the context of aesthetic assessment for graphic design images, the integration of visual and textual features is crucial to capturing the full spectrum of aesthetic attributes. Although visual features provide information about the color, texture, and layout of an image, textual descriptions can convey abstract concepts, thematic elements, and stylistic nuances that are challenging to quantify visually. To effectively combine these two modalities, we introduce the FSB mechanism. This mechanism aligns and blends features from both visual and textual inputs, creating a unified representation that enhances the aesthetic evaluation process.

The FSB is designed to align and merge features from visual and textual modalities, ensuring a comprehensive representation of the aesthetic attributes of graphic design images. The process, as illustrated in Fig. 2-NEW2, involves several key steps:

First, we extract features from both the image and textual descriptions using respective encoders. For each type of description, we employ a text encoder (here, we use the text encoder of CLIP[40]) to transform the textual information into feature vectors. Let $\mathbf{T}_h$ and $\mathbf{T}_d$ represent the holistic and detailed descriptions, respectively. The encoded feature vectors are obtained as follows:

For holistic descriptions

$$f_{\text{txt}}^{\text{theme}} = \text{TxtEn}(\mathbf{T}_h), \tag{1}$$

where $\text{TxtEn}(\cdot)$ is the text encoder and $f_{\text{txt}}^{\text{theme}}$ is the textural feature of the holistic descriptions.

For detailed descriptions

$$\begin{aligned} f_{\text{txt}}^{\text{comp}} &= \text{TxtEn}(\mathbf{T}_d^{\text{comp}}), \\ f_{\text{txt}}^{\text{color}} &= \text{TxtEn}(\mathbf{T}_d^{\text{color}}), \\ f_{\text{txt}}^{\text{detail}} &= \text{TxtEn}(\mathbf{T}_d^{\text{detail}}), \\ f_{\text{txt}}^{\text{atm}} &= \text{TxtEn}(\mathbf{T}_d^{\text{atm}}), \end{aligned} \tag{2}$$

where $f_{\text{txt}}^{\text{comp}}$, $f_{\text{txt}}^{\text{color}}$, $f_{\text{txt}}^{\text{detail}}$, and $f_{\text{txt}}^{\text{atm}}$ are the textural features of the detailed descriptions for composition, color, detail, and atmosphere, respectively. $\mathbf{T}_d^{\text{comp}}$, $\mathbf{T}_d^{\text{color}}$, $\mathbf{T}_d^{\text{detail}}$, and $\mathbf{T}_d^{\text{atm}}$ represent the detailed descriptions for composition, color, detail, and atmosphere, respectively.

Let $\mathbf{I}$ represent the image and f_{img} denote the visual features extracted using an image encoder

$$f_{\text{img}} = \text{ImgEn}(\mathbf{I}), \tag{3}$$

where $\text{TxtEn}(\cdot)$ is the image encoder. Here, we use the image encoder of CLIP.

The encoded textual features are integrated with the visual features extracted from the image. To ensure compatibility and effective blending, both visual and textual feature vectors are normalized (Norm). Normalized features from both modalities are then combined using the Hadamard product ($\odot$) to obtain a similarity matrix (Sim)

$$\text{Sim} = \text{Norm}(f_{\text{img}}) \odot \text{Norm}(f_{\text{txt}}^n), \tag{4}$$

where f_{txt}^n denotes the textural features of the holistic and detailed descriptions. The term is $n \in \{\text{theme}, \text{comp}, \text{color}, \text{detail}, \text{atm}\}$.

Then, the similarity matrix is multiplied with normalized features from both modalities using the tensor product ($\otimes$). These operations enhance the interaction between the features, capturing the synergy between visual and textual attributes

$$F_i = \text{CA}(\text{Sim} \otimes \text{Norm}(f_{\text{img}}), \text{Sim} \otimes \text{Norm}(f_{\text{txt}}^n)), \tag{5}$$

where F_i ($i \in \{0, 1, 2, 3, 4\}$) means the blended feature (F_0 for holistic descriptions, F_1, F_2, F_3, and F_4 for detailed descriptions) applied to each set of textual features and their corresponding visual features. This results in multiple fused feature vectors, each representing different aspects of the graphic design image's aesthetic attributes. CA is the cross-attention operation, which is used to enhance the interaction between the visual and textual features. This mechanism allows the model to focus on relevant parts of the image when considering textual information and vice versa, thereby improving the quality of the fused features.

3.5 Score Bagging

In the context of aesthetic assessment for graphic design images, ensuring robust and reliable evaluations is crucial. The aesthetic qualities of an image can be multi-faceted, incorporating various aspects such as composition, color, detail, and atmosphere. To achieve a comprehensive and accurate assessment, it is essential to aggregate the scores derived from multiple fused features effectively. The SG mechanism is introduced to address this need, as illustrated in Fig. 2-NEW3. This mechanism aggregates aesthetic scores from different blended feature vectors, reducing the impact of potential biases and inconsistencies, and thus ensuring a more stable and reliable assessment.

For each blended feature vector F_i (where $i = 0, 1, 2, 3, 4$), a score S_i is predicted using two multi-layer perceptrons (MLPs). The MLPs take the fused feature vector as input and output a predicted aesthetic score. This step ensures that each aspect of the image (captured by different textual descriptions) contributes to the final aesthetic score. Each score S_i is assigned a weight α_i based on the importance of the corresponding textual description in the context of aesthetic evaluation. These weights are not learned but predefined to reflect the relative significance of different aspects of the graphic design image, according to the Gestalt principles of perception[33] and color theory.[34] The rationale behind the specific values of α_i is grounded in the understanding of how different design elements contribute to the overall aesthetic quality. Thus, we set the weights of these textual descriptions as $\alpha_0 = 0.4$ (for holistic descriptions), $\alpha_1 = 0.2$, $\alpha_2 = 0.1$, $\alpha_3 = 0.2$, and $\alpha_4 = 0.1$ (for detailed descriptions), to ensure that the final aesthetic score is a balanced and comprehensive reflection of the various contributing factors. The weighted scores are then aggregated to produce a final aesthetic score. This aggregation process ensures that the final score reflects a balanced combination of all individual assessments, reducing the impact of any single biased or inconsistent prediction. The final aesthetic score (score) is computed as

$$S_i = \text{MLP}(F_i), \tag{6}$$

$$\text{score} = \sum_{i=0}^{4} S_i \times \alpha_i. \tag{7}$$

To ensure that the final score is within a reasonable and interpretable range, a normalization step is applied. This step adjusts the final score based on predefined criteria or learned parameters, ensuring consistency across different images and assessments

$$\text{Score} = \text{Norm}(\text{score}). \tag{8}$$

3.6 Loss Function

During the training phase, we employ L1 loss to optimize the parameters of the whole model, which is defined as

$$\mathcal{L}_1 = \frac{1}{N_z} \sum_{i=1}^{N_z} |y_i - \hat{y}_i|, \tag{9}$$

where y_i denotes the ground truth (GT) aesthetic score of the graphic design image x_i ($i = 1, 2, \ldots, N_z$) and N_z represents the number of samples in the training dataset. $\hat{y}_i$ is the predicted aesthetic scores. This loss function measures the absolute differences between the predicted aesthetic scores $\hat{y}_i$ and the GT scores y_i, aiming to minimize these differences during the training process.

4 Experiments

4.1 Implementation Details

We implement our approach in Python with the Pytorch toolbox on an NVIDIA GeForce RTX 3090 (with 24G RAM). We optimize the network via stochastic gradient descent (SGD) with a momentum of 0.9 and weight decay of 10^{-4}. The learning rate is set to 0.001 and exponentially decayed by 0.1 after each five epoch. During training, we resize the resolution to 384 and random cropping to 256, adding random image flipping for data augmentation.

4.2 Datasets

In this work, the experiments are primarily carried out on the available graphic design image dataset HDDI. The HDDI dataset[5] is the first IAA dataset entirely composed of human-designed digital images. It contains 140 images created using modern design tools such as Photoshop and is categorized into four themes: font design, card design, logo design, and poster design. It scales the number of votes into the [0.5, 5.5] score range, where 0.5 means the worst and 5.5 represents the best. We use k-fold cross-validation to train our model. The dataset is randomly divided into k disjoint subsets of the same size, where the k-1 subset is used as the training set to train the model, and the remaining subset is used as the test set to test the model, and we set $k = 5$ here.

Meanwhile, we also conduct experiments on artistic IAA datasets BAID and TAD66K. BAID[30] consists of 60,337 artistic images covering various art forms, with more than 360,000 votes from online users. It scales the number of votes into the [0, 10] score range, where 0 means the worst and 10 represents the best. Following the common practice of Yi et al.,[30] we split the 60,337 images in BAID into 53,937:6400 for training and 6400 for testing. TAD66K[10] includes 1431 artistic images. We keep the same split with the original dataset for selecting 289 images as test and the remaining 1142 images for training.

4.3 Evaluation Metrics

To evaluate score regression performance, following Ref. 30, we use two popular metrics: Spearman's rank correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC). SRCC measures the monotonic relationship between the ground truth and predicted scores, ranging from -1 to 1. PLCC measures the linear correlation between the ground truth and predicted scores, also ranging from -1 to 1. In addition, we convert both predicted and ground-truth scores into binary class labels using a threshold of 3 (midpoint of the 0.5 to

1 Natural Image Aesthetic Assessment (IAA) 2 Artistic IAA 3 Graphic Design IAA

	Dataset	HDDI			BAID			TAD66K			Parameters	FLOPs
	Method	SRCC↑	PLCC↑	ACC↑	SRCC↑	PLCC↑	ACC↑	SRCC↑	PLCC↑	ACC↑	(M)	(G)
1	$NIMA_{18}$	0.287	0.281	81.78	0.393	0.382	71.01	0.383	0.408	60.90	23.51	4.14
1	$MLSP_{19}$	0.312	0.295	84.65	0.441	0.430	74.92	0.418	0.422	63.58	73.97	32.02
1	$TANet_{22}$	0.326	0.306	85.70	0.453	0.437	75.45	0.349	0.357	45.32	13.88	**2.01**
1	EAT_{23}	0.398	0.351	86.27	0.486	0.495	77.23	-	-	-	**87.65**	140
1	$KZIAA_{24}$	0.405	0.352	84.67	0.523	0.567	78.24	0.471	0.481	66.56	-	-
1	$HyperEmo_{24}$	0.419	0.362	85.24	0.531	0.573	78.98	0.485	0.492	67.34	-	-
2	$SAAN_{23}$	0.385	0.343	82.11	0.473	0.467	76.80	0.425	0.440	65.01	30.81	18.25
2	$SSMRL_{24}$	-	-	-	0.508	0.558	77.72	0.452	0.475	65.03	25.62	4.52
3	$TAHF_{23}$	0.406	0.357	86.42	-	-	-	-	-	-	-	-
3	Ours	**0.423**	**0.375**	**87.68**	**0.539**	**0.581**	**79.76**	**0.491**	**0.499**	**68.05**	32.11	21.64

Fig. 3 Comparison with state-of-the-art open-source six NIAA methods, two AIAA methods, and one GDIAA method on AIAA datasets BAID and TAD66k, and GDIAA dataset HDDI. ACC means accuracy. "↑": the higher the values, the better. The best results are marked in bold.

5.5 range) for HDDI and 5 (midpoint of the 0 to 10 range) for BAID and TAD66K to calculate accuracy (ACC).

4.4 Performance Comparison

4.4.1 *Quantitative comparison*

We compare our proposed model with one state-of-the-art (SOTA) GDIAA method (TAHF),[5] two SOTA artistic IAA (AIAA) methods (SAAN[30] and SSMRL[31]), and six SOTA NIAA methods (NIMA,[16] MLSP,[41] TANet,[10] EAT,[42] KZIAA,[43] and HyperEmo[44]) on the SOTA open-source AIAA datasets BAID and TAD66K, and GDIAA dataset HDDI. Note that most NIAA methods are trained using EMD loss, which requires GT score distributions rather than only mean scores for training. Therefore, we modify the code provided by the researchers of AIAA and GDIAA methods and make them trainable on BAID and TAD66K.

From the experimental results in Fig. 3, it can be seen that our proposed method exceeds all the NIAA, AIAA, and GDIAA method competitors in terms of all three metrics regarding the three benchmark datasets, which demonstrates the superiority of our method. Traditional natural IAA methods such as EAT, TANet, and HyperEmo perform reasonably well but fall short compared with our proposed method, especially on the HDDI dataset. For instance, compared with the SOTA NIAA method HyperEmo and AIAA method SAAN, in terms of the PLCC metric, our proposed method has improved by 1.3%, 0.8%, and 0.7% (HyperEmo) and 3.2%, 11.4%, and 5.9% (SAAN) on the HDDI, BAID, and TAD66K datasets, respectively. These results suggest the effectiveness of integrating image content with detailed textual descriptions to overcome the limitations of traditional methods. By employing holistic and detailed descriptions, our method captures both the overall theme and specific design elements, providing a comprehensive evaluation framework.

Furthermore, the proposed method balances performance with computational efficiency. Although some methods such as HyperEmo have high parameter counts (87.65M) and FLOPs (140G), our method maintains a reasonable parameter count (32.11M) and FLOPs (21.64G) while delivering superior performance.

4.4.2 *Qualitative comparison*

Figure 4 illustrates the qualitative performance of our proposed aesthetic assessment method by comparing the GT and predicted (Pred) scores for a variety of graphic design and artistic images. The quality is highlighted in different colors, where red denotes low-level and green denotes

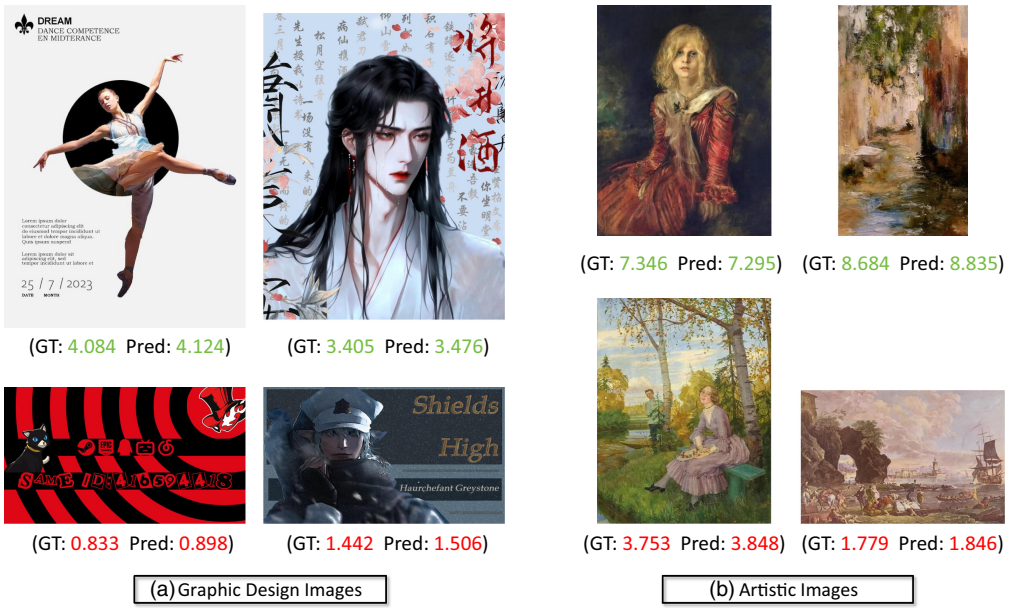

Fig. 4 Graphic design image (a) and artistic image (b) aesthetic assessment prediction results. The GT and predicted (Pred) scores are shown underneath each image. The quality is highlighted in different colors; red and green denote low-level and high-level artistic aesthetics, respectively.

high-level artistic aesthetics. The predicted scores closely match the GT scores across various graphic design and artistic images, indicating the model's accuracy in capturing the aesthetic quality and stylistic details of both graphic designs and artistic images.

4.4.3 *Cross-dataset evaluation*

The generalization ability and robustness of IAA models are crucial for their practical application. Therefore, in this experiment, we conducted cross-dataset validations by training the NIAA, AIAA, and GDIAA models on the BAID dataset and testing it on the HDDI dataset without fine-tuning, and vice versa. We selected several top-performing NIAA, AIAA, and GDIAA methods for performance comparison. The results, shown in Table 1, demonstrate that the proposed method maintains excellent generalization.

4.5 Component Evaluation

We conducted an extensive component evaluation to confirm the effectiveness of the major components in our approach, as shown in Table 2. The results indicate that all components of our

Table 1 Performance results of cross-dataset evaluation.

	Train on BAID and test on HDDI			Train on HDDI and test on BAID		
Method	SRCC↑	PLCC↑	ACC↑	SRCC↑	PLCC↑	ACC↑
NIMA	0.173	0.168	54.67	0.285	0.231	48.26
TANet	0.187	0.175	56.23	0.293	0.256	49.76
HyperEmo	0.236	0.209	60.62	0.337	0.306	51.53
KZIAA	0.218	0.195	58.87	0.316	0.283	50.95
SAAN	0.192	0.179	56.36	0.308	0.264	48.57
Ours	**0.264**	**0.217**	**62.21**	**0.375**	**0.312**	**52.53**

"↑": the higher the values, the better. The best results are marked in bold.

Table 2 Quantitative evidence of component studies.

Method	HDDI			BAID		
	SRCC↑	PLCC↑	ACC↑	SRCC↑	PLCC↑	ACC↑
Baseline	0.322	0.293	70.57	0.456	0.418	69.42
$Txt_h + FSB + SB$	0.374	0.326	79.76	0.486	0.522	72.47
$Txt_t + FSB + SB$	0.382	0.331	80.37	0.493	0.536	73.46
$Txt_h + Txt_t + FSB$	0.411	0.349	84.27	0.514	0.562	75.63
$Txt_h + Txt_d + SB$	0.418	0.356	85.49	0.521	0.571	77.46
$Txt_h + Txt_d + FSB + SB$	**0.423**	**0.375**	**87.68**	**0.539**	**0.581**	**79.76**

The best results are marked in bold.
"↑": the higher the values, the better.

proposed method contribute to improving the graphic design image assessment performance. We directly input the images into an image encoder and then predict aesthetic scores as a baseline.

Line 2 shows that the combination of holistic textual descriptions [Eq. (1), Txt_h] with FSB (Sec. 3.4) and SB (Sec. 3.5) significantly improves performance. For instance, on the HDDI dataset, this combination increases the SRCC metric from 0.322 to 0.374 and the PLCC metric from 0.293 to 0.326. This improvement highlights the importance of blending holistic textual features with visual features to enhance the model's understanding of aesthetic qualities. When using detailed textual descriptions [Eq. (2), Txt_d] with FSB and SB (line 3), there is a further improvement in performance. On the HDDI dataset, the SRCC metric increases to 0.382, and the PLCC metric rises to 0.331. This demonstrates that detailed descriptions, which capture specific aspects of the design, contribute more effectively to the aesthetic assessment.

Integrating both holistic and detailed textual descriptions with FSB (line 4) shows substantial performance gains. The SRCC and PLCC metrics on the HDDI dataset increase to 0.411 and 0.349, respectively. This indicates that combining both types of textual descriptions provides a more comprehensive understanding of the aesthetic qualities, leading to better performance.

The combination of holistic and detailed textual descriptions with SB (line 5) also enhances performance, with SRCC reaching 0.418 and PLCC achieving 0.356 on the HDDI dataset. This further confirms the utility of using both types of descriptions and the robustness of the SB technique.

The full integration of holistic and detailed textual descriptions, FSB, and SB yields the best performance (line 6). On the HDDI dataset, the SRCC metric reaches 0.423, the PLCC metric hits 0.375, and the accuracy is 87.68%. On the BAID dataset, this combination also shows the highest metrics: SRCC of 0.539, PLCC of 0.581, and accuracy of 79.76%. This comprehensive approach demonstrates the superior effectiveness of our proposed method.

4.6 Ablation Study

4.6.1 *Effectiveness of the number of textural descriptions*

To verify the effectiveness of the number of textual descriptions (Sec. 3.3) in our proposed method, we conducted an extensive ablation study varying the number of textual descriptions from one to four. The detailed results, as shown in Table 3, reveal the impact of the number of textual descriptions on performance metrics across the HDDI and BAID datasets.

When only one textual description is used, the model achieves an SRCC of 0.414, a PLCC of 0.367, and an accuracy of 86.87% on the HDDI dataset. By increasing the number of textual descriptions to two, the performance slightly improves. The SRCC on the HDDI dataset increases to 0.417, PLCC to 0.368, and accuracy to 87.32%. Incorporating three textual descriptions further enhances the model's performance. The best performance is achieved when four textual descriptions are used. The SRCC on the HDDI dataset reaches 0.423, PLCC 0.375, and accuracy 87.68%. This confirms that incorporating multiple textual descriptions provides a

Table 3 Ablation studies on the number of textural descriptions (Sec. 3.3).

	HDDI			BAID		
Method	SRCC↑	PLCC↑	ACC↑	SRCC↑	PLCC↑	ACC↑
1	0.414	0.367	86.87	0.529	0.569	78.53
2	0.417	0.368	87.32	0.533	0.571	78.87
3	0.419	0.372	87.51	0.537	0.576	79.12
4 (Ours)	**0.423**	**0.375**	**87.68**	**0.539**	**0.581**	**79.76**

"↑": the higher the values, the better. The best results are marked in bold.

comprehensive and nuanced analysis of the graphic design images, leading to the highest performance metrics.

4.6.2 *Different feature similarity blending methods*

To verify the effectiveness of different FSB (Sec. 3.4) methods in our proposed approach, we conducted an extensive ablation study. We compared the performance of three alternative blending methods: element-wise addition, element-wise multiplication, and feature concatenation, against our proposed method. The results of these experiments are summarized in Table 4.

Using element-wise addition for FSB, the model achieves an SRCC of 0.415, PLCC of 0.367, and an accuracy of 87.23% on the HDDI dataset. When element-wise multiplication is applied for blending features, the performance slightly decreases. Feature concatenation shows better performance than the previous two methods. Our proposed method, which involves a more sophisticated blending mechanism, achieves the best performance. On the HDDI dataset, the SRCC reaches 0.423, PLCC 0.375, and accuracy 87.68%. The results highlight the importance of an effective feature blending strategy to accurately capture the intricate relationships between visual and textual features, ultimately enhancing the model's ability to assess the aesthetic quality of graphic design images.

4.6.3 *Different score bagging methods*

To verify the effectiveness of different SB methods (Sec. 3.5) in our proposed approach, we compared the performance of three SB methods: averaging the predicted scores (average), adaptively adjusting the weight of each score during the model training process (adaptive), and our proposed method. The results of these experiments are summarized in Table 5.

Using the average method for SB, the model achieves an SRCC of 0.413, PLCC of 0.366, and an accuracy of 86.76% on the HDDI dataset. On the BAID dataset, the SRCC is 0.524, the PLCC is 0.570, and the accuracy is 78.87%. This method, although simple, provides a decent baseline but does not fully capture the nuances of the different feature contributions. The adaptive

Table 4 Ablation studies on different FSB methods (Sec. 3.4).

	HDDI			BAID		
Method	SRCC↑	PLCC↑	ACC↑	SRCC↑	PLCC↑	ACC↑
Element-wise addition	0.415	0.367	87.23	0.528	0.570	79.14
Element-wise multiplication	0.411	0.363	86.92	0.525	0.566	78.41
Feature concatenation	0.418	0.371	87.31	0.533	0.578	79.47
Ours	**0.423**	**0.375**	**87.68**	**0.539**	**0.581**	**79.76**

"↑": the higher the values, the better. The best results are marked in bold.

Table 5 Ablation studies on different SB methods (Sec. 3.5).

	HDDI			BAID		
Method	SRCC↑	PLCC↑	ACC↑	SRCC↑	PLCC↑	ACC↑
Average	0.413	0.366	86.76	0.524	0.570	78.87
Adaptive	0.418	0.371	87.04	0.531	0.575	79.03
Ours	**0.423**	**0.375**	**87.68**	**0.539**	**0.581**	**79.76**

"↑": the higher the values, the better. The best results are marked in bold.

SB method improves performance compared with the average method. The SRCC on the HDDI dataset increases to 0.418, PLCC to 0.371, and accuracy to 87.04%. On the BAID dataset, the SRCC is 0.531, the PLCC is 0.575, and the accuracy is 79.03%. The adaptive method adjusts weights based on feature importance, leading to better performance. Our proposed SB method achieves the best performance. On the HDDI dataset, the SRCC reaches 0.423, PLCC 0.375, and accuracy 87.68%. On the BAID dataset, the SRCC is 0.539, the PLCC is 0.581, and the accuracy is 79.76%. This method effectively combines scores derived from multiple fused features, ensuring robust and reliable assessments.

4.7 Comparison of the Score Distributions

We also provide a detailed comparison of the score distributions generated by our approach against other state-of-the-art methods on popular datasets. Upon analyzing the score distribution on the HDDI dataset, as shown in Fig. 5, we observed that our approach closely aligns with the GT distribution, particularly in the middle score range (approximately between scores 1 and 3). This suggests that our method is more effective at accurately capturing the aesthetic quality of images with medium scores. In contrast, other methods such as TANet and HyperEmo tend to have more concentrated distributions in the high and low score ranges. This concentration could lead to biases, resulting in overestimation or underestimation of certain images' aesthetic scores. Our method, by comparison, demonstrates a more balanced and realistic distribution. On the TAD66K dataset, as shown in Fig. 6, our method also shows a score distribution that is more consistent with the GT distribution across the entire score range, especially in the middle and

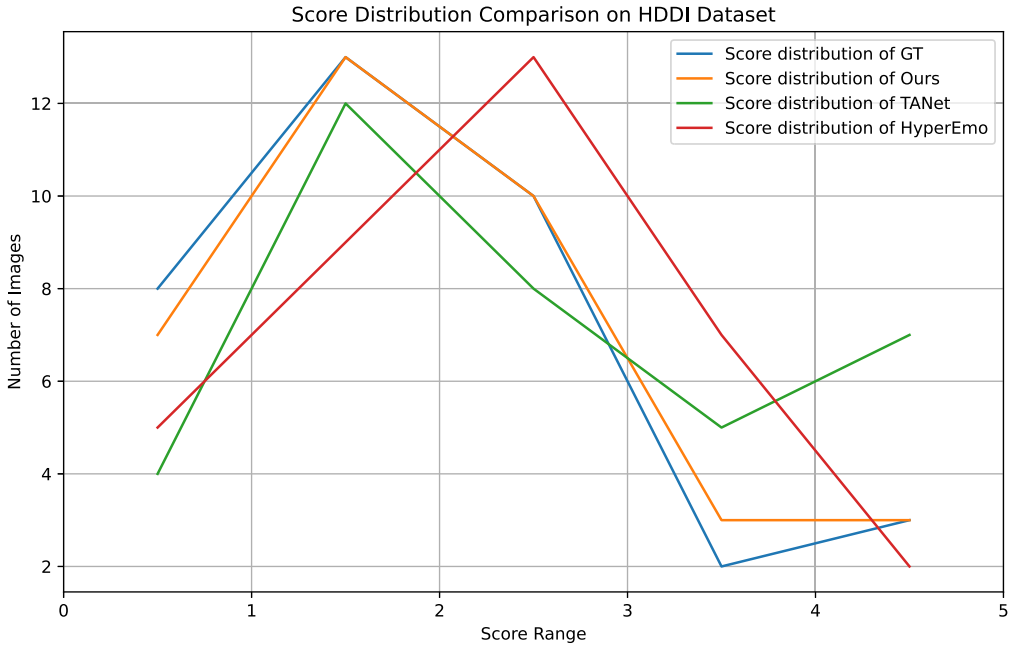

Fig. 5 Score distribution comparison on the HDDI dataset.

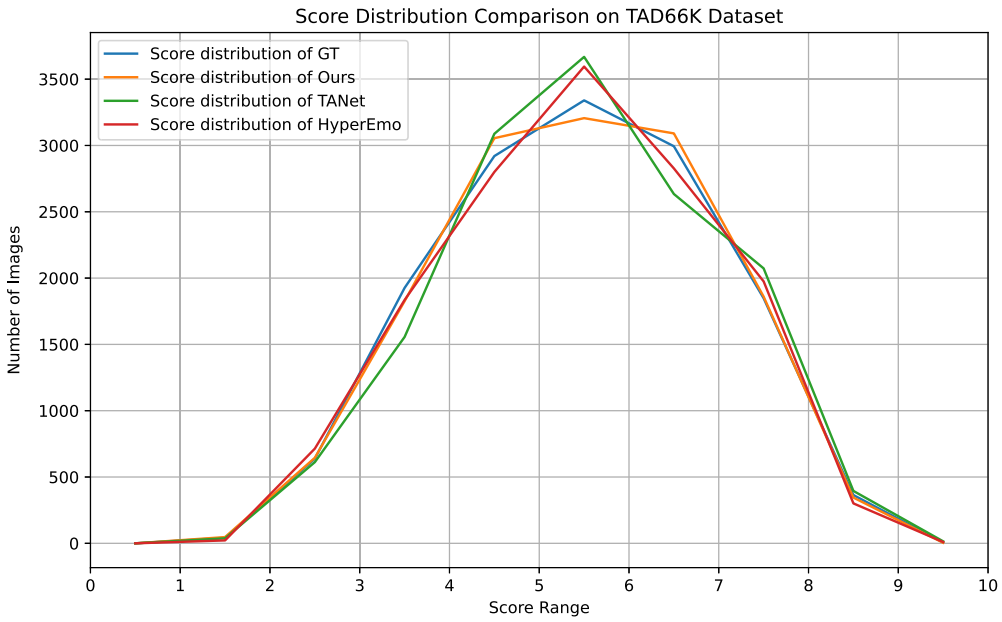

Fig. 6 Score distribution comparison on the TAD66K dataset.

high score regions (scores 5 to 8). Other methods, such as TANet and HyperEmo, exhibit a tendency to cluster scores in the lower ranges, which may distort the overall assessment of image aesthetics. Our approach, with its smoother distribution, better reflects the actual variation in aesthetic quality as observed in the GT distribution. These results underscore the effectiveness of our proposed method in generating score distributions that more accurately represent the true aesthetic qualities of images across different score ranges.

4.8 Bias Analysis of Score Bagging

To demonstrate whether our SB method has a bias (i.e., whether the scores tend to lean toward a certain direction/attribute, e.g., composition, color, detail, atmosphere, main theme), we verify it through the following experimental design. The experiment is mainly divided into two main parts: the comparison between independent scoring and fusion scoring, and the bias analysis. We selected 1000 images from the TAD66K dataset that encompass a variety of styles and aesthetic attributes for testing. For independent scoring, the model independently scores each attribute (composition, color, detail, atmosphere) of each image, obtaining a separate score for each attribute (without integration). Then, the average of all independent scores for that attribute is taken to obtain the average independent score for the attribute. For example, the average independent score for the "color" attribute represents the average of the independent scores for the color attribute across all test images. On the other hand, for fusion scoring, we combine the scores of multiple attributes (e.g., composition, color, and detail) and derive a comprehensive score based on the set weights or integration strategy. Then, this comprehensive score is compared with the image's performance on that attribute. The average integrated score is the average of the integrated scores for all images on that attribute. For example, for the "composition" attribute, the average integrated score represents the average of the comprehensive scores on the "composition" attribute, obtained by integrating the scores of various attributes such as composition, color, detail, atmosphere, and theme through the scoring integration mechanism. To be more specific, the average independent score reflects the result of assessing a single aesthetic attribute without considering other attributes, whereas the average integrated score reflects the performance on that attribute after considering other attributes and combining the scores of various attributes. The comparison between the two can help evaluate whether the scoring integration mechanism has introduced bias toward certain attributes. As shown in Table 6, although a significant positive bias was found in the composition aspect (p-value < 0.05), the bias in other attributes is not significant. Even if there is some slight bias, the overall performance of the integrated score is still highly correlated with the independent score, so the overall accuracy

Table 6 Bias analysis of SB.

Attribute	Mean independent score	Mean fusion score	Difference value (fusion independence)	Bias direction	Bias saliency (*p*-value)
Composition	3.45	3.60	+0.15	Positive bias	0.042
Color	4.10	4.00	−0.10	Negative bias	0.089
Detail	3.80	3.85	+0.05	Positive bias	0.153
Atmosphere	4.25	4.30	+0.05	Positive bias	0.125
Main theme	3.90	3.80	−0.10	Negative bias	0.078

of the scoring system is not greatly affected. It is up to the specific application scenario to decide whether to make minor adjustments to these biases.

4.9 Limitations

Although our proposed method demonstrates significant advancements in the aesthetic assessment of graphic design images, it is not without limitations. First, the reliance on high-quality textual descriptions means that the accuracy of our method heavily depends on the quality and detail of the textual data provided. Inconsistent or vague descriptions can potentially lead to suboptimal performance. Second, although our FSB mechanism effectively combines visual and textual features, it may still struggle with extremely abstract or highly subjective aesthetic elements that are difficult to quantify even with rich descriptions. In addition, our method requires substantial computational resources due to the complexity of the multi-task learning framework and the necessity of processing both image and textual data, which may limit its scalability and real-time application in resource-constrained environments.

5 Conclusions

In this paper, we presented an innovative multimodal learning approach for the aesthetic assessment of graphic design images, integrating image content with detailed textual descriptions to overcome the limitations of traditional methods. By employing holistic and detailed descriptions, our method captures both the overall theme and specific design elements, providing a comprehensive evaluation framework. The FSB mechanism enhances the representation of aesthetic qualities by aligning features from both visual and textual modalities. Experimental results demonstrate that our approach achieves state-of-the-art performance on both graphic design and natural image benchmark datasets, underscoring its effectiveness and robustness. Despite some limitations, for instance, the potential dependence on the quality of textual descriptions generated, the need for large-scale datasets with diverse design elements for better generalization, and the higher computational resources required for multimodal integration, our method sets a new standard for understanding and evaluating the visual aesthetics of graphic design images, offering valuable insights for researchers and practitioners in the field.

Future work will focus on several key areas. Improving scalability will involve optimizing the model architecture to reduce its computational demands, such as by implementing more efficient feature extraction techniques and exploring model compression methods. In addition, we plan to develop strategies for better generalization from smaller, less diverse datasets, potentially through the use of transfer learning or data augmentation techniques.

Disclosures

The authors have no relevant financial interests and no other potential conflicts of interest in the paper.

Code and Data Availability

Code, data, and materials will be made available upon request.

References

1. M. Nishiyama et al., "Aesthetic quality classification of photographs based on color harmony," in *Comput. Vis. and Pattern Recognit.*, pp. 33–40 (2011).
2. J. Zujovic et al., "Classifying paintings by artistic genre: an analysis of features & classifiers," in *IEEE Int. Workshop on Multimedia Signal Process.*, pp. 1–5 (2009).
3. Y. Wang, Y. Gao, and Z. Lian, "Attribute2font: creating fonts you want from attributes," *ACM Trans. Graph.* **39**(4), 1–15 (2020).
4. A. Shirani et al., "Let me choose: from verbal context to font selection," arXiv:2005.01151 (2020).
5. Y. Wan et al., "Automatic image aesthetic assessment for human-designed digital images," in *McGE '23: Proc. 1st Int. Workshop on Multimedia Content Generation and Eval.: New Methods and Pract.*, pp. 1–8 (2023).
6. M. Song et al., "Rethinking object saliency ranking: a novel whole-flow processing paradigm," *IEEE Trans. Image Process.* **33**, 338–353 (2024).
7. Y. Fang et al., "Perceptual quality assessment of smartphone photography," in *IEEE/CVF Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, Seattle, WA, USA, pp. 3677–3686 (2020).
8. V. Hosu et al., "Koniq-10k: an ecologically valid database for deep learning of blind image quality assessment," *IEEE Trans. Image Process.* **29**, 4041–4056 (2020).
9. Y. Yang et al., "Personalized image aesthetics assessment with rich attributes," in *IEEE/CVF Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, New Orleans, LA, USA, pp. 19861–19869 (2022).
10. S. He et al., "Rethinking image aesthetics assessment: models, datasets and benchmarks," in *Proc. Thirty-First Int. Joint Conf. Artif. Intell.*, pp. 942–948 (2022).
11. L. Li et al., "Anchor-based knowledge embedding for image aesthetics assessment," *Neurocomputing* **539**, 126197 (2023).
12. X. Nie et al., "BMI-Net: a brain-inspired multimodal interaction network for image aesthetic assessment," in *MM '23: Proc. 31st ACM Int. Conf. Multimedia*, pp. 5514–5522 (2023).
13. D. Soydaner and J. Wagemans, "Multi-task convolutional neural network for image aesthetic assessment," *IEEE Access* **12**, 4716–4729 (2024).
14. A. Pandit et al., "Image aesthetic score prediction using image captioning," in *ICCCE 2023. Cognitive Science and Technology*, A. Kumar, S. Mozar, and J. Haase, Eds., pp. 413–425, Springer, Singapore (2023).
15. S. Kong et al., "Photo aesthetics ranking network with attributes and content adaptation," *Lect. Notes Comput. Sci.* **9905**, 662–679 (2016).
16. H. Talebi and P. Milanfar, "Nima: neural image assessment," *IEEE Trans. Image Process.* **27**(8), 3998–4011 (2018).
17. H. Zeng et al., "A unified probabilistic formulation of image aesthetic assessment," *IEEE Trans. Image Process.* **29**, 1548–1561 (2019).
18. X. Lu et al., "Rapid: rating pictorial aesthetics using deep learning," in *MM '14: Proc. 22nd ACM Int. Conf. on Multimedia*, pp. 457–466 (2014).
19. S. Ma, J. Liu, and C. Wen Chen, "A-lamp: adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment," in *IEEE Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, pp. 4535–4544 (2017).
20. K. Sheng et al., "Attention-based multi-patch aggregation for image aesthetic assessment," in *MM '18: Proc. 26th ACM Int. Conf. Multimedia*, pp. 879–886 (2018).
21. J. Ke et al., "VILA: learning image aesthetics from user comments with vision-language pretraining," in *IEEE Conf. Comput. Vis. and Pattern Recognit.*, pp. 10041–10051 (2023).
22. L. Li et al., "Theme-aware visual attribute reasoning for image aesthetics assessment," *IEEE Trans. Circuits Syst. Video Technol.* **33**(9), 4798–4811 (2023).
23. S. A. Amirshahi and J. Denzler, "Judging aesthetic quality in paintings based on artistic inspired color features," in *Int. Conf. Digit. Image Comput.: Tech. and Appl. (DICTA)*, Sydney, NSW, Australia, pp. 1–8 (2017).
24. X. Guo et al., "Visual complexity assessment of painting images," in *IEEE Int. Conf. Image Process.*, Melbourne, VIC, Australia, pp. 388–392 (2013).
25. C. Li and T. Chen, "Aesthetic visual quality assessment of paintings," *IEEE J. Sel. Top. Signal Process.* **3**(2), 236–252 (2009).
26. W. Luo, X. Wang, and X. Tang, "Content-based photo quality assessment," in *Int. Conf. Comput. Vis.*, Barcelona, Spain, pp. 2206–2213 (2011).
27. T. Chen et al., "Large-scale tag-based font retrieval with generative feature learning," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, pp. 9116–9125 (2019).
28. Q. Dou et al., "Webthetics: quantifying webpage aesthetics with deep learning," *Int. J. Hum.-Comput. Stud.* **124**, 56–66 (2019).
29. M. Zen, N. Burny, and J. Vanderdonckt, "A quality model-based approach for measuring user interface aesthetics with grace," in *Proc. ACM on Hum.-Comput. Interaction*, Vol. 7, pp. 1–47 (2023).

30. R. Yi et al., "Towards artistic image aesthetics assessment: a large-scale dataset and a new method," in *IEEE/CVF Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, pp. 22388–22397 (2023).
31. T. Shi et al., "Semantic and style based multiple reference learning for artistic and general image aesthetic assessment," *Neurocomputing* **582**, 127434 (2024).
32. L. Marchesotti et al., "Assessing the aesthetic quality of photographs using generic image descriptors," in *Int. Conf. Comput. Vis.*, Barcelona, Spain, pp. 1784–1791 (2011).
33. M. Wertheimer, *Investigations on Gestalt Principles*, MIT Press, London, UK (2012).
34. G. A. Agoston, *Color Theory and its Application in Art and Design*, Vol. **19**, Springer (2013).
35. R. Reber, N. Schwarz, and P. Winkielman, "Processing fluency and aesthetic pleasure: is beauty in the perceiver's processing experience?" *Personality Soc. Psychol. Rev.* **8**(4), 364–382 (2004).
36. J. P. Forgas, "Mood and judgment: the affect infusion model (aim)," *Psychol. Bull.* **117**(1), 39 (1995).
37. J. Li et al., "BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. 39th Int. Conf. Mach. Learn.*, Baltimore, Maryland, USA, PMLR, pp. 12888–12900 (2022).
38. W. Kim, B. Son, and I. Kim, "ViLT: vision-and-language transformer without convolution or region supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, Baltimore, Maryland, USA, PMLR, pp. 5583–5594 (2021).
39. T. Brown et al., "Language models are few-shot learners," in *Adv. in Neural Inf. Process. Syst. (NeurIPS 2020)*, Vol. 33, pp. 1877–1901 (2020).
40. A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, PMLR, pp. 8748–8763 (2021).
41. V. Hosu, B. Goldlucke, and D. Saupe, "Effective aesthetics prediction with multi-level spatially pooled features," in *IEEE/CVF Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, Long Beach, CA, USA, pp. 9375–9383 (2019).
42. S. He et al., "EAT: an enhancer for aesthetics-oriented transformers," in *MM '23: Proc. 31st ACM Int. Conf. Multimedia*, pp. 1023–1032 (2023).
43. G. Wang et al., "Keep knowledge in perception: zero-shot image aesthetic assessment," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Seoul, Korea, pp. 8311–8315 (2024).
44. G. Lan et al., "Image aesthetics assessment based on hypernetwork of emotion fusion," *IEEE Trans. Multimedia* **26**, 3640–3650 (2024).

Biographies of the authors are not available.