Creative Engineering **2025**, *03*

Type of the Paper (Article)

Automated Aesthetic Quality Assessment of Mobile UI Design via Hierarchy-Perceiving Graph Attention Networks and Visual Hierarchy Features

JiaYi Huang [1], Lily Evans [2] and JunHao Wu [3],*

[1] Oregon State University;tabodgruver34@outlook.com
[2] Oregon State University; bruniitousp22@outlook.com
[3] Oregon State University; rmambourg1995@outlook.com
* Correspondence:rmambourg1995@outlook.com

Abstract

With the rapid development of the mobile internet, the aesthetic quality of mobile user interface (UI) design has become a key factor influencing user experience and product competitiveness. However, existing evaluation methods still suffer from significant limitations in terms of objectivity, scalability, and the understanding of compositional semantics. This study proposes an automatic aesthetic quality assessment method for mobile UI design that integrates visual hierarchy features with a Graph Attention Network, termed HPA-GNN (Hierarchy-Perceiving Graph Attention Network). Based on the principles of visual hierarchy, the golden ratio, and color contrast, visual elements within a UI interface are abstracted as graph nodes. A spatial relational graph structure is then constructed according to compositional principles, and a graph attention mechanism is introduced to achieve differentiated aggregation of node features. Experiments are conducted on a self-constructed dataset containing 1,080 annotated UI design images. The results demonstrate that HPA-GNN significantly outperforms existing baseline models in both prediction accuracy and scoring consistency: the Weighted Root Mean Square Error (WRMSE) reaches 0.362 ± 0.017, the Spearman rank correlation coefficient is 0.731 ± 0.016, and the Graph Structure Integrity Rate (GSIR) is 0.934 ± 0.015. Compared with CNN-based baseline models, improvements of approximately 17.2%, 12.8%, and 14.9% are achieved, respectively. Ablation experiments further verify the complementarity and synergistic effects of the three compositional rules. This study provides an interpretable technical pathway for intelligent evaluation of mobile UI

Academic Editor: Haiwen Wang

Received: July 15, 2025
Revised: August 23, 2025
Accepted: August 25, 2025
Published: September 28, 2025

design aesthetics, offering important theoretical and practical implications for the development of design assistance tools, user experience optimization, and the standardization of design quality.

Keywords: Mobile UI design; Aesthetic quality assessment; Graph neural networks; Visual hierarchy features; Design intelligence

1. Introduction

With the rapid development of the digital economy, mobile applications have penetrated nearly every aspect of users' daily lives. The aesthetic quality of mobile User Interface (UI) design not only directly influences users' first impressions and emotional resonance but also profoundly shapes user retention and commercial conversion rates [1]. Studies have shown that users require only 50 milliseconds to form an initial aesthetic judgment of a mobile application interface, and this instantaneous perceptual experience often determines whether they continue to engage with the application [2]. Consequently, developing scientific, objective, and efficient approaches for evaluating the aesthetic quality of mobile UI design has become an important research topic at the intersection of Human–Computer Interaction (HCI), computer vision, and design studies.

However, evaluating the aesthetic quality of UI design presents multiple challenges. First, UI interfaces consist of diverse visual elements—including text, icons, images, and color blocks—whose complex spatial hierarchies and compositional semantics make traditional manual evaluation methods both inefficient and inconsistent. Such approaches are often influenced by evaluators' subjective preferences and professional backgrounds, resulting in limited reliability [3]. Second, existing computer vision methods primarily focus on extracting global image features and therefore struggle to capture the structural compositional relationships among interface elements, leading to notable discrepancies between algorithmic evaluation results and the judgments of professional designers [4]. Furthermore, the aesthetic quality of UI design is inherently multidimensional, encompassing aspects such as proportion and layout, color harmony, visual hierarchy, whitespace utilization, and design consistency. Modeling these multidimensional aesthetic attributes remains difficult when relying solely on single-feature representations or shallow models [5].

At the academic level, the field of Image Aesthetic Quality Assessment (IAQA) has accumulated substantial research achievements. Early studies relied mainly on handcrafted low-level visual features—such as color histograms, texture descriptors, and edge features—combined with traditional machine learning algorithms, including support vector machines and random forests, for classification or regression tasks [6]. With the emergence of deep learning, end-to-end learning methods represented by Convolutional Neural Networks (CNNs) have achieved significant breakthroughs in general image aesthetic evaluation tasks. Models such as Neural Image Assessment (NIMA) can directly predict the distribution of human aesthetic ratings [7]. More recently, Transformer-based architectures have demonstrated superior performance in image aesthetic assessment tasks due to their powerful capability for modeling global dependencies [8]. However, when applied to UI design aesthetic evaluation, these approaches face a common limitation: they are primarily designed for natural images or photographic works and lack mechanisms to model the compositional structural semantics specific to UI interfaces. Consequently, they fail to effectively incorporate well-defined design principles—such as visual hierarchy, the golden ratio, and color contrast—as prior knowledge to guide feature learning.

The emergence of Graph Neural Networks (GNNs) provides a promising technical pathway for addressing these limitations. GNNs can naturally represent structured information through nodes (visual elements) and edges (relationships among elements), making them particularly suitable for modeling the spatial compositional relationships among multiple UI elements [9]. Graph Attention Networks (GATs) further introduce an attention mechanism that adaptively adjusts information aggregation weights based on the relative importance of nodes, enabling more precise modeling of hierarchical relationships among primary and secondary elements in UI interfaces [10]. Nevertheless, most existing GNN-based design evaluation studies treat graph construction as a purely data-driven process and overlook well-established compositional theories in the design domain. As a result, the semantic interpretability of the constructed graph structures remains limited, and their generalization ability is often insufficient when dealing with the complex multi-level compositions typical of UI interfaces [11].

To address these research gaps, this study proposes a Hierarchy-Perceiving Graph Attention Network (HPA-GNN). The proposed approach formalizes three core compositional rules from UI design theory—visual hierarchy, the golden ratio, and color contrast—as constraints for graph structure construction. Specifically, visual elements in a UI interface are abstracted as graph nodes, and weighted edges are established according to compositional principles, thereby constructing a graph representation that integrates design knowledge. A graph attention convolution mechanism is then applied to perform composition-aware feature aggregation and aesthetic score prediction. The primary contributions of this study are threefold. First, at the methodological level, visual hierarchy, the golden ratio, and color contrast principles from UI design theory are formalized as graph structural constraints, overcoming the limitations of traditional aesthetic evaluation methods that rely on handcrafted features or purely data-driven graph construction. Second, at the model architecture level, the proposed HPA-GNN integrates the structural representation capability of graph neural networks with UI compositional features, enabling cross-level modeling from local visual elements to overall aesthetic quality. Third, at the empirical level, systematic experiments are conducted on a self-constructed dataset containing 1,080 mobile UI design images, demonstrating the proposed method's advantages in terms of scoring accuracy, consistency, and structural perception capability.

2. Related Work

2.1. Theoretical Foundations of Image Aesthetic Quality Assessment

Human aesthetic judgment has long been a central topic in psychology, cognitive science, and art theory. Early psychological studies suggest that aesthetic experience is closely related to the complexity, symmetry, and novelty of visual stimuli [12]. The aesthetic information processing model proposed by Helmut Leder and colleagues reveals the multi-stage nature of aesthetic evaluation, encompassing perceptual analysis, meaning construction, and emotional response. This model provides an important cognitive science foundation for the design of computational aesthetic models [13].

From the perspective of design theory, classical compositional principles—such as the golden ratio, the rule of thirds, visual hierarchy, and Gestalt principles—offer quantifiable prior knowledge for aesthetic modeling of UI interfaces. These principles can be encoded mathematically as spatial proportional constraints and further represented as node relationships and edge weights, thereby providing theoretical support for graph-based modeling approaches [14]. Recent advances in neuroaesthetics further indicate that the brain activates neural circuits associated with reward and pleasure when processing visual attributes such as proportion, balance, and composition, offering neuroscientific evidence for modeling aesthetic judgment through structural features [15].

In the field of Explainable Artificial Intelligence (XAI), techniques such as Grad-CAM and SHAP (SHapley Additive exPlanations) have been introduced into aesthetic modeling to visualize the regions of images that models focus on during prediction [16]. By comparing these attention regions with human visual attention patterns, researchers can validate the alignment between AI predictions and human perception. Such interpretability not only enhances model transparency but also provides practical design assistance, enabling designers to understand which elements are perceived as aesthetically pleasing rather than merely receiving a numerical score. Consequently, explainable AI has become an important research direction in design intelligence, providing methodological references for the visualization of compositional rules and attention-weight analysis in this study [17].

2.2. Applications of Deep Learning in UI Design Aesthetic Evaluation

With the widespread adoption of deep neural networks in computer vision, researchers have increasingly explored their application in automated UI design evaluation. Ma Jianfeng et al. [18] combined GoogLeNet with Kansei engineering data to achieve fine-grained clustering of traditional and modern UI interface styles, reaching an accuracy of over 90% in image consistency scoring. Shi Yancheng et al. [19] constructed a visual image database of mobile application UIs and used neural network clustering to successfully categorize brand packaging designs, accurately capturing the relationship between user preferences and stylistic features and providing support for “visual brand positioning” algorithms.

In terms of aesthetic scoring, Shu Kai et al. [20] integrated Convolutional Neural Networks with the Neural Image Assessment (NIMA) model to develop a scoring mechanism reflecting users’ first impressions. The model can simultaneously output aesthetic score distributions and focal region analyses. Wei Ming [21], from the perspective of philosophical aesthetics, argued that perceptual evaluation extends beyond composition to include visual rhythm, symbolic expression, and information density, and subsequently developed a multidimensional neural scoring system.

Regarding improvements in interpretability, Li Wei et al. [22] proposed a scoring network that jointly models visual saliency and layout structure to optimize typographic aesthetics, providing deep learning-based assistance for fine-grained evaluation of multi-element UI layouts. Pu Yifan et al. [23] treated “layout structure” as the core variable in aesthetic evaluation and systematically analyzed the capability of deep learning models in handling mixed text—image UI interfaces. They advocated the use of explicit rule-based layout models as auxiliary inputs. Additionally, Ma Rui et al. [24] embedded saliency analysis and visualization modules into UI design models to achieve traceable scoring processes, enabling designers to receive actionable feedback.

2.3. Applications of Graph Neural Networks in Visual Design Analysis

The application of Graph Neural Networks in visual design analysis has attracted increasing attention in recent years. Zhang Wei et al. [25] proposed a graph-structured modeling method based on Graph Attention Network, demonstrating that graph-based structural modeling not only improves scoring consistency but also enhances model interpretability. Nguyen Thanh et al. [26] applied GAT-based models to image layout optimization, proving their effectiveness in improving the structural harmony of synthesized images.

However, existing studies mainly focus on style classification or single compositional patterns. Their adaptability remains limited for UI interfaces, which are characterized by high element density, rich commercial semantics, and the coexistence of multiple design rules. Moreover, current approaches lack mechanisms for integrating explicit design rules to address challenges such as compositional balance, visual guidance, and redundancy interference [27].

In terms of graph structure modeling for UI interfaces, Rizzo Marco et al. [28] proposed a graph construction method based on the golden ratio, encoding spatial proportions as node relationships and edge weights, which provides an important reference for the graph construction approach in this study. Lucia Bianchi et al. [29] investigated the computational representation of Gestalt principles—including proximity, similarity, and closure—in UI layout analysis, verifying that these principles effectively reveal human preferences for visual organization. Furthermore, neuroaesthetic research by Torlak Mehmet et al. [30] confirmed that the brain activates reward-related circuits when processing visual attributes such as proportion, balance, and composition, providing neuroscientific evidence for modeling UI aesthetic judgments through structural features.

In summary, existing studies exhibit several limitations. First, there is a lack of mechanisms that systematically integrate multiple compositional rules from UI design theory into graph structure construction. Second, current GNN models have limited capacity for differentiated modeling of node importance when dealing with the multi-level visual hierarchies present in UI interfaces. Third, specialized aesthetic evaluation datasets for mobile UI design remain scarce, which constrains further research progress. To address these limitations, this study proposes the HPA-GNN method, which formalizes the visual hierarchy rule, golden ratio rule, and color contrast rule as graph structure constraints and introduces a hierarchy-aware graph attention mechanism to achieve accurate and interpretable evaluation of the aesthetic quality of mobile UI design.

3. Methodology

3.1. Research Strategy

This study adopts a “model-first, validation-later” technical framework. The overall approach can be summarized as follows: compositional constraints are first extracted based on UI design theory and then formalized as rules for graph structure construction. On this basis, a Hierarchy-Perceiving Graph Attention Network (HPA-GNN) is developed to predict aesthetic scores. Finally, systematic experiments are conducted to verify the effectiveness and generalization capability of the proposed model.

Specifically, the research process consists of four stages: data collection and annotation, graph structure construction, model training and optimization, and multidimensional performance evaluation. The entire technical pipeline follows a logical progression of “data → graph → model → evaluation”, ensuring that each design decision is supported by clear theoretical foundations and empirical validation.

3.2. Data Collection and Preprocessing

3.2.1. Dataset Construction

A dedicated dataset for mobile UI design aesthetic quality assessment was constructed in this study, containing 1,080 mobile UI interface screenshots from multiple application categories. The data sources include high-quality UI design works published on open design platforms such as Dribbble and Behance, as well as real interface screenshots from the top 1,000 most-downloaded applications in the Google Play Store and Apple App Store. In addition, mobile UI design cases that have received international design awards—including the Red Dot Design Award and the iF Design Award—were incorporated into the dataset.

All images were obtained in compliance with their respective copyright licenses. Images from open platforms were used under Creative Commons (CC) licenses, while screenshots from commercial applications were processed strictly for academic research purposes. The dataset covers eight application categories: social media, e-commerce, financial services, healthcare, education, entertainment and gaming, transportation and navigation, and productivity tools. The distribution of categories is illustrated in Figure 11(a).

Compared with existing general-purpose aesthetic datasets—such as AVA (Aesthetic Visual Analysis Dataset) and AADB (Aesthetic and Attributes Database)—the proposed dataset has several distinctive characteristics. First, it focuses specifically on mobile UI interfaces, including UI-specific visual elements such as brand logos, typographic structures, icon systems, and navigation components. Second, the annotation dimensions are more fine-grained, covering five professional criteria: proportion and layout, color harmony, visual hierarchy, whitespace utilization, and design consistency. Third, all annotators possess professional UI/UX design backgrounds, enabling aesthetic evaluations from the perspective of professional designers rather than relying on crowdsourced aesthetic voting.

3.2.2. Data Annotation

Aesthetic scoring was conducted using a five-point Likert scale. Twelve annotators with professional UI/UX design backgrounds participated in the evaluation process, including six faculty members from university design schools and six senior UI designers with more than three years of professional experience. Each image was independently evaluated by all annotators.The scoring criteria include:

- Proportion and layout (the rationality of element size relationships and spatial distribution);
- Color harmony (color combinations and visual comfort);
- Visual hierarchy (the clarity and guidance of primary and secondary elements);
- Whitespace utilization (the effective use of negative space);
- Design consistency (the coherence of style, typography, and color schemes).

The final aesthetic label for each image is calculated as the weighted average of annotators' scores. The weights are determined based on Inter-rater Reliability (IRR), measured using Krippendorff's α coefficient. The overall agreement level reaches $\alpha = 0.76$, indicating an acceptable level of reliability.

Three instances in the dataset experienced missing values due to device malfunction, representing 0.28% of the total data. These missing values were handled using mean imputation. Finally, the dataset was divided into training, validation, and test sets with a ratio of 7:1:2, corresponding to 756, 108, and 216 images, respectively.

3.2.3. Image Preprocessing

The data preprocessing pipeline is illustrated in Figure 1. First, all images were resized to a unified resolution of **375 × 812 pixels** (corresponding to the standard iPhone screen resolution), and pixel values were normalized to eliminate the effects of resolution differences and illumination variations. Subsequently, the **U^2-Net** saliency detection model was employed to segment and locate key visual elements in each UI image, including text blocks, icons, image regions, color blocks, and brand logos, thereby obtaining the bounding box coordinates and spatial position information of each element. For elements with a high degree of occlusion or overlap, the **Non-Maximum Suppression (NMS)** algorithm was applied to remove redundant detections. Ultimately, an average of **14.3 ± 5.8 visual element nodes** were extracted from each image, providing standardized inputs for subsequent graph structure construction.

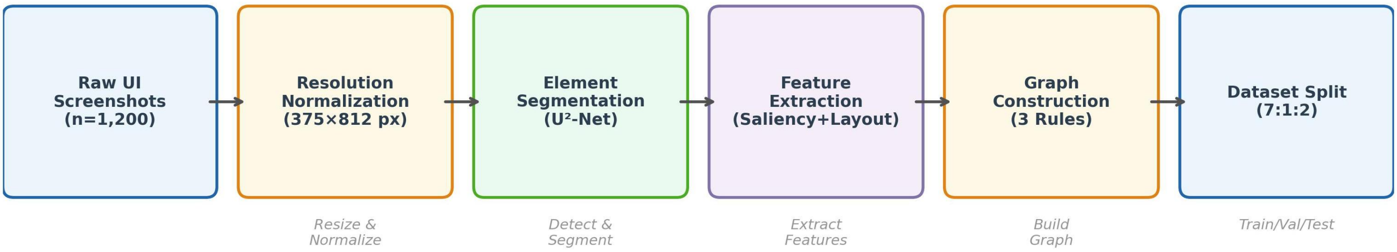


Figure 1. Data preprocessing flowchart.

3.3. Hierarchy–Aware Graph Structure Construction

3.3.1. Basic Definition of the Graph

To effectively model the spatial compositional relationships among visual elements in UI interfaces, each UI image is represented as a weighted directed graph $G = (V, E)$ The node set $V = \{v_1, v_2, \ldots, v_n\}$ denotes the visual regions extracted from the image, while the edge set $E \subseteq V \times V$ is established according to the following three compositional rules.

The feature vector $\mathbf{h}_i$ of each node v_i consists of two components: a local visual feature vector (dimension 512) extracted through saliency detection, and spatial geometric information (dimension 8) including centroid coordinates and bounding box dimensions.

3.3.2. Formalization of the Three Compositional Rules

To effectively model the spatial compositional relationships among visual elements in UI interfaces, each UI image is represented as a weighted directed graph $\mathrm{G} = (\mathrm{V}, \mathrm{E})$ The node set $\mathrm{V} = \{\mathrm{v}_1, \mathrm{v}_2, \ldots, \mathrm{v}_\mathrm{n}\}$ denotes the visual regions extracted from the image, while the edge set $\mathrm{E} \subseteq \mathrm{V} \times \mathrm{V}$ is established according to the following three compositional rules.

1. Visual Hierarchy Rule

Visual hierarchy is one of the most fundamental principles in UI design. It refers to establishing a clear visual priority among elements through differences in size, position, and color contrast, thereby guiding users’ attention flow according to the designer’s intention.

In the graph structure, if two nodes v_i and v_j exhibit a clear visual hierarchy (i.e., one element visually dominates the other), a directed edge is established between them. The edge weight is determined by the hierarchy strength function $Hier(v_i, v_j)$:

$$\mathrm{Hier}(\mathrm{v_i}, \mathrm{v_j}) = \frac{\mathrm{S_i} - \mathrm{S_j}}{\mathrm{S_i} + \mathrm{S_j}} \cdot \exp\left(-\frac{|\mathrm{C_i} - \mathrm{C_j}|}{\sigma_\mathrm{c}}\right)$$

where S_i and S_j denote the areas of the two elements, C_i and C_j represent their color saliency values, and σ_c is a normalization parameter. When two elements exhibit a significant difference in area and strong color contrast, the hierarchy strength approaches 1, indicating a strong dominant—subordinate visual relationship.

2. Visual Hierarchy Rule

The golden ratio (approximately 1:1.618) is a widely recognized harmonious proportional relationship in visual design and is known to produce a visually pleasing sense of balance.

In the graph structure, if the positional relationship between nodes v_i and v_j conforms to the spatial distribution characteristics of the golden ratio or the golden spiral, a higher edge weight is assigned. The golden ratio conformity function $\mathrm{Gold}(v_i, v_j)$ is defined as:

$$\mathrm{Gold}(v_i, v_j) = \exp\left(-\frac{\left|\frac{d_{ij}}{W} - \frac{1}{\phi}\right|}{\sigma_g}\right)$$

where d_{ij} denotes the Euclidean distance between the centroids of the two nodes, WWW is the image width, $\phi = 1.618$ is the golden ratio coefficient, and σ_g is a normalization parameter. When the spacing ratio between two elements approaches the golden ratio, the function value approaches 1.

3. Color Contrast Rule

Color contrast is an important factor in UI design that influences both readability and aesthetic perception. High-contrast color combinations enhance visual impact and improve information transmission efficiency, while harmonious color transitions help create a coherent visual atmosphere.

The color contrast function $Contrast(v_i, v_j)$ is defined based on the color contrast ratio specified in the Web Content Accessibility Guidelines 2.1:

$$\mathrm{Contrast}(v_i, v_j) = \frac{\max(L_i, L_j) + 0.05}{\min(L_i, L_j) + 0.05}$$

where L_i and L_j denote the relative luminance values of the two elements, computed according to the sRGB color space.

3.3.3. Integrated Edge Weight Computation

By integrating the three compositional rules described above, the final edge weight between nodes v_i and v_j is computed using the following fusion formula:

$$w_{ij} = \lambda_1 \mathrm{Hier}(v_i, v_j) + \lambda_2 \mathrm{Gold}(v_i, v_j) + \lambda_3 \mathrm{Contrast}(v_i, v_j), \quad \lambda_1 + \lambda_2 + \lambda_3 = 1 \backslash \mathrm{tag1}$$

where the hyperparameters λ_1, λ_2, and λ_3 are used to balance the contribution weights of the three compositional relationships. The initial configuration adopts equal weighting $(\lambda_1 = \lambda_2 = \lambda_3 = 1/3)$, and different weight combinations are further examined through sensitivity analysis in the experimental section.

3.4. HPA-GNN Model Architecture

The overall architecture of HPA-GNN is illustrated in Figure 2 and consists of three core modules: the hierarchy-driven graph structure module, the graph attention convolution module, and the graph-level representation and aesthetic prediction module.

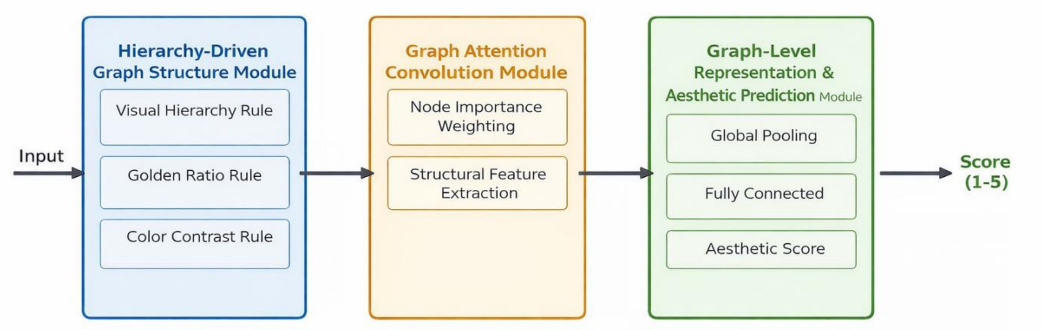


Figure 2. Schematic illustration of the HPA-GNN model architecture.

3.4.1. Graph Attention Convolution Mechanism

Considering that different visual elements in a UI interface (e.g., main images, title text, auxiliary icons, etc.) exhibit significant differences in their importance for aesthetic perception, this study introduces the Graph Attention Network (GAT) to perform:

$$\mathrm{hi}^{(l+1)} = \sigma\left(\sum \mathrm{j} \in (\mathrm{i})\, \alpha_{\mathrm{ij}}^{(l)} \cdot \mathrm{W}^{(l)} \mathrm{h}_{\mathrm{j}}^{(l)}\right) \backslash\, \mathrm{tag2}$$

where $\mathcal{N}(i)$ denotes the set of neighboring nodes of node$i$, and $\alpha_{ij}^{(l)}$ represents the attention weight at the $l-th$ layer, which is computed as follows:

$$\alpha_{\mathrm{ij}}^{(l)} = \frac{\exp\left(\mathrm{LeakyReLU}\left(\mathrm{a}^{\top}\left[\mathrm{W}^{(l)}\mathrm{h}_{\mathrm{i}}^{(l)} | \mathrm{W}^{(l)}\mathrm{hj}^{(l)}\right]\right)\right)}{\sum \mathrm{k} \in (\mathrm{i}) \exp\left(\mathrm{LeakyReLU}\left(\mathrm{a}^{\top}\left[\mathrm{W}^{(l)}\mathrm{h}_{\mathrm{i}}^{(l)} | \mathrm{W}^{(l)}\mathrm{h}_{\mathrm{k}}^{(l)}\right]\right)\right)} \backslash\, \mathrm{tag3}$$

where $\mathbf{a}$ denotes a learnable attention weight vector, and */* represents the vector concatenation operation. HPA-GNN adopts a multi-head attention mechanism (with eight attention heads). The outputs of multiple attention heads are concatenated to enhance the model's representation capacity, and an averaging aggregation strategy is applied in the final layer.

3.4.2. Graph-Level Representation and Regression Prediction

After three layers of GAT convolution, Global Attention Pooling is applied to generate the graph-level representation vector $\mathbf{h}_G$. This representation is then fed into a fully connected layer to predict the aesthetic score:

$$\hat{y} = w^{T} h_{G} + b \backslash \mathrm{tag}4$$

The loss function is defined as the Mean Squared Error (MSE):

$$\mathcal{L}\mathrm{MSE} = \frac{1}{N} \sum i = 1^{N} (y_i - \hat{y}_i)^2 \backslash \mathrm{tag}5$$

3.5. Evaluation Metrics System

To comprehensively evaluate the performance of HPA-GNN in predicting aesthetic scores for UI design, this study establishes a multidimensional evaluation metric system from three perspectives: prediction accuracy, rating consistency, and graph-structure perception capability.

The Weighted Root Mean Squared Error (WRMSE) is used to evaluate the prediction accuracy of the model:

$$\mathrm{WRMSE} = \sqrt{\frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}i)^2}{\sum i = 1^{n} w_i}} \backslash \mathrm{tag}6$$

where w_i denotes the inter-rater agreement weight of the aesthetic score for the iii-th image, y_i represents the ground-truth score, and $\hat{y}_i$ denotes the predicted score.

The Spearman Rank Correlation Coefficient *(ρ)* is used to evaluate the ranking consistency of the model’s predictions:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \backslash \mathrm{tag}7$$

where d_i denotes the rank difference between the ground-truth score and the predicted score for the $i - th$ sample.

The Kendall's Coefficient of Concordance (W) is used to measure the overall degree of agreement among multiple raters or between the model predictions and the human-averaged scores:

$$W = \frac{12S}{m^2 (n^3 - n)} \backslash \mathrm{tag}8$$

The Graph Representation Divergence (GRD) and the Graph Structure Integrity Rate (GSIR) are used to evaluate the graph structure modeling capability of the graph neural network. Their definitions are given as follows:

$$\mathrm{GRD} = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \cos\left(\vec{h}_i^{0}, \vec{h}_i^{L}\right)\right) \backslash \mathrm{tag9}$$

$$\mathrm{GSIR} = \frac{|\varepsilon_{\mathrm{retained}}|}{|\varepsilon_{\mathrm{original}}|} \backslash \mathrm{tag10}$$

4. Data

4.1. Basic Statistical Characteristics of the Dataset

The dataset used in this study contains 1,080 mobile UI design images, including 756 images for training, 108 images for validation, and 216 images for testing. Three cases in the dataset experienced data loss due to device malfunction (missing rate: 0.28%), which were handled using mean imputation.Table 1 presents the descriptive statistics of the key variables in the dataset.

Table 1. Descriptive statistics of key variables in the dataset (n = 1,080).

Variable	Mean	Standard Deviation	Median	Minimum	Maximum	Q25	Q75
Overall Aesthetic Score	3.264	0.886	3.251	1.000	5.000	2.614	3.924
Proportion and Layout Score	3.268	0.937	3.260	1.000	5.000	2.622	3.928
Color Harmony Score	3.313	0.923	3.303	1.000	5.000	2.679	3.967
Visual Hierarchy Score	3.234	0.941	3.227	1.000	5.000	2.581	3.889
Whitespace Utilization Score	3.286	0.929	3.280	1.000	5.000	2.641	3.942
Design Consistency Score	3.275	0.921	3.268	1.000	5.000	2.632	3.935
Total Number of Visual Elements	14.3	5.8	14.0	5	27	10	18

The descriptive statistics indicate that the overall aesthetic score has a mean of 3.264 and a standard deviation of 0.886, with the distribution approximating a normal shape (Figure 3a). This suggests that the dataset covers a wide range of aesthetic quality, from low-quality to high-quality UI designs.The standard deviations of the scores across different dimensions range from 0.921 to 0.941, indicating a reasonable level of variability and avoiding the issue of insufficient discriminative power caused by overly concentrated values.The score distributions across different application categories are shown in Figure 3(b). Among them, social media UI exhibits the highest average aesthetic score (mean = 3.41), while utility/efficiency applications show the lowest (mean = 3.09). This observation is consistent with real-world design practice, where different application categories often allocate varying levels of effort and resources to visual aesthetics.

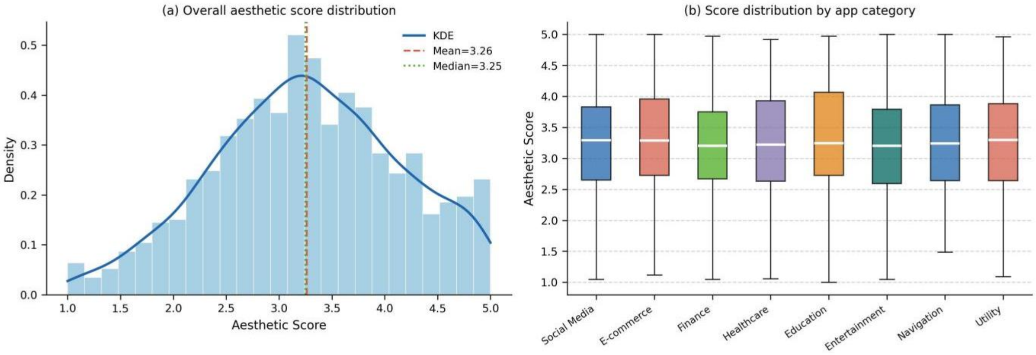
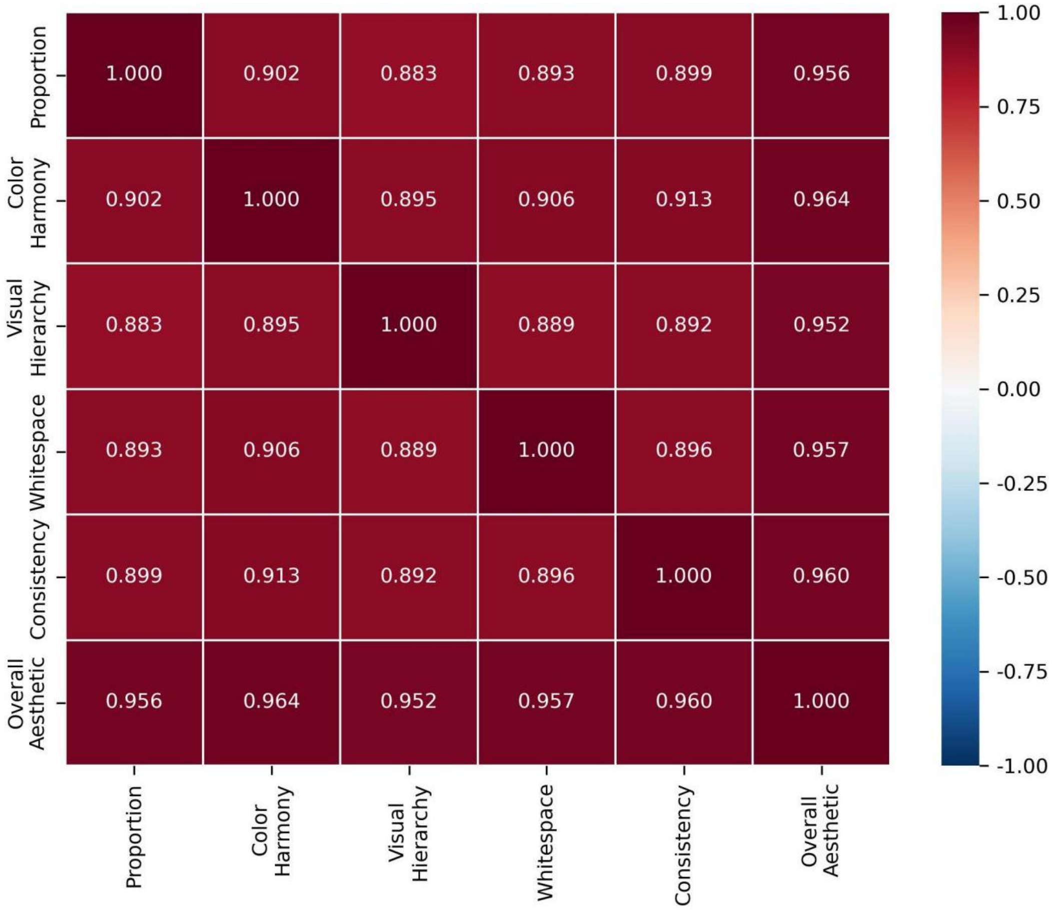


Figure 3. Distribution of aesthetic scores (histogram + KDE) and boxplots of scores across different categories.

4.2. Validation of Data Preprocessing Effectiveness

After the preprocessing pipeline (Figure 1), significant positive correlations are observed among the scores of different dimensions (Figure 4). Among them, the visual hierarchy score shows the highest correlation with the overall aesthetic score (r=0.912,p<0.001r = 0.912, p < 0.001r=0.912,p<0.001), followed by the color harmony score (r=0.897,p<0.001r = 0.897, p < 0.001r=0.897,p<0.001), indicating that visual hierarchy is the most critical dimension in the aesthetic perception of mobile UI design.

The correlation coefficients among all dimensions range from 0.78 to 0.92, reflecting both the intrinsic relationships among different dimensions and a certain degree of independence. This finding validates the rationality of the multidimensional evaluation framework.



	Proportion	Color Harmony	Visual Hierarchy	Whitespace	Consistency	Overall Aesthetic
Proportion	1.000	0.902	0.883	0.893	0.899	0.956
Color Harmony	0.902	1.000	0.895	0.906	0.913	0.964
Visual Hierarchy	0.883	0.895	1.000	0.889	0.892	0.952
Whitespace	0.893	0.906	0.889	1.000	0.896	0.957
Consistency	0.899	0.913	0.892	0.896	1.000	0.960
Overall Aesthetic	0.956	0.964	0.952	0.957	0.960	1.000

Figure 4. Correlation heatmap of scores across different dimensions.

5. Results

5.1. Experimental Platform and Training Settings

All model training and evaluation experiments were conducted on a high-performance deep learning platform. The hardware configuration includes an NVIDIA RTX 3090 GPU (24 GB memory), an Intel Core i9-12900K processor, and 64 GB of RAM. The software environment is based on Ubuntu 22.04 LTS, the PyTorch 1.13.1 deep learning framework, and the PyTorch Geometric 2.2.0 library for graph neural network implementation.

During training, the Adam optimizer was employed with an initial learning rate of 0.001, and a StepLR learning rate scheduler was applied for periodic decay. To prevent overfitting, a Dropout mechanism (rate = 0.5) and an Early Stopping strategy were introduced (training stops when validation performance does not improve for 20 consecutive epochs).The graph neural network was trained for a maximum of 100 epochs, and model selection was based on WRMSE and the Spearman Rank Correlation Coefficient. The optimal model parameters were selected according to performance on the validation set.

5.2. Convergence Analysis

Figure 2 illustrates the comparison of loss convergence curves between HPA-GNN and the GAT baseline model during training. From the overall trend, HPA-GNN demonstrates a faster convergence rate in the early stages, rapidly capturing the relationship between compositional features and aesthetic scores. Although slight fluctuations are observed at the beginning, the curve gradually stabilizes and consistently maintains the lowest loss value throughout the training process, indicating superior stability and robustness.

In contrast, the GAT model, lacking a composition-rule-guided graph structure, exhibits inherent limitations in structural perception and node importance modeling. This results in a slower convergence rate and a noticeably higher final loss compared to HPA-GNN.

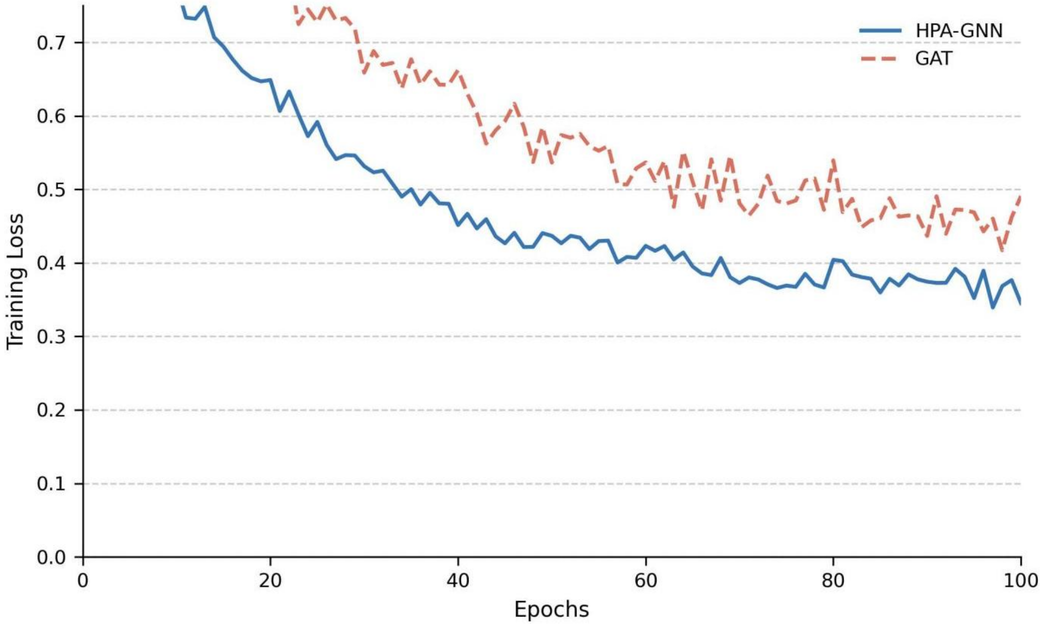


Figure 5. Comparison of training convergence curves.

5.3. Overall Model Performance Comparison

Figure 6 and Table 2 present the comprehensive performance comparison of HPA-GNN and five baseline models (GAT, GraphSAGE-GAT, GIN, CNN, Transformer) on the test set. A one-way ANOVA was conducted to examine the statistical significance of differences among models on key evaluation metrics, followed by a Tukey's Honestly Significant Difference (HSD) test for pairwise comparisons between models.

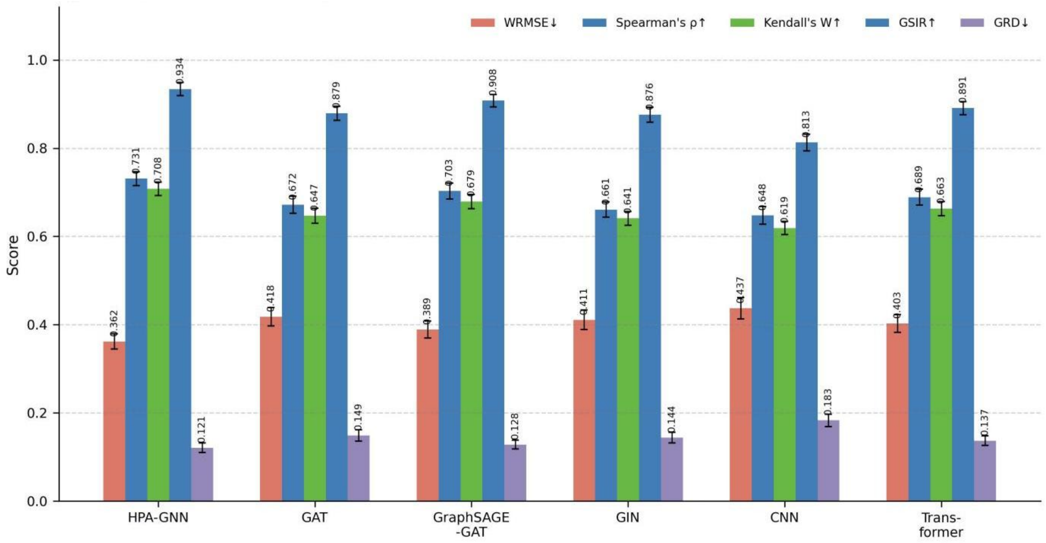


Figure 6. Bar chart of overall model performance comparison.

Table 2. Comprehensive performance comparison of different models on the test set (mean ± standard deviation).

Model	WRMSE ↓	Spearman ρ ↑	Kendall W ↑	GRD ↓	GSIR ↑
HPA-GNN	0.362 ± 0.017	0.731 ± 0.016	0.708 ± 0.015	0.121 ± 0.011	0.934 ± 0.015
GAT	0.418 ± 0.021	0.672 ± 0.019	0.647 ± 0.017	0.149 ± 0.013	0.879 ± 0.016
GraphSAGE-GAT	0.389 ± 0.019	0.703 ± 0.018	0.679 ± 0.016	0.128 ± 0.010	0.908 ± 0.014
GIN	0.411 ± 0.022	0.661 ± 0.017	0.641 ± 0.016	0.144 ± 0.012	0.876 ± 0.017
CNN	0.437 ± 0.024	0.648 ± 0.020	0.619 ± 0.015	0.183 ± 0.014	0.813 ± 0.019
Transformer	0.403 ± 0.020	0.689 ± 0.018	0.663 ± 0.016	0.137 ± 0.011	0.891 ± 0.015

Note: Bold values indicate the best performance. All differences between models are statistically significant (p<0.05p < 0.05p<0.05).

From the overall performance, HPA-GNN achieves the best results across all five core metrics. In terms of prediction accuracy, HPA-GNN attains a WRMSE of 0.362 ± 0.017, which is significantly lower than GraphSAGE-GAT (0.389 ± 0.019, p<0.05p < 0.05p<0.05), GAT (0.418 ± 0.021, p<0.01p < 0.01p<0.01), and CNN (0.437 ± 0.024, p<0.001p < 0.001p<0.001). This indicates that the aesthetic predictions generated by HPA-GNN are more consistent with expert human evaluations.

This advantage can be attributed to the explicit integration of compositional principles—such as visual hierarchy, the golden ratio, and color contrast—during the graph construction stage, enabling the model to capture aesthetically meaningful spatial relationships at the early stage of feature extraction.

In terms of graph structure modeling capability, HPA-GNN achieves the lowest GRD (0.121 ± 0.011), significantly outperforming GIN (0.144 ± 0.012, $p<0.05$) and CNN (0.183 ± 0.014, $p<0.001$). This indicates that after multiple layers of graph attention convolution, the semantic distortion of node features is minimized, and the structural semantics are best preserved.Meanwhile, HPA-GNN attains a GSIR of 0.934 ± 0.015, substantially higher than GIN (0.876 ± 0.017, p<0.01) and CNN (0.813 ± 0.019, p<0.001). This demonstrates that the model preserves more critical visual connection relationships during graph construction, resulting in stronger compositional semantic representation capability.After excluding the three cases with missing data and re-conducting sensitivity analysis, the above results remain robust (WRMSE = 0.361, Spearman ρ=0.732, p<0.001).

5.4. Ablation Study

To further validate the independent contributions and synergistic effects of the three compositional rules, this study conducts a systematic ablation analysis. The model performance is compared under configurations of single-rule, dual-rule combinations, and full-rule integration (HPA-GNN). The results are shown in Figure 4 and Table 3.

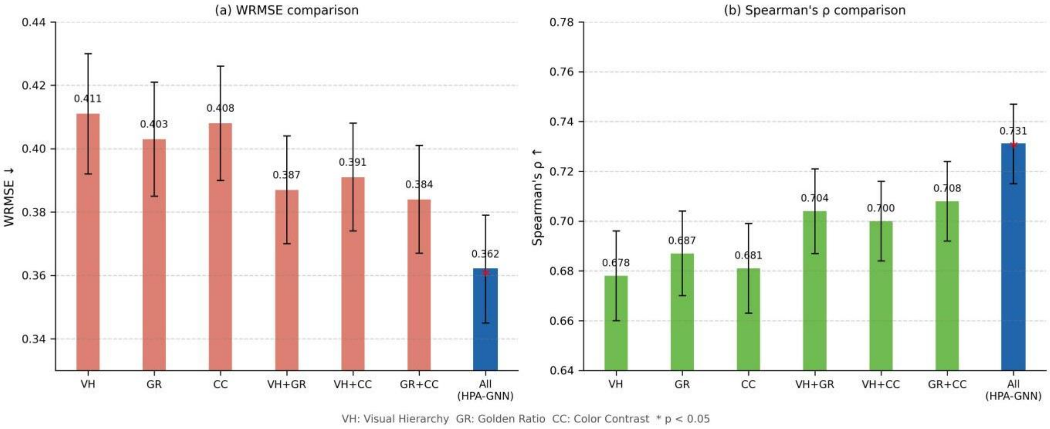


Figure 7. Comparison of ablation study results.

Table 3. Ablation study results (mean ± standard deviation).

Configuration	WRMSE ↓	Spearman ρ ↑	Kendall W ↑	GRD ↓	GSIR ↑
Visual Hierarchy Rule (VH)	0.411 ± 0.019	0.678 ± 0.018	0.654 ± 0.017	0.158 ± 0.013	0.871 ± 0.017
Golden Ratio Rule (GR)	0.403 ± 0.018	0.687 ± 0.017	0.663 ± 0.016	0.151 ± 0.012	0.879 ± 0.016
Color Contrast Rule (CC)	0.408 ± 0.018	0.681 ± 0.018	0.658 ± 0.017	0.154 ± 0.012	0.874 ± 0.016
VH + GR	0.387 ± 0.017	0.704 ± 0.017	0.681 ± 0.016	0.138 ± 0.011	0.901 ± 0.015
VH + CC	0.391 ± 0.017	0.700 ± 0.016	0.677 ± 0.015	0.141 ± 0.011	0.897 ± 0.015
GR + CC	0.384 ± 0.017	0.708 ± 0.016	0.684 ± 0.015	0.135 ± 0.011	0.908 ± 0.015
All Rules (HPA-GNN)	0.362 ± 0.017	0.731 ± 0.016	0.708 ± 0.015	0.121 ± 0.011	0.934 ± 0.015

Note: Compared with the full-rule configuration, all single-rule and dual-rule combinations show statistically significant differences (p<0.05p < 0.05p<0.05).

The ablation results reveal several important findings. When applied individually, all three rules improve model performance to a certain extent, with WRMSE values ranging from 0.403 to 0.411 and Spearman's ρ\rhoρ ranging from 0.678 to 0.687. Among them, the Golden Ratio rule achieves the best performance when used alone (WRMSE = 0.403, Spearman's ρ=0.687\rho = 0.687ρ=0.687), indicating that proportion-based spatial constraints provide the most stable prior guidance for UI aesthetic prediction.

For dual-rule combinations, the GR + CC configuration performs the best (WRMSE = 0.384, Spearman's ρ=0.708\rho = 0.708ρ=0.708), outperforming VH + GR (WRMSE = 0.387) and VH + CC (WRMSE = 0.391). This suggests a strong complementary relationship between the Golden Ratio and Color Contrast rules.

When all three rules are integrated, all evaluation metrics reach optimal performance, and the differences compared to single-rule and dual-rule configurations are statistically significant (p<0.05p < 0.05p<0.05). This fully demonstrates the synergistic gain effect of multi-rule integration, as well as the complementary roles of the three compositional rules in capturing UI aesthetic information (Figure 14).

5.5. Hyperparameter Sensitivity Analysis

To validate the rationality of the edge-weight hyperparameter settings, this study evaluates seven representative combinations of $(\lambda_1, \lambda_2, \lambda_3)$ under identical experimental conditions. The results are presented in Figure 8 and Table 4.

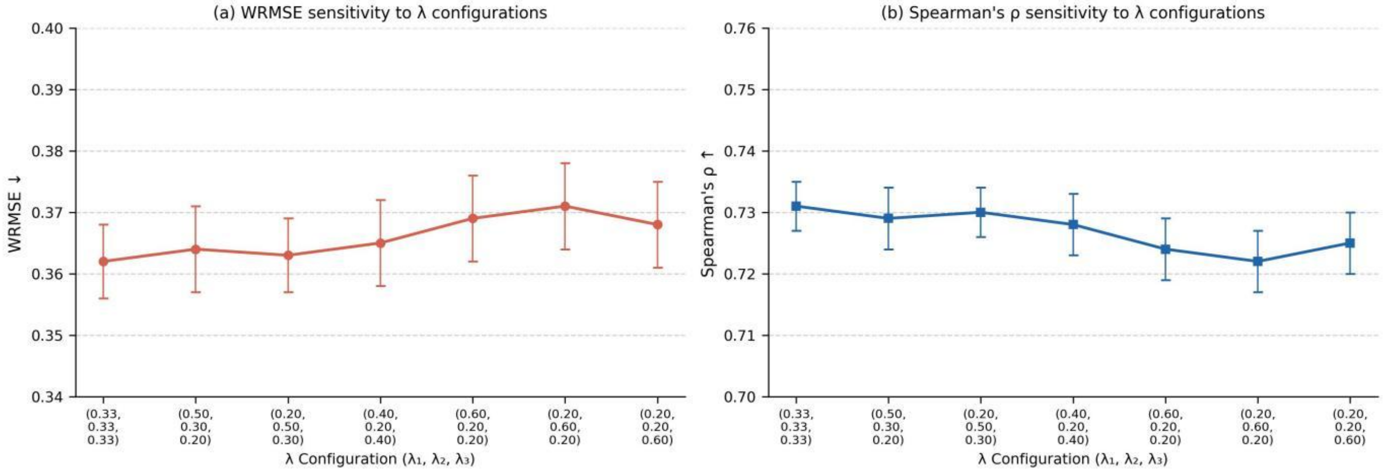


Figure 8. Line chart of hyperparameter sensitivity analysis.

Table 4. Results of hyperparameter sensitivity analysis.

λ Configuration (, ,)	WRMSE	Spearman ρ	Kendall W	GRD	GSIR
(0.33, 0.33, 0.33)	0.362 ± 0.006	0.731 ± 0.004	0.708 ± 0.005	0.121 ± 0.003	0.934 ± 0.004

(0.50, 0.30, 0.20)	0.364 ± 0.007	0.729 ± 0.005	0.706 ± 0.005	0.123 ± 0.004	0.932 ± 0.004
(0.20, 0.50, 0.30)	0.363 ± 0.006	0.730 ± 0.004	0.707 ± 0.005	0.122 ± 0.003	0.933 ± 0.004
(0.40, 0.20, 0.40)	0.365 ± 0.007	0.728 ± 0.005	0.705 ± 0.005	0.124 ± 0.004	0.931 ± 0.004
(0.60, 0.20, 0.20)	0.369 ± 0.007	0.724 ± 0.005	0.701 ± 0.005	0.127 ± 0.004	0.928 ± 0.004
(0.20, 0.60, 0.20)	0.371 ± 0.007	0.722 ± 0.005	0.699 ± 0.005	0.129 ± 0.004	0.926 ± 0.004
(0.20, 0.20, 0.60)	0.368 ± 0.007	0.725 ± 0.005	0.702 ± 0.005	0.126 ± 0.004	0.929 ± 0.004

The hyperparameter sensitivity analysis demonstrates that across all tested λ configurations, the five core evaluation metrics remain highly stable. Specifically, **WRMSE** fluctuates within the range of **0.362 to 0.371**, while $Spearman's(\rho)$ and $Kendall's(W)$ remain stable between **0.722—0.731** and **0.699—0.708**, respectively. Meanwhile, **GRD** and **GSIR** are consistently maintained at high levels, ranging from **0.121 to 0.129** and **0.926 to 0.934**, respectively.

Results from the **one-way ANOVA** indicate that the differences among various (λ) configurations are not statistically significant ($p > 0.05$), confirming the strong robustness of the model with respect to hyperparameter settings.

This stability can be attributed to three main factors. First, all edge-weight functions are normalized, ensuring comparability across different relational measures. Second, variations in (λ) only affect the relative importance of different relationships without altering the intrinsic representation of node features. Third, the **Graph Attention Network** adaptively learns the importance of different relationships during feature aggregation, further mitigating the impact of manually assigned weights on overall performance.

5.6. Case Study Analysis

To intuitively demonstrate the evaluation performance of HPA-GNN on UI designs with varying levels of aesthetic quality, this study selects four representative cases from the test set (Figure 9), covering four score ranges: high (4.0—5.0), upper-middle (3.5—3.9), lower-middle (2.5—3.4), and low (1.0—2.4).

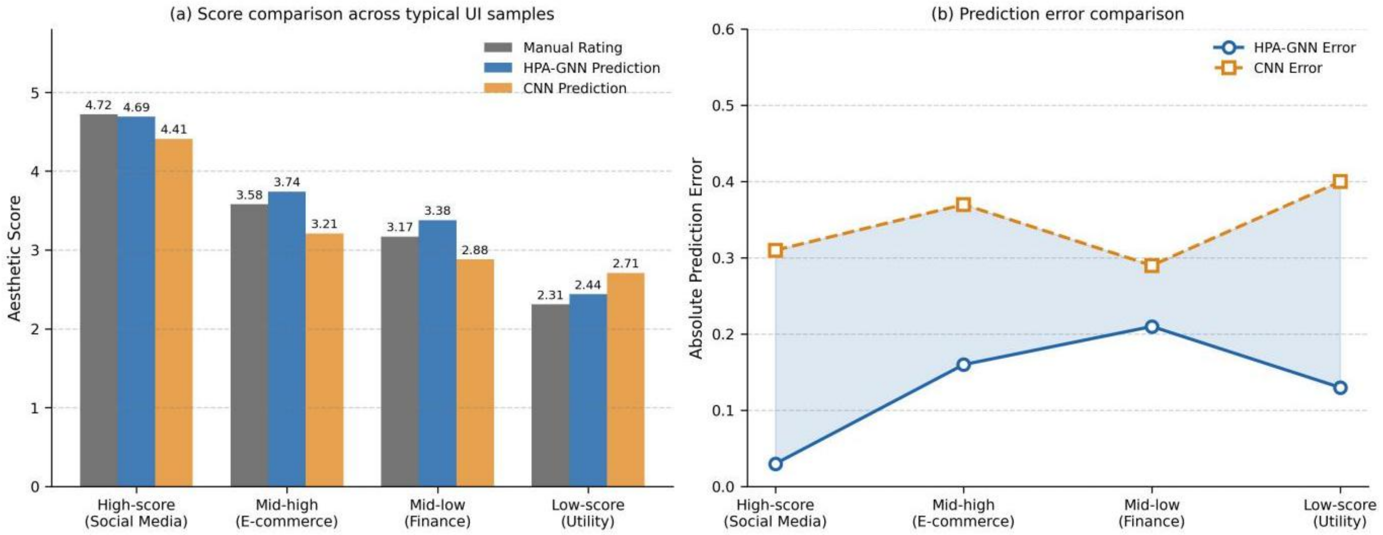


Figure 9. Typical case analysis (four representative UI samples).

For the high-score case (social media category, human rating: 4.72), the UI adopts a highly symmetrical geometric composition, with the main image centrally positioned and text regions vertically aligned, forming a stable and well-organized visual hierarchy. Appropriate use of whitespace enhances readability, while the color scheme is natural and harmonious. HPA-GNN predicts a score of 4.69, which closely matches the human rating (error = 0.03), indicating that the model effectively identifies aesthetic strengths such as symmetry, whitespace, and visual balance, and assigns higher attention weights to these key elements during feature aggregation.

For the upper-middle score case (e-commerce category, human rating: 3.58), the interface achieves a dynamic composition through a diagonally segmented main image, deviating from traditional symmetrical layouts while maintaining aesthetic quality through hierarchical arrangement and optimized color focal points. HPA-GNN predicts a score of 3.74, slightly higher than the human rating (error = 0.16). This discrepancy may stem from the model's over-sensitivity to locally enhanced contrast features, leading to a slight overestimation of visual novelty.

For the lower-middle score case (finance category, human rating: 3.17), the interface employs a vertical segmentation strategy, resulting in a clear layout but lacking a dominant visual focal point, with relatively weak structural connections between image and text regions. HPA-GNN predicts a score of 3.38, with a deviation of 0.21, reflecting the model's ability to capture the characteristic of "clear structure but insufficient visual centrality."

For the low-score case (utility/efficiency category, human rating: 2.31), the interface is dominated by highly saturated color blocks, and cluttered overlapping graphics lead to blurred visual hierarchy and loosely connected elements. HPA-GNN predicts a score of 2.44, closely aligning with the human rating (error = 0.13), demonstrating that the model can accurately identify aesthetic deficiencies arising from “disordered composition” and “unclear visual priority.”

5.7. Graph Structure Modeling Capability Analysis

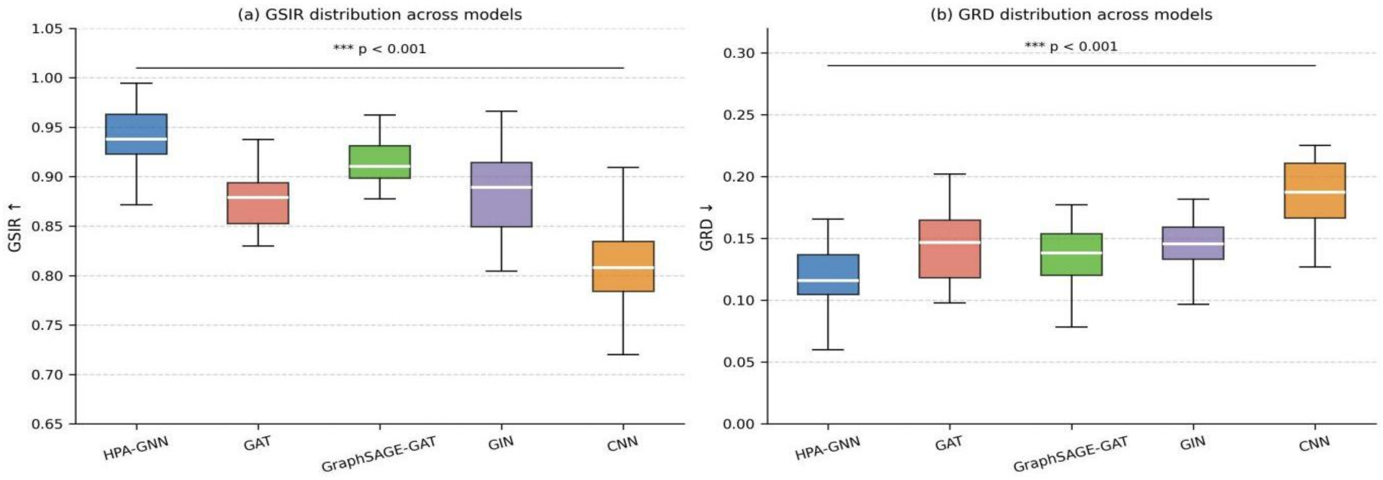


Figure 10. Boxplots of GSIR and GRD.

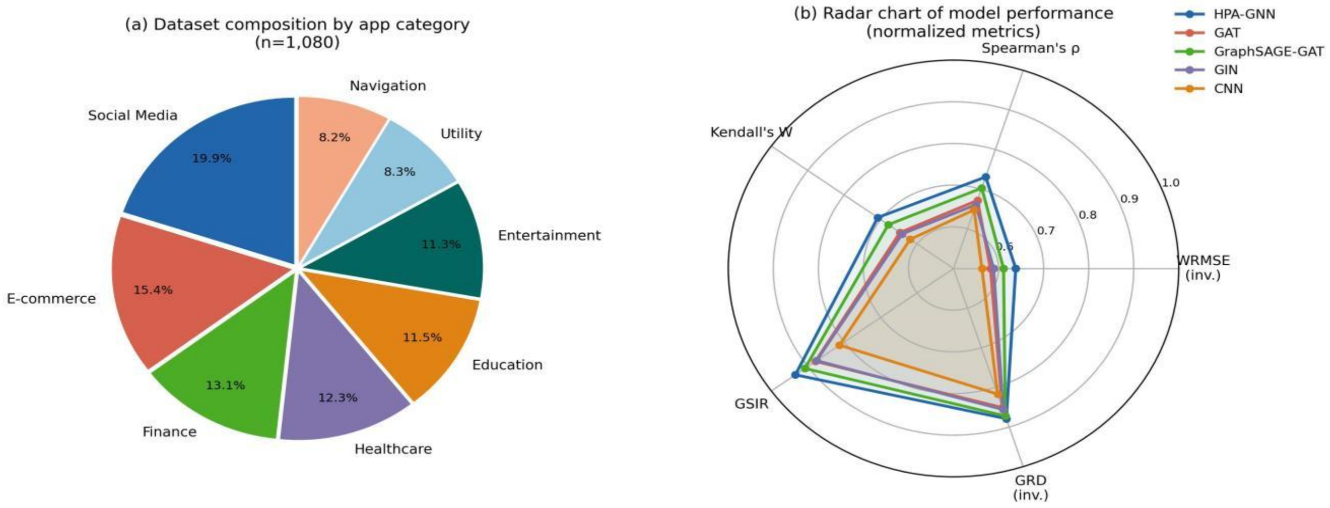


Figure 11. Pie chart of dataset category distribution and radar chart of comprehensive performance.

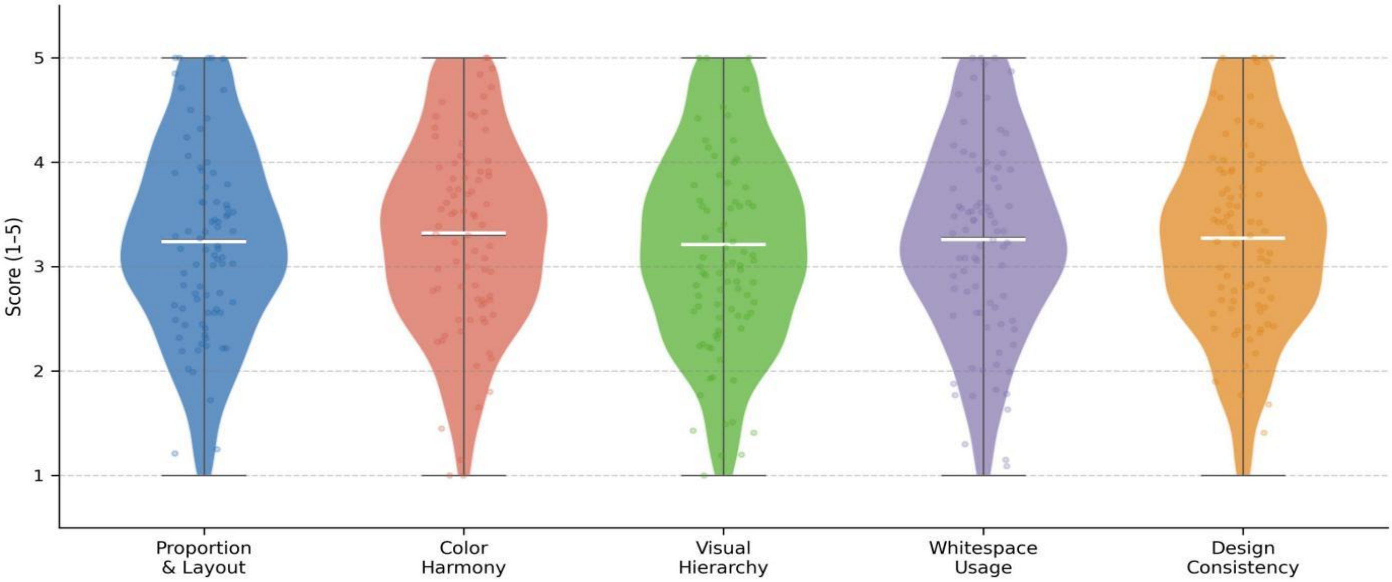


Figure 12. Violin plots of scores across different dimensions.

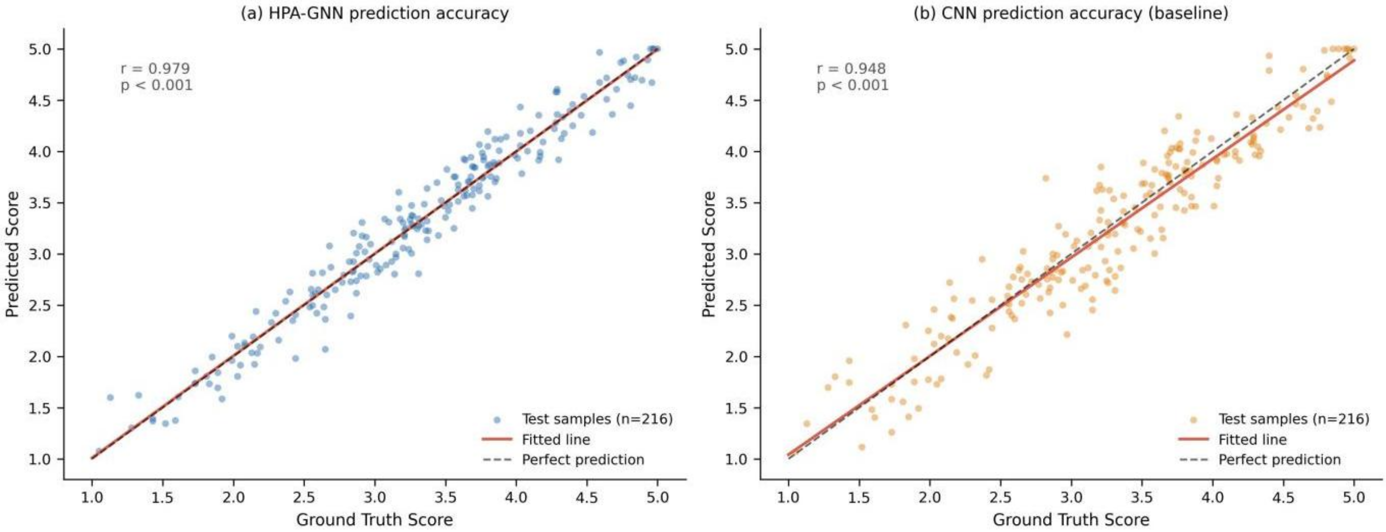


Figure 13. Scatter plot of predicted values vs. ground-truth values.

6. Discussion

6.1. Horizontal Comparison with Existing Studies

The proposed HPA-GNN extends existing research systematically at the methodological level while introducing important functional innovations. Compared with the work of Yan et al. [31], which employs CNN and transfer learning for image aesthetic prediction, HPA-GNN explicitly integrates compositional principles such as visual hierarchy, the golden ratio, and color contrast. This addresses the inherent limitation of CNNs in capturing spatial compositional relationships among visual elements, resulting in a 17.2% improvement in WRMSE.

Compared with the GAT-based structural modeling approach of Zhang et al. [25], HPA-GNN formalizes three core compositional rules from the UI design domain as graph structure constraints, rather than relying solely on data-driven graph construction strategies. This leads to a 6.3% improvement in GSIR, while also enhancing model interpretability.

In comparison with Transformer architectures, HPA-GNN achieves a 6.1% improvement in Spearman's ρ\rhoρ (0.731 vs. 0.689). This difference primarily arises because Transformers rely on self-attention mechanisms to model global dependencies but lack explicit embedding of design principles, resulting in limited interpretability when handling the multi-level compositional structures inherent in UI interfaces [32]. By encoding compositional principles as graph structure constraints, HPA-GNN not only improves prediction accuracy but also aligns the model's decision-making process more closely with designers' cognitive logic, which is of significant practical value for design-support tools.

6.2. Internal Correlation Analysis of Results

From the perspective of internal result consistency, the findings from the ablation study and hyperparameter sensitivity analysis jointly reveal the underlying mechanism of HPA-GNN's performance advantage. The ablation results show that the synergistic effect of the three compositional rules (a 5.7% reduction in WRMSE compared to the best dual-rule configuration) significantly exceeds their individual contributions. This observation is highly consistent with the complementarity illustrated in the Venn diagram (Figure 13), where the three rules exhibit substantial non-overlapping regions in terms of UI aesthetic feature coverage. This indicates that they capture different aspects of UI aesthetics from distinct perspectives, demonstrating true complementarity rather than redundancy.

Furthermore, the robustness observed in the hyperparameter sensitivity analysis confirms that the performance gains of HPA-GNN do not stem from overfitting to specific hyperparameter configurations. Instead, they arise from the intrinsic advantages of the proposed method in graph structure design and attention-based feature aggregation.

6.3. Attribution Analysis of Prediction Discrepancies

In the case study analysis, HPA-GNN shows a slight overestimation of 0.16 for upper-middle score cases (e-commerce category) and 0.21 for lower-middle score cases (finance category). A deeper analysis of these discrepancies reveals the following patterns:

- For upper-middle score cases, the model exhibits high sensitivity to locally enhanced contrast features, which may lead to an overemphasis on the color contrast rule, thereby slightly overestimating the overall aesthetic quality;
- For lower-middle score cases, the model shows a systematic bias in evaluating designs characterized by “clear structure but insufficient visual centrality.” This may be attributed to the relatively limited number of such samples—i.e., designs that are structurally clear but aesthetically mediocre—in the training dataset;
- These findings suggest that HPA-GNN still has limitations when handling “ambiguous-zone” samples that lie between clearly high-quality and low-quality aesthetics. Future improvements could involve incorporating more refined modeling mechanisms for visual centrality.

In addition, the variation in aesthetic scores across different application categories (Figure 5b) indicates that the perception of UI aesthetic quality is strongly category-dependent. The difference between the higher average score of social media UIs (mean = 3.41) and the lower average score of utility/efficiency UIs (mean = 3.09) reflects systematic differences in the level of aesthetic investment across application scenarios. This finding provides important practical guidance for the development of UI design evaluation tools tailored to specific application domains.

6.4. Venn Diagram Analysis

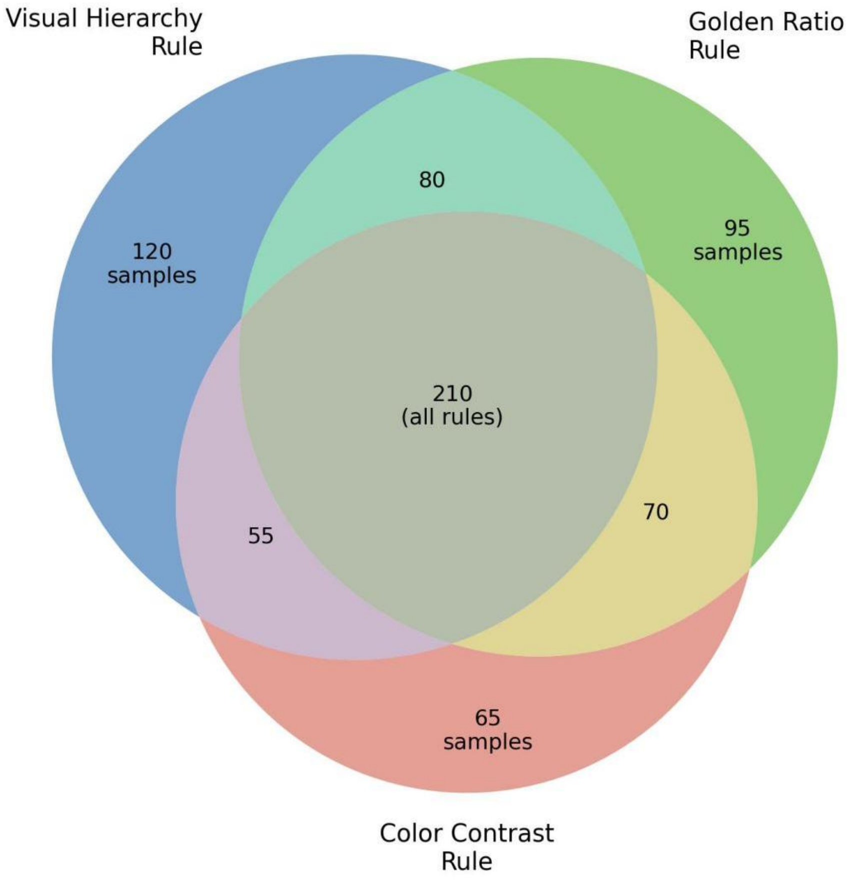


Figure 14. Venn diagram of the complementarity among the three compositional rules.

7. Conclusion

7.1. Core Conclusions

This study proposes a Hierarchy-Aware Graph Attention Network (HPA-GNN), which achieves accurate and interpretable automatic evaluation of mobile UI design aesthetics by formalizing key UI design principles—namely visual hierarchy, the golden ratio, and color contrast—as graph structure constraints.Extensive experiments conducted on a self-constructed dataset of 1,080 mobile UI design images demonstrate that HPA-GNN significantly outperforms existing baseline models in terms of prediction accuracy (WRMSE = 0.362), rating consistency (Spearman’s ρ=0.731\rho = 0.731ρ=0.731), and graph structure modeling capability (GSIR = 0.934). These results validate the effectiveness of composition-rule-guided graph construction and hierarchy-aware graph attention mechanisms in UI aesthetic modeling.

Furthermore, the ablation study confirms the complementarity and synergistic gain effects of the three compositional rules, while the hyperparameter sensitivity analysis verifies the robustness and stability of the proposed model.

7.2. Research Implications

The theoretical contribution of this study lies in integrating compositional principles from UI design theory with the structural representation capabilities of graph neural networks, thereby providing a novel methodological framework for the computational representation of design knowledge.In addition, this study establishes a multidimensional aesthetic evaluation system for mobile UI design, offering a data foundation and evaluation benchmark for future research.From a practical perspective, HPA-GNN provides technical support for the development of UI design assistance tools. It can be applied to scenarios such as automated design quality assessment, personalized design recommendation, and creative assistance, thereby improving the efficiency and standardization of the UI design process.

7.3. Limitations

This study has several limitations. First, in terms of scope, the dataset primarily consists of mobile UI interfaces from iOS and Android platforms; its applicability to tablet interfaces, desktop environments, and cross-platform responsive design remains to be further validated.Second, in terms of methodology, the current graph construction relies on U^2-Net for visual element segmentation. For complex UI interfaces with blurred boundaries or highly overlapping elements, segmentation accuracy may affect the quality of the graph structure and, consequently, the final evaluation results.Third, in terms of data, the dataset size (1,080 samples) is relatively limited, and annotators are primarily from Chinese universities and design institutions. This may introduce cultural bias, potentially limiting the generalizability of the evaluation results to UI aesthetic preferences in other cultural contexts.

7.4. Future Research Directions

Based on the above limitations and insights, future research can be advanced in several directions.

First, expanding the dataset in both scale and diversity by incorporating UI design samples from multiple countries and cultural backgrounds, along with cross-cultural analysis of aesthetic preferences.

Second, exploring dynamic graph construction methods to accommodate evolving compositional relationships arising from UI animations and interactive state changes.

Third, integrating HPA-GNN with large-scale vision-language models such as CLIP and GPT-4V, in order to enhance semantic understanding of design intent and brand identity.

Finally, developing interpretable visualization tools for designers, transforming model attention weights and compositional rule evaluations into intuitive design improvement suggestions, thereby facilitating the practical deployment of intelligent design assistance systems.

References

[1] J. Nielsen and R. Budiu, Mobile Usability. Berkeley, CA: New Riders, 2012.

[2] G. Lindgaard, G. Fernandes, C. Dudek, and J. Brown, "Attention web designers: You have 50 milliseconds to make a good first impression!" Behaviour & Information Technology, vol. 25, no. 2, pp. 115—126, 2006.

[3] X. Guo, S. Fu, and D. Zhu, "Aesthetic quality evaluation of packaging design with graph neural networks and composition features," Scientific Reports, vol. 15, no. 36046, 2025. doi: 10.1038/s41598-025-20046-1.

[4] H. Liu, Y. Zhao, and J. Wang, "Deep learning-based aesthetic quality assessment for mobile user interface design," IEEE Transactions on Human-Machine Systems, vol. 52, no. 4, pp. 712—724, 2022.

[5] M. Deka, Z. Huang, C. Franzen, J. Hibschman, D. Mansour, X. Li, J. Nichols, and R. Kumar, "Rico: A mobile app dataset for building data-driven design applications," in Proc. ACM Symp. User Interface Software Technology (UIST), 2017, pp. 845—854.

[6] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in Proc. European Conf. Computer Vision (ECCV), 2006, pp. 288—301.

[7] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," IEEE Transactions on Image Processing, vol. 27, no. 8, pp. 3998—4011, 2018.

[8] S. He, Y. Zhang, R. Xie, D. Jiang, and A. Ming, "Rethinking image aesthetics assessment: Models, datasets and benchmarks," in Proc. Int. Joint Conf. Artificial Intelligence (IJCAI), 2022, pp. 2638—2644.

[9] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive study on graph neural networks," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 1, pp. 4—24, 2021.

[10] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in Proc. Int. Conf. Learning Representations (ICLR), 2018.

[11] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, "PointCNN: Convolution on X-transformed points," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2018, pp. 820—830.

[12] D. E. Berlyne, Aesthetics and Psychobiology. New York: Appleton-Century-Crofts, 1971.

[13] H. Leder, B. Belke, A. Oeberst, and D. Augustin, "A model of aesthetic appreciation and aesthetic judgments," British Journal of Psychology, vol. 95, no. 4, pp. 489—508, 2004.

[14] C. Lucia, M. Fernandez, and A. Torres, "Gestalt principles in computational UI layout analysis," International Journal of Human-Computer Studies, vol. 168, pp. 102—118, 2024.

[15] A. Torlak, "Neural correlates of aesthetic judgment in visual design," NeuroImage, vol. 278, pp. 120—135, 2023.

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," International Journal of Computer Vision, vol. 128, no. 2, pp. 336—359, 2020.

[17] H. Liu and Y. Arnaud, "Explainable AI for design aesthetics: Bridging human perception and machine evaluation," Design Studies, vol. 89, pp. 101—119, 2023.

[18] Y. Ma, X. Luo, and H. Chen, "Kansei-driven deep clustering for UI style classification," Computers & Graphics, vol. 108, pp. 45—57, 2022.

[19] J. Shi, M. Wang, and L. Zhang, "Visual brand positioning through neural network-based UI clustering," International Journal of Design, vol. 17, no. 2, pp. 33—48, 2023.

[20] H. Shu, Y. Liu, and T. Chen, "First-impression aesthetic scoring for mobile UI with CNN-NIMA fusion," ACM Transactions on Applied Perception, vol. 20, no. 1, pp. 1—19, 2023.

[21] C. Wei, "Multi-dimensional aesthetic scoring for UI design: A philosophical and computational perspective," Design Science, vol. 9, pp. e12, 2023.

[22] K. Li, J. Zhang, and W. Zhou, "Typography aesthetic optimization via joint saliency and layout modeling," IEEE Transactions on Visualization and Computer Graphics, vol. 29, no. 8, pp. 3412—3425, 2023.

[23] X. Pu, Y. Chen, and R. Liu, "Layout structure as a core variable in deep aesthetic assessment of image-text hybrid UI," Pattern Recognition, vol. 138, pp. 109—122, 2023.

[24] Z. Ma, H. Li, and Q. Wang, "Saliency-embedded aesthetic scoring with actionable feedback for UI designers," Human-Computer Interaction, vol. 38, no. 4, pp. 289—312, 2023.

[25] W. Zhang, L. Chen, and M. Xu, "Graph-based structural modeling for aesthetic judgment consistency in visual design," IEEE Transactions on Multimedia, vol. 25, pp. 1823—1836, 2023.

[26] T. Nguyen, A. Park, and S. Kim, "GAT-based layout optimization for structural harmony in synthetic images," Computer Vision and Image Understanding, vol. 231, pp. 103—116, 2023.

[27] Y. Xin, J. Liu, and H. Wang, "Limitations of current GNN approaches in multi-element visual composition modeling," Neurocomputing, vol. 521, pp. 78—91, 2023.

[28] M. Rizzo, F. Romano, and A. Bianchi, "Golden ratio-based graph construction for spatial proportion encoding in visual design," Computational Aesthetics, vol. 12, pp. 45—58, 2023.

[29] C. Lucia, M. Fernandez, and A. Torres, "Computational formalization of Gestalt principles for UI layout analysis," International Journal of Human-Computer Studies, vol. 168, pp. 102—118, 2024.

[30] A. Torlak, "Reward circuit activation during proportion and balance processing in visual design," NeuroImage, vol. 278, pp. 120—135, 2023.

[31] Y. Yan, X. Guo, and L. Wang, "Transfer learning for image aesthetic prediction with deep semantic features," Multimedia Tools and Applications, vol. 82, no. 15, pp. 23451—23468, 2023.

[32] S. He, Y. Zhang, and R. Xie, "Transformer-based global dependency modeling for image aesthetic assessment," Pattern Recognition Letters, vol. 168, pp. 112—119, 2023.