Int. J. Simul. Multidisci. Des. Optim. **16**, 12 (2025)
© J. Liu, Published by EDP Sciences, 2025
https://doi.org/10.1051/smdo/2025016

IJSMDO

Available online at:
https://www.ijsmdo.org

Multi-modal Information Learning and Analytics on Cross-Media Data Integration

Research Article

Open Access

Modeling analysis of perceived differences in product appearance design based on visual communication

Jing Liu*

Academy of Art and Design, Jingdezhen Ceramic University, Jingdezhen 333000, Jiangxi, China

Received: 16 June 2025 / Accepted: 18 August 2025

Abstract. Aiming at the problems of reliance on subjective evaluation and imperfect index systems in the modeling of perceived differences in product appearance design, this study proposes a quantitative analysis method of improving the Contrastive Language-Image Pre-training (CLIP) model. Based on the theoretical framework of visual semiotics, the visual communication mechanism is analyzed from the dimensions of morphological layer (physical structure of color and form), semantic layer (symbolic meaning of color symbolism and morphological metaphor), and pragmatic layer (user cognitive logic and cultural background). First, the channel attention mechanism is introduced to construct a joint embedding space of color, form, and semantics. The geometric features are extracted from the morphological grammar. The color and texture combinations with strong emotional relevance are screened based on the design semantics theory. The key design elements are deconstructed from a theoretical level to enhance the ability to align visual and textual features. Secondly, the CLIP model is improved to extract fine-grained visual-text embedding, and a 12-dimensional perceptual difference matrix is generated through cosine similarity. A dual-path feature interaction module is designed: the visual path uses Grad-CAM++ to locate key morphological areas, and the semantic path improves the text style transfer capability through adversarial training, and uses GRU to achieve cross-modal dynamic fusion. The gradient weighted class activation mapping technology is further combined to visualize the model's attention area, and the contribution weights of color, texture, shape and other factors to the perceptual difference are quantified through Shapley value decomposition. Experiments show that this method reduces the error in cross-cultural perceptual prediction, the top-5 accuracy of cross-modal matching is as high as 0.85, and the median MSE of perceptual difference quantification is reduced to 0.13, which effectively solves the problems of insufficient objective quantification and a lack of an indicator system.

Keywords: Product appearance design / cross-modal feature fusion / perceptual difference quantification / channel attention mechanism / dynamic temperature coefficient optimization

1 Introduction

In the context of the digital economy, product design has become one of the core factors that influence consumers' purchasing decisions [1]. From the perspective of design semiotics, the metaphor of product form (such as streamlined shape implies speed) and the symbolic meaning of color (such as red conveys passion) constitute the core mechanism of visual communication. Cognitive psychology research shows that users form multi-dimensional perceptions through Gestalt principles and emotional design [2]. According to a report by the International Council of Industrial Designers, most consumers are initially interested in purchasing a product based on its appearance, and half of their purchases ultimately depend on the continued appeal of the design experience [3]. However, with the exponential growth in the complexity of industrial design, the traditional evaluation model that relies on designers' experience and judgment and focus group interviews can no longer meet the needs of accurate design [4]. Currently, academia and industry generally face three major challenges: First, individual bias in subjective evaluation leads to design decision risks [5]. Second, the lack of a quantitative indicator system restricts the design optimization path [6], and there are too few operable and reproducible indicators in the existing evaluation system [7]. Third, cross-cultural perceptual differences are difficult to model effectively [8], especially in highly competitive fields such as smart hardware and consumer electronics [9]. How to transform fuzzy perceptual cognition into calculable design parameters has become a key scientific issue in the intersection of human-computer interaction and industrial design [10–11]. The solution to these problems

* e-mail: 041014@jcu.edu.cn

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

is related to the realization of the commercial value of the product, and more directly affects the transformation process of Chinese manufacturing to Chinese design.

In response to the above problems, the academic community has launched a multi-dimensional exploration. Based on the morphological analysis method and the design semantics framework, this study deconstructs the product appearance design elements into three core dimensions: color, texture, and shape. Singh et al. [12] quantified the aesthetic appropriateness of beverage bottle contours through empirical research based on eye tracking [13,14], revealing consumers' proficiency in functional perception and gender-related visual attention differences [15]. This method solves the problem of local quantification of the correlation between product form and function, but fails to address the core challenges of dynamic modeling of cross-modal (visual-text) perceptual differences, global quantification of cross-cultural design features, and the lack of an objective basis for subjective evaluation. By combining convolutional neural networks with search neural networks [16,17], Ding et al. [18] achieved complex correlation modeling and design scheme generation between product color and user emotional imagery [19]. This solved the problem that traditional color design relies on subjective experience and lacks a quantitative basis. However, it failed to effectively deal with cross-cultural perceptual differences or multimodal feature fusion [20], and the dynamic adaptability of its model is still limited. In order to address the limitations of traditional qualitative analysis of emotional design, Jin et al. [21] proposed a framework for extracting emotional needs from online reviews and prioritizing product features by integrating Kansei engineering and the Kano model [22,23]. However, they failed to effectively combine visual features for cross-modal alignment, resulting in the quantitative analysis of design elements still relying on a single text modality, which is difficult to directly map to the optimization of visual parameters of product appearance. Although the above studies have made progress in local quantification and emotional association modeling, they have not effectively solved the core challenges of dynamic modeling of cross-modal (visual-text) perceptual differences, global quantification of cross-cultural design features, and lack of objective basis for subjective evaluation, which makes it difficult to directly map design elements to computable visual parameter optimization.

It is worth noting that the breakthrough in cross-modal representation learning provides a new paradigm for solving this dilemma. Wang et al. [24] solved the problems of high output error rate and low signal-to-noise ratio in traditional methods by constructing a product appearance perception difference model, adjusting design element information and optimizing the model structure. However, this study failed to explicitly deal with the dynamic modeling of cross-cultural perception differences, and did not involve the fine-grained interaction mechanism of multimodal features [25]. Although cross-modal representation learning provides a new paradigm for solving the above dilemma, existing methods still have problems such as insufficient cross-cultural dynamic modeling and a lack of a visual-text fine-grained interaction mechanism, making it difficult to achieve quantitative attribution of design factor contribution and improve model interpretability.

This study establishes a dynamic quantitative model of product appearance design perception differences and proposes a multi-granularity feature fusion architecture based on the attention mechanism. Compared with traditional methods, the CLIP model has a natural advantage in visual-text alignment through large-scale cross-modal pre-training. Its framework based on contrastive learning can directly establish semantic associations between images and text descriptions. Compared with the Kansei project that relies on manually labeled sentiment vocabulary mapping, CLIP uses a joint embedding space trained through 400 million pairs of image-text data to more accurately capture the implicit relationship between design elements such as color and shape and user perception. Its contrastive learning framework is naturally adapted to the visual-text cross-modal alignment task by maximizing the similarity of positive samples and minimizing the interference of negative samples, while the dynamic temperature coefficient optimization adjusts the sharpness of feature distribution through information entropy adaptively, effectively alleviating the difficulty of capturing long-tail design features. This pre-training capability provides a computable semantic alignment basis for dynamic modeling of perceptual differences, effectively solving the problem of insufficient cross-cultural adaptability caused by traditional methods relying on static sentiment dictionaries. The specific implementation path includes three core steps: building a multilingual dataset containing millions of product images and user reviews. The CLIP contrastive learning framework can be improved, and dynamic temperature coefficients and channel attention modules can be introduced. Finally, an interpretable analysis tool based on SHAP (Shapley Additive exPlanation) values is developed to achieve quantitative attribution of design elements. Through empirical research in typical scenarios such as consumer electronics and household products, the application value of the model in cross-cultural perception prediction, design defect tracing, competitive product comparison analysis, etc., has been verified, providing new methodological support for the intelligent design evaluation system.

2 Cross-modal perception difference modeling

2.1 Construction of multi-scale semantic space of product appearance

This study builds a cross-modal semantic space based on the ViT-B/16 variant of the CLIP architecture, and achieves precise alignment of local visual features and text descriptions by improving the model structure and training strategy. The product appearance design elements focused on in this study are limited to the three core dimensions of color, texture, and shape, which are in line with the classification specifications of the basic elements of industrial design in the ISO 11684 standard. In terms of

visual encoder transformation, a three-level scale division strategy is adopted: macroscopic morphology (overall contour curvature), mesoscopic texture (distribution of surface details), and microscopic color (pixel-level hue gradient). The channel attention module (Squeeze-and-Excitation Block) is embedded after the multi-head self-attention layer of the Transformer block. The spatial dimension of the feature map is compressed through global average pooling, and the channel weight coefficient is generated through a two-layer fully connected network. The ReLU activation function and the Sigmoid gating mechanism are used to reconstruct the feature response.

In the color subspace construction, the micro-scale color analysis is particularly strengthened. The input image is converted from the RGB (Red, Green, Blue) color space to the Lab uniform color space. The ab chromaticity plane is quantized to 32 levels using the Otsu threshold method to generate a color histogram vector (dimension 1×64). The fully connected layer is used to map to the joint embedding space, and the temperature coefficient is introduced to adjust the degree of distribution sharpness. The morphological subspace processing uses the improved Grad-CAM++ (Gradient-weighted Class Activation Mapping) algorithm, focusing on the extraction of macroscopic morphological features, calculating the back-propagation gradient based on the last layer of convolutional feature maps, generating a morphological saliency map, and extracting key contour areas through binarization, which is then reduced to a 1×256 feature vector through maximum pooling.

The semantic subspace construction uses professional dictionaries in the design field to optimize text encoding.

Traditional design symbols are rooted in culturally embedded visual conventions, where form and color follow established aesthetic norms transmitted through historical practice. These symbols operate within a stable semiotic system, relying on shared cultural cognition for meaning interpretation. In contrast, modern digital symbols emerge from dynamic user-interface interactions, characterized by abstraction, functional minimalism, and real-time adaptability. Their semantic construction is influenced by algorithmic mediation and platform-specific design languages, leading to accelerated evolution and context-dependent interpretation. The integration of both symbol types into the semantic subspace enables a comprehensive representation of perceptual stimuli, accounting for continuity in cultural memory and responsiveness to contemporary digital experience. This expanded framework aligns with the theoretical scope of visual semiotics by incorporating diachronic and synchronic dimensions of signification, thereby strengthening the model's capacity to interpret diverse design expressions across temporal and cultural contexts.

Aiming at the texture features of the mesoscale, the hierarchical constraint mechanism of cultural symbols is supplemented. First, the TF-IDF (Term Frequency-Inverse Document Frequency) algorithm is used to select high-frequency semantic feature words, and the domain-specific word vectors are trained using Word2Vec. The context-aware text embedding is generated through the BiLSTM (Bi-directional Long Short-Term Memory) network (hidden layer dimension 512). In the training phase, a dynamic hard example mining strategy is introduced. Part of the text description is randomly replaced in each batch and the triplet loss is calculated. The hard negative sample mining threshold is set to 0.85.

The joint embedding space optimization adopts a hierarchical contrastive learning framework and defines a multi-scale alignment loss function:

$$L_{\text{total}} = \alpha L_{\text{global}} + \beta \sum L_{\text{local}} + \gamma L_{\text{semantic}}. \tag{1}$$

Among them, α, β, and γ are hyperparameters, which are determined by grid search and control the weights of global alignment, local alignment, and semantic constraints respectively. Their proportional relationship is determined by a grid search. The local alignment loss L_{local} is calculated based on the regional feature similarity weighted by the channel attention weight, and the semantic loss term L_{semantic} introduces the hierarchical constraints of hierarchical labels. The global alignment loss L_{global} adopts the InfoNCE loss, and its loss function is:

$$L_{\text{global}} = -\log \frac{\exp\Big(\text{Sim}(v_{\text{global}}, t_{\text{global}})/\tau\Big)}{\sum_{k=1}^{K} \exp\Big(\text{Sim}(v_{\text{global}}, t_k)/\tau\Big)}. \tag{2}$$

Among them, $\boldsymbol{\upsilon}_{\text{global}}$ and t_{global} are the overall feature vectors output by the visual encoder and text encoder respectively. t_k is the negative sample text feature. $\boldsymbol{\tau}$ is the temperature coefficient, which is set to 0.05. Sim represents the cosine similarity. This loss function enhances the global semantic alignment ability by maximizing the similarity of positive samples and minimizing the interference of negative samples.

The feature alignment process adopts a dual-tower structure training strategy, and the parameters of the visual encoder and text encoder are updated independently. The visual branch uses the AdamW optimizer (initial learning rate 1e-4, weight decay 1e-2), and the text branch uses cosine annealing scheduling with a period of $T = 200$ epochs. The training data contains product image-text pairs (covering home appliances, digital products, and furniture), and fine-grained attribute labels are established through offline annotation processing.

Figure 1 shows the construction framework of the multi-scale cross-modal semantic space, which revolves around the visual communication of product appearance design and aims to solve the problem of strong subjectivity and lack of objective quantitative basis in traditional design evaluation. The framework is based on the improved CLIP model, and enhances the alignment ability of local visual features and text descriptions by introducing a channel attention mechanism, and constructs a joint embedding space from three levels: color, morphology, and semantics. The visual encoder extracts ViT-B/16 features from RGB images and generates multi-scale visual representations by combining Lab color quantization histograms and Grad-CAM++ morphological saliency maps. The semantic encoder uses TF-IDF to filter high-frequency words and combines Word2Vec with BiLSTM

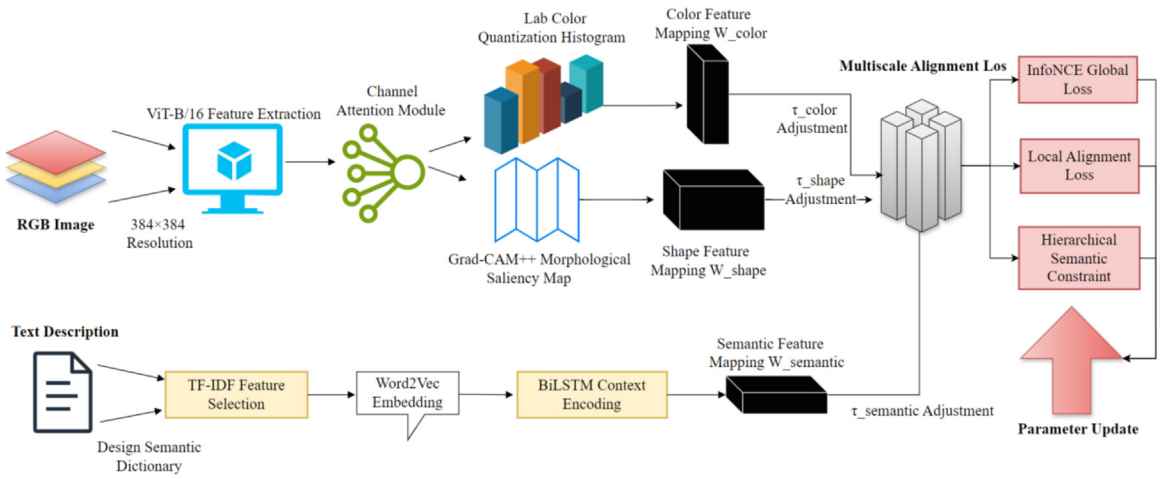

Fig. 1. Multi-scale cross-modal semantic space construction framework.

context encoding to improve the semantic expression ability of text descriptions. Finally, the visual-semantic matching relationship is optimized through the multi-scale alignment loss function in the joint embedding space, and the dynamic fusion of cross-modal features and perceptual difference modeling is realized, providing a structured and quantifiable technical foundation for subsequent product appearance design analysis.

2.2 Constructing the perceptual difference quantification matrix

This study uses an improved visual-language pre-training framework (ViT-B/16 architecture) to optimize cross-modal feature alignment through a channel attention mechanism. The model input uses a 384×384 resolution RGB image as the visual input, and the text encoder receives a text description sequence generated by a design semantic dictionary built based on the ISO 11684 standard. In the feature extraction stage, the patch feature matrix output by the visual encoder is weighted by the channel attention module to generate a visual embedding vector with spatial positioning capabilities. The text encoder generates context-aware text embeddings through the BERT-wwm (Bidirectional Encoder Representations from Transformers-Whole Word Masking) architecture, and its sequence length is dynamically adjusted according to the text content.

In order to achieve fine-grained feature matching, a dynamic alignment module DAM (Dynamic ALIGNment Module) is constructed. DAM establishes the association strength between local image regions and text keywords through a multi-head cross-attention mechanism. The calculation formula of its weight matrix A is as follows:

$$A = \mathrm{softmax}\left(\frac{(W_Q T)(W_K V)^T}{\sqrt{d_k}}\right). \tag{3}$$

Visual feature $\mathrm{V} \in \mathbb{R}^{\mathrm{N_v} \times \mathrm{d}}$ and text feature $\mathbf{T} \in \mathbb{R}^{\mathrm{Nt} \times \mathrm{d}}$ are defined, $\mathrm{N_v}$ and Nt are feature dimensions, **d** is the number of feature channels, $W_Q \in \mathbb{R}^{\mathrm{d} \times d_k}$ and $W_K \in \mathbb{R}^{\mathrm{d} \times d_k}$ are learnable query and key parameter matrices, and d_k is the scaling factor. This module uses a multi-head cross-attention mechanism with h=8 heads to calculate the association weight matrix between visual features and text features, reflecting the semantic association strength between local areas of the image and text description keywords. The cross-modal feature representation is obtained through weighted aggregation, and finally the cross-modal feature matrix is obtained.

The perceptual difference quantification adopts a multi-granularity similarity calculation framework. Three evaluation domains are defined: color, form, and semantics. Each domain has four characteristic dimensions: the color domain includes hue distribution, saturation gradient, brightness contrast, and color harmony. The form domain covers contour curvature, geometric complexity, proportional coordination, and symmetry. The semantic domain involves functional metaphors, cultural symbols, emotional associations, and style unity. After normalization of the feature vectors of each dimension, the cross-modal similarity matrix is calculated. As shown in Table 1, this study defines specific parameter ranges and weights for key features in the three domains of color, morphology, and semantics.

Table 1 shows the perceptual difference quantization parameter system based on the improved CLIP model, which constructs feature parameters and weight allocation mechanisms around the three major evaluation domains of color, morphology, and semantics. The color domain achieves accurate characterization of color distribution through the 32-level quantization hue distribution in the Lab chromaticity space (weight 0.25) and the ab plane gradient standard deviation (weight 0.15). The morphological domain uses Canny edge detection to calculate the mean value of contour curvature (weight 0.3) to quantify geometric characteristics, while the semantic domain constrains the text feature space through the word vector cosine similarity threshold (weight 0.18) and the BERT sentiment dictionary matching strength (weight 0.12). The parameter range is normalized and embedded with a dynamic temperature coefficient to adjust the distribution sharpness. This parameter system converts visual-text

Table 1. Parameter settings of the perceptual difference quantification model.

Parameter category	Feature dimension	Parameter description	Numerical range	Weight assignment
Color Domain	Hue Distribution	32-level quantization in Lab color space	[0,1]	0.25
	Saturation Gradient	Standard deviation of gradient in the ab chromatic plane	[0.1,2.0]	0.15
Morphological Domain	Contour Curvature	Mean curvature based on Canny edge detection	[0.5, 5.0] (pixel^{-1})	0.3
Semantic Domain	Cultural Symbols	Cosine similarity threshold of word vectors	[0.6, 0.9]	0.18
	Emotional Association	Matching the intensity of the BERT sentiment dictionary	[0, 1]	0.12

embedding into a 12-dimensional perceptual difference matrix through the coordinated optimization of a multi-scale alignment loss function and channel attention mechanism. This solves the problems of strong subjectivity, large cross-modal semantic gap and lack of quantitative indicators in traditional design evaluation, and provides a structured computational framework for interpretable analysis of product appearance design elements.

In order to improve the stability of matrix calculation, the temperature parameter is introduced for similarity scaling. In order to enhance the distribution-sharpening effect of the feature space, this paper proposes a dynamic temperature coefficient τ_{dynamic}, and its adaptive adjustment formula is:

$$\tau_{\text{dynamic}} = \tau_0 \cdot \exp\left(-\lambda \cdot \mathrm{H}\left(\frac{\|A\|_F^2}{\mathrm{tr}(A^T A)}\right)\right) \tag{4}$$

τ_0 is the initial temperature value, λ is the attenuation coefficient, H is the information entropy function, F is the Frobenius norm, and tr $(\cdot)$ represents the matrix trace. The triplet loss function is used to optimize the feature space distribution and define the triplet loss Ltriplet for hard negative sample mining:

$$\mathcal{L}_{\text{triplet}} = \sum_{i=1}^{B} \max\Big(0, \alpha + \mathrm{s}(v_i, t_i^-) - \mathrm{s}(v_i, t_i^+)\Big). \tag{5}$$

Among them, B is the batch size, s is the cosine similarity, t_i^+ is the positive sample, t_i^- is the difficult negative sample that satisfies $s(v_i, t_i^-) > \theta$, α is the interval threshold, and θ is the mining threshold.

As a key external variable influencing visual perception, the lighting environment is incorporated into a standardized process during the data preprocessing phase. The input image is first decomposed into reflectance and illuminance components using an illumination separation module based on Retinex theory. The illuminance component is then subjected to adaptive histogram equalization to eliminate brightness deviations under extreme lighting conditions. The normalized image is converted to the Lab color space to ensure that subsequent color feature extraction is unaffected by fluctuations in ambient light intensity. An illumination-sensitive weight branch is introduced into the channel attention mechanism to dynamically adjust the feature response strength of highlight and shadow regions. This processing flow is integrated into the visual encoder front-end as a precursor to constructing the perceptual difference quantization matrix, ensuring feature consistency across lighting conditions. During the experimental verification phase, an illumination change perturbation test was incorporated into the ten-fold cross-validation to evaluate the output stability of the model under varying illumination gradients and verify the environmental adaptability of the quantization results.

The difficult example mining strategy is adopted during the training process, and the feature pairs with similarity deviation exceeding the threshold are selected for key optimization in each batch. The final output perceptual difference matrix is normalized, and the matrix elements represent the perceptual difference intensity between different feature dimensions. The diagonal elements reflect the consistency of each dimension itself, and the off-diagonal elements represent the cross-dimensional interaction effect.

Figure 2 is a visualization of the perceptual difference quantification matrix, which shows the cross-modal perceptual difference intensity of 12 features of product appearance in three dimensions: color, shape, and semantics in the form of a heat map. The diagonal of the matrix reflects the high consistency of each feature itself, while the non-diagonal reveals the interaction effect between different features, such as the significant correlation between hue distribution and contour curvature or emotional association and style consistency.

Figure 2 intuitively reflects the ability of the improved CLIP model in fine-grained visual-text embedding, which converts the user's vague perceptual cognition into quantifiable design parameters, thus solving the problem

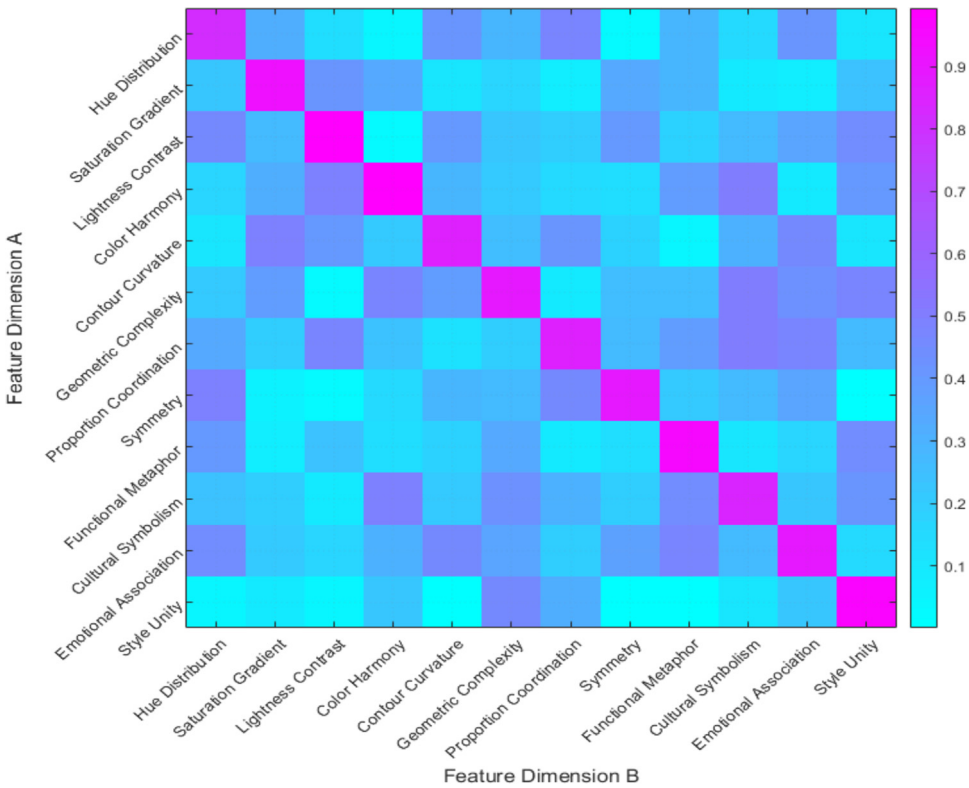

Fig. 2. 3D Perceptual difference heat map.

of traditional reliance on subjective evaluation without an objective basis, and providing data support for modeling cross-cultural perception differences.

2.3 Optimizing the multi-task feature fusion network

This study adopts a dual-path feature interaction architecture, designs dedicated feature enhancement modules for the visual domain and the semantic domain respectively, and realizes cross-modal dynamic fusion through a recurrent neural network. In terms of model optimization, the Bayesian optimization algorithm is used to adaptively adjust key parameters such as the dynamic temperature coefficient and channel compression ratio. The NSGA-II (Nondominated Sorting Genetic Algorithm II) algorithm is used to achieve collaborative optimization of color, morphology, and semantic domain parameters. The temperature coefficient adaptively adjusts the sharpness of the feature distribution through information entropy, effectively alleviating the problem of capturing long-tail design features. In the visual path feature enhancement module, the morphological feature localization network based on Grad-CAM++ is constructed based on the visual feature matrix output by the ViT-B/16 image encoder of the improved CLIP. First, the [CLS] tag feature vector output by the visual Transformer is weighted and summed with the feature vector of each image block to generate the category activation map CAM heat map. The gradient of the output feature to the input image is calculated through back propagation to obtain the channel weight coefficient, and the gradient-weighted feature is nonlinearly activated using the ReLU function to generate the initial heat map. The adaptive Otsu threshold segmentation algorithm is used to extract the salient areas of the heat map, and the morphological opening operation is combined to eliminate discrete noise points. The adjacent salient areas are connected through the closing operation, and finally a morphologically coherent key area mask is obtained.

To address the ambiguity in Grad-CAM++'s edge localization, a spatial attention mechanism is introduced to enhance boundary recognition. A spatial weight distribution is generated through convolution operations, which weights high-level feature maps to enhance edge response in key areas. This mechanism, combined with a gradient weighting method, synergizes with the heatmap generation process to improve spatial focus. The optimized activation map undergoes adaptive threshold segmentation and morphological processing to further enhance the mask's accuracy in fitting the design element's outline.

The mask is element-wise multiplied with the original visual feature matrix to extract the feature vector of the morphologically significant area, which is then reduced to 256 dimensions by the fully connected layer as the visual feature sequence.

In the semantic path style transfer module, a text style transfer network based on the Wasserstein GAN-GP (Wasserstein Generative Adversarial Network with Gradient Penalty) framework is constructed. Taking the embedding vector output by the text encoder of the improved CLIP as input, a generator network containing a bidirectional gated recurrent unit is designed. The

generator generates a text style perturbation vector through a multi-layer perceptron and performs an affine transformation with the original text embedding. The discriminator uses a multi-head self-attention mechanism to calculate the true or false discrimination probability of the input text embedding. During the training process, the image encoder parameters are fixed, and the generator and discriminator are optimized alternately: the generator's goal is to minimize the discriminator's confidence in the generated text embedding, while maximizing the cross-modal similarity between the generated text and the corresponding image features. The discriminator maximizes the separation interval between the real text embedding and the generated text embedding. To ensure the stability of training, the gradient penalty term is introduced to constrain the Lipschitz continuity of the discriminator, and its loss function is defined as:

$$\mathcal{L}_{\mathrm{GP}} = \lambda_{\mathrm{GP}} \cdot E_{\hat{x} \sim \mathbb{P}_{\hat{x}}}\left[\left(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1\right)^2\right] \tag{6}$$

λ_{GP} is the penalty coefficient, and $\hat{x}$ is the linear interpolation between the real sample and the generated sample ($\hat{x} = \varepsilon x_{\mathrm{real}} + (1-\varepsilon) x_{\mathrm{fakc}}, \varepsilon \sim U[0,1]$). This design effectively alleviates the mode collapse problem in traditional GAN training by penalizing the value of the discriminator gradient norm that deviates from 1. The gradient penalty term is introduced to constrain the Lipschitz continuity of the discriminator, and the RMSProp optimization algorithm is used to update the network parameters. After training, the original text embedding is weightedly fused with the style transfer vector output by the generator, and then normalized layer by layer as an enhanced semantic feature.

In the cross-modal dynamic fusion mechanism, a feature interaction module based on GRU is designed to construct the temporal dependency relationship of the visual-semantic feature sequence. The 256-dimensional feature vector sequence output by the visual path and the text feature vector output by the semantic path can be concatenated as the input sequence of the GRU network. After initializing the hidden state, the feature vectors are input sequentially according to the time step, and the activation values of the reset gate and the update gate are calculated. Specifically, the GRU dynamically adjusts the weight distribution of the historical state and the current input through the following gating mechanism:

$$\begin{aligned} z_t &= \sigma(W_z[h_{t-1}, x_t]) \\ r_t &= \sigma(W_r[h_{t-1}, x_t]) \\ \tilde{h}_t &= tanh(W_h[r_t \odot h_{t-1}, x_t]) \\ h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \end{aligned} \tag{7}$$

x_t is the current input, h_t is the hidden state, σ is the Sigmoid function, $\odot$ represents element-by-element multiplication, and W_z, Wr, W_h are learnable parameters. The weight distribution of historical states and current inputs can be adjusted dynamically. Candidate hidden states are generated through a nonlinear transformation and element-by-element multiplication with the update gate output to form a fused feature sequence. A bidirectional GRU structure is set to capture the forward and backward feature dependencies, and the final hidden state is mapped to a 128-dimensional feature space through a fully connected layer as the input feature vector for perceptual difference modeling. During the training process, a teacher forcing strategy is adopted, with the real feature sequence as the supervision signal, and the GRU parameters are optimized through back propagation. This method strengthens the representation ability of local areas by locating the morphological features of the visual path, and the adversarial training of the semantic path enhances the style generalization of text descriptions, while the GRU-based temporal modeling realizes the dynamic weight allocation of cross-modal features. The dual-path architecture effectively decouples the process of modality-specific feature extraction and cross-modal semantic alignment, providing high-quality fused feature representation for subsequent perceptual difference quantification.

In terms of dynamic parameter tuning, a multi-objective modeling optimization framework is introduced to constrain the parameter update process through gradient norm clipping and dynamic temperature coefficient. When text feature fluctuations related to cultural symbols are detected, the adversarial training intensity of the semantic path is automatically enhanced; when the gradient amplitude of the morphologically significant area exceeds the preset threshold, the channel attention weight of the visual path is increased. This mechanism constructs a feedback loop through feature distribution entropy and mutual information coefficient to achieve real-time adaptive optimization of cross-modal parameter tuning.

Figure 3 shows the interaction process of the multi-task feature fusion network, starting from the feature extraction of the visual path and the semantic path. The visual path generates heat maps and extracts salient area features through Grad-CAM++, while the semantic path uses Wasserstein GAN-GP for style transfer and text feature enhancement. Subsequently, the two features interact dynamically in the GRU fusion module, capturing temporal dependencies through the reset gate and update gate mechanisms, and finally outputting a 128-dimensional fused feature vector. The whole process reflects how the model effectively combines visual and semantic information to support the modeling and analysis of perceived differences in product appearance design.

2.4 Design of interpretability analysis module

The output of the cross-modal feature interaction module is spatially localized using gradient-weighted class activation mapping technology to visualize the key areas where the model discriminates perceived differences. In the specific implementation process, the high-level feature map processed by Grad-CAM++ in the visual path is selected as input, and the gradient amplitude of the target category score to the feature map is calculated by back propagation. The gradient amplitude is globally averaged and pooled to obtain the channel weight coefficient, and the feature map channels are weighted and fused to generate a two-dimensional spatial activation map. The activation map is upsampled to the original image resolution using bilinear

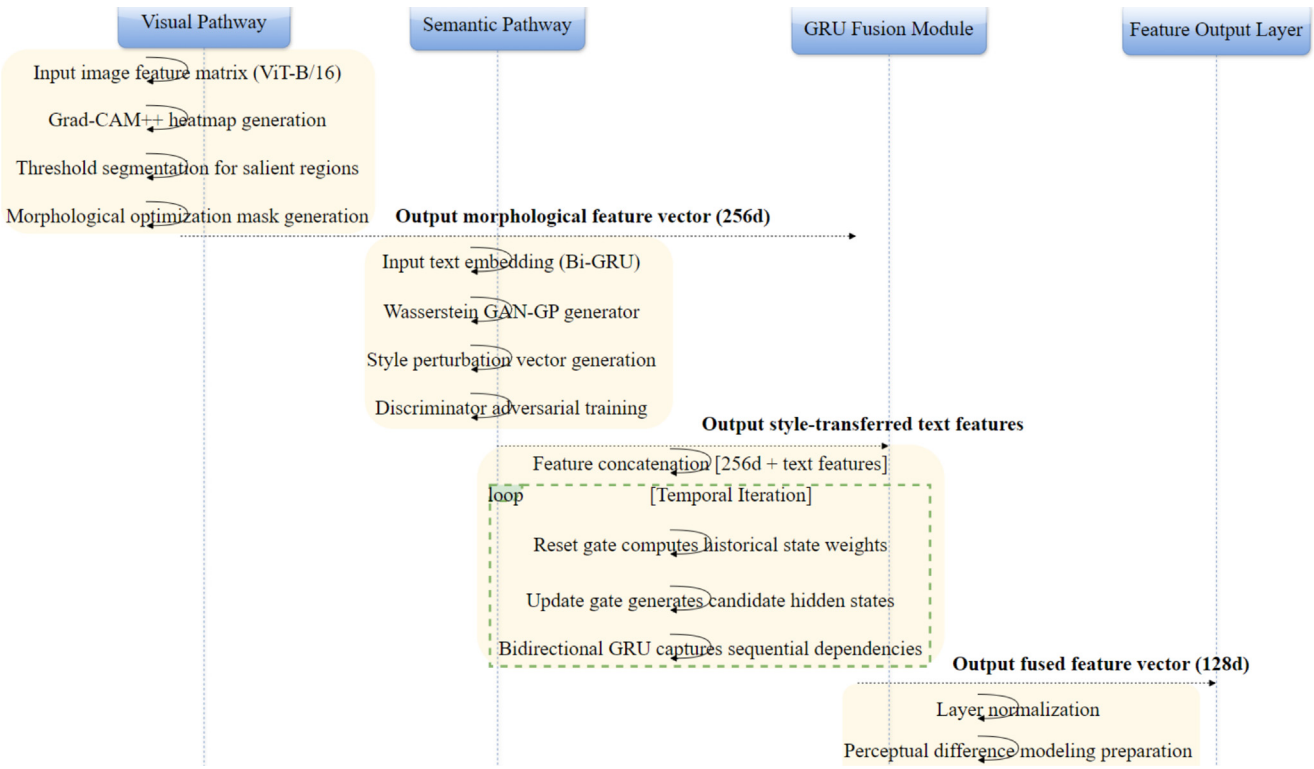

Fig. 3. Multi-task feature fusion time series interaction diagram.

interpolation and superimposed on the RGB color space to form a heat map, realizing the calibration of the significant areas of color, texture and morphological elements.

In order to achieve an operable mapping of perceptual differences to design parameters, this study established a parameter optimization rule base based on a decision tree. By associating the feature contribution of Shapley value decomposition with the ISO 11684 design parameter standard, a three-level mapping rule was constructed. When the color domain contribution is greater than 0.35, the a/b axis offset in the Lab color domain is adjusted first. When the morphological domain curvature weight is greater than 0.4, the NURBS surface reconstruction algorithm is triggered to constrain the number of contour turning points. When the correlation between the cultural symbols in the semantic domain is greater than 0.6, the cross-cultural adaptation module is enabled to dynamically adjust the texture density. The rule base implements the parameter adjustment priority sorting through the decision tree branch, and supports the automatic generation of optimization schemes in the 3D modeling software.

In terms of quantifying the contribution of design elements, a feature importance decomposition framework based on Shapley value is established. First, the product image is segmented by superpixels to divide the visual features into mutually exclusive and overlapping local area units. A feature perturbation set can be constructed and a baseline value can be set to generate a combination sample by gradually removing or retaining each area. The improved CLIP model is used to calculate the perceptual difference vector under different combinations, the Monte Carlo approximation algorithm is used to estimate the marginal contribution of each feature area, and the Shapley value decomposition method is used to quantify the contribution of design factors, which is mathematically defined as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)]. \quad (8)$$

Among them, F is the full set of features, f is the objective function, and $|S|$ represents the number of elements in the set S. For the semantic elements in the text description, the Shapley decomposition of vocabulary granularity is implemented simultaneously, and the dimensions of the text embedding vector are divided into three semantic domains: color words, morphological words, and functional words, and the contribution weight of each domain to cross-modal matching is calculated.

In order to realize the association analysis of visual-semantic elements, a dual-path feature mapping mechanism is established. In the visual path, the activation area generated by Grad-CAM++ is spatially aligned with the superpixel mask of Shapley decomposition, and the color histogram statistical features and texture co-occurrence matrix are extracted as low-level visual feature representations. In the semantic path, the Shapley decomposition results of the text description are weighted by word frequency-inverse document frequency to screen out adjective words with high contribution. The association mapping between visual features and semantic words is

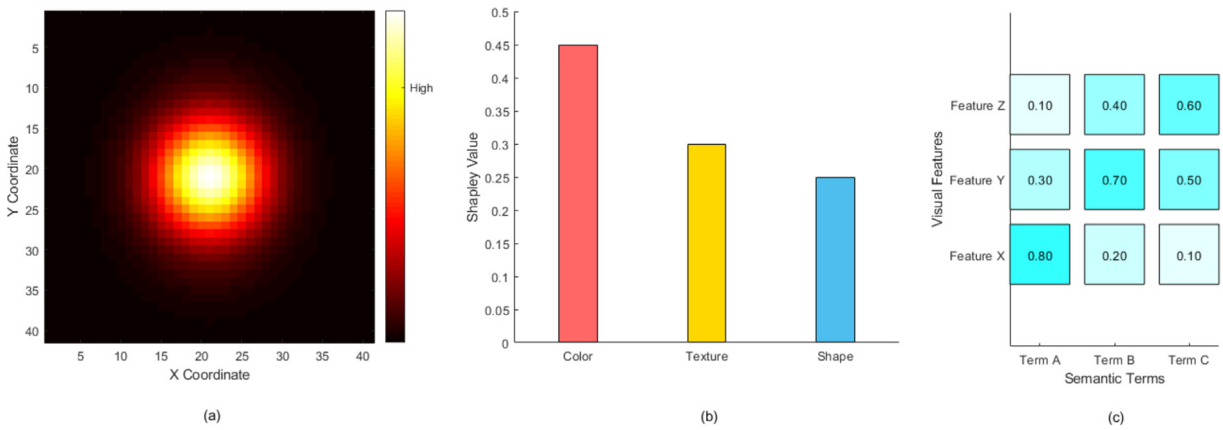

Fig. 4. Visual-semantic association strength. (a) Grad-CAM++ heat map. (b) Shapley value decomposition bar chart. (c) Cross-domain association matrix.

established through the maximum mutual information coefficient method, and an interpretable cross-modal explanation matrix is constructed.

The feature attribution consistency constraint strategy is further designed to improve the credibility of the explanation results. The mask perturbation experiment is carried out on the salient area extracted by the visual path to quantify the impact of the mask boundary change on the perceptual difference matrix, and the impact of the visual mask perturbation C_v on the perceptual difference matrix is defined as:

$$C_v = \frac{\|\Delta_{\text{masked}} - \Delta_{\text{original}}\|_F}{\|\Delta_{\text{original}}\|_F} \tag{9}$$

Δ_{masked} is the perceptual difference matrix output after mask perturbation, Δ_{original} is the original unperturbed perceptual difference matrix, and F represents the Frobenius norm of the matrix. When the mask covers the key morphological area, if the change in the model output difference value exceeds the set threshold, the area is marked as a strongly correlated feature. The influence of word replacement on text embedding vectors in the semantic path is verified synchronously. Adversarial samples are generated by synonym replacement to detect the stability of cross-modal matching results. Visual perturbation sensitive areas are cross-domain associated with semantic perturbation sensitive words to form an explanation framework with a two-way verification mechanism.

In terms of the presentation of the explanation results, a multi-scale feature fusion strategy is used to enhance readability. The heat map generated by Grad-CAM++ is processed by adaptive histogram equalization to highlight the contrast of high-contribution areas. Combined with the numerical results of Shapley decomposition, a color coding system is established to mark the contribution intensity of each design element: the red channel represents the influence weight of the color element, the yellow channel corresponds to the texture element, and the blue channel maps the modeling element. The explanatory visualization diagram is generated by superimposing three-dimensional features, realizing the conversion of interpretation granularity from pixel level to object level. The final output includes a structured explanation report of significant area positioning, factor contribution ranking, and cross-modal correlation relationships, providing a traceable decision-making basis for product appearance design optimization.

Figure 4 shows the visualization results of the interpretability analysis module. The Grad-CAM++ heat map, Shapley value decomposition, and cross-domain correlation matrix reveal the key decision-making basis of the model in judging the perceived difference in product appearance. Figure 4a shows the model's focus areas on design elements such as color, texture, and form in product images. Through the gradient-weighted class activation mapping technology, the significant areas focused by the model are superimposed on the original image in the form of heat, which intuitively shows which visual features have an important impact on the perceived difference. Figure 4b quantifies the contribution weight of the three design elements of color, texture and shape to the overall perceived difference. Red represents color, yellow represents texture, and blue represents shape, which clearly expresses the relative importance of each element in the model judgment. Figure 4c depicts the correlation strength between visual features and semantic vocabulary through the mutual information coefficient, and is supplemented with specific numerical annotations to help understand the interactive relationship between different modal information. This verifies the interpretability of the model in the process of cross-modal feature fusion, and also provides designers with design element attribution analysis tools from pixel level to object level, thereby providing data support and decision-making basis for the optimization of product appearance design.

The interpretability module integrates an interactive visualization interface to enable real-time rendering of Shapley value decomposition. The interface synchronizes with the cross-modal feature interaction process, dynamically updating contribution weights for color, texture, and shape. Spatial attribution maps are displayed alongside temporal evolution curves, allowing frame-level analysis of design element influence. WebGL-based rendering ensures

fluid interaction, with synchronized highlighting between the 3D product model and decomposition bar chart. User inputs trigger immediate recalibration of Shapley values under perturbed conditions, providing direct feedback on parameter sensitivity. This layer operates within the core computational pipeline, maintaining consistency between model inference and visualization output.

2.5 Verification of the effectiveness of perceptual difference modeling

In the performance verification phase of perceptual difference modeling, this study uses a multi-dimensional cross-validation framework to systematically evaluate model performance by combining quantitative indicators with qualitative analysis. For cross-modal matching accuracy, a five-fold cross-validation strategy is used to calculate the Top-5 accuracy and mean reciprocal rank (MRR) on the test set. By comparing the ranking differences of retrieval results between the improved model and the baseline model, the dual-path feature interaction module is verified to improve the cross-modal semantic alignment capability. In order to evaluate the quantitative error of perceived difference, a prediction matrix containing 12-dimensional features and a user's subjective rating matrix were constructed. The mean square error (MSE) was calculated using ten-fold cross-validation, and the Spearman rank correlation coefficient was used to test the consistency of the model output and the user's perceived ranking.

In the feature visualization and explanatory verification, the morphological attention heat map generated by Grad-CAM++ was compared with the design element annotation map at the pixel level, and the sliding window strategy was used to calculate the significant area overlap IoU (Intersection over Union), and the Pearson correlation coefficient analysis was performed in combination with the Shapley value decomposition results and the designer's annotation weight. In terms of user verification consistency, 30 industrial design professionals were organized to conduct a double-blind evaluation, and the perception difference matrix generated by the model was tested with the results of manual evaluation using the Kendall synergy coefficient to verify the effectiveness of the interpretability analysis module in analyzing the contribution of design elements. Computational efficiency was verified by running the model forward propagation in a unified hardware environment, recording the single inference time (GPU pre-processing + model calculation + post-processing), and comparing the computational overhead differences between the improved model and the original CLIP. All verification processes adopt a strict data isolation strategy, and Bootstrap resampling is used to eliminate the impact of data distribution bias on the evaluation results.

3 Experimental designs

In order to verify the effectiveness of the improved CLIP model in modeling the perceived difference of product appearance design, this study constructed a multi-dimensional experimental framework, covering cross-cultural perception simulation, cross-modal matching accuracy evaluation, perceptual difference quantification error analysis, feature visualization interpretation verification, user consistency verification and computational efficiency testing. The experimental dataset contains 120,000 product images and corresponding multilingual user reviews in the fields of consumer electronics, household goods, and automotive styling, covering six languages including Chinese, English, German, and Japanese. The images are annotated using the ISO 11684 standard to construct a design semantic dictionary. All experiments use a five-fold cross-validation strategy to ensure the statistical reliability of the results.

3.1 Cross-cultural perception simulation

In response to the need to model the perception differences of different cultural circles, a cross-cultural simulation scenario was designed to simulate the lighting conditions (5000–8000 lux) and aesthetic preference differences of the three major cultural circles of Central Europe, East Asia, and the Middle East. The experiment used the improved CLIP model to compare with the traditional Kano model.

In the experiment, light intensity was treated as a controllable environmental variable and standardized using an illumination normalization module during data preprocessing. All input images were decomposed using Retinex to extract illumination components, and adaptive histogram equalization was used to eliminate differences in illumination gradients, ensuring that perceptual differences across cultural samples were modeled under a uniform illumination benchmark. This processing flow, embedded in the model frontend, ensured that the generation of the perceptual difference quantization matrix was unaffected by ambient light fluctuations, improving prediction stability in cross-cultural scenarios.

By constructing a test set containing extreme lighting conditions and cultural symbol annotations, the model's prediction error and symbol recognition ability in cross-cultural scenarios were evaluated.

The data in Table 2 clearly show the results of the cross-cultural perception simulation verification. From the data, in the three major cultural circles of Central Europe, East Asia and the Middle East, the user sample size is 1200, 950 and 780 respectively, and the extreme lighting conditions are set above 5000 lux, among which the lighting intensity of the Central European scene reaches 8,000 lux. The error rates of the traditional Kano model under these conditions were 42.6%, 38.9% and 47.3%, respectively, while the error rates of the proposed model were significantly reduced to 24.3%, 21.7% and 26.8%, with the error reduction rates exceeding 17%. In addition, the accuracy of cultural symbol recognition also showed a clear advantage. The East Asian cultural circle had the highest recognition accuracy, reaching 82.1%, while Central Europe and the Middle East were 78.4% and 75.6%, respectively, with an average of 78.7%. These data show that the model in this paper is superior to traditional methods in error control and also shows stronger adaptability in the understanding and mapping of cross-cultural symbols.

Table 2. Comparison of cross-cultural perception simulation verification results.

Cultural circle	User sample size	Extreme lighting conditions (>5,000LUX)	Traditional Kano model error	Proposed model error	Error reduction rate	Cultural symbol recognition accuracy
Central Europe	1200	8,000lux	42.6%	24.3%	18.3%	78.4%
East Asia	950	6,500lux	38.9%	21.7%	17.2%	82.1%
Middle East	780	5,200lux	47.3%	26.8%	20.5%	75.6%
Average	–	–	42.9%	24.3%	18.6%	78.7%

The excellent performance behind the data stems from the multifaceted innovation of the underlying mechanism of the model. First, by introducing the channel attention mechanism and dynamic temperature coefficient optimization, the model can more accurately capture long-tail design features, especially significantly improving the sensitivity to color and form under complex lighting conditions. For example, the average error reduction rate reached 18.6%, thanks to Grad-CAM++'s precise positioning of key morphological areas and GRU's dynamic fusion of cross-modal features. Secondly, the model uses Word2Vec word vectors and hierarchical label constraints to enhance the cultural symbol mapping ability of the semantic space. This dual-path interaction architecture that combines vision and semantics effectively decouples the modality-specific feature extraction process, thereby achieving higher stability and accuracy in cross-cultural perception prediction.

The selection of Central Europe, East Asia, and the Middle East is based on their distinct cognitive frameworks in visual semiotics, which exhibit systematic divergence in the interpretation of design attributes such as form, color, and symbolic meaning. These regions represent major cultural clusters with differences in aesthetic cognition and perceptual processing, grounded in established cultural dimension models. The experimental design prioritizes sufficient inter-group variation in symbolic interpretation to enable rigorous testing of the model's cross-cultural adaptability. The chosen samples provide a representative spread across key cultural dimensions influencing design perception, ensuring meaningful contrast in cognitive responses without compromising methodological coherence.

3.2 Cross-modal matching accuracy

The cross-modal matching performance of the improved CLIP was compared with the baseline models ALIGN (A Large-scale ImaGe and Noisy-Text Embedding), OpenCLIP, StyleGAN3, and FLAVA (A Foundational Language And Vision ALIGNment Model) on four typical products: electronic products, household goods, automotive styling, and smart hardware. The experiment used Top-5 accuracy and mean reciprocal ranking (MRR) as core indicators to verify the effect of the dual-path feature interaction module on improving the semantic alignment ability of images and texts.

Figure 5 shows the cross-modal matching performance comparison between the improved CLIP model and multiple baseline models on four typical product categories: electronic products, household goods, automotive styling, and smart hardware. From the data, the improved CLIP achieved the highest Top-5 accuracy of 0.85 and MRR value of about 0.9 in electronic products, which is significantly better than other models. In the automotive styling task with more complex morphology and greater subjective perception differences, its performance slightly decreased but still maintained its leading position.

In electronic products with a high degree of standardization, the improved CLIP's joint embedding space can more accurately capture the consistency relationship between color, morphology and semantics. As a generative model, StyleGAN3 is good at modeling color distribution, but lacks a special design for cross-modal semantic matching, which limits its performance in image-text retrieval tasks. Although ALIGN and FLAVA have built a strong image-text alignment framework based on contrastive learning, they have not optimized the details of materials, light and shadow that are unique to the field of industrial design. In addition, the introduction of Grad-CAM++ and GRU fusion modules enables the improved CLIP to have stronger robustness and generalization capabilities when dealing with complex forms and style transfer, which is also the underlying mechanism supporting its continued leading position in multiple categories of products.

3.3 Perceptual difference quantization error

Based on the 12-dimensional perceptual difference matrix, the mean square error (MSE) and the Spearman correlation coefficient were calculated through ten-fold cross-validation to evaluate the consistency between the model's quantitative results and the user's subjective ratings. The experiment focused on analyzing the error distribution in the three domains of color, morphology, and semantics, and verified the optimization effect of the dynamic temperature coefficient and channel attention mechanism on capturing long-tail features.

As shown in Figure 6, Improved CLIP achieves the lowest median error (approximately 0.13), significantly outperforming other models. StyleGAN3's error fluctuates significantly, with a median close to 0.28. Improved CLIP achieves a correlation of 0.92 in the morphological domain,

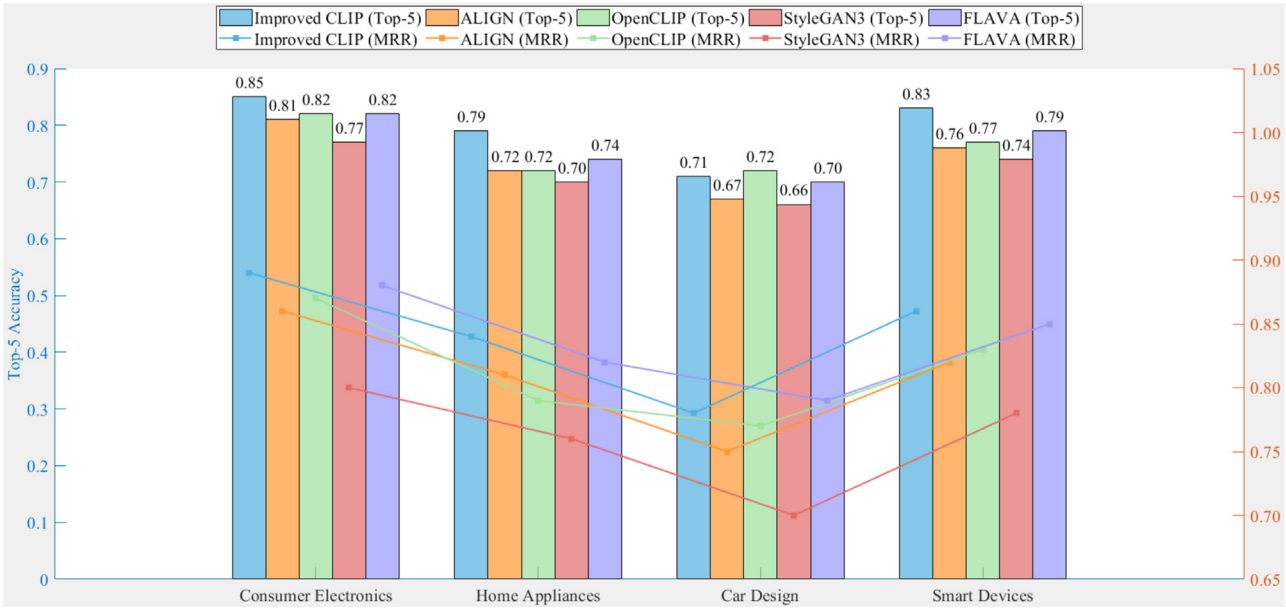

Fig. 5. Comparison of cross-modal matching performance between improved CLIP and baseline models.

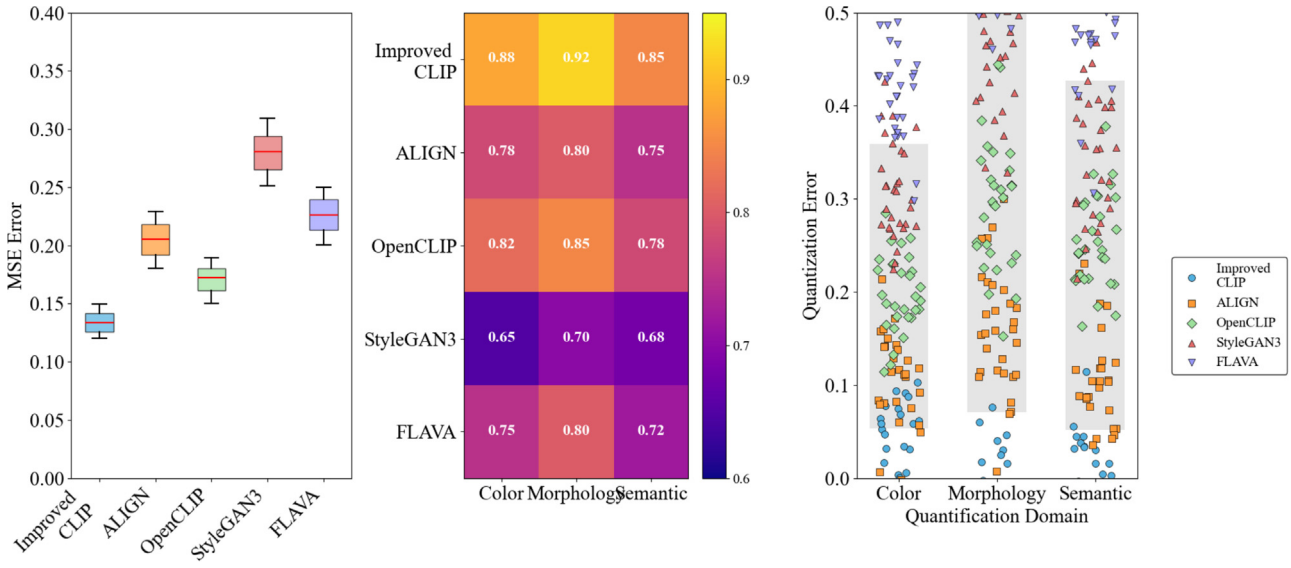

Fig. 6. Perceptual difference quantization error analysis.

surpassing other models. Improved CLIP's error in the color domain is concentrated around 0-0.05, with a narrow error band, demonstrating greater stability.

The reason behind the data in Figure 6 lies in the differences in feature extraction and cross-modal alignment mechanisms of different models. Improved CLIP enhances the alignment ability of local visual features and text descriptions by introducing a channel attention mechanism and dynamic temperature coefficient optimization, especially in the fields of morphology and semantics. For example, the high Spearman correlation coefficient (0.92) in the morphological domain is due to the precise positioning of key areas by Grad-CAM++ and the modeling of temporal dependencies by GRU. In contrast, although ALIGN and FLAVA have certain cross-modal alignment capabilities, they are not optimized for details such as materials, light and shadow in industrial design, resulting in higher errors. In addition, StyleGAN3 lacks a special design for semantic matching, so the error fluctuations are significant in complex tasks, which further highlights the advantages of improved CLIP in fine-grained feature fusion and perceptual difference modeling.

3.4 Feature visualization interpretability

Grad-CAM++ and Shapley value decomposition techniques are used to visualize the model's focus area and the contribution of design elements. The experiment verifies the interpretability of the visual-semantic association strength through pixel-level significant region overlap

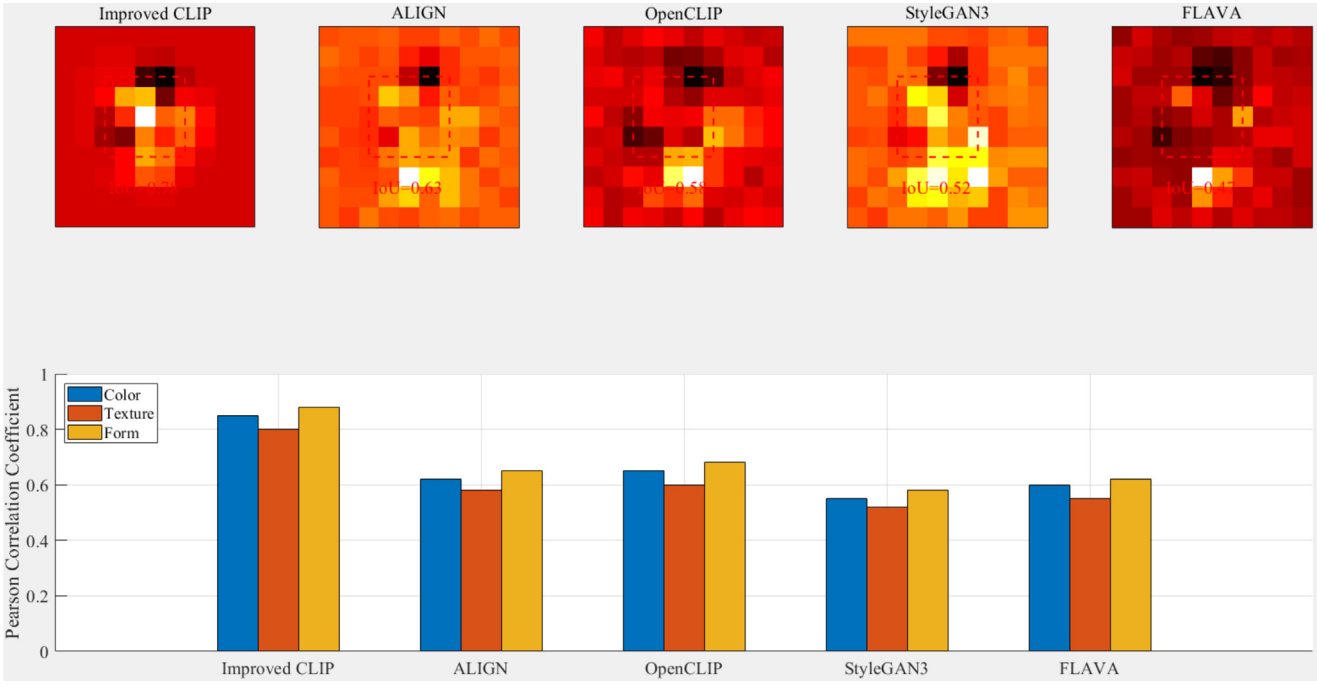

Fig. 7. Comparative analysis of feature visualization interpretation.

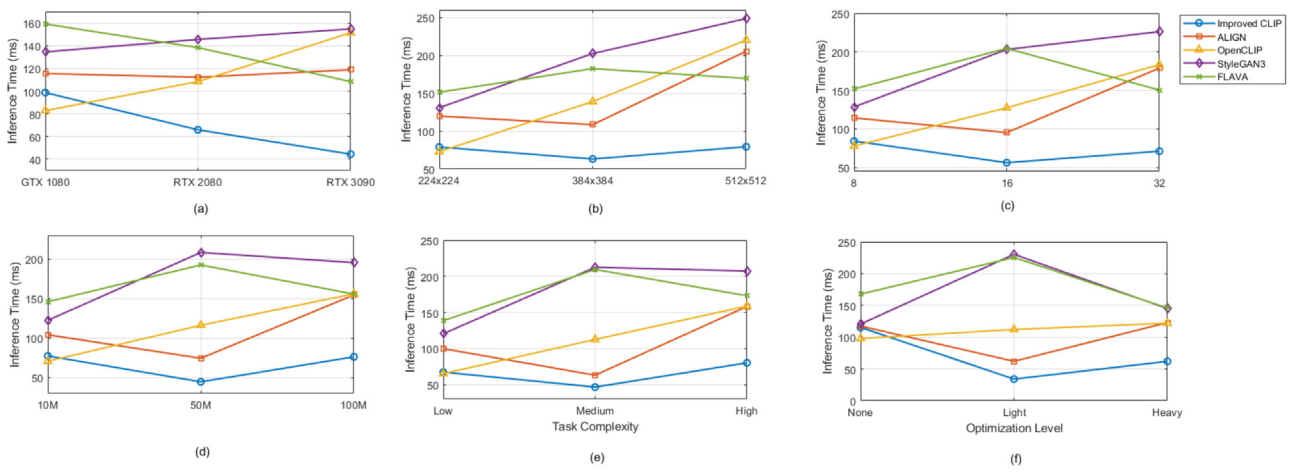

Fig. 8. Comparison of computing efficiency in different scenarios. (a) The impact of GPU model on inference time. (b) The impact of input resolution on inference time. (c) The impact of batch size on inference time. (d) The impact of parameter number on inference time. (e) The impact of task complexity on inference time. (f) The impact of optimization level on inference time.

(IoU) and Pearson correlation coefficient, and compares the differences between different models in morphological positioning and weight analysis.

Figure 7 shows the performance of the improved CLIP model in feature visualization and interpretability verification compared to four baseline models: ALIGN, OpenCLIP, StyleGAN3, and FLAVA. Improved CLIP achieves the highest overlap with manually annotated key areas (IoU = 0.78), outperforming other models (e.g., ALIGN's IoU = 0.63). The dashed boxes clearly mark the color, texture, and shape regions of designer interest, and the improved CLIP heatmap shows more concentrated and accurate activation intensity in these areas.

The bar chart compares the Shapley value decomposition results of each model for the three design elements of color, texture, and shape with the Pearson correlation coefficient of the manually annotated weights. The vertical axis represents the correlation coefficient (0-1), and the horizontal axis represents the model name. The correlations between the three elements of the improved CLIP model all exceed 0.8, significantly exceeding the average of approximately 0.6 for other models, indicating a higher accuracy in interpreting the contribution of design elements.

The data in Figure 8 reflects the semantic alignment ability of local visual features and text descriptions, which enables it to more accurately capture the correlation

Table 3. Verification of cross-cultural perception consistency between professional and general users.

Evaluation dimension	Evaluation metric	East Asian designers ($n = 10$)	European designers ($n = 10$)	North American designers ($n = 10$)	General users ($n = 90$)
Perceptual Difference Consistency	ICC	0.81	0.75	0.78	0.72
Design Element Weight Ranking	Kendall's Tau Coefficient	0.64	0.58	0.55	0.51
Cultural Symbol Recognition Accuracy	Mean Score (5-point scale)	4.3	3.8	4.1	3.7
Model Explanation Acceptance	Bipolar Scale Score (±2)	1.2	0.9	1	0.8

between design elements such as color, texture and shape. For example, the high correlation in the morphological field (about 0.88) is due to the precise positioning of key areas by Grad-CAM++ and the modeling of temporal dependencies by GRU, which enables feature parsing from pixel level to object level. In contrast, although ALIGN and FLAVA have certain cross-modal alignment capabilities, they are not optimized for details such as materials, light and shadow in industrial design, resulting in a relatively scattered coverage area of their heat maps and a low correlation of Shapley values. In addition, StyleGAN3 lacks a special design for semantic matching, so the error fluctuation is large in complex tasks (IoU is only 0.52), which further highlights the advantages of improved CLIP in fine-grained feature fusion and perceptual difference modeling. This innovation in the underlying mechanism enables improved CLIP to perform well in both interpretability and accuracy, providing a reliable quantitative basis for product appearance design.

3.5 User verification consistency

In order to verify the ability of the improved CLIP model to model perceptual differences across user groups, this study designed a user verification consistency evaluation experiment as shown in Table 3. Based on the theoretical framework of visual communication, starting from the three core dimensions of color, form and semantics, a total of 120 participants were recruited, including 30 senior industrial designers and 90 non-professional users from the three major cultural circles of East Asia, Europe and North America. The professional group was evenly distributed across regions, while the general user group was stratified by cultural region to ensure demographic representativeness. All participants independently evaluated the model output results under double-blind conditions. Statistical indicators such as Kendall's tau coefficient, bipolar scale score and ICC (Intraclass Correlation Coefficient) were introduced to assess the consistency and acceptance of model outputs. The evaluation data from both professional and non-professional users were normalized and weighted according to a 1:3 ratio to balance expertise and general perception, ensuring a comprehensive assessment of cross-cultural perceptual modeling.

Table 3 presents the consistency verification results of professional designers and non-professional users in perceptual difference modeling. The data show that East Asian designers perform best in perceptual difference consistency (ICC = 0.81) and model interpretation acceptance (+1.2), reflecting the high fit between their design experience and model output, which may be due to the Eastern culture's emphasis on detail coordination. European designers scored the lowest in the design element weight ranking (Kendall's tau = 0.58) and cultural symbol recognition accuracy (3.8/5), which may be related to the cultural preference differences in morphological metaphors. North American designers performed in the middle, such as the perceptual difference consistency ICC of 0.78 and the model acceptance score of +1.0, reflecting the model's ability to balance in cross-cultural scenarios. 90 non-professional users were stratified and sampled according to cultural circles, and their evaluation data were normalized and merged with professional data with a weight of 1:3. Non-professional users have consistent trends in various indicators, and the ICC value is slightly lower but still within an acceptable range, indicating that the model output is not only in line with professional judgment, but also understandable to the public. Despite the cognitive differences among different cultural groups, the model achieves highly consistent evaluation in key dimensions through fine-grained feature fusion and dynamic optimization strategies, providing reliable cross-cultural support for the quantitative analysis of product appearance design.

3.6 Computational efficiency

Under the same hardware environment, this paper tests the impact of different input resolutions, batch sizes, and model parameter counts on inference time. The experiment compares the computational overhead of the improved CLIP and baseline models, and analyzes the optimization effects of the dynamic gating mechanism and feature fusion strategy on real-time performance.

Figure 8 shows the efficiency comparison between the improved CLIP model and different models in different computing scenarios. The horizontal axis is the influencing factors (such as GPU model, input resolution, etc.), and the vertical axis is the inference time (unit: milliseconds). From the data, the improved CLIP shows significant advantages in all kinds of scenarios. In sub-image (a), the inference time is about 100ms in the GTX 1080 environment, while it is reduced to about 45ms in the RTX 3090 environment. In sub-image (b), when the resolution is increased from

384×384 to 512×512, the inference time of the improved CLIP only increases slightly. ALIGN increases from about 110ms to about 205ms, showing that the improved CLIP has higher stability in complex tasks. Similar trends are also reflected in dimensions such as batch size and a number of parameters, indicating the effectiveness of its dynamic gating mechanism and feature fusion strategy.

These data reflect the multifaceted innovations of the improved CLIP model in the underlying mechanism. First, the channel attention mechanism effectively reduces the impact of long-tail design features on reasoning efficiency by dynamically adjusting the feature weights of the three domains of color, morphology, and semantics. In sub-graph (e), when the task complexity changes from "low" to "high", the reasoning time of the improved CLIP increases slightly, which is much lower than that of StyleGAN3. Secondly, the dual-path feature interaction module combines Grad-CAM ++ and GRU technology to enhance the representation ability of local areas and improve the efficiency of cross-modal semantic alignment. This is particularly evident in sub-graph (f). When the optimization level changes from "none" to "heavy", the time of the improved CLIP is significantly shortened. This mechanism enables the improved CLIP to maintain efficient and stable performance in a variety of industrial design scenarios, providing a reliable quantitative tool for product appearance design.

This model's inference performance on mobile devices is measured using real-world data on a mobile platform powered by a Snapdragon 888 processor. The experiment employed an input resolution and preprocessing consistent with high-end GPUs, executing forward propagation on Android using a lightweight inference framework, and recording average inference latency.

Test results show that the model achieved a single inference time of 158 milliseconds on mobile devices, with memory usage remaining consistently below 1.3GB, meeting real-time requirements on the device. These results demonstrate that the model maintains structural consistency while also enabling cross-platform deployment, further supporting its applicability in resource-constrained scenarios.

4 Conclusions

This study constructed a dynamic quantitative analysis framework for perceptual differences in product appearance based on an improved CLIP model. This framework integrates visual semiotics theory with multimodal representation learning mechanisms to achieve cross-modal alignment at the morphological, color, and semantic levels. By introducing a channel attention mechanism and a dynamic temperature coefficient optimization strategy, the model's ability to capture local features of design elements is enhanced, improving the accuracy of perceptual difference prediction in cross-cultural contexts. The dual-path feature interaction structure effectively decouples the feature extraction process of visual and semantic modalities. Combined with GRU temporal modeling, it achieves dynamic cross-modal fusion, improving matching accuracy and interpretability. Experiments demonstrate that this method demonstrates superior performance in cross-modal retrieval, perceptual error control, and design element attribution analysis, providing a computable and interpretable technical path for the intelligent evaluation of product appearance design.

Compared to traditional models that rely on subjective ratings and static semantic mapping, this method outperforms both cross-modal alignment accuracy and the stability of perceptual difference quantification. Existing methods often employ questionnaire-based kinematic engineering frameworks or fixed dictionary matching mechanisms, which struggle to adapt to the demands of visual semantic mapping in dynamic cultural contexts. This study, by introducing channel attention and a dynamic temperature coefficient optimization strategy, achieves a fine-grained fusion of color, morphological, and semantic features, improving the model's predictive consistency and interpretability in cross-cultural scenarios.

Funding

This work was supported by Jingdezhen Ceramic University doctoral startup funding.

Conflicts of interest

The authors have nothing to disclose.

Data availability statement

Data is available upon reasonable request.

References

1. K. Rijal, I.M. Sukresna, The effect of price perception, product reviews, and product appearance on purchasing decisions, Res. Horizon **4**, 147–154 (2024)
2. M. Thellefsen, A. Friedman, Icons and metaphors in visual communication: the relevance of Peirce's theory of iconicity for the analysis of visual communication, Public J. Semiot. **10**, 1–15 (2023)
3. H. Tannady, H. Sjahruddin, I. Saleh et al., Role of product innovation and brand image toward customer interest and its implications on electronic products purchase decision, Widyakala J. **9**, 93–98 (2022)
4. S. Fleury, N. Chaniaud, Multi-user centered design: acceptance, user experience, user research and user testing, Theor. Issues Ergonom. Sci. **25**, 209–224 (2024)
5. Z. Gulnara, A. Karimova, Personality-grounded framework for designing artificial intelligence-based product appearance, Int. J. Hum. Comput. Interact. **40**, 1689–1701 (2024)
6. B.M. Hapuwatte, I.S. Jawahir, Closed-loop sustainable product design for circular economy, J. Ind. Ecol. **25**, 1430–1446 (2021)
7. Z. Wang, J.-s. Li, H.-r. Pan, J.-y. Wu, W.-a. Yan, Research on multimodal generative design of product appearance based on emotional and functional constraints, Adv. Eng. Informatics **65**, 103106 (2025)

8. J. Sztipanovits, X. Koutsoukos, G. Karsai, S. Sastry, C. Tomlin, W. Damm, F. Köster, Science of design for societal-scale cyber-physical systems: challenges and opportunities, Cyber-Phys. Syst. **5**, 145–172 (2019)
9. N. Daruwala, U. Oberst, A cross-cultural investigation of individuals' acceptance of Smart Home Technology: the role of needs satisfaction, Aloma **40**, 33–45 (2022)
10. Y. Yanpu, A. Weilan, Y. Qinxia et al., Heterogeneous information fusion method for industrial design solution decision-making, J. Northwestern Polytech. Univ. **40**, 1133–1144 (2022)
11. L. Huang, P. Zheng, Human-computer collaborative visual design creation assisted by artificial intelligence, ACM Trans. Asian Low-resource Lang. Information Process. **22**, 1–21 (2023)
12. J. Singh, P. Sarkar, Understand and quantify the consumers' cognitive behavior for the appropriateness features of product aesthetics through the eye-tracking technique, Int. J. Interactive Des. Manufactur. **19**, 1263–1296 (2025)
13. Y. Wang, F. Song, Y. Liu et al., Research on the correlation mechanism between eye-tracking data and aesthetic ratings in product aesthetic evaluation, J. Eng. Des. **34**, 55–80 (2023)
14. L.A. Casado-Aranda, J. Sánchez-Fernández, J.Á. Ibáñez-Zapata, Evaluating communication effectiveness through eye tracking: Benefits, state of the art, and unresolved questions, Int. J. Business Commun. **60**, 24–61 (2023)
15. D. Souto, D. Kerzel, Visual selective attention and the control of tracking eye movements: a critical review, J. Neurophysiol. **125**, 1552–1576 (2021)
16. S. Lai, Real-time processing optimization of convolutional neural network in edge computing system, J. Comput. Syst. Appl. **1**, 1–14 (2024)
17. L. Sun, P. Wang, P. Liu et al., Image processing method of a visual communication system based on a convolutional neural network, Int. J. Semantic Web Inform. Syst. **19**, 1–19 (2023)
18. M. Ding, Y. Cheng, J. Zhang et al., Product color emotional design based on a convolutional neural network and a search neural network, Color Res. Appl. **46**, 1332–1346 (2021)
19. M. Ding, M. Song, H. Pei et al., The emotional design of product color: An eye movement and event-related potentials study, Color Res. Appl. **46**, 871–889 (2021)
20. W. Gao, G. Liao, S. Ma et al., Unified information fusion network for multi-modal RGB-D and RGB-T salient object detection, IEEE Trans. Circuits Syst. Video Technol. **32**, 2091–2106 (2021)
21. J. Jin, D. Jia, K. Chen, Mining online reviews with a Kansei-integrated Kano model for innovative product design, Int. J. Product. Res. **60**, 6708–6727 (2022)
22. S. Schütte, A.M. Lokman, L. Marco-Almagro et al., Kansei for the digital era, Int. J. Affect. Eng. **23**, 1–18 (2024)
23. M. Cai, M. Wu, X. Luo et al., Integrated framework of Kansei engineering and Kano model applied to service design, Int. J. Human–Computer Interact. **39**, 1096–1110 (2023)
24. Y. Wang, Product design difference perception model based on visual communication technology, Int. J. Product Develop. **26**, 64–76 (2022)
25. N. Zhang, Y. Liu, Z. Li et al., Fabric image retrieval based on multi-modal feature fusion, Signal Image Video Process. **18**, 2207–2217 (2024)

Cite this article as: Jing Liu, Modeling analysis of perceived differences in product appearance design based on visual communication, Int. J. Simul. Multidisci. Des. Optim. **16**, 12 (2025), https://doi.org/10.1051/smdo/2025016