Automatic Generation and Optimization Algorithm for Brand Visual Identity System Based on Multimodal Generative Adversarial Network

Yunfei Gong[1]*, Xiaoyue Cao[2]
Design College, Zhoukou Normal University, Zhoukou 466000, China[1]
Academy of Fine Arts, Zhoukou Normal University, Zhoukou 466000, China[2]

***Abstract*—A Brand Visual Identity System (BVIS) is a crucial carrier for conveying a brand's core values, but traditional manual design struggles to meet the demands of efficient and personalized design in the digital age. Addressing the shortcomings of existing Multimodal Generative Adversarial Networks (MM-GANs) in BVIS design, such as semantic-visual mapping bias, fragmented element styles, and a lack of quantified iteration mechanisms, this study integrates attention mechanisms, perceptual loss, style consistency constraints, and adaptive optimization strategies to construct an improved automatic BVIS generation and optimization framework. The model utilizes publicly available datasets for preprocessing and alignment of text, color, and graphic multimodal data. Text semantic encoding is achieved through Bidirectional Encoder Representations from Transformers (BERT), and multimodal feature fusion is accomplished using an attention matrix. Adversarial training is conducted using an improved U-Net and PatchGAN. The model is then compared and tested with Generative Adversarial Network (GAN) series models, diffusion models, and Contrastive Language-Image Pre-Training (CLIP) guided generative models. Experiments show that the model outperforms the control model in metrics such as Fréchet Inception Distance (FID), Structural Similarity Index Measure (SSIM), and Semantic Matching (SM), demonstrating excellent robustness and generalization ability in scenarios with noise interference, cross-industry collaboration, and modality loss. Ablation experiments prove that multimodal feature alignment is key to improving model performance. The generated visual elements exhibit a consistent style and good semantic matching. This study presents a practical intelligent brand design solution and analyzes the ethical risks of copyright infringement, counterfeiting, and misuse of AI-generated logos, providing insights for compliant technology applications and expanding the engineering application of multimodal generation models in brand visual design.**

***Keywords*—Brand digitalization; visual identity system; multimodal; generative adversarial network; attention mechanism**

I. Introduction

In the context of global brand digital transformation, Brand Visual Identity System (BVIS) is a core intangible asset for enterprises to convey brand connotation and establish consumer awareness. The design quality directly affects the brand's market competitiveness [1]. Traditional BVIS design relies on the experience of designers and manual iteration, which has problems such as long cycles, homogenization of style, and lagging market response [2]. With the rapid development of Artificial Intelligence (AI) in the field of computer vision, Generative Adversarial Network (GAN) has become the mainstream tool for automatic generation of visual content due to its unsupervised learning and visual synthesis capabilities [3]. Multimodal Generative Adversarial Network (MM-GAN), as an extension of unimodal GAN, can integrate text, graphics, and color information, adapt to the collaborative design needs of multiple elements such as the BVIS logo, font, and color system, and is expected to break through the limitations of traditional manual design [4]. Currently, GAN-derived models have been used in graphic design and style transfer, but there are still shortcomings in the specific research on automatic generation optimization for BVIS [5]. Most existing models only generate single elements such as logos, making it difficult to model the style and semantic relationship between components and output a complete and coordinated visual system [6]. The multimodal semantic gap is prominent, the mapping accuracy from text semantics to visual elements is insufficient, and it is difficult to restore the brand tone [7]. Optimization relies heavily on subjective human judgment, lacking quantitative indicators and adaptive iteration strategies for visual communication effects, which limits practical implementation [8]. Although diffusion models, Contrastive Language-Image Pre-Training (CLIP) guidance frameworks, and multimodal Transformers have excellent semantic understanding and high-definition generation capabilities, they have high computational costs, and weak style customization constraints, making it difficult to deploy in a lightweight manner in professional design scenarios. The GAN series has more advantages in training efficiency, style controllability, and industry adaptability, but few studies have combined BVIS business characteristics to optimize MM-GAN, making its technical position in the generative AI system relatively backward [9]. At the application level, practical obstacles remain: training datasets generally suffer from bias and class imbalance, reducing the model's cross-scenario generalization ability. AI-generated brand logos are prone to copyright infringement and malicious counterfeiting risks, and the fairness and transparency of the model lack sufficient discussion, hindering the large-scale implementation of BVIS intelligent design. To address these research gaps, this study, based on MM-GAN, introduces an attention mechanism and improves the U-Net generator and PatchGAN discriminator for BVIS generation tasks, constructing a technical framework that includes multimodal semantic fusion, style consistency constraints, and quantized adaptive optimization. A cross-industry multimodal dataset is constructed and preprocessed for alignment. The attention mechanism is used to enhance

*Corresponding author.

semantic-visual accuracy mapping, ensuring style consistency across all visual components. Dynamic iteration of results is achieved based on quantitative indicators of dissemination effectiveness. This study compensates for the deficiencies in MM-GAN semantic mapping and style control, supplements a quantified optimization scheme for design practice, expands the application of multimodal generation models in brand design, and provides theoretical and technical references for digital intelligent brand design. This study also discusses dataset bias, intellectual property, and ethical risks, promoting the compliant implementation of generation technology.

II. Literature Review

As a key vehicle for building brand awareness and conveying brand value, brand visual identity (BVI) has become a core issue of academic concern. Existing research mainly focuses on different dimensions of BVI, providing rich theoretical and empirical support for understanding the intrinsic connection between visual elements and brand development. The following analysis will examine each dimension in conjunction with relevant literature.

In the research on specific forms and effects of BVI, Vinitha et al. [10] explored the brand biomorphic visual identity and its effects from a holistic perspective, revealed the application value and influence path of biomorphic elements in the brand visual system, and provided a new perspective for the differentiated design of BVI. Then, Erjansola et al. [11] focused on the core visual element of the brand logo, taking the university merger scene as the research object, and analyzed the transmission mechanism from brand logo to brand association and corporate identity. They made clear the key role of identity-based logo association in the process of organizational integration, and deepened the understanding of the function of logo in brand identity construction. On this basis, Singla and Sharma [12] further refined the research dimension of visual elements, and discussed the intermediary role of fonts in connecting brand identity and brand perception. They confirmed that the matching of font style and brand personality transmission had a significant impact on consumers' brand cognition, and enriched the research system of micro-elements of BVI.

In the study of cross-scenario application and technology integration of visual recognition, Tan et al. [13] broke through the scope of traditional commercial brands and applied the integration of visual recognition and text recognition to the field of network security, and proposed a phishing detection method based on hybrid visual recognition and text recognition. They verified the feasibility of applying the authentication logic related to visual recognition in non-commercial scenarios. Turoê et al. [14] focused on the shared mobility system of smart cities. They pointed out that visual communication can be used as an important entry point to improve the system's recognition and competitiveness, and constructed a framework for the connection between visual recognition and smart city service scenarios. In the study of dynamic adaptation of visual elements, Lelis et al. [15] regarded typography as a constant vector of dynamic signage and analyzed the core supporting role of typography in the evolution of dynamic signage. Their research provides a theoretical basis for the dynamic development of visual recognition.

In the research on the theoretical framework and effect extension of BVI, Andrade et al. [16] constructed a theoretical framework for the development of visual elements based on the dimension of brand personality, clarified the matching mechanism between visual element personality and brand personality, and provided methodological guidance for the systematic construction of BVI. Lusianti et al. [17], from the perspective of brand resonance, discussed the promotion of brand resonance related to BVI on the sustainable marketing performance of national medical insurance projects, and extended the research of BVI to the field of public services. Zha et al. [18] further expanded the research boundary and brought visual identity into the category of sensory brand experience, empirically tested the influence of sensory brand experience on brand loyalty, and revealed the long-term value of visual elements as the core of sensory experience.

Combined with the development trend of AI, Cai [19] discussed the innovative design idea of a brand vision system under the empowerment of visual communication design, and analyzed the application path of intelligent technology in the process of brand vision landing. It provided practical reference for the research of intelligent visual design. Hwang [20] explored the method of using generative AI for brand design through case experiments, and pointed out that AI generation had the defect of insufficient semantic-brand tone matching, which required human intervention to correct and optimize. Ho-dac and Mulder-Nijkamp [21] focused on the field of sustainable packaging, explored the balance strategy between brand differentiation and standardization, and expanded the research boundary of brand visual system in the green design scene. Chulayevska [22] studied the maintenance strategy of brand visual image in the process of large-scale development in view of the brand development status of small and medium-sized enterprises in the era of AI. It provided a marketing reference for the intelligent visual construction of small and medium-sized brands.

To sum up, the existing literature has studied brand visual identity from many dimensions, such as form design, factor function, scene application, effect communication, etc., and made clear the important role of visual elements in building brand image, shaping consumers' cognition, and improving marketing effectiveness. Related research has gradually formed a development vein from micro-visual elements to macro-visual systems, and from commercial scenes to the public domain. However, the existing research mostly focuses on the static analysis and effect verification of brand visual recognition. On the one hand, it lacks the exploration of cutting-edge generation technologies such as diffusion models, multimodal Transformers, and CLIP-guided generation, and the application of the current mainstream text-to-image generation model is not discussed enough. On the other hand, there is no mature technical scheme for automatic generation and dynamic optimization of a brand visual identity system. Especially, there is little systematic research on the application of a multimodal generation confrontation network in BVIS construction. The above research gap has become the core breakthrough point of this study to carry out research on an automatic generation and

optimization algorithm of BVIS based on MM-GAN. It has also supplemented new theoretical and technical directions in this field.

III. Construction of BVI Model

A. Model Design

The automatic generation and optimization model of BVIS constructed in this study takes MM-GAN as the core architecture and integrates multi-source modal data such as text style description, color parameters, and graphic structure features. Fig. 1 shows the structure of the GAN.

In Fig. 1, the model construction first carries out multimodal data preprocessing and alignment operations, and maps heterogeneous modal data to a unified feature space, laying a solid foundation for subsequent generation tasks. The multimodal input set is defined as $X = \{X_t, X_c, X_g\}$. X_t stands for the text style description sequence, X_c is the color parameter matrix, and X_g is the basic graphic structure feature. To eliminate the dimensional heterogeneity between different modal data, the modal data are standardized. Taking the normalization of the color parameter matrix X_c as an example, the calculation is shown in Eq. (1), and the values are mapped to the interval [0,1]:

$$X_c^* = \frac{X_c - min(X_c)}{max(X_c) - min(X_c)} \quad (1)$$

Aiming at the text mode X_t, the pre-trained language model Bidirectional Encoder Representations from Transformers (BERT) is used to carry out semantic encoding, and the text semantic feature vector F_t is obtained, as shown in Eq. (2):

$$F_t = \text{BERT}(X_t; \theta_b) \quad (2)$$

θ_b is the trainable parameter set of the BERT model [23]. Multimodal feature alignment is realized by constructing a modal attention matrix, the core of which is to calculate the semantic association weight between text features and color and graphic features. The calculation of text-color feature alignment weight is shown in Eq. (3):

$$W_{tc} = \frac{\exp(F_t \cdot F_c^T / \sqrt{d_k})}{\sum_{i=1}^{n} \exp(F_t \cdot F_{c,i}^T / \sqrt{d_k})} \quad (3)$$

F_c is the color feature vector, d_k is the feature dimension, and W_{tc} is the text-color alignment weight matrix [24]. Based on the alignment weights, the multimodal feature fusion is completed, and the fusion feature F_{fusion} is obtained. The calculation process is shown in Eq. (4):

$$F_{fusion} = W_{tc} \cdot F_c + W_{tg} \cdot F_g + F_t \quad (4)$$

W_{tg} is the weight matrix of text-graphic feature alignment. The generator module adopts the improved U-Net architecture, and takes the fusion feature F_{fusion} as the input to generate the initial BVIS visual sample $G(F_{fusion})$. The generation process can be represented by Eq. (5):

$$G(F_{fusion}) = \text{U-Net}(F_{fusion}; \theta_g) \quad (5)$$

θ_g is the trainable parameter set of the generator [25]. The discriminator module adopts a PatchGAN structure, and its core function is to distinguish generated samples from real BVIS samples. Its discriminant probability output is shown in Eq. (6):

$$D(Y; \theta_d) = \sigma(\text{PatchGAN}(Y; \theta_d)) \quad (6)$$

Y is the input sample (including real samples and generated samples), θ_d is the discriminator parameter set, and σ stands for the Sigmoid activation function [26]. The core loss function of the model countermeasure training includes the generator loss L_g and discriminator loss L_d, in which discriminator loss consists of real sample loss and generated sample loss, as shown in Eq. (7):

$$L_d = -\mathbb{E}_{Y \sim P_{real}}[\log D(Y)] - \mathbb{E}_{Z \sim P_z}[\log(1 - D(G(Z)))] \quad (7)$$

P_{real} is the true BVIS sample distribution, P_z is the potential distribution of fusion features, and Z is the potential feature vector [27]. In addition to counteracting the loss, the generator loss introduces additional perceptual loss to improve the visual quality of generated samples, and its expression is shown in Eq. (8):

$$L_g = -\mathbb{E}_{Z \sim P_z}[\log D(G(Z))] + \lambda_p \cdot L_{percept}(G(Z), Y_{real}) \quad (8)$$

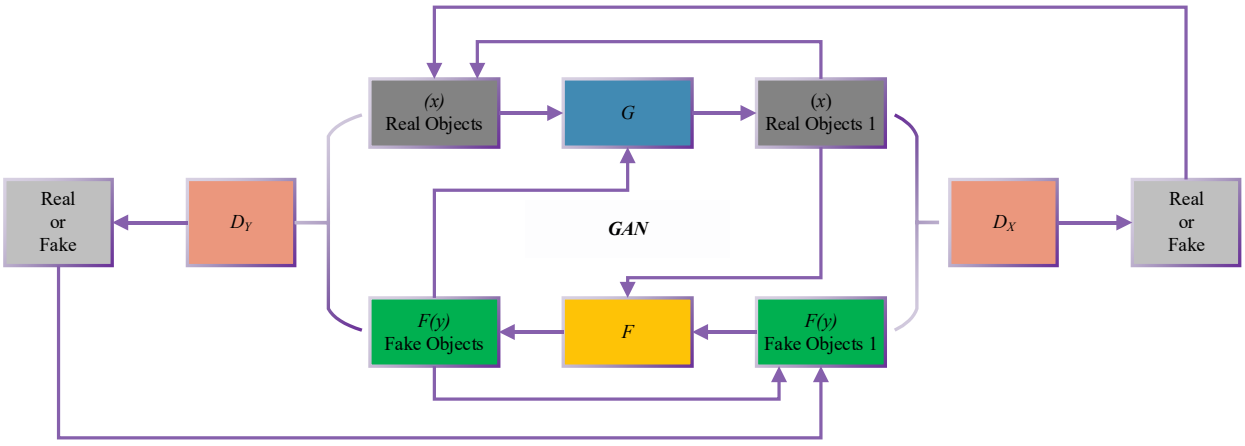

Fig. 1. Structure of GAN.

λ_p is the weight of perceptual loss, and $L_{percept}$ is the function of perceptual loss. Visual Geometry Group 16-Layer (VGG16) model is used to extract features and calculate Euclidean distance, as shown in Eq. (9):

$$L_{percept}(G(Z), Y_{real}) = \sum_{l=1}^{L} \frac{1}{H_l W_l C_l} \| \phi_l(G(Z)) - \phi_l(Y_{real}) \|_2^2 (9)$$

$\phi_l(\cdot)$ is the feature extraction function of the lth layer of VGG16. H_l, W_l, C_l respectively correspond to the height, width, and number of channels of the l −th layer feature map [28]. To ensure the style consistency of each visual element (logo, font, and color) of BVIS, a style consistency constraint loss L_{style} is introduced. Its calculation is shown in Eq. (10):

$$L_{style} = \sum_{i=1}^{m} \sum_{j=i+1}^{m} \mathrm{Sim}(S_i, S_j) \quad (10)$$

m is the number of BVIS visual elements, S_i is the style feature vector of the i th element, and $\mathrm{Sim}(\cdot)$ is the cosine similarity function, as shown in Eq. (11):

$$\mathrm{Sim}(S_i, S_j) = 1 - \frac{S_i \cdot S_j}{\|S_i\|_2 \cdot \|S_j\|_2} \quad (11)$$

The total loss function of the model is the weighted sum of each loss term, and the expression is shown in Eq. (12):

$$L_{total} = L_d + \alpha \cdot L_g + \beta \cdot L_{style} \quad (12)$$

α, β are the weight coefficients of the generator loss and style consistency loss, respectively [29]. In the model optimization stage, the quantitative index of brand visual communication effect is taken as the guide, and the communication effect score S_{effect} is defined as shown in Eq. (13):

$$S_{effect} = \gamma \cdot R + (1 - \gamma) \cdot A \quad (13)$$

R is the identity score, A is the audience acceptance score, and γ is the weight coefficient [30]. An adaptive optimization function based on the propagation effect score is constructed to realize the dynamic adjustment of model parameters, as shown in Eq. (14):

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla_{\theta_t}(1 - S_{effect}) \cdot L_{total} \quad (14)$$

θ_t is the model parameter set of the t round of training. η is the learning rate. Through iterative training, the model converges to a stable state, and the BVIS sample output by the generator meets the convergence condition shown in Eq. (15):

$$\| L_{total}(t+1) - L_{total}(t) \| < \varepsilon \quad (15)$$

ε is the preset convergence threshold [31]. The model solves the problem of inaccurate semantic-visual mapping through the multimodal feature alignment and fusion mechanism [32].

B. Experimental Design

1) Model training environment, optimization strategy, and hyperparameter setting: This study conducts model training based on the PyTorch 1.12.1 deep learning framework. The hardware used is a parallel training cluster composed of multiple NVIDIA Tesla V100 32GB graphics cards. The model parameters are initialized using a standard Gaussian distribution to avoid training oscillations caused by extreme initial parameter values. The training process employs the Adam optimizer, with the first-order moment estimation decay coefficient β_1=0.9 and the second-order moment estimation decay coefficient $\beta_2 = 0.999$. The initial learning rate is set at 1×10^{-4}, and a piecewise learning rate decay scheme is adopted. After every 20 complete iterations, the learning rate is reduced to 0.8 times the current value, balancing the convergence speed in the early training stage and the accuracy of parameter fine-tuning in the later stage. The batch size is set to 16, and the total number of global iterations is 200. Based on the characteristics of the model task, the weight coefficients of the loss functions were determined: the generator loss weight $\alpha = 0.6$, the style consistency loss weight β =0.4, the perceptual loss weight $\lambda_p = 0.5$, and the visual propagation effect weight $\gamma = 0.7$. The convergence determination threshold ε is set at 1×10^{-5}. When the loss function fluctuation is less than this threshold for three consecutive rounds, the model is judged to have converged, and training is terminated. The improved U-Net network embeds three groups of serial dilated convolution modules in the encoding stage to expand the feature receptive field. The attention module adopts an 8-head multi-head attention mechanism, with a unified feature dimension $d_k = 128$, ensuring the integrity and computational efficiency of multimodal feature interaction. According to statistics, the average time per model iteration is 45 minutes, and the cumulative running time of the complete training process is approximately 150 hours.

2) Dataset preprocessing, modal matching, and quality control: In this study, the brand's industry is taken as the core association logic, and the strong binding matching rules of text-color-graphic multimodal data are established. The style description text and color parameter matrix corresponding to the same brand are matched one by one to the logo graphic samples one by one, ensuring the semantic homology of multimodal data from the source. Aiming at the original image data, a variety of online data augmentation strategies are introduced, including random cropping, horizontal flipping, and adaptive fine-tuning of brightness and contrast to enrich the diversity of samples. At the same time, the median filtering algorithm is used to denoise the image to eliminate the salt-and-pepper noise and clutter interference generated in the process of acquisition and transmission. Aiming at the problems of category imbalance and domain deviation in the dataset, the hierarchical resampling algorithm is used to balance the sample set. The sample distribution ratio is adjusted according to the industry and style dimensions to weaken the negative impact of data deviation on the generalization ability of the model. All labels of the datasets are labeled by professionals in the field of visual design according to the unified labeling specifications. After the labeling work is completed, two rounds of cross-checking and sampling verification are carried out to eliminate invalid and mislabelled samples, thus fully ensuring the standardization, accuracy, and credibility of labels.

3) Multi-modal feature alignment, fusion and complete operation process: After the normalized pretreatment of different modal data, the text semantic coding is completed by the BERT model, and the text feature vector with a fixed dimension is generated. The modal features of color and graphics are extracted synchronously, and the semantic association weights between text-color and text-graphics are calculated based on the modal attention matrix to achieve accurate alignment of cross-modal features. Based on the calculated attention weights, the multi-modal feature weighted fusion is completed, and the fusion feature vector is generated and input into the improved U-Net generator, and the initial brand visual identification system sample is output. The samples are sent to the PatchGAN discriminator to complete the authenticity discrimination, and the network parameters are updated by combining the resistance loss, perception loss, and style consistency loss. After each round of training, adaptive parameter optimization is performed based on the comprehensive score of visual communication effect, and the network weights are dynamically adjusted. The entire iterative process continues until the model loss function meets the convergence condition, ultimately outputting a brand visual recognition system solution with a unified style and semantic matching.

IV. Data Source

All the experimental data in this study are taken from public authoritative datasets, and the reliability of the data has been verified in many studies in related fields, which can effectively support the related research work of brand visual identity. Logo Detection 3 K dataset (LOGO DET-3K) is selected as the image graphic data, which is divided into 9 categories and 3,000 sub-categories, including 158,652 brand logo images and 200,000 labeled data. It is widely used in tasks such as brand logo detection and visual generation, and many related research studies in the industry have used this dataset to carry out experiments. The dataset is accessed at https://blog.51cto.com/u_16099190.

The color modal data is taken from the audio-visual cross-modal correlation dataset of the National Basic Science Public Science Data Center. The dataset is constructed based on the hue, saturation, and lightness color space, covering 8 hues, 32 standard color cards, and 50 groups of three-color color matching samples. It has complete color parameters and color scheme data, and can support color feature extraction, color semantic association analysis, and other work. The dataset can be accessed at https://www.nbsdc.cn/general/datalinks/16666.11.nbsdc.

The text style data adopts the Chinese text style classification test set of Magic Community. The dataset contains four typical text styles: news, science and technology, spoken language, and e-commerce, which can meet the model training needs of brand text semantic coding and style matching. The access address of the dataset is https://www.modelscope.cn/datasets/IIC/NLP_style_classification_Chinese_testset.

In the data usage stage, this study is based on the industry of the brand and describes the graphic samples, color parameters, and texts corresponding to the same brand. Binding and pairing to realize semantic homology of multimodal data. To solve the problems of class imbalance and data deviation in the dataset, this study uses the stratified resampling method to balance the number of samples, and optimizes the original images through image augmentation operations such as random cropping, horizontal flipping, brightness fine-tuning, and the median filtering algorithm to improve data quality. All data labels are uniformly marked by professionals, and two rounds of cross-checking are completed to ensure accurate and reliable labeling results. In Tables I to IV, the characteristics of the dataset, experimental hardware environment, experimental software environment, and model core parameters used in this study are shown in turn.

TABLE I. Statistics of Multimodal Experimental Datasets

Dataset Name	Data Type	Data Scale
LogoDet-3K	Graphic mode	9 categories, 3,000 subcategories, 158,652 images, and 200,000 labels.
Audio-visual cross-modal correlation dataset	Color mode	8 tones, 32 standard color blocks, 50 sets of three-color color matching.
Chinese text style classification test set	Text mode	Four styles: news, science and technology, spoken language and e-commerce

Data source: Compiled by the author.

TABLE II. Experimental Hardware Configuration

Hardware Type	Specific Configuration
CPU	Intel Xeon Gold 6248R 2.5GHz
GPU	NVIDIA Tesla V100 32GB
RAM	128GB DDR4 2933MHz
SSD	2TB SSD+10TB HDD
Graphics memory	A single card is 32GB, which supports multi-card collaborative training.

Data source: Compiled by the author.

TABLE III. Experimental Software Environment and Dependent Libraries

Software Category	Specific Version/Tool
Operating system	Ubuntu 20.04 LTS
Deep learning framework	PyTorch 1.12.1
Programming language	Python 3.8.16
Visual processing library	OpenCV 4.5.5

Data source: Compiled by the author.

TABLE IV. Dataset Partition and Repeated-Experiment Parameters

Parameter Category	Parameter Name	Value/ Configuration
Data partitioning	Training set proportion	80%
Data partitioning	Verification set proportion	10%
Data partitioning	Test set proportion	10%
Training parameters	Number of repeated experiments	10 rounds

Data source: Compiled by the author.

V. Evaluation of Multimodal Brand Identity Model

A. Comparative Evaluation of Core Performance Indicators

Based on the standard data division (80% of the training set, 10% of the verification set, and 10% of the test set), three core dimensions, namely, generation quality, semantic mapping accuracy, and style consistency, are selected and compared with mainstream models. After 10 rounds of repeated experiments, the random error is reduced, and the analysis results of the mean value, standard deviation, and significant differences between models are presented below. Fig. 2 shows the results of comparative evaluation of core performance indicators.

In Fig. 2, the lower Fréchet Inception Distance (FID) and Style Feature Distance (SFD) represent the higher quality of generated samples and the better style consistency. The higher Structural Similarity Index (SSIM) and Semantic Matching (SM) indicate that semantic mapping accuracy and visual structure similarity are better. The data show that the proposed model is significantly superior to the traditional GAN, DCGAN (Deep Convolutional Generative Adversarial Network), and mainstream multimodal generation models in all core indicators: the FID average of 16.9 is 24.9% lower than StyleGAN2, and the SFD average of 0.82 is 40.6% lower than MM-GAN. SSIM and SM are 0.92 and 0.89, respectively, which are 27.8% and 30.9% higher than the basic GAN. At the same time, the standard deviation of each index of the proposed model is the smallest, which shows that the training stability of the model is stronger. This verifies the effectiveness of the multimodal feature alignment mechanism and style consistency constraint. It can effectively eliminate the barriers of semantic-visual mapping and improve the accuracy and coordination of BVIS generation.

B. Quantitative Evaluation of Model Robustness and Generalization Abilities

Aiming at the heterogeneity of data and the complexity of scenes in practical applications, a multi-scenario quantitative test is designed. It includes three types of test scenarios: data noise interference (Gaussian noise, salt-and-pepper noise), cross-industry data adaptation (beauty, science and technology, catering), and modality missing adaptation (single-modal input, dual-modal input) to quantify the stable performance of the model under complex conditions. Fig. 3 shows the results of quantitative evaluation of model robustness and generalization ability.

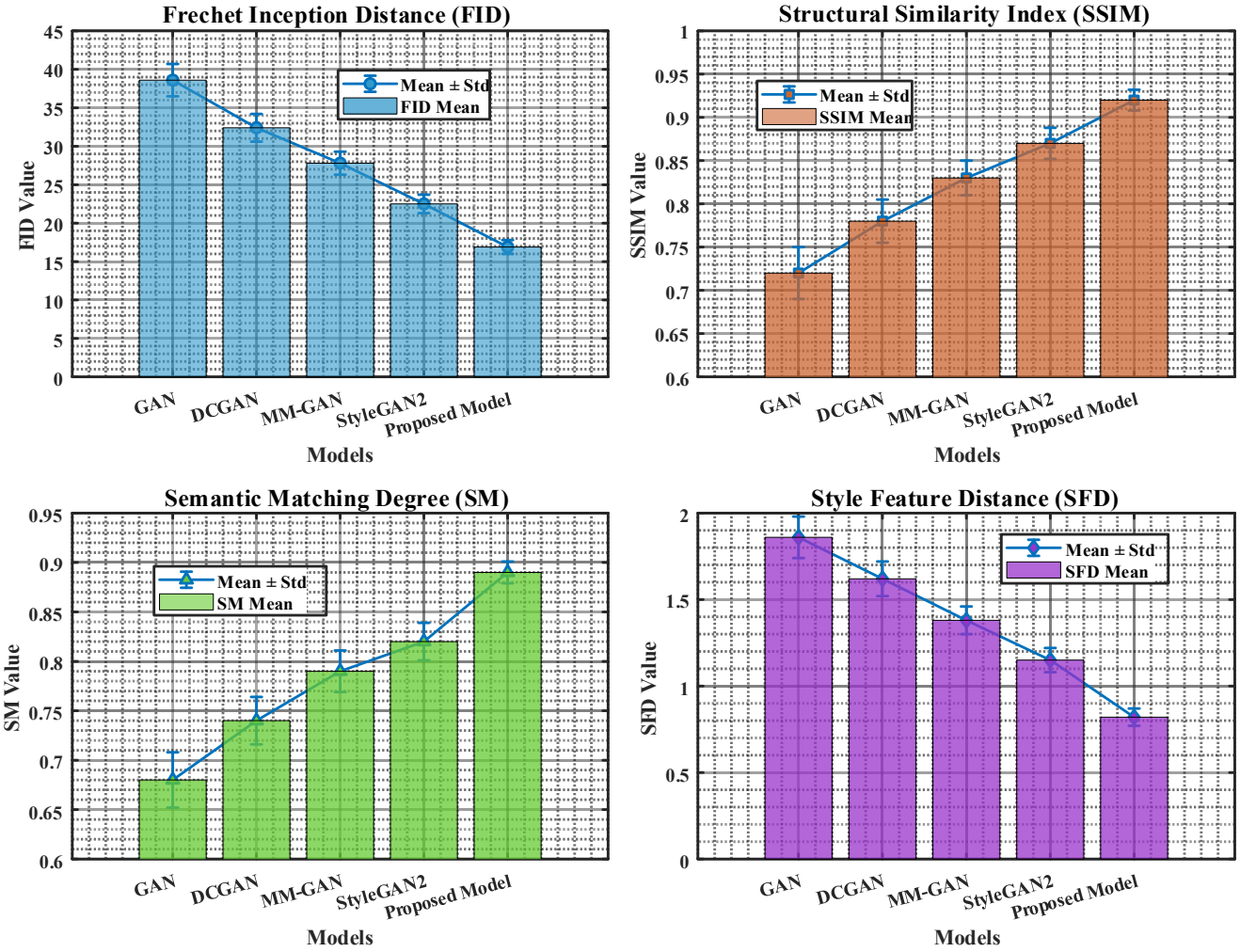

Fig. 2. Results of comparative evaluation of core performance indicators.

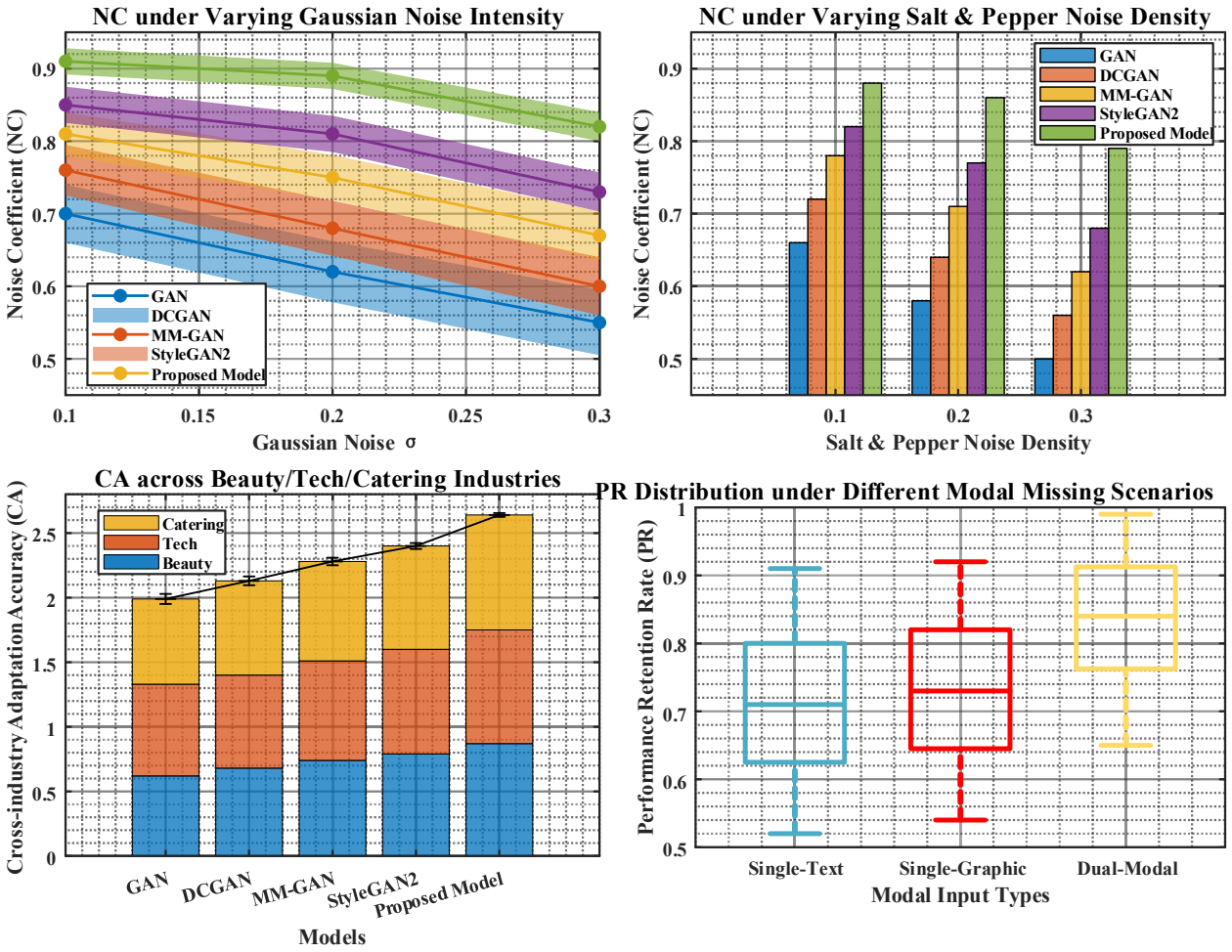

Fig. 3. Results of quantitative evaluation of model robustness and generalization ability.

In Fig. 3, the proposed model shows significant advantages in the noise-interference scene: the value of Noise Coefficient (NC) reaches 0.82 under σ=0.3 Gaussian noise, which is 12.3% higher than that of StyleGAN2. When the salt-and-pepper noise density is 0.3, the NC value is 0.79. It is 27.4% higher than that of MM-GAN, which verifies the strong anti-noise robustness. In cross-industry adaptation, the average value of total CA is 2.64, which is 32.7% higher than that of basic GAN. The beauty, technology, and catering industries all maintain high accuracy and outstanding generalization ability. In the scene of modal absence, the single-text-input Performance Retention (PR) reaches 0.84, and the dual-modal-input PR reaches 0.93. It is much higher than the contrast model, reflecting the effectiveness of the modal adaptation mechanism. On the whole, the proposed model is more stable under data heterogeneity and scene complexity. Its robustness and generalization ability can meet the practical application requirements of brand visual identification.

C. Evaluation of Ablation Experiment

To verify the independent contribution and synergistic effect of the core modules of the proposed model, an ablation experiment is designed to remove the multimodal feature alignment mechanism, style consistency constraints, and adaptive optimization modules one by one. The core performance indicators (FID, SSIM, SM, NC) are used as the evaluation basis to compare with the complete model. The experiment is based on the same dataset and training parameters. The effects of each module are isolated by the control variable method. The following is a quantitative analysis of the influence of each module on the model generation quality, robustness, and generalization ability, and the necessity of key technologies is clarified. Fig. 4 shows the results of the ablation experiment.

In Fig. 4, the ablation experimental results show that the complete model is the best in all core indicators: the lowest FID value (16.9) and the highest SSIM, SM, and NC values, which verifies the effectiveness of the synergy of each module. After removing the multimodal feature alignment mechanism, each index degrades most significantly, with FID rising by 69.8% and SM dropping by 19.1%, which shows that this module is the core to ensure semantic matching and generation quality. The lack of a style consistency constraint leads to a reduction in the structural similarity of generated content, while the adaptive optimization module mainly affects the anti-noise ability of the model (NC decreases by 7.9%). On the whole, the three modules are irreplaceable, and the contribution of the feature alignment mechanism is the most critical. Together, the three modules improve the generation quality, robustness, and semantic matching ability of the model.

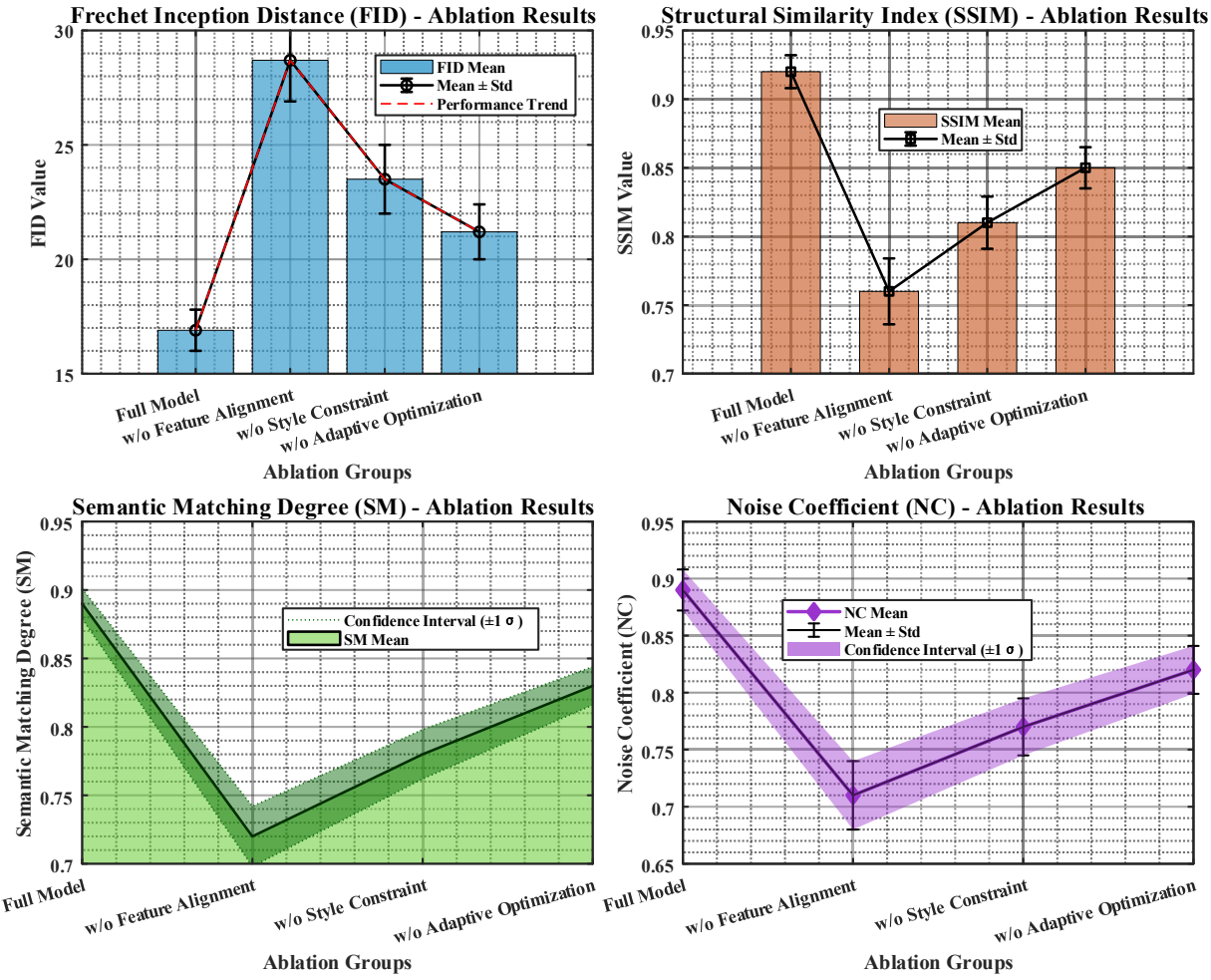

Fig. 4. Results of ablation experiment.

D. Comprehensive Evaluation Results

To highlight the comprehensive performance of this model, this study comprehensively analyzes the model through deeper comparative evaluation. Tables V-IX show the comprehensive evaluation results of the model.

In Tables V-IX, combined with the experimental data of each group, it shows that the comprehensive performance advantage of this model is remarkable. The data in Table V show that the FID value of this model is 16.9, which is the lowest among all comparison models, representing the best image generation quality. SSIM reaches 0.92. The semantic matching degree is 0.89, which is far superior to traditional models such as basic GAN and DCGAN, and superior to the CLIP guidance model and diffusion model. It proves that the multi-modal fusion scheme in this study can effectively improve the visual restoration degree and the text-visual semantic matching ability. The statistical results in Table VI show that the standard deviation of each index is small, the 95% confidence interval is narrow, and the p-values are all less than 0.05. This shows that the model performance is stable and the experimental results are statistically significant. The Kappa coefficients of all indicators in Table VII are higher than 0.8, which proves that the evaluation criteria of self-defined indicators such as semantic matching degree and communication effect score are scientific and the results are credible. From Table VIII, the parameters and reasoning time of this model are much lower than those of CLIP and diffusion models, which ensures the generation effect and gives consideration to the operation efficiency. In Table IX, in the cross-dataset test, the indicators of the model only drop slightly. There is no obvious performance decline, which shows excellent generalization ability. On the whole, this model is balanced in terms of generation quality, stability, operation efficiency, and cross-scene adaptability, and can adapt to the actual design requirements of a brand visual identification system.

TABLE V. COMPARATIVE RESULTS OF CORE PERFORMANCE INDICATORS FOR DIFFERENT GENERATION MODELS

Model	FID	SSIM	Semantic Matching Degree
Basic GAN	22.5	0.72	0.68
DCGAN	21.1	0.71	0.67
StyleGAN2	22.6	0.79	0.72
CLIP-guided generation model	19.3	0.83	0.77
Diffusion model	18.2	0.85	0.80
The proposed model	16.9	0.92	0.89

Data source: Compiled by the author.

TABLE VI. VERIFICATION RESULTS OF INDICATOR CONSISTENCY

Evaluation Indicator	Mean Value	Standard Deviation	95% Confidence Interval	Significance (p-Value)
FID	16.9	0.42	[16.48,17.32]	p<0.05
SSIM	0.92	0.018	[0.902,0.938]	p<0.05
Semantic matching degree	0.89	0.021	[0.869,0.911]	p<0.05
Noise figure	0.82	0.015	[0.805,0.835]	p<0.05
Cross-industry adaptation accuracy	2.64	0.033	[2.607,2.673]	p<0.05
Performance retention rate	0.84	0.024	[0.816,0.864]	p<0.05

Data source: Compiled by the author.

TABLE VII. CONSISTENCY VERIFICATION BETWEEN MANUAL SCORING AND ALGORITHM SCORING

Evaluation Indicator	Algorithm Average Score	Artificial Average Score	Kappa Coefficient
Semantic matching degree (SM)	0.89	0.87	0.86
Communication effect score	0.86	0.85	0.84
Performance retention rate (PR)	0.84	0.83	0.82

Data source: Compiled by the author.

TABLE VIII. PARAMETER SCALE AND INFERENCE EFFICIENCY COMPARISON OF VARIOUS MODELS

Model	Parameter Quantity (Ten Thousand)	Single Sample Inference Time (ms)
Basic GAN	426	28.6
DCGAN	512	32.1
StyleGAN2	1286	65.3
CLIP-guided generation model	2153	92.7
Diffusion model	1869	84.2
The proposed model	675	36.5

Data source: Compiled by the author.

TABLE IX. CROSS-DATASET GENERALIZATION PERFORMANCE TEST OF THE PROPOSED MODEL

Dataset	FID	SSIM	Semantic Matching Degree (SM)
Original dataset	16.9	0.92	0.89
Logo-2K+dataset	17.5	0.90	0.87

Data source: Compiled by the author.

VI. CONCLUSION

In the context of global brand digital transformation, traditional BVIS design heavily relies on the subjective experience of designers, resulting in problems such as long design cycles, style homogenization, and delayed market response. Multimodal generative adversarial networks offer a new approach to intelligent BVIS generation, but still face technical bottlenecks such as insufficient accuracy in multimodal semantic mapping, inconsistent visual element styles, and a lack of quantitative optimization mechanisms. This study uses MM-GAN as the basic architecture, integrating attention mechanisms and adaptive optimization strategies to complete multimodal data preprocessing alignment, feature fusion, generation constraints, and dynamic optimization, constructing an automatic BVIS generation and optimization algorithm. Experimental results show that the proposed model achieves a Fraser Inception distance of 16.9, significantly outperforming StyleGAN2. The structural similarity index and semantic matching (SM) reach 0.92 and 0.89, respectively. The model also demonstrates good anti-interference ability and generalization performance in scenarios with noise interference, cross-industry adaptation, and modality loss. Ablation experiments confirm that the multimodal feature alignment mechanism is the core module of the model, style consistency constraints ensure the internal coordination of the visual system, and adaptive optimization improves scene adaptability. These three elements synergistically enhance generation quality, semantic matching accuracy, and model robustness, providing technical support for BVIS intelligent design. Compared to diffusion and CLIP models, the proposed method has advantages in training efficiency and industry customization, but there is still room for improvement in detail generation capabilities. This study also discusses ethical application risks such as dataset bias, copyright of AI-generated content, and brand counterfeiting. Due to data limitations, the current dataset has limited industry coverage. Future research could expand the dataset with more perceptual modal information to extend the dynamic brand visual recognition system generation task, further enhancing the algorithm's engineering application value.

ACKNOWLEDGMENT

This work was supported by Zhoukou Normal University 2024 High-Level Talent Research Start-up Funding Project: “A Study on Narrative Design Strategies of Regional Public Brands for Agricultural Products in Central China from the Perspective of Cultural Memory” (No. ZKNUC2024053).

REFERENCES

[1] J. Suriadi, M. Mardiyana, and B. Reza, “The concept of color psychology and logos to strengthen brand personality of local products,” Linguistics and Culture Review, vol. 6, no. S1, pp. 839–856, 2022.

[2] M. Tourky, P. Foroudi, S. Gupta, and A. Shaalan, “Conceptualizing corporate identity in a dynamic environment,” Qualitative Market Research: An International Journal, vol. 24, no. 2, pp. 113–142, 2021.

[3] M. Ngoc N and H. Tien N, “Branding strategy for gamuda land real estate developer in ho chi minh city vietnam celadon city project,” Psychology and Education, vol. 58, no. 5, pp. 3308–3316, 2021.

[4] P. Srivastava, D. Ramakanth, K. Akhila, and K. K. Gaikwad, “Package design as a branding tool in the cosmetic industry: consumers’ perception vs reality,” SN Business & Economics, vol. 2, no. 6, p. 58, 2022.

[5] Y. G. Cui, P. van Esch, S. Phelan, “How to build a competitive advantage for your brand using generative AI,” Business Horizons, vol. 67, no. 5, pp. 583-594, 2024.

[6] J. Butcher and F. Pecot, “Visually communicating brand heritage on social media: champagne on Instagram,” Journal of Product & Brand Management, vol. 31, no. 4, pp. 654–670, 2022.

[7] A. Tzotzis, A. Tsagaris, and P. Kyratsis, “A novel computational-based visual brand identity (CBVBI) product design methodology,” Machines, vol. 10, no. 11, pp. 1065, 2022.

[8] A. Adamus-Matuszyńska, P. Dzik, J. Michnik, and G. Polok, “Visual component of destination brands as a tool for communicating sustainable tourism offers,” Sustainability, vol. 13, no. 2, p. 731, 2021.

[9] J. J. Yoo, "Visual strategies of luxury and fast fashion brands on Instagram and their effects on user engagement," Journal of Retailing and Consumer Services, vol. 75, p. 103517, 2023.

[10] U. Vinitha V, S. Kumar D, and K. Purani, "Biomorphic visual identity of a brand and its effects: a holistic perspective," Journal of Brand Management, vol. 28, no. 3, pp. 272–290, 2021.

[11] A. M. Erjansola, J. Lipponen, K. Vehkalahti, H.-M. Aula, and A.-M. Pirttilä-Backman, "From the brand logo to brand associations and the corporate identity: visual and identity-based logo associations in a university merger," Journal of Brand Management, vol. 28, no. 3, pp. 241–253, 2021.

[12] V. Singla and N. Sharma, "Understanding role of fonts in linking brand identity to brand perception," Corporate Reputation Review, vol. 25, no. 4, pp. 272–286, 2022.

[13] C. C. L. Tan, K. L. Chiew, K. S. C. Yong, Y. Sebastian, J. C. M. Than, W. K. Tiong, "Hybrid phishing detection using joint visual and textual identity," Expert Systems with Applications, vol. 220, p. 119723, 2023

[14] K. Turoń, A. Kubik, M. Ševčovič, J. Tóth, and A. Lakatos, "Visual communication in shared mobility systems as an opportunity for recognition and competitiveness in smart cities," Smart Cities, vol. 5, no. 3, pp. 802–818, 2022.

[15] C. Lelis, S. Leitao, Ó. Mealha, and B. Dunning, "Typography: the constant vector of dynamic logos," Visual Communication, vol. 21, no. 1, pp. 146–170, 2022.

[16] B. Andrade, R. Morais, and E. S. de Lima, "The personality of visual elements: A framework for the development of visual identity based on brand personality dimensions," The International Journal of Visual Design, vol. 18, no. 1, p. 67, 2024.

[17] D. Lusianti, W. Widodo, and M. Mulyana, "Mas'uliyah society brand resonance: enhancing sustainable marketing performance of the national health insurance program," Qubahan Academic Journal, vol. 4, no. 3, pp. 619–637, 2024.

[18] D. Zha, P. Foroudi, T. C. Melewar, and Z. Jin, "Examining the impact of sensory brand experience on brand loyalty," Corporate Reputation Review, vol. 28, no. 1, pp. 14–42, 2025.

[19] Y. Cai, "Innovative design and development of brand visual system under the background of artificial intelligence-assisted visual communication design," Discover Applied Sciences, vol. 7, no. 8, p. 864, 2025.

[20] Y. Hwang, "A study on brand design methodology using generative AI," Journal of Digital Design & Culture, vol. 14, no. 2, pp. 41-58, 2025.

[21] L. M. Ho-dac and M. Mulder-Nijkamp, "Brands in transition: balancing brand differentiation and standardization in sustainable packaging," Sustainability, vol. 17, no. 6, p. 2381, 2025.

[22] L. Chulayevska, "Safeguarding brand identity while scaling: strategic marketing for small and medium companies in the AI age," Актуальні питання економічних наук, vol. 4, no. 21, pp. 207–218, 2026.

[23] W. Lim, K. S. C. Yong, B. T. Lau, and C. C. L. Tan, "Future of generative adversarial networks (GAN) for anomaly detection in network security: a review," Computers & Security, vol. 139, pp. 103733, 2024.

[24] H. Wang, H. Qian, and S. Feng, "GAN-STD: small target detection based on generative adversarial network," Journal of Real-Time Image Processing, vol. 21, no. 3, p. 65, 2024.

[25] H. Westerbeek and T. Van Schaik, "Platform power, athlete branding, generative AI, and the future of sport governance–a systematic review," Frontiers in Sports and Active Living, vol. 7, p. 1642180, 2025.

[26] P. Yuan, "A research on the dynamization effect of brand visual identity design: Mediated by digital information smart media," Journal of Information Systems Engineering and Management, vol. 9, no. 1, p. 24153, 2024.

[27] Y. Zhang, "Generational perspectives on logo complexity: Influencing luxury perception and purchase intention," Frontiers in Communication, vol. 10, p. 1475326, 2025.

[28] W. Liu, X. Cun, and C. M. Pun, "DH-GAN: image manipulation localization via a dual homology-aware generative adversarial network," Pattern Recognition, vol. 155, p. 110658, 2024.

[29] S. P. Scott, D. Sheinin, and L. I. Labrecque, "Small sounds, big impact: sonic logos and their effect on consumer attitudes, emotions, brands and advertising placement," Journal of Product & Brand Management, vol. 31, no. 7, pp. 1091–1103, 2022.

[30] B. Sunarso and F. Mustafa, "Analysing the role of visual content in increasing attraction and conversion in MSME digital marketing," Journal of Contemporary Administration and Management (ADMAN), vol. 1, no. 3, pp. 193–200, 2023.

[31] K. G. Yulius, N. Yuliantoro, and Y. D. Timba, "Kota Rempah: strengthening ternate's brand identity through gastronomic souvenirs for a sustainable tourism," Journal of Research on Business and Tourism, vol. 5, no. 1, pp. 16–33, 2025.

[32] S. Presutti, "Authentic Roman type: historical legacies in contemporary Rome's city brand," Visual Communication, vol. 24, no. 4, pp. 790–807, 2025.