ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Research on Packaging Graphic Design Methods for Regional Specialty Agricultural Products Based on Multimodal AI Generation Technology

Xiyao Jiang[1,*]

[1] Shanghai Art & Design Academy School of Product Design 201808

SUMMARY: *In order to solve the problems of long artificial conception cycle, unstable regional semantic expression and lack of consistency of series styles in the packaging graphics design of regional agricultural products, this paper constructs a packaging graphics design method based on multimodal AI generation technology. In this study, the product name, origin description, cultural keywords, regional pattern images, product photos and layout constraints are integrated into the unified computing framework. Through multi-modal feature extraction, cross-modal alignment, condition generation and discriminant evaluation, the intelligent generation of packaging graphics is realized. The experiment was carried out based on 4260 groups of multimodal samples, and the training set, validation set and test set were divided by 7:2:1. The results show that the SSIM of the proposed model reaches 0.871, the PSNR reaches 31.84 dB, the text-image semantic consistency reaches 0.903, and the comprehensive color coordination score is 8.74, which are better than those of the template stitching method, the single-modal text-generated image model and the basic condition generation model. In the tea packaging task, SSIM is further improved to 0.884 and PSNR reaches 32.41 dB. Research shows that this method can better coordinate the relationship between regional culture translation, commodity recognition and visual organization, and provide a more operational implementation path for the combination of agricultural product packaging graphic design and computer generation technology.*

KEYWORDS: *Multi-modal AI generation; Regional characteristic agricultural products; Packaging graphic design; Cross-modal feature fusion*

1 Introduction

Under the background of the continuous integration of digital economy, platform retail and intelligent design tools, the market competition of regional characteristics of agricultural products is no longer solely determined by origin resources or processing quality. The identification, narrative and communication functions of packaging graphics are becoming more and more critical. For tea, dried fruits, cereals, mushrooms, honey and local snack food, consumers often cannot directly contact the product ontology before purchase. Therefore, packaging graphics have become an important medium to connect the origin image, product attributes and consumption judgment. In reality, although many local agricultural products have distinct regional culture and product characteristics, the packaging design is still in the stage of experience collage, symbol stacking or template application, and the problems of graphic style dispersion, insufficient recognition and superficial cultural expression are relatively common. This kind of design method is not only difficult to adapt to the fast-paced

*jiangxiyao369@126.com
https://doi.org/10.65102/is2026709 Published: 30 April, 2026

communication environment of e-commerce platform, but also difficult to form a stable, transferable and extensible visual system [1].

Taking the actual design process of several regional agricultural products brands as an example, the development of traditional packaging graphics usually relies on designers to repeatedly retrieve materials, manually refine local elements and complete multiple rounds of scheme adjustment. From the preliminary research to the completion of the draft, the cycle is often measured in weeks, and different designers have different understandings of "regionality", "modern sense" and "consumer friendliness", which leads to obvious fluctuations in the quality of the scheme. It is worth noting that agricultural product packaging does not only pursue "good looking", but also takes into account category identification, information hierarchy, green perception, cultural credibility and visual grasping ability in online thumbnail scenes [2]. Purely relying on human experience to deal with these entangled goals is often inefficient and expensive to modify. When the product line is extended to different specifications, different seasons or different marketing scenarios, traditional methods are more prone to style breakage and visual distortion.

From the existing research, the field of packaging design has begun to pay attention to the relationship between visual elements and consumer cognition, and certain progress has been made in color, material, complexity, natural sense and regional symbol composition. At the same time, the development of computer vision, text-generated images, multi-modal large models and controllable diffusion models also provide new technical entry for packaging graphics design [3]. Some studies attempt to improve creative efficiency with image generation models, while others focus on user preference prediction, visual attention distribution, and packaging evaluation mechanisms. However, there are still several shortcomings in the existing results. Firstly, many researches use the generative model for general visual creation, but few systematically model the specific scene of regional characteristic agricultural products, which has both cultural attributes and commodity attributes. Second, some methods focus on the unidirectional generation from text to image, and do not consider the collaborative relationship between heterogeneous information such as local patterns, color semantics, product categories, and consumption scenes. Third, the generated results often emphasize novelty, but ignore the actual requirements of packaging graphics in brand consistency, recognition stability and market suitability [4].

At present, the multimodal AI generation technology with more explanatory and control power provides a new solution to the above problems. In this technology, packaging graphics are no longer regarded as isolated images, but textual descriptions of origin, product attribute words, regional cultural symbols, historical image samples, consumer preference labels and existing packaging styles are integrated into a unified modeling framework, and the computational reconstruction of graphic design process is realized through cross-modal alignment, feature coding, condition control and generation feedback [5]. Based on this understanding, this paper intends to introduce the multimodal AI generation technology into the packaging graphics design of regional agricultural products, and build a method system for "element extraction, semantic representation, graphics generation and result evaluation", so that the model can improve the efficiency, stability and adaptability of packaging graphics generation while retaining the regional cultural identification [6].

The core question this paper tries to answer is: whether multi-modal feature fusion can express the local characteristics of regional agricultural products more effectively than single image generation or empirical design; After the introduction of text, image and style constraints, whether the packaging graphics generation results can achieve synchronous improvement in visual coordination, cultural fit and consumption recognition efficiency; Whether the constructed model can maintain a good migration ability between different types

of regional agricultural products without obvious distortion due to large sample differences. The research value of this paper is to promote the multi-modal generation model from general visual generation to the design scene of agricultural product packaging graphics with more application constraints. It not only discusses "whether the generation can be generated", but also discusses "whether the generation is accurate, usable, and consistent with the regional brand communication logic". At the theoretical level, this paper is helpful to expand the application boundary of multi-modal generation technology in packaging visual design. In practice, it can provide an intelligent graphic design path for regional characteristic agricultural products that takes into account cultural expression and design efficiency.

2 Related Research

2.1 Research progress on graphic design of regional characteristic agricultural products packaging

The early research on the graphic design of regional agricultural product packaging mainly focused on the refinement of regional cultural symbols, color image matching and the translation of traditional visual elements. Related work generally believes that local food packaging not only bears the function of identifying commodity categories, but also carries the construction tasks of origin memory, cultural attribution and consumption association [7, 8]. At this stage, the design methods mostly rely on manual experience, and organize folk patterns, architectural contours, landscape images or manual fonts into packaging graphics language. Although it can strengthen regional identification to a certain extent, it is prone to problems such as symbol stacking, style convergence and information hierarchy confusion. Especially in the scenarios of e-commerce thumbnails, social communication graphs and serial extension, it is often difficult to balance the identification efficiency and visual consistency for graphic schemes formed solely by static experience [9, 10].

With the advancement of computer-aided design and user perception research, packaging graphics research has begun to introduce more quantitative characteristics of the analysis path. Some scholars have discussed the impact of packaging forms on consumer evaluation, natural perception and purchase intention from the perspective of visual complexity, color and material perception [11, 12]. Some studies have also analyzed the effect of external packaging cues on attention allocation and decision-making behavior by combining eye movement and expression feedback data [13]. This shows that the packaging graphics of regional characteristic agricultural products are no longer just an aesthetic problem, but gradually enter the computational research framework of "visual generation, user perception, and market feedback". However, the existing research still stays at the level of design evaluation or element analysis, and the linkage modeling between regional cultural semantics, product attribute information and graphics generation process is still insufficient, which also leaves a clear space for the subsequent introduction of multimodal AI generation methods.

2.2 Application Development of multi-modal AI generation technology in packaging visual design

With the continuous evolution of deep learning generative models, multimodal AI has begun to provide new technical fulcrums for packaging visual design. Compared with the early computer-aided design methods that rely on single image retrieval, template stitching or parametric typesetting, the text-image joint generation model can understand the correspondence between product attributes, style descriptions and visual intention at the

semantic level, thus significantly expanding the generation space of packaging graphics [14, 15]. In this process, diffusion model, image prompt adapter and instruction editing model gradually become the core technology path. Related studies show that the model can not only generate image results with high integrity according to text prompts, but also complete style transfer, structure adjustment and detail redrawing by combining reference images, local constraints and editing instructions, which provides a computable basis for fast iteration of packaging graphics [16, 17].

However, early generation methods mainly focus on general image creation, and still have obvious limitations in packaging design scenarios. First, the model often emphasizes the visibility of the picture, but it is difficult to stably control the brand identity, information hierarchy and seriation consistency. Second, the single text-driven mechanism is still coarse in the depiction of complex information such as local patterns, agricultural product forms and cultural semantics of origin, and the generated results are prone to the problem of "good-looking images but not suitable for commodities" [18]. In order to alleviate these shortcomings, in recent years, multi-modal generation research has begun to emphasize conditional control, contextual learning and interactive design exploration. By introducing structural conditions, image prompts, semantic constraints and designer feedback, the model has stronger controllable generation ability [19, 20]. At the same time, research on generative AI in the field of visualization also points out that the value of a generative system is not only reflected in the output images, but also in the compression of multi-source information into actionable design variables, and the selection and optimization of schemes through human-machine collaboration [12]. In order to more clearly present the technological progressive relationship between the research on regional characteristic agricultural product packaging graphics and the application of multimodal AI, this paper summarizes the related research paths, as shown in Table 1. As can be seen in the table, packaging graphics research has been organized from regional elements dominated by human experience, and gradually shifted to a comprehensive research framework combining computer-aided analysis, user-perceived quantification, and multimodal controllable generation. The research focus also extends from "whether the graphics have local characteristics" to "whether the graphics can be stably generated, accurately identified and effectively transformed into market communication capabilities". This means that the multi-modal AI generation technology is promoting the packaging visual design from "manual drawing" to a new stage of "semantic modeling-conditional generation-interactive correction", and also provides a method basis for the intelligent construction of regional characteristics of agricultural products packaging graphics.

Table 1: Comparison of regional characteristic agricultural product packaging graphics research and multimodal AI application paths

Research Stage	Main Focus	Computer Technology Support	Role in Packaging Graphic Design	Main Limitations
Experience-driven stage of regional symbolic design	Extraction of folk patterns, local architecture, landscape imagery, and handcrafted typography	Basic graphic software and layout tools	Strengthens regional identity and establishes an initial visual style	Relies heavily on designers' experience, with a tendency toward stylistic homogeneity and limited scalability for series development
Computer-aided stage of user perception analysis	Relationship among color, material, visual complexity, and consumer perception	CAD tools, eye-tracking, facial expression recognition, and statistical data analysis	Shifts packaging evaluation from subjective judgment to quantitative analysis	Places greater emphasis on outcome evaluation, with insufficient attention to generation mechanisms
Unimodal generative design stage	Generation of packaging visual schemes based on text or image prompts	Text-to-image models and image editing models	Improves design generation efficiency and expands the creative space	Regional semantic expression is unstable, and brand consistency control is relatively weak
Multimodal controllable generation stage	Joint generation by integrating text, images, style constraints, and design feedback	Diffusion models, conditional control modules, image prompt adapters, and multimodal large models	Enhances controllability of generation and the precision of local cultural expression, making it suitable for serialized design	Data construction is costly, and the evaluation system still needs further improvement

2.3 Co-evolution of multimodal generative models and design optimization methods

With the parallel development of generative models and design optimization methods, the research on packaging graphics has gradually moved away from the path of relying solely on artificial conception and static template call. In the early stage, related research mostly uses generative adversarial networks to realize packaging image synthesis. The core mechanism is that the generator outputs the candidate graphics, the discriminator determines the difference between them and the real packaging samples, and continuously compreses the distribution distance between them through adversarial training. For the packaging of regional characteristic agricultural products, the value of this mechanism is that it can encode origin patterns, product appearance, main colors and decorative styles into learnable features, so as to improve the overall integrity and visual coordination of graphics generation [21]. However,

GAN-like methods still have shortcomings in complex text control, local structure stability and seriation style consistency, and are prone to the problem that the graphics seem complete but are not tightly combined with the real regional context.

With the introduction of diffusion model, conditional control module and multi-modal large language model, the generation of packaging graphics has entered the stage of "generation-constraint-feedback" linkage optimization. Textual conditional image generation model can map product name, regional narrative and style description into visual semantic space, and conditional control technology further emplaced contour sketches, layout boundaries, color constraints and reference images into the generation process, thus enhancing the controllability of output results [22, 23]. On this basis, multimodal instruction editing and agent-based generation methods incorporate designers' language feedback, local modification requirements and historical scheme preferences into the unified optimization link, so that the model is not only responsible for "generating graphs", but also capable of performing "modifying graphs" and "aligning design intentions" [24, 25]. This kind of method is particularly important for regional characteristic agricultural products, because its packaging graphics should not only reflect local recognition, but also adapt to commercial communication scenarios, and a single generation ability is not enough to solve the design adaptation problem.

In order to more clearly distinguish the applicable boundaries of different technical paths in packaging graphic design, this paper summarizes and compares the traditional rule method, GAN generation method, diffusion model, conditional control diffusion model and multi-modal surrogate generation method in Table 2. As can be seen from Table 2, although the rule or template method has the characteristics of fast writing speed and low operation threshold, its presentation of regional cultural connotation is still limited, and the generated results often stay at the surface symbol call level, which is difficult to form a deeper visual translation. The GAN method has significantly improved the graphics realism, local pattern learning and overall completion compared with traditional methods, but there are still shortcomings in complex text constraints, structural stability and seriation control. The introduction of diffusion model significantly broadens the creative generation space of packaging graphics, so that the visual output can more fully respond to the semantic and style description of the text. Furthermore, when the conditional control mechanism and multimodal interaction strategy are incorporated into the generation process, the model truly has the ability to be close to the actual design task, and can achieve a relatively stable balance between structure control, style unification and regional semantic preservation. It can be seen that the co-evolution of multimodal generation model and design optimization method is not only to improve the delicacy of image surface, but also to promote the generation of packaging graphics from visual output to applicable and extensible design implementation.

Table 2: Comparison of multimodal generative models and packaging design optimization methods

Method Type	Main Technical Characteristics	Role in Packaging Graphic Design	Support for Regional Feature Expression	Main Limitations
Traditional Rule-/Template-Based Methods	Fixed layouts and manual invocation of graphic elements	Facilitate rapid draft production	Low; expression often remains at the level of symbolic collage	Single style and weak scalability
GAN-Based Generative Methods	Adversarial training between generator and discriminator	Improve graphic integrity and visual realism	Moderate; capable of learning local pattern features	Weak text control and insufficient stability
Text-to-Image Diffusion Models	Visual generation based on text prompts	Expand the space for creative exploration	Relatively strong; capable of expressing product semantics	Regional symbols are prone to overgeneralization
Conditional Controlled Diffusion Models	Introduction of contours, reference images, and local constraints	Improve layout and structural controllability	Strong; suitable for serialized packaging generation	High training cost
Multimodal Agent-Based Generation and Editing	Integration of language instructions, image prompts, and feedback optimization	Support generation, editing, and iterative scheme refinement	Very strong; closer to the actual design workflow	Evaluation standards have not yet been fully unified

3 Research Methods

3.1 Packaging graphics design framework driven by multimodal AI generation technology

In order to get rid of the limitations of experience splicing and isolated drawing modification in the generation process of regional agricultural product packaging graphics, this paper constructs a packaging graphics design framework composed of multi-source input, cross-modal representation, condition generation and result evaluation. The framework does not regard packaging graphics as a single image output result, but integrates product names, origin narratives, regional cultural symbols, existing packaging samples, color preferences, consumption scene labels and layout constraints into the unified computing link to establish a stable mapping relationship between the semantic layer and the visual layer. The purpose is to

make the graphics generated by the model not only have visual integrity, but also maintain a high degree of coordination in regional identification, category communication and series consistency.

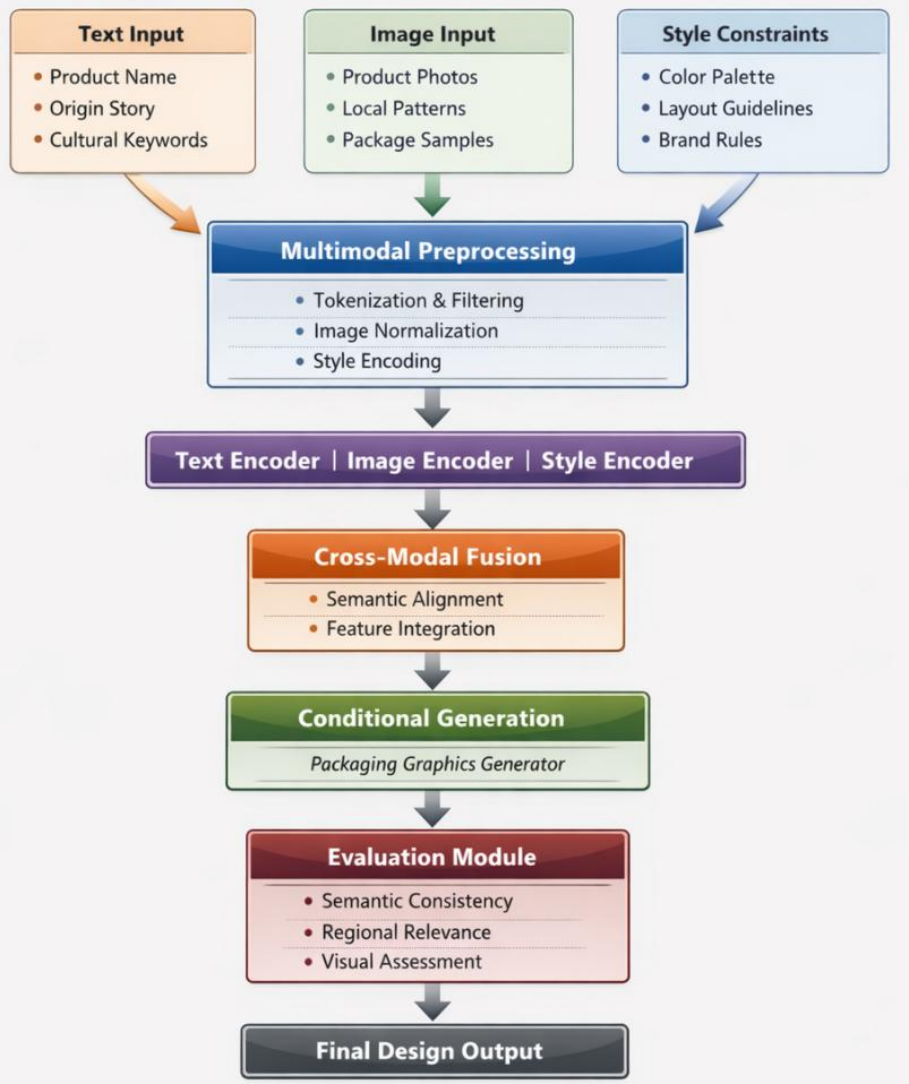

Figure 1: Framework for packaging graphics design driven by multimodal AI generation techniques

As shown in Figure 1, the overall framework consists of four levels. The input layer is responsible for receiving text descriptions, reference images and style control information. The text end contains agricultural product categories, origin characteristics, functional selling points and cultural keywords, the image end contains local patterns, physical photos, traditional packaging samples and visual cases of competitive products, and the control end gives page boundaries, dominant color ranges, complexity thresholds and brand identification requirements. After entering the preprocessing stage, the system performs word segmentation, keyword screening and semantic embedding on the text, normalizes the size of the image, converts the color space, cuts the pattern area and suppresses noise, and reduces the irrelevant interference through label cleaning and style coding. After normalization, the multi-modal data are sent to the text encoder, image encoder and style encoder respectively to form feature vectors that can be used for joint learning.

In the feature modeling link, this paper uses a cross-modal fusion module to align

multi-source features. Let the text input be T, the image input be I, the style constraint input be S, and the corresponding encoding results be $E_t(T)$, $E_i(I)$ and $E_s(S)$ respectively, then the joint representation vector H can be written as follows.

$$H = \phi(W_t E_t(T) + W_i E_i(I) + W_s E_s(S)) \quad (1)$$

where, W_t, W_i and W_s are the weight matrices of different modes, and $\phi(\cdot)$ represents the nonlinear mapping function. The significance of this formula does not lie in the simple superposition of features, but in the learnable weighted coupling of text semantics, image morphology and style preference according to packaging design tasks, so that the model can perceive "what to sell", "what to look like" and "what visual way to present" at the same time.

In the generation stage, the joint representation H is input into the graph generator together with the layout condition C, and the packing graph result Y is output. Its generation relation can be expressed as follows.

$$Y = G(z, H, C) \quad (2)$$

Here, z is a random perturbation variable and G represents the generative network. The purpose of introducing C is to include the main visual area, text blank area, label information area and auxiliary decoration area in the packaging layout into the condition control, so as to avoid the problems such as graphics crowding, information gland or visual center deviation in the generated results. After the generation is completed, the candidate graphics are sent to the evaluation terminal, and the comprehensive judgment of semantic consistency, regional fit, color coordination and recognition clarity is carried out, and the feedback results are sent back to the generator to form a closed-loop optimization.

3.2 Multi-modal feature extraction and representation of regional characteristic agricultural product packaging elements

In the generation task of regional characteristic agricultural product packaging graphics, what really affects the output quality is not only the model scale, but whether the input elements can be accurately disassembled, stably encoded, and formed a joint representation with design significance. Different from general image generation, packaging design involves more complex sources of elements, including textual information such as product names, origin descriptions, process descriptions, and cultural keywords, as well as visual constraints such as physical images of agricultural products, local patterns, traditional vessel forms, existing packaging samples, and target color systems, layout boundaries, and style labels. Therefore, this paper divides the extraction of packaging elements into three branches: text modality, image modality and style modality, and completes the cross-modal aggregation through a unified representation space, so that the subsequent generation process can simultaneously retain regional identification, category features and visual order.

The text modality is mainly responsible for carrying the semantic information of "what to say". The system segments and cleans the product name, regional source, cultural image, functional selling point and consumption scene description, and constructs the word element sequence $X_t = \{x_1, x_2, \ldots, x_n\}$. After being mapped by the text encoder, the semantic feature matrix is obtained:

$$F_t = E_t(X_t) \quad (3)$$

where $E_t(\cdot)$ represents the text encoding function and F_t corresponds to the base

representation of the wrapper semantic layer. This feature can reflect the differences in category attributes and regional narrative of different products such as "high mountain tea", "ancient honey" and "mountain mushroom", and provide a conditional basis for the selection of graphic themes.

The image modality is responsible for extracting visual cues of "what looks like". In this paper, agricultural product photos, local pattern images, traditional packaging patterns and auxiliary reference images are uniformly scaled to standard sizes, and color normalization, edge enhancement and region cropping are performed to reduce the interference of background noise on recognition results. Let the input image be Xi, and the image features are obtained after the visual encoder:

$$F_i = E_i(X_i) \tag{4}$$

On this basis, in order to highlight the local regions with recognition value, we introduce an attention-based feature enhancement mechanism. Global average pooling and Max pooling are performed on image feature F_i respectively to obtain statistical vectors $P_{avg}$and P_{max}, which are shared to generate visual attention weights A_i:

$$A_i = \sigma(W_2\delta(W_1(P_{avg} + P_{max}))) \tag{5}$$

where σ is the Sigmoid function and δ is the nonlinear activation function. Through this weight, the system can strengthen key graphic areas such as mountain contour, grain texture, fruit section, and folk pattern boundary, avoiding the model from evenly distributing attention in complex background.

Style modalities are used to express design constraints on how it should be presented. In this paper, comprehensive color vectors, composition density parameters, decorative complexity indicators, and style labels are extracted from the reference packaging samples, which are denoted as X_s. After processing by the style encoder, the style feature $F_s = E_s(X_s)$ can be obtained. Then, the three types of modal features are fed into the fusion module to form a unified representation:

$$H = \mathrm{Concat}(W_t F_t, W_i(F_i \odot A_i), W_s F_s) \tag{6}$$

where, W_t, W_i and W_s are learnable mapping matrices, $\odot$ represents element-wise weighting. In order to avoid information redundancy caused by simple stitching, this paper further uses linear projection and normalization operations to compress and align the joint features, and obtains the final packaging element representation Z:

$$Z = \mathrm{Norm}(W_h H + b_h) \tag{7}$$

This representation has both semantic interpretation and visual operability: text features ensure that the graphics theme does not deviate from the product and regional context, image features maintain the recognition basis of packaging graphics, and style features provide uniform aesthetic boundaries and layout tendencies for subsequent generation. Compared with the methods that only rely on a single text prompt or a single image reference, the proposed multi-modal representation method can more completely extract the key elements required for regional characteristic agricultural products packaging, and provide a more stable input basis for the next graphics generation and discriminant evaluation.

3.3 Packaging graphics generation method based on multi-modal AI generation technology

3.3.1 Design and Implementation of multimodal graphics generation module

The generator module takes the joint representation Z obtained in Section 3.2 as the core input, and introduces the layout condition vector C and the random disturbance vector r to form the generator input set. Considering that the packaging graphics need to preserve stable structural boundaries and have certain creative changes, this paper adopts a two-level structure of "cross-modal fusion layer + conditional decoding layer". The cross-modal fusion layer is responsible for mapping text semantics, image patterns and style constraints into a unified latent space, and the conditional decoding layer reconstructs the output wrapper graph Y_g by layer-by-layer upsampling and convolution. Its generation process can be expressed as follows.

$$Y_g = G(Z, C, r) \tag{8}$$

Here, G(·) denotes the multi-modal conditional generator. In order to enhance the retention of regional visual elements in the generated graphics, we introduce a cross-modal attention mechanism in the decoding stage, so that the text keywords and the response strength of the local image patterns on the feature map are dynamically aligned. Let the decoding feature of layer l be U_l, the text attention matrix and the image attention matrix be A_t^l and A_i^l respectively, then the enhanced feature can be written as follows.

$$\widetilde{U}_l = U_l + \alpha A_t^l U_l + \beta A_i^l U_l \tag{9}$$

where, α and β are the modal weight coefficients. Through this process, the model can more stably retain the key design elements in the output, such as terrace contour, mountain texture, fruit cut surface, traditional pattern edge and main color level, and reduce the deviation of the generated results from the regional context.

3.3.2 Design and implementation of multimodal graphics discrimination and evaluation module

If the generator only pursues image integrity, it is often easy to get the result of "smooth screen but not applicable packaging". Therefore, this paper constructs a composite module consisting of an authenticity discrimination unit and a design evaluation unit. The authenticity discrimination unit is responsible for distinguishing the distribution difference between the generated graphics and the real packaging samples. The design evaluation unit further examines the correspondence between the graphics and the input semantics, style constraints and layout requirements.

Let the true packing graph be Y_r, the generated graph be Y_g, and the discriminator be denoted D(·), then its output can be expressed as:

$$p = D(Y) \tag{10}$$

Here, $p \in [0,1]$ represents the probability that the input graph belongs to the true sample distribution. Different from general image tasks, the discriminator in this paper does not only read the image itself, but also jointly receives the text semantic vector and the style control vector to test the two dimensions of "whether the image is real" and "whether the image is right". In addition to the discrimination results, the evaluation module calculates S_{sem}

(semantic consistency), S_{sty} (style fit) and S_{lay} (format clarity), and weights them to obtain a comprehensive evaluation score:

$$S = \lambda_1 S_{sem} + \lambda_2 S_{sty} + \lambda_3 S_{lay} \tag{11}$$

Here, λ_1, λ_2, and λ_3 are the weight parameters. The significance of this design is that the model does not regard the packaging graphics as common visual output, but as the result of a task with both communication function and design constraints. The discriminator is responsible for constrinting the authenticity boundary, and the evaluation module is responsible for constrinting the usability boundary. The combination of the two modules can effectively suppress the problems such as graphic drift, information congestion and regional symbol weakening.

3.3.3 Multimodal joint training process

In the training phase, the parameters of the generator and discriminator are updated by alternating optimization. In each round of training, the system firstly extracts text description, regional reference map, style label and real packaging graphics from the sample library, and forms multi-modal conditional input after coding. Subsequently, the generator outputs candidate wrapper graphs, and the discriminator receives both real and generated graphs and compares them. The joint training objective can be written as follows:

$$\min_G \max_D \ \mathcal{L}_{adv}(G, D) = \mathbb{E}_{Y_r \sim p_{data}}[\log D\,(Y_r)] + \mathbb{E}_{Z \sim p_Z}[\log(\,1 - D(G(Z, C, r)))] \tag{12}$$

This equation describes the basic adversarial relationship between the generator and the discriminator. During training, the generator tries to improve the realism and fitness of the generated graphics, which makes it difficult for the discriminator to distinguish. The discriminator continues to improve its ability to recognize real packaging distribution and pseudo-generated graphics. Considering that the packaging task has strong semantic constraints, we synchronously introduce semantic consistency feedback and structure evaluation feedback in the adversarial training process, so that the gradient update not only depends on true or false discrimination, but is also driven by the design goal.

Figure 2: Process of packaging graphics generation and training based on multimodal AI generation technology

The specific process is shown in Figure 2: Joint features are formed after input samples enter the encoder. The generator outputs the wrapper graph according to the conditional constraints. The discrimination and evaluation module outputs the authenticity probability and the comprehensive design score respectively. After the loss is backpropagated, the system updates the generator and discriminator parameters and enters the next round of training. After multiple iterations, the generated results gradually transitioned from "capable of graphing" to "stable graphics, accurate semantics, and usable structure". The advantage of this joint training method is that it transforms the generation of packaging graphics from a single visual fitting to a multi-objective joint correction process, which is more suitable for the design tasks with both commodity attributes and cultural attributes such as regional characteristic agricultural products.

3.4 Model Training and optimization

After the construction of multi-modal packaging element coding, graphics generation module and discriminant evaluation module, the key of model training is no longer the improvement of single image quality, but how to achieve a stable balance between authenticity, semantic consistency and design usability. Regional characteristics of agricultural product packaging graphics have strong task constraints. If the model only pursues realistic images, it is easy to weaken regional semantics or the information area is crowded by decorative elements. If the text correspondence is overemphasized, the graphics may be stiff and the visual hierarchy is

insufficient. Therefore, in this paper, the generator and discriminator are alternately updated in the training stage, and a multimodal generative loss is introduced on the basis of adversarial learning to ensure that the output results are not only close to the real packaging distribution, but also able to respond to product attributes, regional culture and layout control requirements.

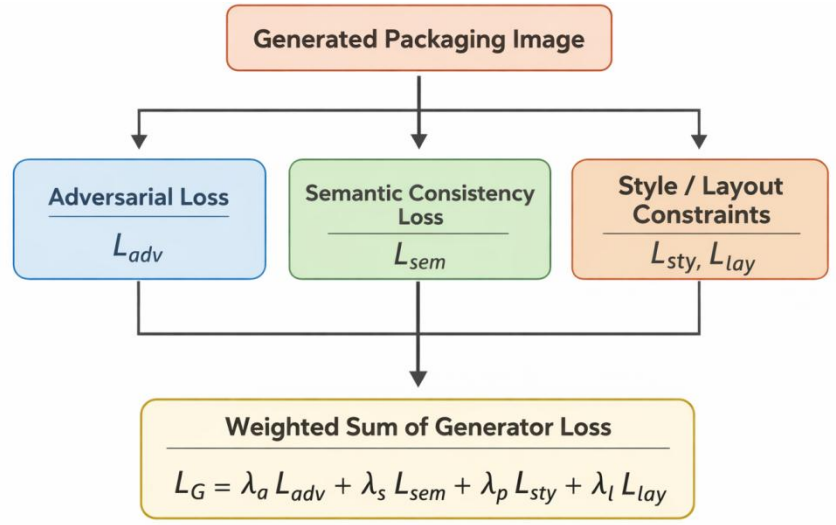

Figure 3: Schematic representation of the composition of the multi-modal generative loss function

As shown in Figure 3, the generation loss designed in this paper consists of four parts: the adversarial loss is used to constrain the generated graph to be close to the true packaging sample distribution; Semantic consistency loss is used to ensure a stable correspondence between graphic topics and textual descriptions. Style preserving loss is used to maintain the consistency of target color, pattern temperament and overall visual tone. The layout constraint loss is used to limit the spatial conflict between the main visual area, the information blank area and the auxiliary decoration area. During training, the Adam optimizer was used to update the parameters of the model, the learning rate was set to 2×10^{-4}, the batch size was set to 32, and the generator and discriminator were alternately iterated in a 1:1 ratio to avoid training imbalance caused by too fast one-sided convergence.

3.4.1 Multi-modal generative loss function design

In order to make the packaging graphics output by the generator meet the three requirements of "visible, relevant and usable" at the same time, the total loss of the generator is defined as the weighted sum of multiple losses:

$$\mathcal{L}_{\mathrm{G}} = \lambda_{\mathrm{a}}\mathcal{L}_{\mathrm{adv}} + \lambda_{\mathrm{s}}\mathcal{L}_{\mathrm{sem}} + \lambda_{\mathrm{p}}\mathcal{L}_{\mathrm{sty}} + \lambda_{\mathrm{l}}\mathcal{L}_{\mathrm{lay}} \tag{13}$$

Here $\mathcal{L}_{\mathrm{adv}}$ is the adversarial loss, $\mathcal{L}_{\mathrm{sem}}$ is the semantic consistency loss, $\mathcal{L}_{\mathrm{sty}}$ is the style preserving loss, $\mathcal{L}_{\mathrm{lay}}$ is the layout constraint loss, and $\lambda_{\mathrm{a}}, \lambda_{\mathrm{s}}, \lambda_{\mathrm{p}}$ and λ_{l} are the corresponding weight coefficients. Rather than simply stacking metrics, this design attempts to transform different design requirements in packaging graphics generation into computable objectives, so that the model can automatically reconcile multiple constraints during training. Here, the adversarial loss is written as follows.

$$\mathcal{L}_{\mathrm{adv}} = -\mathbb{E}_{\mathrm{Z}\sim \mathrm{p_Z}}[\log \mathrm{D}\,(\mathrm{G}(\mathrm{Z}, \mathrm{C}, \mathrm{r}))] \tag{14}$$

This term is used to promote the generation result to approximate the real packaging sample distribution, and avoid the broken edge, loose texture or unbalanced color patch relationship of the graphics. The semantic consistency loss is achieved by comparing the distance between the generated graphical features and the semantic features of the text:

$$\mathcal{L}_{\mathrm{sem}} = 1 - \cos(\mathrm{E_t}(\mathrm{T}), \mathrm{E_v}(\mathrm{Y_g})) \tag{15}$$

where,$\mathrm{E_t}(\mathrm{T})$ represents the text encoding feature and $\mathrm{E_v}(\mathrm{Y_g})$ represents the visual semantic feature of the generated graph. This can reduce the semantic offset phenomenon such as "the text says mountain black tea, but the graphics are more like fruit and vegetable gift boxes". The style-preserving loss is used to maintain the comprehensive color and decorative features of the target package, while the typography constraint loss focuses on penalizing the problems of image-text overlap, visual center drift and information area congestion. Therefore, the generator does not obtain a single true or false feedback during training, but a set of optimization signals directly corresponding to the packaging design task, which also provides a clear basis for the establishment of subsequent comprehensive optimization objectives.

3.4.2 Comprehensive optimization objective design

In the generation task of regional agricultural product packaging graphics, a single loss often can only constrain a certain side of the model, and it is difficult to take into account authenticity, regional semantics, style unity and layout usability at the same time. Based on this, this paper further constructs a comprehensive optimization objective on the basis of multimodal generative loss, so that the model training shifts from "generating visible images" to "generating usable packaging solutions". On the one hand, adversarial learning is used to approximate the distribution of real packaging samples, and on the other hand, feature matching and semantic alignment mechanisms are used to ensure that the cultural expression and visual organization of the generated graphics do not deviate from the design task itself.

In order to reduce the drift of the generated results in local patterns and overall style, we introduce feature matching loss. Let the features extracted by the KTH layer of the discriminator be $\mathrm{F_k^r}$ for the real packaging figure and $\mathrm{F_k^g}$ for the generated packaging figure, then the feature matching loss can be expressed as follows.

$$\mathcal{L}_{\mathrm{fm}} = \sum_{\mathrm{k=1}}^{\mathrm{K}} \left\| \mathrm{F_k^r} - \mathrm{F_k^g} \right\|_1 \tag{16}$$

The term does not directly compare pixels, but rather constrains the distribution distance between true and generated samples in the multi-layer feature space. The significance of this process is that the model optimization can pay attention to the main visual contour, pattern density, color patch relationship and local decoration details at the same time, avoiding the problem of realistic surface and loose internal structure caused by only adversarial feedback. This is particularly important for the packaging of regional characteristic agricultural products, because local patterns, morphological characteristics and visual order of agricultural products are often reflected in multi-scale features.

On this basis, the adversarial objective, semantic consistency objective, style preservation objective, layout constraint objective and feature matching objective are integrated into the total loss function to form a comprehensive optimization expression:

$$\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{a}}\mathcal{L}_{\mathrm{adv}} + \lambda_{\mathrm{s}}\mathcal{L}_{\mathrm{sem}} + \lambda_{\mathrm{p}}\mathcal{L}_{\mathrm{sty}} + \lambda_{\mathrm{l}}\mathcal{L}_{\mathrm{lay}} + \lambda_{\mathrm{f}}\mathcal{L}_{\mathrm{fm}} \tag{17}$$

Here, $\lambda_a, \lambda_s, \lambda_p, \lambda_l$ and λ_f represent the weight coefficients of each loss. By adjusting these parameters, the balance between "realistic generation" and "design adaptation" can be controlled. If the proportion of adversarial loss is too high, the model is more likely to pursue visual realism and ignore the packaging purpose. If the proportion of semantic and layout constraints is too large, the generated results may become rigid and lack necessary visual tension. Therefore, the value of the comprehensive optimization objective does not lie in stacking more constraints, but in making all kinds of constraints jointly serve the core requirements of the packaging design task.

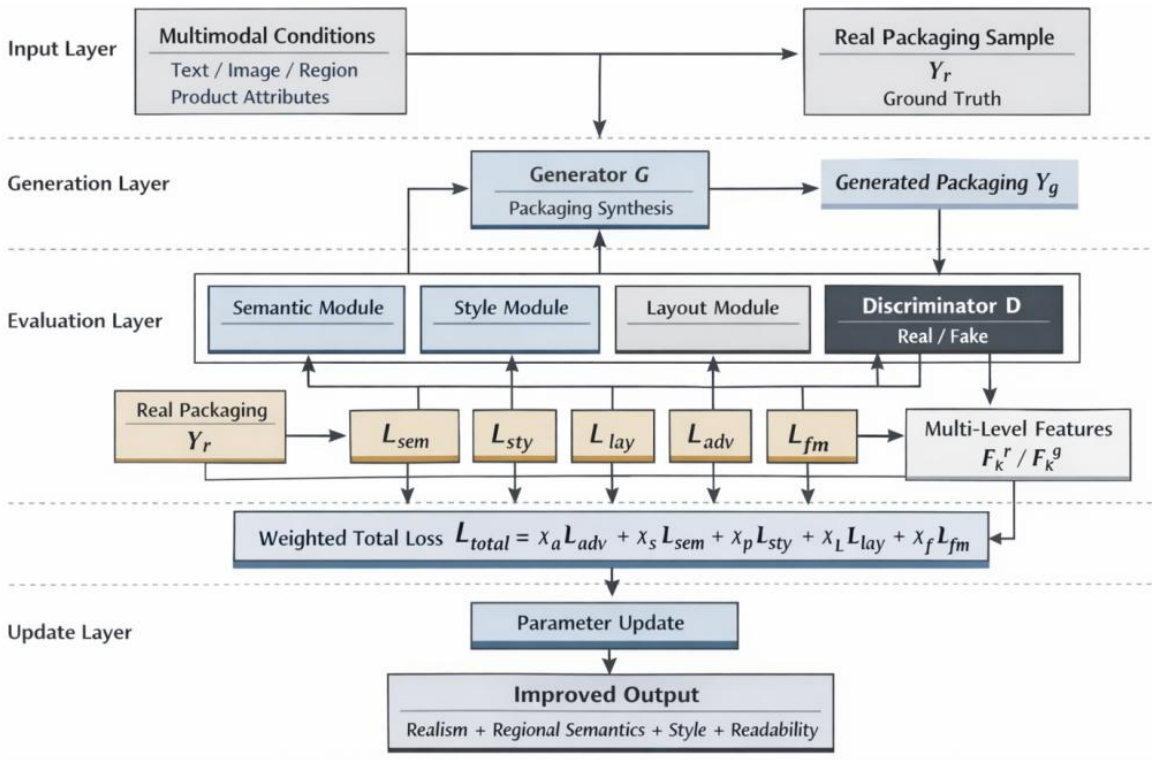

Figure 4: Schematic diagram of the integrated optimization objective design

As shown in Figure 4, the integrated optimization objective unifies the generator output, the discriminator feedback, and the multimodal evaluation results into the same update loop. In the training process, the system iteratively updated the parameters according to the change of the total loss, so that the generated graphics gradually approached the real packaging samples, while maintaining stability in regional cultural expression, product attribute correspondence and readability of information area. Therefore, the model optimization no longer stops at the "true or false judgment" in the sense of general image generation, but enters the multi-objective collaboration stage for packaging design landing, which also provides a quantifiable training basis for subsequent experimental evaluation.

4 Experimental Evaluation

4.1 Experimental Design

In order to verify the effectiveness of the multimodal AI generation method proposed in this paper in the graphic design of regional characteristic agricultural products packaging, experiments are carried out around the generation quality, semantic fit and design usability.

The dataset is composed of physical images of regional agricultural products, local pattern images, existing packaging samples and corresponding text descriptions, covering five categories of products such as tea, honey, mushroom, dried fruits and miscellaneous grains, and a total of 4260 groups of multimodal samples are organized. The text content includes product name, origin description, cultural keywords, selling point labels and style tips, and the image part includes product appearance, regional visual elements and packaging reference images. After unified cleaning, labeling and size normalization, the image part is divided into training set, validation set and test set, and the ratio is set to 7:2:1. The experimental group uses the multi-modal conditional generation model constructed in this paper, and the control group is set as the template stitching method, the single-modal text generation image model and the basic generation model without style constraints, to compare the performance differences of different technical paths in packaging graphics generation. The evaluation indexes include structural similarity index, peak signal-to-noise ratio, text-image semantic consistency and comprehensive color harmony score to measure the integrity, visual quality and task suitability of the graphics. The experimental environment is configured with NVIDIA RTX 3090 GPU, Python 3.10 and PyTorch 2.1, Adam is used as the optimizer, the initial learning rate is set to 2×10^{-4}, the batch size is 32, and the number of training rounds is 150. By unifying the training conditions and testing process, the results of different models are comparable.

4.2 Experimental Results

The experimental results show that the multi-modal AI generation method proposed in this paper shows stable comprehensive advantages in the task of regional characteristic agricultural product packaging graphics. The proposed model is superior to the control group in terms of structure preservation, image clarity, semantic correspondence and color harmony. The reason is not only that the generator itself is more complex, but more crucially, the model introduces text semantics, region images, style constraints and layout control during the training process, so that the output can be consistent at the "like package", "like the product" and "like the region" levels.

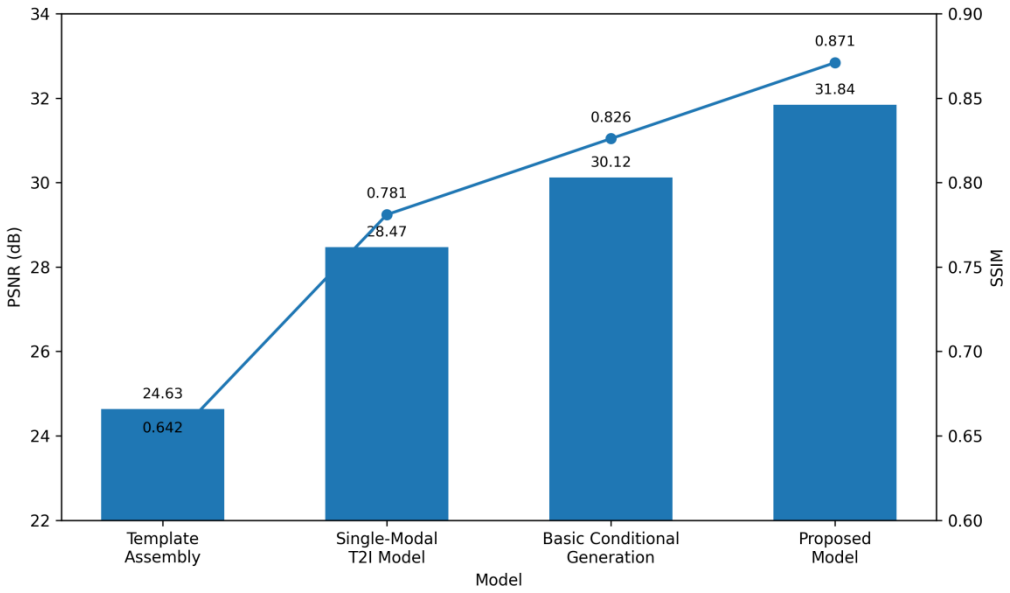

Figure 5: Comparison of SSIM and PSNR for different models

As shown in Figure 5, the proposed model achieves the highest values in SSIM and PSNR, reaching 0.871 and 31.84 dB, respectively, which are significantly higher than 0.642 and 24.63 dB of the template stitching method, and 0.781 and 28.47 dB of the single-modal text-generated image model. Although the basic conditional generative model has been able to output relatively complete packaging graphics, due to the lack of style constraints and multi-modal alignment, its local pattern edges and main visual areas still have some drift, so it still lags behind the proposed model in terms of structural similarity and image quality. The results show that multimodal joint modeling not only improves the overall stability of image generation, but also enhances the learning ability of packaging graphics to the structure law of real design samples.

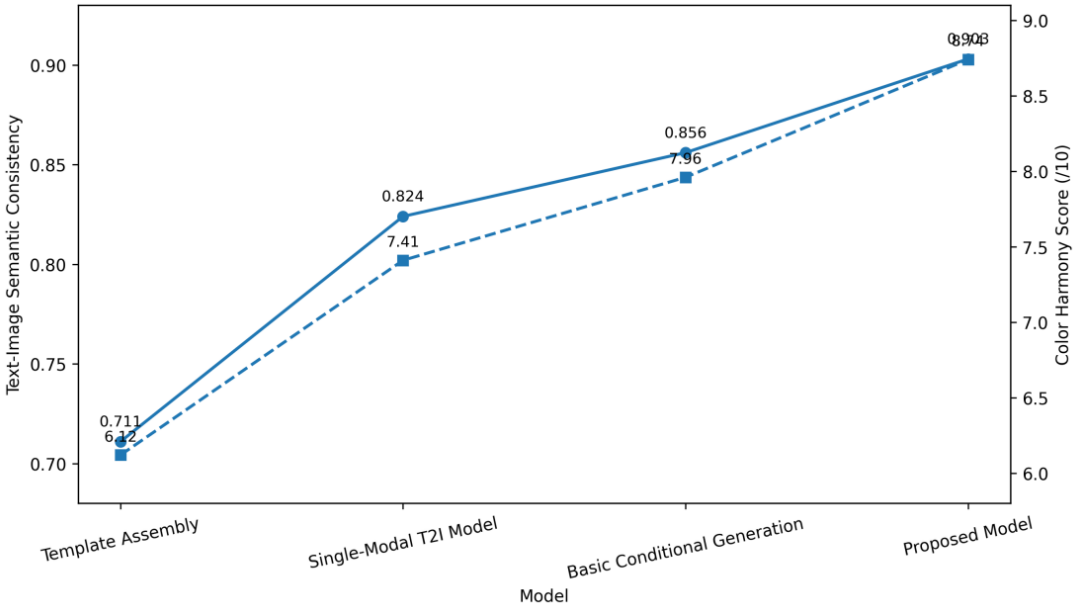

Figure 6: Comparison of text-image semantic consistency and comprehensive color harmony score of different models

In terms of semantic fit, the text-image semantic consistency of the proposed model reaches 0.903, which is higher than 0.711 of the template stitching method, 0.824 of the single-modal text-generated image model, and 0.856 of the basic conditional generation model. Figure 6 further shows that the proposed model also keeps the lead in the comprehensive color harmony score, with a score of 8.74, while the other three methods score 6.12, 7.41, and 7.96, respectively. Combined with the observation of the generated samples, it can be found that although the template splicing method is more direct in calling local elements, it is prone to the problem of disconnection between regional elements and product attributes. The single-modal text-generated image model can respond to text prompts, but often treats "regionality" as a broad decorative style, which is difficult to accurately distinguish the differences in visual temperament between mountain tea, honey or mushroom packaging. The basic conditional generation model has some control ability, but it is still not stable enough in comprehensive color unification and auxiliary pattern restraint. In contrast, the proposed model synchronizes product semantics, local patterns and style boundaries in the cross-modal fusion stage, so the generated graphics are closer to the visual logic in real packaging design.

Table 2: Experimental results of the proposed model in different regional characteristic agricultural product categories

Product Category	SSIM	PSNR (dB)	Semantic Consistency	Color Harmony Score (/10)
Tea	0.884	32.41	0.918	8.93
Honey	0.879	32.06	0.911	8.87
Mushrooms	0.862	31.35	0.896	8.56
Dried Fruits	0.868	31.67	0.901	8.71
Coarse Grains	0.857	31.12	0.889	8.44

To further test the adaptability of the model in different product categories, Table 2 shows the breakdown results on five types of regional characteristic agricultural products. It can be seen that the proposed model maintains a high level in the five categories of tea, honey, mushroom, dried fruit and miscellaneous grains, especially in the packaging of tea and honey. This is related to the fact that the two types of products themselves have stronger cultural narrative and color style discrimination, and the multi-modal model is easier to extract stable features from text descriptions and reference images. Relatively, the visual form of the packaging of multigrain and mushroom is closer to that of daily food packaging, and the style boundary is less distinct than that of tea and honey. Therefore, the gap between different models is slightly narrowed, but the proposed method still maintains the highest score. This shows that the model is not only effective for a few high identification categories, but has good generalization ability in different types of regional agricultural products.

On the whole, the improvement of the method in this paper is not a local uplift on a single index, but a synchronous improvement of multiple results. The improvement of SSIM and PSNR indicates that the generated graphics are more stable in structure and clarity. The improvement of semantic consistency indicates that the deviation between graphic content and text intention is effectively compressed. The rise of the comprehensive color coordination score reflects that the model in packaging design no longer stops at "can generate", but begins to approach the level of "can design". Because of this, the proposed method shows stronger application potential in the graphics task of regional agricultural product packaging.

4.3 Discussion

The experimental results show that the multi-modal AI generation method proposed in this paper shows stable comprehensive advantages in the graphic design of regional characteristic agricultural products packaging. Compared with the template stitching method, the single-modal text-generated image model and the basic conditional generation model, the proposed model achieves better results in SSIM, PSNR, semantic consistency and comprehensive color harmony scores. This shows that the model improvement is not only reflected in the surface quality of the image, but also reflected in the collaborative improvement of structure maintenance, cultural fit and design usability. The key to this result is that the model integrates text semantics, regional images, style constraints and layout conditions into a unified computing framework. Through cross-modal alignment and joint optimization, the generation process of packaging graphics is changed from "visual output" to "design solution". From the perspective of method mechanism, the multimodal fusion module enhances the model's ability to identify the relationship between regional patterns, product appearance and cultural keywords, and the conditional generation module improves the coordination degree between the main visual organization, decorative density and information area whiteness. For categories with strong cultural narratives such as tea and honey, this advantage is more obvious, indicating that multimodal modeling has high suitability for

extracting visual cues with local characteristics. In other words, the improvement in model performance does not come from simply scaling up the size of the parameters, but from a closer correspondence between design goals and computational mechanisms.

The method in this paper still has some limitations. On the one hand, the style distribution of regional agricultural product packaging samples is still not balanced, and the visual features of some low-resource categories are weak, which may affect the generalization ability of the model to subdivision scenes. On the other hand, although multimodal conditional constraints and joint training improve the generation stability, they also increase the training cost and inference complexity, and there is still room for optimization in lightweight deployment. At the same time, current evaluation systems pay more attention to image quality and semantic fit, and lack of continuous verification of recognition efficiency, brand memory and purchase conversion in real consumption situations. Subsequent research can further expand the type and granularity of regional packaging data, introduce more detailed user perception indicators and market feedback data, and combine lightweight generation network and interactive design mechanism to promote this method from experimental verification to practical design application.

5 Conclusion

Focusing on the problems of unstable regional expression, low generation efficiency and easy drift of visual style in the graphic design of regional agricultural products packaging, this paper constructs a design method based on multi-modal AI generation technology. In this study, the product text, regional image, style constraints and layout conditions are integrated into the unified computing framework. Through multi-modal feature extraction, condition generation, discriminant evaluation and comprehensive optimization, the packaging graphics are transformed from experience splicing to intelligent generation. Experimental results show that the proposed model is superior to the control methods in many indicators, SSIM reaches 0.871, PSNR reaches 31.84 dB, text-image semantic consistency reaches 0.903, and the comprehensive color harmony score reaches 8.74. In the tea packaging task, SSIM is further improved to 0.884 and PSNR reaches 32.41 dB, which shows that the method has good adaptability to categories with strong regional cultural characteristics. From the perspective of research significance, the generation of packaging graphics is not understood as general image synthesis, but as a composite design task including product identification, cultural translation and visual organization. The established method not only improves the structural stability and design usability of the generated results, but also provides a more specific technical path for the combination of packaging visual design and computer generation model. Of course, there are still some problems in the current model, such as high training cost and the generalization ability of low-resource categories needs to be improved. In the future, the regional packaging data types can be expanded, user perception and market feedback indicators can be introduced, and lightweight deployment and human-machine collaborative optimization can be promoted.

References

[1] Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with clip latents[J]. arXiv preprint arXiv:2204.06125, 2022, 1(2): 3.

[2] Saharia C, Chan W, Saxena S, et al. Photorealistic text-to-image diffusion models with deep language understanding[J]. Advances in neural information processing systems,

2022, 35: 36479-36494.

[3] Ruiz N, Li Y, Jampani V, et al. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 22500-22510.

[4] Brooks T, Holynski A, Efros A A. Instructpix2pix: Learning to follow image editing instructions[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 18392-18402.

[5] Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 3836-3847.

[6] Ye H, Zhang J, Liu S, et al. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models[J]. arXiv preprint arXiv:2308.06721, 2023.

[7] Hu H, Chan K C K, Su Y C, et al. Instruct-imagen: Image generation with multi-modal instruction[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 4754-4763.

[8] Huang Y, Xie L, Wang X, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 8362-8371.

[9] Sun Q, Cui Y, Zhang X, et al. Generative multimodal models are in-context learners[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 14398-14409.

[10] Wang Z, Li A, Li Z, et al. Genartist: Multimodal llm as an agent for unified image generation and editing[J]. Advances in Neural Information Processing Systems, 2024, 37: 128374-128395.

[11] Peng X, Koch J, Mackay W E. Designprompt: Using multimodal interaction for design exploration with generative ai[C]//Proceedings of the 2024 ACM Designing Interactive Systems Conference. 2024: 804-818.

[12] Ye Y, Hao J, Hou Y, et al. Generative AI for visualization: State of the art and future directions[J]. Visual Informatics, 2024, 8(2): 43-66.

[13] Li R, Li H. The impact of food packaging design on users' perception of green awareness[J]. Sustainability, 2024, 16(18): 8205.

[14] Mehta A, Serventi L, Kumar L, et al. Packaging, perception, and acceptability: a comprehensive exploration of extrinsic attributes and consumer behaviours in novel food product systems[J]. International Journal of Food Science and Technology, 2024, 59(10): 6725-6745.

[15] Mehta A, Serventi L, Kumar L, et al. Exploring the effects of packaging on consumer experience and purchase behaviour: insights from eye tracking and facial expressions on orange juice[J]. International Journal of Food Science and Technology, 2024, 59(11):

8445-8460.

[16] Berthold A, Guion S, Siegrist M. The influence of material and color of food packaging on consumers' perception and consumption willingness[J]. Food and Humanity, 2024, 2: 100265.

[17] Guo X, Huang J, Wan X. Influence of exposure to novel food packaging on consumers' adoption of innovative products[J]. Food Quality and Preference, 2024, 119: 105230.

[18] Turan D, Keukens B M, Schifferstein H N J. Food packaging technology considerations for designers: Attending to food, consumer, manufacturer, and environmental issues[J]. Comprehensive Reviews in Food Science and Food Safety, 2024, 23(6): e70058.

[19] Herbes C, Mielinger E, Krauter V, et al. Company views of consumers regarding sustainable packaging[J]. Sustainable production and consumption, 2024, 52: 136-150.

[20] Saintives C, Meral H. Is it really natural? How minimalist food packaging influences consumers' perception of product naturalness[J]. British Food Journal, 2024, 126(11): 3888-3905.

[21] Li X, Wang S, Ruan Y, et al. Taste or health: The impact of packaging cues on consumer decision-making in healthy foods[J]. Appetite, 2024, 203: 107636.

[22] Baek E, Huang Z, Lee S S. Visual complexity= hedonic? Effects of visually complex packages on consumer perceptions and evaluations of products[J]. Journal of Retailing and Consumer Services, 2023, 74: 103435.

[23] Hidayanto A F, Hamat B, Ariff N S B N A. A systematic literature review on the development of traditional regional food packaging as regional identity[C]//3rd Borobudur International Symposium on Humanities and Social Science 2021 (BIS-HSS 2021). Atlantis Press, 2022: 793-799.

[24] Hidayanto A F, Hamat B, Ariff N S B N A. Study of Visual Elements of Traditional Food Packaging Designs Typical of Heritage as Regional Identity[C]//International Conference on Applied Science and Technology on Social Science 2022 (iCAST-SS 2022). Atlantis Press, 2022: 348-353.

[25] Hidayanto A F, Hamat B, Ahmad N S B N, et al. The Conceptual Framework for Developing Packaging Designs of Traditional Heritage Foods from the Visual Aspect[J]. Pakistan Journal of Life & Social Sciences, 2024, 22(2).