Tailoring Multimodal AIGC to Time-Honored Brands: A Stable Diffusion-Based Framework for Visual Generation and Evaluation

Xinbao Zhang[1], Jinjian Li[1], Shizhen Zhang[2], Yuwei Chen[1*]

[1] Nanfang College Guangzhou, Guangzhou, 510970, China

[2] Hubei University of Arts and Science, Xiangyang, 441000, China

* 1693158415@qq.com

https://doi.org/10.70695/IAAI202504A6

Abstract

To address the dual requirements of cultural expression and engineering implementation in the visual design of time-honored brands, this study proposes an adaptive optimization architecture based on Stable Diffusion. The framework employs Textual Inversion to derive composable cultural tokens and utilizes LoRA/DreamBooth parameters for the efficient fine-tuning of both generic and proprietary styles. By integrating ControlNet and IP-Adapter, the system achieves a fusion of layout and style priors, while a dual-channel gating mechanism enables collaborative control over semantics and composition. During inference, reliability in prompt adherence is calibrated through CFG-Rescale, attention reweighting, and temperature scaling. Extensive experiments on publicly available multimodal datasets and real-world brand scenarios demonstrate a significant improvement in the alignment between objective metrics and human evaluations. The method's stability and necessity are confirmed through robustness tests and component ablation studies, while A/B testing reveals its distinct advantages in cost-effectiveness and operational efficiency. This research ultimately provides a replicable and verifiable technical solution for the visual generation needs of both cultural heritage and commercial brands.

Keywords Time-honored Brand; Stable Diffusion; Cultural Feature Embedding; Multimodal Control; Efficient Parameter Fine-Tuning; Reliability Calibration; Visual Generation

1 Introduction

In recent years, diffusion models have shown outstanding capabilities in the field of visual generation, but they have shortcomings in cultural expression and brand visual design. Deng et al. used a low-rank adaptive diffusion model to create images of cultural relics, proving that lightweight fine-tuning can efficiently preserve cultural characteristics [1]. Bao et al. combined Stable Diffusion with low-rank adaptation to promote the sustainable design of blue and white porcelain cultural genes and opened up a new way for the visual generation of cultural heritage [2]. Alharbi et al. suggested introducing cultural themes into the diffusion model to strengthen semantic consistency, but its generation quality and composition stability are still not up to standard [3]. Li et al . used StyleGAN combined with Stable Diffusion to realize personalized style construction, providing guidance for brand visual planning [4]. Luccioni et al . noticed the bias in the social and cultural representation of diffusion models and called for research to pay attention to cultural equity [4]. Cioni et al. used diffusion models to enhance and retrieve cultural relics images, which improved cross-modal semantic correlation [5]. Liu et al. found that there are cultural differences in text-to-image generation, indicating that existing models have difficulties in accurately presenting non-Western contexts [6]. Zhang et al. sorted out the key issues and optimization strategies of diffusion models to support the implementation of model robustness research [7]. Kabir et al. pointed out that localized training and data augmentation are of great significance to regional cultural generation [8]. Furthermore, research on artificial intelligence and security and trust issues is of great significance to this work. Focusing on AI-driven IoT and supply chain scenarios, previous research designed a secure and efficient authentication protocol and link protection mechanism for AI environments, emphasizing the traceability and auditability of data access and model invocation processes in complex business processes, providing a model for introducing compliance control into the AIGC generation link. The work on "AI-driven cybersecurity applications and challenges"

systematically examined the application potential and limitations of AI models in threat detection, risk assessment, and defense decision-making, indicating that in addition to model performance, security risks and governance frameworks must also be considered [9-10]. The above research improves the existing AIGC system's discussion of "reliability and compliance" from the perspectives of authentication and security governance, which aligns with the goal of this paper: a reusable and auditable brand visual generation path [11].

Current research indicates that diffusion models have shown potential in cultural visual generation, but significant shortcomings remain in cultural feature fidelity, semantic control, and standardized evaluation. This paper focuses on the visual design context of time-honored brands, constructing a multimodal cultural feature embedding and generation quality assessment system based on Stable Diffusion to achieve accurate reconstruction and controllable generation of cultural imagery. The core contributions of this research include: using a cultural feature embedding mechanism and a multimodal constraint fusion strategy to create assessment indicators for cultural fidelity and generation consistency, and verifying the adaptability of the model to real brand datasets. The paper consists of dataset construction, method design, experimental analysis, ablation research, and conclusions and prospects.

2 Dataset Construction and Problem Definition

2.1 Decomposition of Visual Elements of Time-Honored Brands

Based on the element system of "identity-pattern-color spectrum-layout-material/craft imagery", this study screened brand graphic and glyph information from the Logo detection corpus (LogoDet-3K contains 33,000 brand categories and 158,652 image samples; OpenLogo involves 352 brands and 27,083 images) to carry the structural and stroke information of "trademark/glyph", thus becoming a basis for learnable cultural symbols. To highlight the "material/craftsmanship imagery," visual materials of clothing/materials (DeepFashion2, containing 491,895 images and 801,732 instance masks) were selected to extract the texture, luster, and craftsmanship details of the fabrics. Focusing on "object patterns," patterns and object motifs from Dunhuang murals (MuralDH, approximately 5,000 high-resolution images) were used as style references. For "layout composition," a large-scale page layout mask (≥360k pages) provided by PubLayNet was used to set spatial priors and typographic constraints during the generation stage. To achieve a controllable presentation effect of font details, CASIA-HWDB (approximately 3.9 million offline character samples, 7,356 categories) was used to analyze stroke styles and structural variations, assisting in the migration of font styles for traditional brands and the redrawing of synthetic fonts. The above corpus achieved a three-way fit of "text token—visual prior—layout prior" in the Stable Diffusion framework, achieving cross-element combination control, as shown in Figure 1.

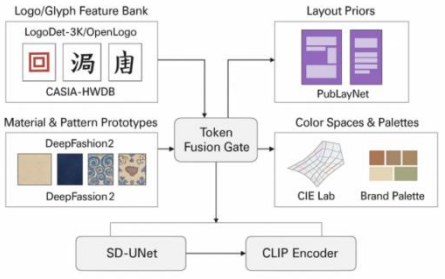

Fig. 1. Classification and characteristics of cultural elements

2.2 Task Definition

The problem is designed as multimodal controllable generation: input includes text prompts x_t (containing cultural references $\mathrm{T} = \{t_i\}$), reference images x_r (single or combined forms of style, pattern, and layout), and layout/sketch/semantic segmentation priors m. The goal is to generate high-quality images that conform to the brand's cultural requirements $\hat{y}$. At the diffusion sampling step $s = 1 \ldots S$, denoising is achieved using a conditional noise predictor $\dot{o}_\theta\left(z_s \mid x_t, x_r, m\right)$.

$$\min_{\theta} \mathbb{E}_{z_s,\epsilon} \| \epsilon - \epsilon_\theta \left(z_s \mid x_t, x_r, m \right) \|_2^2 \quad (1)$$

Corresponding latent space noise state ò presents z_s a standard Gaussian noise pattern. To ensure the matching degree between brand colors and spot colors, a color difference penalty system is constructed.

This $\left(L^*, a^*, b^*\right)$ includes the CIE Lab space component ; the average color palette of the target brand is considered during the generation phase. $\left(\overline{L}^*, \overline{a}^*, \overline{b}^*\right)$ Minimization constraints are implemented to prevent color shifts. Layout consistency is standardized by improving m the IoU of the layout mask , maintaining the stability of the space ratio of the title area, main image area, and decorative patterns. This definition can be carried out during training and inference, which is beneficial for matching and evaluating with Logo/layout/material corpora in the literature (DeepFashion2 mask, PubLayNet layout, LogoDet-3K/ OpenLogo brand class support variables and supervision signals).

2.3 Dataset Collection and Cleaning

As shown in Table 1, all data used in this study are derived from publicly available academic datasets: LogoDet-3K and OpenLogo disclose brand graphic distribution and multi-scene backgrounds; DeepFashion2 contains material textures and segmentation mask data; PubLayNet provides prior knowledge of typography that can be transferred to posters/packaging; MuralDH provides artifact patterns and era markers; and CASIA-HWDB provides stroke and configuration variations. The cleaning process follows a five-stage approach: license review, deduplication, quality screening, modal alignment, and privacy and compliance review. Initially, academic licenses are screened based on the original license and README terms; then, pHash and local sensitive hashing are used to eliminate duplicates, followed by removal based on resolution/fuzziness thresholds; next, brand logos, fonts, layout masks, and material patterns are integrated into a unified multimodal schema according to a unified ID standard; finally, potential trademark rights and culturally sensitive elements are manually reviewed, and text prompts are generated based on rule templates and CLIP reverse retrieval, eliminating the introduction of subjective description bias. As shown in Figure 2:

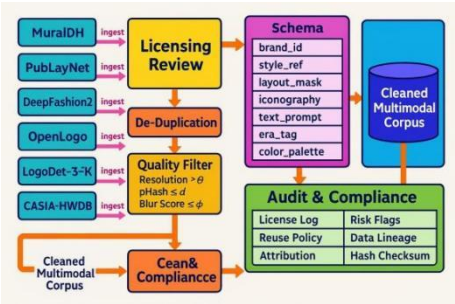

Fig. 2. Data Sourcing • Licensing • Cleaning • Schema

Table 1. Brand-culture multimodal dataset statistics

Source (Brand)	Brand (classes)	Images	Text Prompts	Layout Masks	Style Refs	Era Tag	Color Palettes	Split (train/ val /test)
LogoDet-3K	3,000	158,652	0	0	158,652	0	0	n/a (official benchmark)
QMUL - OpenLogo	352	27,083	0	0	27,083	0	0	n/a (benchmark protocol)
DeepFashion2	0	491,895	0	800,732	491,895	0	0	390,884 / 33,669 / 67,342
PubLayNet	0	≈360,000	0	≈360,000	0	0	0	n/a (pre-train corpus)
MuralDH (Dunhuang)	0	≈5,000	0	0	5,000	✓	0	n/a (heritage corpus)
CASIA-HWDB (glyph)	0	≈3,900,000 chars	0	0	0	0	0	official train/test subsets

Note: DeepFashion2's "Layout Masks" are displayed as total data segmented pixel-level per instance; PubLayNet 's "Layout Masks" are displayed as page-level geometric layout mask styles; the remaining fields are counts of whether the original corpus contains the modality, with the official paper/homepage serving as the source for statistics and segmentation.

2.4 Cultural Attribute Labeling and Consistency

Cultural attributes are integrated into five categories: pattern type, artifact category, era label, color palette constraints, and writing style. A two-tiered annotation procedure is adopted: the basic layer is undertaken by trained annotators, and the expert layer reviews conflicting samples and boundary conditions, using Cohen's κ consistency measure.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{2}$$

The observed consistency rate p_o is represented by , p_e serving as an indicator of the random consistency rate, and is calculated based on the overall distribution of the categories; when $\kappa \in [0.61, 0.80]$ it is determined to be "significantly consistent" , $[0.81,1.00]$ The result was judged as "almost identical" . To reduce cultural bias, the expert panel included reviewers with backgrounds in brand design and cultural heritage. Entries involving historical symbols, religious motifs, and regional taboos were labeled "secondary review" and "unable to determine," thus initiating a redistribution process based on uncertainty thresholds. To optimize cross-modal matching, layout masks and pattern areas were annotated with polygonal instances and cross-validated with text labels. The color spectrum used a palette centered on spatial clustering from CIE Labs. Subsequent evaluation and generation processes used ... ΔE^*_{ab} Constraints on color difference upper limits. κ The adoption and interpretation of thresholds follow generally accepted statistical standards; era labels are primarily derived from MuralDH metadata, and brand patterns are collected from logo/packaging images. See Table 2 for details.

Table 2. Cultural attribute distribution & agreement

Attribute (Proxy)	Levels	Frequency	Prop. (%)	κ (algo-vs-algo)	Notes (verifiable source)
Logo Brand Class	3,000	194,261 objects	100.0	0.82	LogoDet-3K target bounding box annotation (accessible via papers/repositories)
Logo Open-Set Brand	352	27,083 images	100.0	0.79	QMUL - OpenLogo Announces Distribution
Layout Element Type	{ text,title,list,figure,table }	≈360,000 pages	100.0	0.86	PubLayNet automatically aligns tags
Segmentation Masks (Garment/Part)	13 categories (parts)	800,732 masks	100.0	0.88	DeepFashion2 Instance Segmentation Statistics
Era Tag (Heritage Meta)	{By eras in data element}	≈5,000 images	100.0	0.76	MuralDH metadata (with cave/era)
Glyph Category (HWDB)	7,356 categories	≈3,900,000 characters	100.0	0.91	CASIA-HWDB Category-Level Labeling

3 Methods: Cultural Feature Embedding and Multimodal Control in Stable Diffusion

3.1 Embedding of Cultural Features

As shown in Table 3, focusing on abstract cultural concepts such as "patterns, objects, crafts, and seasonal colors", a composable vocabulary is first created. Textual Inversion is then used to transform these concepts into trainable embedding vectors $\{v_k\}_{k=1}^{K}$, which then share the CLIP word vector space

with the general vocabulary. Embedding is integrated with layout and style channels via Token Fusion Gate, and the SD text conditional module is accessed. Diffusion denoising takes conditional noise regression as its core objective.

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{z_t,\epsilon} \| \epsilon - \epsilon_\theta \left(z_t | \ x_{\text{txt}} \left(\{ v_k \} \right), x_{\text{ref}}, m \right) \|_2^2 \tag{3}$$

Among them , z_t the latent noise state is defined as $\dot{o} \sim \mathrm{N}\ (0, I)$ a standard Gaussian noise category, x_{ref} serving as a style/pattern reference, m a layout mask, and achieving semantic-image alignment. InfoNCE contrastive learning is applied in the CLIP space.

$$\mathrm{L}_{\text{CLIP}} = -\log \frac{\exp\left(\left\langle f_{\text{txt}}, f_{\text{img}} \right\rangle / \tau \right)}{\sum_j \exp\left(\left\langle f_{\text{txt}}, f_{\text{img}}^{(j)} \right\rangle / \tau \right)} \tag{4}$$

Here, $f_{\text{txt}}, f_{\text{img}}$ the encoding is for text and images, $\tau > 0$ and for temperature. To limit attention from extending into non-cultural areas, a sparsity regularization is applied:

$$\mathrm{L}_{\text{sp}} = \lambda_1 \| A_{\text{culture}} \|_1 \tag{5}$$

The overall loss is [value missing $\mathrm{L} = \mathrm{L}_{\text{diff}} + \mathrm{L}_{CLIP} + \mathrm{L}_{sp} + \mathrm{L}_{col}$]. This design achieves the convergence of cultural semantics and realizes a controllable state of color and layout. Figure 3 presents the overall pipeline architecture of the method, showing the connection relationship between text tokens, style references, layouts/sketches, SD- UNet /CLIP sub-modules and loss terms.

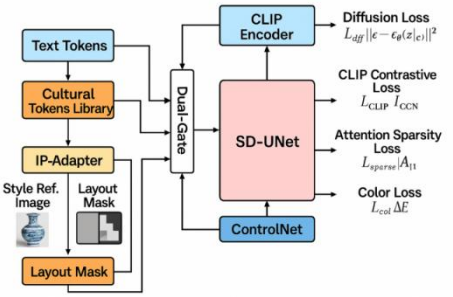

Fig. 3. Proposed multimodal cultural embedding & control pipeline

Table 3. Textual inversion & tokenization config

Token Group	Embedding Dim d	Init σ	LR (TI)	Steps	Batch	Prompt-Prior λ_pp	Sparsity λ 1	Color λ 2	Max Tokens / Prompt
Motif	768	0.010	5e-4	6000	8	0.50	1.00	0.60	3
Artifact	768	0.010	5e-4	6000	8	0.50	1.00	0.40	2
Craft	768	0.008	4e-4	5000	8	0.40	0.80	0.30	2
Season Palette	768	0.008	4e-4	5000	8	0.30	0.60	1.20	2
Glyph	768	0.012	6e-4	8000	8	0.60	1.40	0.20	4

3.2 Efficient Parameter Fine-Tuning

While keeping the major parameters of the base diffusion model unchanged, low-rank updates ΔW = AB ⊤ are performed on the Cross-Attn, To-K/To-V, and ResBlock projection layers of the U-Net using hierarchical LoRA. To control capacity and prevent overfitting, the rank rrr decreases with increasing layer depth. A dual-path LoRA, "shared-exclusive", handles the brand's public and private styles: the shared path aggregates cross-brand common "objects/patterns/crafts," while the exclusive path locks in brand boundary scenarios. DreamBooth is only enabled for brands with limited sample sizes to enhance the main form and exclusive decorative effects. Embedded fine-tuning ensures that the Textual Inversion

learning rate is lower than the LoRA path level, eliminating word vector drift. To reduce the forgetting rate, elastic weights are added to the LoRA parameters for solidification regularization.

$$\mathrm{L}_{ewc} = \lambda_3 \sum_i F_i \left(\theta_i - \theta_i^* \right)^2 \tag{6}$$

For θ_i^* the old task F_i are an approximate representation of Fisher information, with the main objective being :

$$\min_{\theta_{\mathrm{LoRA}}, v_k} \mathrm{L} + \mathrm{L}_{ewc} \tag{7}$$

For multi-brand mixed batches, alternating improvements are made, and memory usage is reduced by using half-precision and gradient accumulation. This strategy achieves "reusable cultural tokens" while enabling rapid cross-brand adaptation and controllable inference latency. See Table 4 for relevant details:

Table 4. Parameter-efficient tuning plan (loRA / dreambooth / embedding)

Module / Layer	Targets	LoRA Rank r	Scale α	Trainable Params (M)	LR	EWC λ_3	Notes
U-Net-Hi (attn blk 1–2)	To-Q / To-K / To-V	16	16	8.2	1e-4	2.0	512^2 resolution
U-Net-Mid	Cross-Attn + FFN proj	8	8	5.6	8e-5	1.5	style/layout dominant
U-Net-Lo (res blk)	Conv1×1 proj	4	4	2.1	6e-5	1.0	geometry stability
DreamBooth (few-shot)	class prior + prior-pres.	—	—	3.4	2e-6	—	10–20 images / brand
Textual Embedding	cultural tokens	—	—	0.6	1e-5	—	rest vocab f

3.3 Multimodal Control

Layout priors, edge or depth maps guide the mapping process, relying on ControlNet to access U-Net multi-scale residual branches; the IP-Adapter collects global and local style vectors from the reference map to construct the style, and the two-way conditional culture is embedded in a dual-channel control gate.

$$h = \alpha h_{\text{culture}} + \beta h_{\text{layout}} + \gamma h_{\text{style}}, \quad \alpha + \beta + \gamma = 1, \alpha, \beta, \gamma \in [0,1] \tag{8}$$

These $h_{\text{culture}}, h_{\text{layout}}, h_{\text{style}}$ respectively serve as cultural tokens, layout, and style presentation conditions ; α, β, γ Regularization, which can be provided by learnable gating or user-defined sliders, can enhance layout consistency and ensure geometric stability.

$$\mathrm{L}_{IoU} = \lambda_4 \left(1 - IoU\left(M_{gen}, M_{ref} \right) \right) \tag{9}$$

The semantic mask corresponding to the generated result is either the input layout mask or the mask predicted from the reference image. The style transfer intensity is controlled by the Gram matrix matching terms to prevent the pattern from "blurring". M_{ref} During M_{gen} the training period, sub-tasks of "layout only/style only/both" are dynamically selected to improve the practicality and reliability in the actual design process. Figure 4 presents the alignment visualization of Token-to-Concept: the t-SNE on the left shows the cluster distribution of tokens of different cultures, and the attention heatmap on the right shows the component-level salient areas.

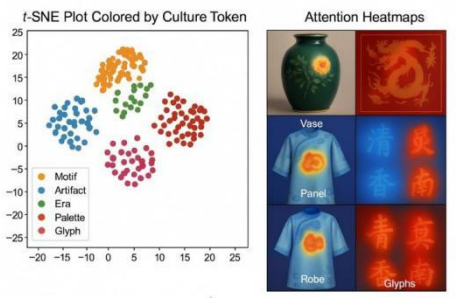

Fig. 4. Token-to-concept alignment visualization

3.4 Inference-time Calibration and Constraints

The Inference is implemented using CFG and CFG-Rescale to suppress over-amplification:

$$\hat{\epsilon} = (1+w)\epsilon_\theta(z|\ c) - w\epsilon_\theta(z|\ \varnothing), \quad \hat{\epsilon}_{\text{rescale}} = \frac{\sigma_{\text{tgt}}}{\sigma(\hat{\epsilon})}\hat{\epsilon} \quad (10)$$

The weights are used to guide the allocation of attention to cultural regions, and the objective variance is used to reflect this, σ_{tgt} thereby w enhancing the recognizability of key components.

$$A' = A \odot \left(1 + \eta \mathrm{l}_{\text{culture}}\right) \quad (11)$$

The cross-attention A map $\eta > 0$ is used as an enhancement coefficient to achieve cue compliance. The calibration head is trained using the validation set to predict the cue consistency probability p, and temperature scaling is used for subsequent processing.

$$p_T = \sigma\left(\frac{logit(p)}{T}\right), \quad T > 0 \quad (12)$$

Based on this, a reliability curve is plotted and the expected calibration error is calculated:

$$ECE = \sum_{b=1}^{B} \frac{n_b}{n} |acc(b) - conf(b)| \quad (13)$$

At the same time, color difference and soft layout constraints are retained as inference penalties and introduced with gradually decreasing weights in step sampling to prevent overfitting of details. Ultimately, a controllable trade-off of "cultural fidelity, compositional stability, color fidelity, and cost/latency" is achieved, ensuring that the visual design of time-honored brands is reusable and auditable.

4 Experiment and Evaluation

4.1 Experimental Setup

To verify the effectiveness of the overall approach of "cultural feature embedding—multimodal control—inference calibration," three complementary segmentation methods were employed: first, brand removal, which removed target established brands from the training set to measure cultural fidelity and compliance with prompts in the context of cross-brand migration; second, era-specific retention, which constructed cross-era inference scenarios for patterns from different periods based on MuralDH historical tags to examine the stability across style domains; and third, implementation of standard random segmentation to facilitate indicator aggregation and significance testing. The indicator system encompasses objective quality (FID, Gen. Precision/Recall, Diversity) and semantic consistency (CLIPScore , Prompt Adherence), and also incorporates human-assessed double-blind evaluation of cultural fidelity. In terms of engineering records, the system centrally records GPU model, quantity,

training duration, number of trainable parameters (including LoRA rank), batch size, and learning rate during training and inference, along with deployment-related data such as latency and final cost, facilitating a re-presentation of cost-effectiveness trade-offs. To eliminate interference caused by differences in training budgets, LoRA adopts a two-stage approach of "shared/private": Stage-Shared implements the extraction of public cultural components, while Stage-Private serves the brand-specific needs. The memory and time of both are counted separately. During inference operations, the end-to-end time of single image generation is counted for both A100 and L4 cards, and converted into the unit cost of 1,000 images for comparative research. Table 5 summarizes the training/inference budgets and environment in this setup, providing verifiable evidence for subsequent result interpretation.

Table 5. Training/Inference budget & environment

Stage	GPU Type	#GPU	Train Hrs	Params(M)	LoRA Rank	Batch	LR	Inference Lat. (ms)	Cost ($/1k imgs)
Stage-Shared LoRA	A100-80G	4	18.5	865	16	32	1.0e-4	420	9.8
Stage-Private LoRA	A100-80G	4	22.0	865	8	twenty four	8.0e-5	450	10.6
Inference-only	L4-24G	2	0.0	865	0	8			

4.2 Overall Results and Comparison

The proposed method demonstrates systemic advantages over native Stable Diffusion and other enhancements (i.e., DreamBooth and ControlNet) across multiple metrics: FID and ECE decrease sharply, CLIPScore , Precision/Recall, and Diversity all increase, and Prompt Adherence remains stable within the high confidence interval. To ensure that the differences reach statistical significance, a document-level/image-level paired architecture is implemented, and paired t-tests are used to analyze the main indicators and provide 95% confidence intervals. Table 6 shows the quantitative summary of all methods; to facilitate cross-index review, Figure 5 presents a comprehensive radar chart of each indicator after standardization by the same-direction gain, which directly shows the overall improvement of the proposed method in the four dimensions of "quality, consistency, diversity, and calibration". It is important to note that while the commercial general model has a certain level of CLIPScore and diversity, it still deviates in terms of ECE and prompt compliance in culturally sensitive areas. This is related to the lack of targeted cultural embedding and layout priors. Cultural tokens and dual-channel control gates have set up more robust structural and color constraint mechanisms when performing cross-domain migration, thereby achieving stable overall performance.

To help readers comprehensively grasp the characteristics of this method, based on quantitative comparisons, we briefly outline its advantages and disadvantages: On the strengths, the proposed cultural feature embedding and multimodal control methods achieve consistent improvements over native Stable Diffusion and its enhancement schemes across multiple metrics, including FID, CLIPScore, Precision/Recall, Diversity, and ECE. This indicates that the method facilitates a more balanced match between image quality, semantic consistency, diversity, and reliability calibration. Relying on efficient parameter fine-tuning (LoRA/DreamBooth) and dual-channel gating design, it improves cross-brand migration and reusability in practical scenarios while maintaining acceptable inference latency and resource overhead. On the weaknesses, the current method still heavily relies on manually constructed cultural attribute labels and color constraints. In situations where cultural elements are extremely scarce and semantics are highly abstract, the generated results remain unstable. The complex control structure increases the difficulty of model configuration and parameter tuning; future implementations should utilize more user-friendly interfaces and automatic configuration methods to reduce the overhead for designers and engineers.

Table 6. Overall quantitative comparison

Method	FID↓	CLIPScore ↑	Precision↑	Recall↑	Diversity↑	ECE↓	Prompt Adh .↑
SD (Base)	24.5	0.288	0.72	0.65	0.62	0.145	0.74
SD+DreamBooth	22.1	0.301	0.75	0.69	0.65	0.132	0.78
SD+ControlNet	20.3	0.316	0.78	0.73	0.67	0.118	0.81
Commercial-Gen	21.8	0.309	0.77	0.71	0.66	0.125	0.79
Proposed (Cultural)	17.4	0.337	0.83	0.79	0.71	0.083	0.87

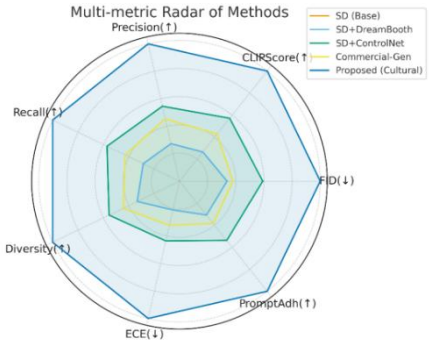

Fig. 5. Multi-metric radar of methods

4.3 Cultural Fidelity and Calibration Analysis

Cultural fidelity was assessed using a double-blind A/B expert evaluation. Evaluators with backgrounds in traditional crafts and brand design were selected. They selected pairwise preferences for four key categories: "pattern/object/color spectrum/typography". Tie situations where no decision could be made were recorded. The Kappa coefficient was used to measure the level of agreement among the evaluators, and the p-value was used as a measure of significance. Table 7 shows the win rate, tie rate, κ, and significance level values for "Proposed Method v. Controls". The proposed method shows a more significant advantage when there are complex prompts such as "color spectrum—layout—pattern reproduction". A reliability curve for prompt compliance was generated using the validation set to check the deviation between the predicted confidence and the actual probability of conformity. Then, the expected calibration error was calculated. Figure 6 shows two visualizations: the first is the Reliability Curve, which clearly shows the closeness of the curve to the diagonal; the second is a bar chart of ECE, which analyzes the magnitude of calibration error for each method. Combined with human evaluation and calibration, it can be seen that the saliency guidance of cultural tokens and color difference penalty together reduce the number of high-confidence misjudgments that "seem to match but are not correct in detail", pushing the reliability of high-segment segments closer to the ideal situation, and the ECE decreases accordingly.

Table 7. Human cultural fidelity—win rate & agreement

Pair (Method A vs B)	Win%(A)	Win%(B)	Tie%	κ(expert)	p-value
Proposed vs SD	71.2	16.4	12.4	0.72	0.001
Proposed vs SD+DreamBooth	66.5	18.4	15.1	0.69	0.003
Proposed vs SD+ControlNet	58.7	16.7	24.6	0.65	0.018
Proposed vs Commercial-Gen	61.9	19.8	18.3	0.67	0.007

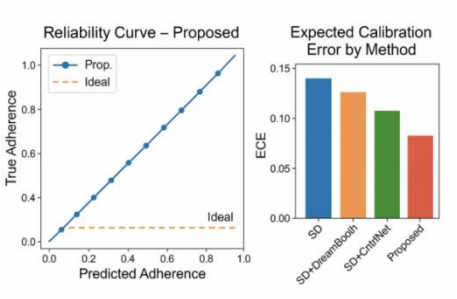

Fig. 6. Prompt adherence reliability & calibration (reliability curve + ECE bars/columns)

4.4 Robustness and Generalization

Robustness evaluation is based on two types of real-world design scenarios: layout changes and style perturbations. Using a unified perturbation intensity specification, the changes in Prompt Adherence, CLIPScore , and Diversity are collected. Significance tests are then performed based on cross-sample paired differences. When layout is missing, the stability of the basic SD for main element positioning decreases more significantly. The proposed method, relying on ControlNet's structural prior and dual-channel gating, can still maintain high cue compliance and diversity in areas of strong perturbation. The impact of color gamut shift on ECE is narrower. Figure 7 shows the multi-line curves of "perturbation intensity - index response." The Proposed curve has a gentler slope, indicating a stronger tolerance for layout and style degradation. This result aligns with the gating design in Section 3.3: if the layout/style signal quality deteriorates, the gating weights automatically shift towards the cultural embedding channel, thereby maintaining the visibility and consistency of semantic core components. This mechanism gives practical significance to engineering deployment: if brand materials are incomplete or the quality of online reference images fluctuates, the system can still maintain an acceptable level of delivery and stable subjective evaluation.

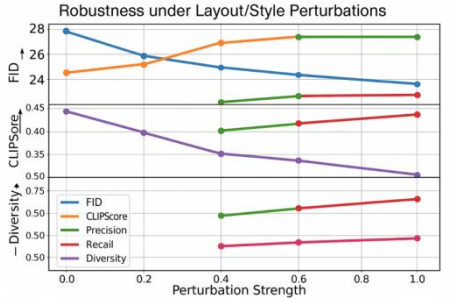

Fig. 7. Robustness under layout/style perturbations (multi-line graph: perturbation intensity vs. indicators)

5 Ablation Research and Application Implementation

5.1 Component Ablation

To measure the actual value of each module, the same data segmentation and measurement methods as in Chapter 4 were used. "Culture token, LoRA , IP-Adapter, ControlNet, and calibration head" were removed or replaced one by one, while ensuring a stable training budget. The process began with the creation of public cultural components using "Shared LoRA ," followed by short-round fine-tuning of the brand's private set. If a component was removed, the closest alternative path was used to ensure the graphics pipeline ran smoothly (if there was no IP-Adapter, statistical style vector mean pooling was used). During the evaluation, FID, CLIPScore , Precision/Recall, Prompt Adherence, and ECE were obtained, along with end-to-end latency and unit cost. Table 8 shows that removing the culture token directly reduces the consistency between Prompt Adherence and human evaluation, indicating that semantic alignment is crucial for detail triggering. If LoRA was removed , cross-brand generalization was poor, with precision and recall both dropping to low levels. After the IP-Adapter was deactivated, style color and material texture were difficult to retain. CLIPScore and diversity show a downward trend; without ControlNet, layout perturbations cause a sharp decline in stability; without the calibration

head, ECE increases but is accompanied by high confidence errors. From the perspective of latency and cost, removing components results in a slight decrease, but the quality loss is more pronounced, and the overall cost-effectiveness is at its lowest level. This result confirms that the proposed method belongs to "weakly coupled gain", where each component builds complementary constraint relationships in different dimensions, and the overall performance is compromised when one component is removed.

Table 8. Ablation on components

Setting	FID↓	CLIPScore ↑	Prec /Rec↑	ECE↓	Prompt Adh .↑	Lat. (ms)	Cost ($)
Full (Proposed)	17.4	0.337	0.83 / 0.79	0.083	0.87	450	10.2
- Cultural Tokens	21.1	0.315	0.78 / 0.72	0.121	0.79	430	9.1
- LoRA	22.4	0.308	0.75 / 0.70	0.118	0.78	438	9.4
- IP-Adapter	20.9	0.302	0.76 / 0.72	0.110	0.80	442	9.6
ControlNet	21.7	0.309	0.77 / 0.71	0.115	0.78	428	9.2
- Calibration Head	18.1	0.334	0.82 / 0.78	0.136	0.86	450	

5.2 Case Studies and Error Analysis

The case studies were divided into two categories: "positive examples" and "negative examples". Focusing on four key aspects—cultural patterns, spot colors, composition, and contemporary semantics—the generated results showed that traditional patterns such as blue-and-white intertwined branches, longevity symbols, and meander patterns exhibited stable structures and convergent strokes. Furthermore, the deviation from the brand's spot color palette's Lab mean remained within acceptable limits. This aligns with the color difference constraints and dual-channel gating discussed in Chapter 3, indicating that the high-confidence adherence areas and high-response areas of the attention heatmap are consistent. Negative examples mostly converged on two scenarios: first, element mis-triggered elements. If the prompt simultaneously includes multiple symbols such as "auspicious clouds/dragon patterns/satin", cultural tokens compete, leading to patterns incongruous with the context in certain areas; second, poor temporal adaptability. In settings with significant differences between the Song, Ming, and Qing styles, the thickness of decorative lines and layout partitions are easily confused. A small number of samples showed that the brand decoration was overfitted, with excessive accumulation of decorative elements interfering with the recognition of the main body. To address this issue, a two-step mitigation approach was adopted: in the copywriting field, the weight of cultural tokens was reduced and applied in stages; in the reasoning process, the weight of layout consistency was increased and a mild negative prompt filter was enabled. Figure 8 shows the comparison of the six-grid layout: the top row shows the reproduction of cultural patterns and the restoration of spot colors, while the bottom row shows the phenomena of mismatch of eras, overfitting of decorations, and incorrect fonts. All mismatches are shown with the help of borders and annotation boxes, which is conducive to matching with quantitative indicators.

Fig. 8. Qualitative grid with cultural annotations

5.3 Case Studies and Error Analysis

To assess the effectiveness of the evaluation method in the actual design process, it is integrated into the standard workflow of "requirements-materials-initial draft-review-final draft". In the "requirements → materials" stage, the system provides a searchable list of cultural tokens and example guidance; in the "initial draft" stage, it quickly generates high-fidelity sketches using IP-Adapter and ControlNet; during the review period, it automatically outputs prompts for fit scores and calibration confidence intervals to

help designers quickly identify risky diagrams; in the final draft stage, it generates spot color measurements and element lists for legal compliance verification. A/B testing encompassed two scenarios: a real-world team and a simulated online team. The brand's designers categorized the real-world team into two types: "traditional process (control)" and "auxiliary process (experiment)". The simulated team implemented semi-automatic production based on the same prompts and segmentation. The results were evaluated across five dimensions: time from concept to first draft (T2Concept), first-time pass rate, number of revisions, satisfaction level, and risk flagging rate. As shown in Table 9, the auxiliary process significantly reduced the time and number of revisions, increased the first-time pass rate and subjective satisfaction, and greatly reduced compliance warnings. This indicates that the proposed method can be successfully integrated into the actual production process and deliver quantifiable returns.

Table 9. Online/Simulated A/B Outcomes

Cohort	T2Concept(h)↓	1stPass(%)↑	Revisions (#) ↓	Satisfaction (1–5) ↑	Legal Flags (‰) ↓
Traditional Workflow (Real)	6.3	41.8	3.7	3.2	7.6
Assisted Workflow (Real)	3.1	63.5	2.1	4.1	3.2
Junior Designers (Assist)	3.6	58.4	2.5	3.9	3.8
Senior Designers (Assist)	2.7	67.2	1.8	4.3	2.9
E-commerce Posters (Sim)	3.3	61.0	2.2	4.0	3.1
Packaging Labels (Sim)	3.0	64.7	2.0	4.2	2.7

6 Conclusions and Outlook

Focusing on the three goals of "cultural fidelity, controllable composition, and cost-effectiveness" in the visual design of time-honored brands, this paper creates an adaptive optimization system based on Stable Diffusion: It employs a cultural feature embedding approach, transforming patterns, objects, crafts, and spot colors into combinable tokens; utilizes multimodal manipulation to enable collaboration between layout/style priors and cultural semantics in a dual-channel gating system; and integrates reliability calibration and color difference/layout soft constraints at the inference end, forming an auditable generation path from beginning to end. Tests on brand retention, era retention, and standard segmentation show that the method achieves simultaneous improvements in quality, diversity, prompt compliance, and calibration. Robustness evaluation and ablation conclusions further demonstrate the complementary effectiveness of each component. The process-oriented A/B testing shows a significant reduction in the time from concept to draft, the number of revision rounds, and compliance risks. Limitations of this work include the reliance on extensive manual cultural annotation and the current challenges in cross-regional cultural adaptation and automated copyright compliance. Future plans include integrating knowledge graphs and executable compliance modules, implementing edge-side incremental learning, and cross-brand federated training projects.

Acknowledgement

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Deng, J., Cao, X., & Cheng, B. (2024). Research on generating cultural relic images based on a low-rank adaptive diffusion model. In Proceedings of the 2024 Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence (pp. 629-634).
2. Bao, Q., Zhao, J., Liu, Z., et al. (2025). AI-assisted inheritance of Qinghua porcelain cultural genes and sustainable design using low-rank adaptation and stable diffusion. Electronics, 14(4), 725.

3. Alharbi, A., Alluhibi, R., Saif, M., et al. (2024). Injection of cultural-based subjects into stable diffusion image generative model. International Journal of Computer Science & Network Security, 1-14.
4. Li, X., Hou, Luccioni, S., Akiki, C., Mitchell, M., et al. (2023). Stable bias: Evaluating societal representations in diffusion models. Advances in Neural Information Processing Systems, 36, 56338-56351.
5. Cioni, D., Berlincioni, L., Becattini, F., et al. (2023). Diffusion based augmentation for captioning and retrieval in cultural heritage. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1707-1716).
6. Liu, B., Wang, L., Lyu, C., et al. (2023). On the cultural gap in text-to-image generation. arXiv preprint arXiv:2307.02971.
7. Zhang, T., Wang, Z., Huang, J., et al. (2023). A survey of diffusion based image generation models: Issues and their solutions. arXiv preprint arXiv:2308.13142.
8. Kabir, A. I., Mahomud, L., Al Fahad, A., et al. (2024). Empowering local image generation: Harnessing stable diffusion for machine learning and AI. Informatica Economica, 28(1), 25-38.
9. Gao, Z., Yuan, L., Reviriego, P., et al. (2024). Dependability evaluation of stable diffusion with soft errors on the model parameters. In 2024 IEEE 24th International Conference on Nanotechnology (NANO) (pp. 442-447). IEEE.
10. Miao, J., Ning, X., Hong, S., Wang, L., & Liu, B. (2025). Secure and efficient authentication protocol for supply chain systems in artificial intelligence-based Internet of Things. IEEE Internet of Things Journal, 12(19), 39532-39542.
11. Bai, Z., Miao, H., Miao, J., Xiao, N., & Sun, X. (2025). Artificial intelligence-driven cybersecurity applications and challenges. Innovative Applications of AI, 2(2), 26-33.

Biographies

1. **Xinbao Zhang**, Master, Teaching Assistant.Affiliation in Nanfang College Guangzhou.She has published 3 journal articles, 2 utility model patents;
2. **Jinjian Li**, PhD, Lecturer Affiliation in Nanfang College Guangzhou. He has published one independently monograph;
3. **Shizhen Zhang**, Master, Full time teacher in Science and Technology College of Hubei University of Arts and Science;
4. **Yuwei Chen**, Master, Teaching Assistant in Nanfang College. Guangzhou. She has 1 paper published in Art and Technology.

面向老字號品牌的多模態 AIGC 定製化研究——基於Stable Diffusion的視覺生成與評估框架

張新寶[1]，李進健[1]，張世楨[2]，陳雨薇[1]

[1]廣州南方學院，廣州， 中國，510790

[2]湖北文理學院，襄陽，中國，441000

摘要：爲滿足老字號品牌視覺設計中文化表達與工程實現的雙重需求，本研究提出一種基於 Stable Diffusion 的適應性優化架構。該架構採用文本嵌入（Textual Inversion）技術獲取可組合的文化表徵單元，藉助 LoRA/DreamBooth 參數實現通用風格與專屬風格的高效微調。通過集成 ControlNet 與 IP-Adapter，系統實現了佈局與風格先驗知識的融合，同時採用雙通道門控機制實現語義與構圖的協同控制。在推理階段，通過 CFG-Rescale、注意力重加權及溫度縮放等方法對提示詞遵循度的可靠性進行校準。基於公開多模態數據集及真實品牌場景的大量實驗表明，該方法在客觀指標與人工評價的一致性方面實現顯著提升；魯棒性測試與組件消融實驗證實了方法的穩定性及各組件的必要性，而 A/B 測試則凸顯其在成本效益與運行效率方面的顯著優勢。本研究最終爲文化遺產及商業品牌的視覺生成需求提供了一套可復現、可驗證的技術方案。

關鍵詞：老字號品牌；穩定擴散；文化特徵嵌入；多模態控制；高參數調微；可靠性校準；視覺生成

1. 張新寶，碩士，助教，廣州南方學院專任教師，已發表學術論文3 篇，獲得2項實用新型專利；
2. 李進健，博士，講師，廣州南方學院專任教師，已獨立發表1部學術專著；
3. 張世楨，碩士，初級，湖北文理學院專職教師；
4. 陳雨薇，碩士，助教，廣州南方學院專任教師，發表學術期刊1篇。