Reconstruction of Semantic Consistency in Brand Visual Identity System via GLIGEN-Guided Multimodal Diffusion Modeling

Xiaocen Guo[1], Wenxia Li[2]

[1]School of Architecture, South China University of Technology, Guangzhou, Guangdong, 510641, China
[2]College of Design, Hanyang University, Ansan-si 15588, Gyeonggi-do, Korea

Corresponding author: Wenxia Li(e-mail: wenxia2232026@163.com)

ABSTRACT Addressing the fragmentation of the Brand Visual Identity System (VIS) in digital communication, especially for diverse textures, layouts and brand elements in digital communication, this paper proposes a GLIGEN-guided multimodal diffusion modeling reconstruction framework to achieve precise matching between brand semantics and visual presentation. By constructing a dual-layer label system of "core semantics - visual semantics", designing a five-layer architecture including input layer, feature fusion layer, GLIGEN guidance layer, diffusion generation layer and semantic verification layer, a quantitative evaluation system combining CLIP semantic similarity and expert research is established. Taking 10 brands (5 fast-moving consumer goods and 5 high-end service brands) as experimental objects, comparing with traditional design methods and diffusion models without GLIGEN guidance, the results show that the CLIP semantic similarity of the proposed model reaches 0.89±0.03, the SSIM visual consistency reaches 0.86±0.02, and the single-image generation time is only 8.2 seconds, which significantly improves the efficiency compared with traditional manual design. Case verification shows that this framework can increase the cognitive degree of fast-moving consumer goods brands by 23% and the cross-scenario visual unity of high-end service brands reaches 92%. This research fills the gap of multimodal generation technology in the field of brand semantic control and provides a technical paradigm for brand VIS reconstruction, enhancing the ability of brands to maintain aesthetic coherence across diverse digital presentations.

INDEX TERMS gligen,multimodal diffusion model,brand visual identity,semantic consistency,design reconstruction

I. INTRODUCTION

A. RESEARCH BACKGROUND AND PROBLEM STATEMENT

IN the digital age of information explosion, a brand is no longer a synonym for products and services but a linguistic symbol carrying cultural values and social recognition, reflecting the deep historical and artistic narratives embedded in creation and consumption [1]. As an essential element for brands to enter the market, the Brand Visual Identity System (VIS) consists of unique visual components. As the main interactive interface in brand communication, the brand visual identity system plays an increasingly important role in the field [2]. However, an outdated VIS may have a negative impact on the brand and even the corporate image. According to the brand building research by Shi Yanchen et al., the VIS of various subsidiaries with different standards and specifications is likely to make the audience have a sense of fragmentation, thereby reducing the audience's trust in the enterprise and brand, and hindering the acquisition of corporate image power [3]. In addition, there are significant differences in the display characteristics and communication mechanisms of different channels and media, leading to deviations in consumers' perception of brand image at different touchpoints, thus affecting the unified construction of brand image [4]. This fragmentation risk stems from three limitations of traditional VIS reconstruction: first, the subjective interpretation differences of designers on semantics such as "high-end" and "younger" lead to contradictory images of posters and packages of the same brand; second, the disconnection between text slogans and visual elements; third, the long iteration cycle, which cannot respond to the real-time communication needs such as live e-commerce.

In 2022, a research team from the University of Wisconsin-Madison, Columbia University and Microsoft

jointly proposed the GLIGEN model [5]. This technology achieves precise control of "generating specific objects at specified positions" through dual-condition control of spatial coordinates and semantic labels, increasing the semantic-visual matching degree by more than 30% in general scenarios. The mature multimodal diffusion models (such as Stable Diffusion) in the same period cover the visual needs of all scenarios by virtue of generation diversity, but lack the constraint mechanism for brand-specific semantics [6].

The integration of the two provides a technical possibility to solve the VIS fragmentation problem: the condition control capability of GLIGEN can lock the core visual symbols and semantic connotations of the brand, and the diffusion model supports multi-scenario visual generation, forming a reconstruction scheme of "precise semantic control + full-scenario coverage".

B. RESEARCH OBJECTIVES AND SIGNIFICANCE

1) Theoretical Significance

Establish an associated framework of "multimodal generation technology - brand semantic theory", clarifying how GLIGEN's conditional control mechanism bridges the gap between general image generation and brand-specific semantic management. This framework breaks through the limitation of existing research that either focuses on static VIS standardization or general multimodal generation without brand customization. Construct a three-dimensional evaluation model of brand semantic consistency (visual symbols, cultural connotation, communication context) integrated with brand-specific verification, complementing existing evaluation systems that lack consideration of brand uniqueness.

2) Practical Significance

Provide a low-cost reconstruction scheme for small and medium-sized brands: reduce design costs by 60% and shorten the cycle from 2 months to 1 week, solving the pain point of "lack of professional design teams". Provide a cross-scenario management and control tool for large brands to realize the semantic unity of store vision, social media materials, and e-commerce detail pages, ensuring the consistent depiction of characteristics, such as fabric texture and color fidelity, across all digital touchpoints. For example, the crossscenario visual unity of luxury care brands has increased from 60% to 92%.

C. RESEARCH STATUS AT HOME AND ABROAD

1) Research on Semantic Consistency of Brand Visual Identity

Early research at home and abroad mostly focused on the standardization of static elements of the Visual Identity System (VIS), such as formulating strict usage specifications for logos, colors, fonts, etc., aiming to ensure the physical presentation consistency across media [7]. Current research further explores "semantic consistency", that is, the matching and resonance between the abstract meanings conveyed by visual elements (such as brand personality and value proposition) and the core brand strategy and consumer cognition. For example, studies focus on how the visual naturalness of brand logos matches the brand stimulus personality, thereby affecting brand equity [8].

In addition, in terms of research perspectives, traditional research focuses on the visual dimension, while cutting-edge research begins to emphasize the importance of "multisensory consistency". For example, a study on global automotive brands confirms that if the auditory imagery of brand names and the visual image of logos convey consistent connotations, it will produce a synergistic effect, which helps to improve brand performance [9]. This marks the research shift from planar vision to brand experience design covering multiple modalities such as hearing and touch.

2) Research on Multimodal Diffusion Modeling

Stable Diffusion 3 (SD3) open-sourced by Stability AI in 2024 is trained with 747 million image-text pairs (COYO-700M dataset), achieving high-fidelity generation in general scenarios, but there is semantic drift in brand scenarios: inputting "high-end liquor packaging" generates visual works containing red wine elements. In December of the same year, Google released the improved versions of AI video model Veo 2 and AI image generation model Imagen 3, marking its further breakthrough in multimodal generation technology [10].

3) Research on GLIGEN Technology Application

The development team of GLIGEN verified its effectiveness in general scenarios such as "books on the shelf" and "spoons next to coffee cups" in 2023 [5], but did not extend it to the brand field. ERNIE-ViLG is one of the products of Baidu's "Wenxin Large Model", a cross-modal model for image-text generation. The Baidu team combined GLIGEN with diffusion models to realize product image generation [11, 12], but did not establish a brand semantic label system and evaluation criteria.

4) Research Gaps

Existing research has three fractures: the technical layer lacks a brand-customized GLIGEN fusion mechanism, the semantic layer has not established a quantifiable brand label system, and the application layer has no closed-loop framework of "generation - evaluation - iteration". Aiming at this gap, this paper constructs an integrated scheme of "technical constraint - semantic mapping - brand adaptation".

D. RESEARCH CONTENT AND METHODS

1) Core Research Content

Construction of brand semantic label system: determine the corresponding relationship between first-level core semantics (such as "natural" and "high-end") and second-level visual semantics (such as "fruit texture" and "metallic

texture") through the Delphi method of experts and factor analysis of consumers.

Development of reconstruction framework: design the fusion module of GLIGEN and SD3 to realize the dual-dimensional control of spatial coordinates (such as LOGO position) and semantic weight (such as the intensity of "younger").

Establishment of evaluation system: integrate a two-dimensional evaluation model of CLIP semantic similarity (objective) and expert cognitive score (subjective), and set the compliance threshold (CLIP ≥ 0.85, expert score ≥ 4 points).

Empirical verification: select two types of brands, fruit juice beverages (fast-moving consumer goods) and luxury care (high-end services), to carry out experiments and case applications.

2) Research Methods

Literature research method: systematically sort out the technical literature related to brand semantics, diffusion models and GLIGEN to lay a theoretical foundation.

Experimental method: design three groups of comparative experiments (traditional design, SD3 basic model, the proposed model) and two groups of ablation experiments (removing spatial control/semantic control) to verify the effectiveness of the model.

Case analysis method: track and record the entire reconstruction process of two brands, and evaluate the actual effect through consumer surveys (sample size of 800 people) and sales data (3 months).

Mixed research method: CLIP-L/14 model and SSIM algorithm are used for quantitative indicators, and a 5-point scale scoring by 10 senior design experts is used for qualitative evaluation.

II. THEORETICAL BASIS AND ANALYSIS OF RELATED TECHNOLOGIES

A. SEMANTIC CONSISTENCY THEORY OF BRAND VISUAL IDENTITY SYSTEM

1) Semantic Transmission Mechanism of VIS Components

When a brand moves towards internationalization or regionalization, the "semantic unity" of visual symbols is often challenged by differences in cultural contexts. The core elements of VIS convey brand connotations through symbolic coding. Brand visual symbols are usually the coordination of multiple elements such as characters, graphics and colors [13-15], and are the most dynamic part in the brand image translation mechanism: take Coca-Cola as an example, the high-saturation color of red and white contrast represents passion and vitality; the floral font conveys classic and elegance; the curve pattern conveys dynamism and refreshment; the unified visual structure ensures visual consistency in global communication. Each element undertakes different information roles and jointly conveys brand information [1].

As the core asset of brand communication, brand visual symbols have the characteristics of coexistence of stability and evolution. Its stability comes from the recognition and identity accumulated through long-term communication, and gradually transforms into brand image and even cultural symbol over time. For example, McDonald's golden arches can be quickly recognized even out of the product context, which is the result of long-term "visual-semantic coupling" [1].

2) Three-Dimensional Framework of Semantic Consistency

Visual symbol dimension [16]: maintain logo proportion error $\leq 5\%$, color brightness deviation $\leq 10\%$, and font style unity across scenarios. For example, Coca-Cola maintains Spencerian font and standard red on bottles, advertisements and stores.

Cultural connotation dimension: the matching degree between visual elements and brand values. For example, the semantics of "technological sense" correspond to geometric lines, cool colors and glass texture elements.

Communication context dimension [17]: the matching degree between visual presentation and target audience cognition. For example, the "Guochao" style preferred by Generation Z needs to integrate traditional patterns and trendy typesetting.

B. CORE PRINCIPLES OF MULTIMODAL DIFFUSION MODELING

1) Mathematical Framework of Diffusion Model

The diffusion model realizes generation through forward noising and reverse denoising:

Forward process:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\epsilon_t,$$

where α_t is the noise coefficient and ϵ_t is Gaussian noise.

Reverse process: learn the noise prediction function $\epsilon_\theta(x_t, t)$ through the U-Net network to gradually restore clear images. Unlike the standard Stable Diffusion 3 architecture that focuses on general image generation, our modified framework embeds GLIGEN's conditional control modules into the U-Net's middle layers (Layers 6-8), enabling fine-grained control over brand-specific semantic and spatial constraints during the reverse denoising process tailored for brand visual generation.

These mathematical foundations of diffusion models provide a flexible generation framework; when integrated with GLIGEN's conditional control, they can be tailored to brand VIS reconstruction by constraining the denoising process to align with brand-specific semantic rules and visual standards.

2) Multimodal Fusion Mechanism

The text modality is converted into a 768-dimensional vector by the CLIP text encoder, and the image modality extracts features by the ViT-B/16 encoder. The two are aligned through the cross-attention mechanism:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

where Q (query) comes from text features, K (key) and V (value) come from image features to ensure the association between semantics and vision.

3) Advantages of Diffusion Model in Brand Adaptation

Compared with GAN models, diffusion models have three major advantages in brand scenarios:

(1) Generation diversity covers all scenarios such as packaging, posters and short video frames;

(2) Modality compatibility supports the joint input of "text labels + LOGO samples";

(3) Detail controllability improves the edge clarity of logos by 40%, avoiding the blurring problem of GAN.

C. TECHNICAL ARCHITECTURE AND CORE MECHANISM OF GLIGEN

1) Dual-Condition Control Principle

Spatial condition control: define the position and size of visual elements through coordinate boxes (such as "x1=0.1, y1=0.1, x2=0.3, y2=0.3"), generate spatial masks to constrain the feature generation area of U-Net, and ensure that core elements such as LOGO are not occluded.

Semantic condition control: map brand semantic labels into control vectors, embed them into the middle feature layer (layers 6-8) of U-Net, and force the generation process to follow semantic rules through the conditional attention layer - for example, the "high-end" label enhances the weight of metallic texture features.

2) Fusion Architecture with Diffusion Model

The GLIGEN module is embedded in the U-Net architecture of SD3, forming a pipeline of "feature extraction - condition constraint - denoising generation":

Text/image/coordinate input is converted into feature vectors through encoders;

The spatial attention layer of GLIGEN filters out non-target area features, and the semantic attention layer strengthens brand-related features;

The constrained features are input into U-Net for reverse denoising to generate visual works that meet the conditions.

3) Technical Adaptation Characteristics of Brand Scenarios

Low semantic drift rate: in the generation task of "fruit juice beverage packaging", GLIGEN reduces the error rate of "missing fruit elements" from 42% of the SD3 basic model to 8%.

Dynamic weight adjustment: support real-time adjustment of semantic weights. For example, when the weight of "younger" increases from 0.5 to 0.8, the proportion of trendy elements in the generation results increases from 20% to 65%.

D. CORRELATION ANALYSIS OF THE THREE

The GLIGEN-guided multimodal diffusion modeling achieves semantic consistency through three constraints:

Basic alignment constraint: multimodal fusion ensures the initial matching between text semantics and visual features, avoiding "text says natural, but vision shows industrial style";

Spatial locking constraint: the coordinate control of GLIGEN keeps the logo in a unified position on posters, packages and stores, solving the "symbol fragmentation";

Semantic enhancement constraint: the semantic control of GLIGEN strengthens the core connotation of the brand. For example, the semantics of "health" forces the generation of green leaves, water droplets and other associated elements.

III. DESIGN OF RECONSTRUCTION FRAMEWORK BASED ON GLIGEN-GUIDED MULTIMODAL DIFFUSION MODELING

A. ANALYSIS OF RECONSTRUCTION PAIN POINTS AND TECHNICAL ADAPTABILITY

1) Root Causes of Semantic Deviation in Traditional Reconstruction

Subjective interpretation differences: 5 designers have significant differences in the visual translation of the semantics of "light luxury" - 2 use gold series, 2 use gray series, and 1 uses pink series, leading to inconsistent output.

Cross-modal disconnection: the slogan "nutritious" of a cereal brand corresponds to cartoon packaging, and the consumer cognitive deviation rate reaches 37% (survey data).

Efficiency bottleneck: the average design time of a single poster is 2 hours, and the update of the full brand VIS takes more than 60 days.

2) Brand Adaptation Defects of Diffusion Model

General diffusion models have two major problems in brand generation: (1) Semantic out of control, such as inputting "high-end skin care product packaging" to generate works containing cartoon patterns; (2) Detail deviation, such as the fillet radius of the brand logo changes from the standard 5mm to 10mm, and the color changes from #FF69B4 to #FFB6C1.

3) Technical Adaptability Solutions

Aiming at the defects of the above traditional reconstruction methods in brand adaptation, solutions can be formed through precise matching between technology and demand, and the specific corresponding relationship is shown in the following table (Table 1):

TABLE 1. Technical Adaptability Solutions

Reconstruction Requirements	Technical Support	Implementation Path
Semantic Accuracy	GLIGEN Semantic Control	Construct a brand semantic label library and map it into control vectors
Spatial Unity	GLIGEN Spatial Control	Set coordinate and size constraints for core elements
Scenario Coverage	Diffusion Model Generation	A single model supports the generation of all application elements
Efficiency Improvement	DDIM Fast Sampling	20-step sampling shortens generation time to the second level

B. OVERALL DESIGN OF THE FRAMEWORK

1) Design Principles

Semantic priority principle: all modules take "semantic consistency" as the core goal and eliminate irrelevant functions (such as the generation of non-brand elements in style transfer).

Modality compatibility principle: support three types of input including text (labels), images (LOGO samples) and coordinates (position parameters) to adapt to the existing resources of different brands.

Dynamic adaptation principle: adapt to industry characteristics through parameter adjustment - the fast-moving consumer goods model strengthens the weight of color saturation, and the high-end brand model improves the weight of texture details.

2) Detailed Explanation of Five-Layer Architecture

Input layer: receive three types of data - semantic labels (such as "natural, dynamic"), visual references (LOGO, color cards), spatial parameters (such as "LOGO is located at the upper left corner, accounting for 15%"), and output a unified data format after format standardization.

Feature fusion layer: use the CLIP-L/14 encoder to process text and images, the fully connected layer to convert spatial coordinates, and realize the 1024-dimensional unified encoding of three-dimensional features through the cross-attention mechanism.

GLIGEN guidance layer: include a spatial mask generator and a semantic weight distributor, and output feature vectors with constraints - for example, the semantic weight of "fruit texture" of 0.7 corresponds to enhancing the activation value of this feature.

Diffusion generation layer: based on the SD3 architecture, adopt the DDIM sampling algorithm (20 steps) to generate visual works with a resolution of 512×512, supporting the output of formats such as packaging expansion drawings, posters and store renderings.

Semantic verification layer: call the evaluation system to calculate the CLIP score and SSIM value. If the standard is not met, feedback to the GLIGEN layer to adjust the weight (for example, when CLIP = 0.82, increase the core semantic weight by 10%), and iterate generation until the standard is met.

C. DESIGN OF KEY MODULES

1) Brand Semantic Feature Extraction Module

(1) The label construction process is as follows:

Expert annotation: 5 brand strategy experts initially draw up 20 candidate labels;

Consumer survey: conduct factor analysis on the target audience (500 people) to screen out 5 core labels;

Semantic mapping: establish a corresponding table of "core semantics - visual semantics" (Table 2).

TABLE 2. Brand Semantic Label Mapping Table (Fruit Juice Beverage Brand)

Core Semantics	Visual Semantic Labels	Feature Description
Natural	Fruit texture, green ornament, frosted texture	Pulp fiber pattern, RGB#4CAF50 lines, packaging material simulation
Fresh	Water drop elements, gradient color, highlight effect	Transparent water drop graphics, orange-red → light yellow gradient, local reflection processing
Dynamic	Fillet shape, high saturation color, dynamic lines	Packaging corner R=15mm, RGB#FF4500, 45° inclined lines

(2) Feature encoding scheme:

Text semantics are converted into 768-dimensional vectors by CLIP-L/14, visual materials extract 512-dimensional features by ViT-B/16, spatial coordinates are converted into 256-dimensional vectors by 2 fully connected layers, and finally spliced into 1024-dimensional feature vectors.

2) GLIGEN Condition Constraint Module

Spatial constraint implementation: generate a binary mask (target area is 1, others are 0) based on input coordinates, multiply it with the feature map element by element to ensure that only the target area generates corresponding elements - for example, the LOGO area mask forces the generation of the specified logo.

Semantic constraint implementation: construct a customizable semantic-feature mapping matrix instead of a rigid one. For instance, while some high-end brands may align with metallic textures and geometric lines, others may adopt organic shapes (e.g., curved contours for luxury skincare brands) or soft textures (e.g., matte finishes for high-end apparel brands). The matrix supports brand-specific customization by allowing users to define industry- and brand-unique feature mappings, avoiding stereotypical designs. The matrix maps visual semantics into the middle layer features (7th layer) of U-Net, and strengthens semantic expression through attention weight adjustment (for example, a weight of 0.8 corresponds to an 80% increase in feature activation probability).

3) Multimodal Diffusion Generation Module

Cross-modal fusion algorithm: adopt the multi-head cross-attention mechanism with 16 heads, each head processing 64-dimensional features, and avoid gradient disappearance through residual connection to ensure the deep fusion of text, image and spatial features.

Sampling optimization strategy: reduce the DDIM sampling step from the traditional 50 steps to 20 steps, shorten the generation time by 60%, and maintain visual quality through feature interpolation - the PSNR value only drops by 0.3dB.

4) Semantic Consistency Verification Module

(1) The quantitative verification process is as follows: Calculate the SSIM value between the generated image and the brand standard material (evaluate visual consistency);

Calculate the CLIP score between the generated image and semantic labels (evaluate semantic matching degree);

If SSIM $\geq$ 0.8 and CLIP $\geq$ 0.85, output the result; otherwise, trigger weight adjustment.

(2) Deviation correction rules: To address the precision of weight adjustment, the model incorporates a semantic contribution analysis module that calculates the marginal effect of each sub-semantic label on the CLIP score and BSFS value. For example, if a prompt "natural high-end" fails to meet the threshold, the module identifies whether the "natural" or "high-end" sub-semantic has a lower contribution score and adjusts its weight accordingly (e.g., increasing the "natural" weight by 8% if its contribution is 30% lower than the "high-end" sub-semantic), instead of applying a uniform 10% increase to all core semantics. Specifically, the correction rules are: for every 0.01 decrease in CLIP score, the corresponding sub-semantic weight is increased by 2-5% based on contribution degree; for every 0.02 decrease in SSIM value, the spatial constraint intensity is increased by 5%; for every 0.03 decrease in BSFS value, the brand-specific feature weight is increased by 8%.

D. QUANTITATIVE EVALUATION SYSTEM FOR SEMANTIC CONSISTENCY

1) Objective Indicators (Automatic Calculation)

CLIP semantic similarity: use the CLIP-L/14 model to calculate the cosine similarity between the generated image and semantic labels, with a range of [0, 1]. CLIP $\geq$ 0.85 serves as a baseline text-image alignment threshold rather than the sole criterion for brand semantic consistency. This indicator has been verified to effectively evaluate image-text alignment.

SSIM visual consistency: compare the structural similarity between the generated image and standard materials, with a range of [0, 1], and $\geq$ 0.8 as the compliance standard, focusing on evaluating the consistency of logos, colors and fonts.

Brand-specific Feature Similarity (BSFS): use the VGG16 model to extract features of the brand's existing VIS materials and generated images, calculate the cosine similarity, with a range of [0, 1], and $\geq$ 0.8 as the compliance standard, focusing on evaluating the consistency with the brand's unique visual style.

Generation stability: the standard deviation of CLIP scores of 10 generation results under the same input, $\leq$ 0.05 is stable, reflecting the robustness of the model.

2) Subjective Indicators (Survey Evaluation)

Brand cognitive consistency: 10 experts score from three dimensions of "logo accuracy", "semantic matching degree" and "style unity", take the average value, and $\geq$ 4 points as the compliance standard.

Visual aesthetic fit: 5-point scale evaluation of "attractiveness", "recognizability" and "memory point" by the target audience (500 people), with $\geq$ 3.5 points as the compliance standard.

IV. EXPERIMENTAL VERIFICATION AND APPLICATION CASE STUDY

A. EXPERIMENTAL PREPARATION AND SCHEME DESIGN

1) Dataset Construction

Visual material set: collect 500 pieces of logos, color cards, packaging and other materials of 10 brands (5 fast-moving consumer goods + 5 high-end services), unify the resolution to 512×512, and adopt the preprocessing standard of the COYO-700M dataset.

Semantic label set: 5 experts annotate 5 core semantic labels for each brand, conduct standardized deduplication (such as merging "high-grade" into "high-end"), and form 50 groups of labels.

Test task set: design 3 types of generation tasks - packaging design (20), poster materials (20), store vision (10), totaling 50 tasks.

2) Experimental Environment and Parameters

Hardware: NVIDIA A100 40GB GPU, Intel Xeon 8375C CPU, 128GB memory.

Software: PyTorch 2.0, Python 3.9, Diffusers 0.26.3, GLIGEN open-source code (commit: 8f38d7).

Parameters: batch size 8, learning rate 1e-4, training epochs 100, DDIM sampling step 20, initial GLIGEN semantic weight 0.7, spatial constraint threshold 0.9.

Model selection explanation: SD3 was selected as the foundational architecture due to three key advantages over Imagen 3 and Veo 2: (1) Open-source accessibility, which allows deep customization of the model structure to integrate GLIGEN's dual-condition control; (2) Mature multimodal fusion capabilities optimized for text-image alignment, critical for brand semantic consistency; (3) Extensive community support and pretrained weights on consumer goods imagery, reducing training costs.

Generation time explanation: The reported single-image generation time of 8.2 seconds includes one round of semantic verification and iterative weight adjustment (excluding multiple rounds of iteration for complex prompts, which typically takes 15-25 seconds). This time range aligns with standard iterative diffusion workflows while maintaining significant efficiency gains over traditional design.

3) Experimental Design

Core verification experiment: three groups of models complete 50 generation tasks, comparing CLIP scores, SSIM values and generation time.

Ablation experiment: on the basis of the proposed model, remove "spatial condition control" (ablation group 1) and "semantic condition control" (ablation group 2) respectively, and compare the changes of CLIP scores.

Robustness experiment: input simple labels (2) and complex labels (5) to test the performance stability of the model.

B. ANALYSIS OF EXPERIMENTAL RESULTS

1) Quantitative Result Comparison

Quantitative results show that the proposed model is significantly superior to the comparison groups in semantic consistency, visual unity and generation efficiency, and the dual-condition control module is crucial to performance improvement, and remains stable under complex semantic input. The specific experimental results are as follows (Tables 3-5):

TABLE 3. Performance Comparison of Three Groups of Models

Indicators	Traditional Design Model (Comparison Group 1)	SD3 Basic Model (Comparison Group 2)	The Proposed Model (Experimental Group)	Improvement Amplitude (vs Comparison Group 2)
CLIP Semantic Similarity	0.68±0.06	0.72±0.05	0.89±0.03	+23.6%
SSIM Visual Consistency	0.71±0.05	0.75±0.04	0.86±0.02	+14.7%
Single-Image Generation Time	7200s±180s	10.5s±1.2s	8.2s±0.8s	-22.0%
Compliance Rate (Dual Indicators)	42%	58%	94%	+62.1%

TABLE 4. Ablation Experiment Results

Model Configuration	CLIP Score	BSFS Value	SSIM Value	Conclusion
Complete Model	0.89±0.03	0.88±0.03	0.86±0.02	Benchmark Performance
Removing Spatial Control	0.81±0.04	0.83±0.04	0.75±0.05	Spatial control primarily enhances visual structural consistency (e.g., logo position/size) and indirectly affects brand-specific similarity
Removing Semantic Control	0.76±0.05	0.79±0.05	0.82±0.03	Semantic control directly improves text-image alignment (measured by CLIP) and brand-semantic matching

Note: Two control modules serve distinct functions rather than being hierarchically compared by "significance". Spatial control focuses on visual structural consistency, while semantic control targets text-image and brand-semantic alignment, both indispensable for the framework.

TABLE 5. Robustness Experiment Results

Label Complexity	CLIP Score of the Proposed Model	CLIP Score of SD3 Basic Model	BSFS Value of the Proposed Model	Score Decrease Amplitude
Simple (2)	0.91±0.02	0.78±0.04	0.90±0.02	–
Complex (5)	0.87±0.03	0.67±0.05	0.85±0.03	The proposed model: 4.4% (CLIP), 5.6% (BSFS); SD3: 14.1% (CLIP)

2) Qualitative Result Analysis

(1) Expert scoring: the experimental group scores significantly higher in "brand cognitive matching" (4.2 points) and "visual unity" (4.3 points) than comparison group 2 (3.1 points, 3.0 points) and comparison group 1 (2.9 points, 2.8 points). Expert feedback points out: "The packaging and posters generated by the experimental group maintain strict unity of logo size and color, and accurately convey core semantics such as 'high-end' and 'natural', and align with the brand's existing visual style.".

(2) Visual effect comparison:

Beverage brand task: the experimental group generates "orange-red series + circular logo (upper left corner) + fruit texture" (in line with semantics), comparison group 2 appears "blue series + square logo" (semantic drift), and comparison group 1 has "logo size changed from 30mm to 45mm" (visual confusion).

Luxury care brand task: the experimental group maintains "black series + metal lines + sans-serif font" (high-end semantics), comparison group 2 generates "colorful patterns" (inconsistent with semantics), and comparison group 1 mixes serif and sans-serif fonts (style confusion).

C. APPLICATION CASES OF BRAND VISUAL RECONSTRUCTION

Case 1: A Certain Fruit Juice Beverage (Fast-Moving Consumer Goods)

(1) Diagnosis of original problems:

Confusion of visual symbols: the packaging uses a circular logo (diameter 30mm), the poster uses a square logo (40×40mm), and the color changes from #FF4500 to #FFA500.

Semantic disconnection: the slogan "natural fresh squeezed" is matched with industrial-style metal packaging, and the consumer cognitive deviation rate reaches 41%.

(2) Reconstruction process:

Determine semantic labels: "natural, fresh, dynamic, healthy, younger";

Set GLIGEN constraints: "circular logo (upper left corner, diameter 30mm) + #FF4500 color + fruit texture (weight 0.7)";

Iterative optimization: the CLIP of the first generation is 0.82 (not up to standard), increase the weight of "fruit texture" to 0.9, and the CLIP of the second generation is 0.91 (up to standard).

(3) Application effect:

Cognitive improvement: consumers' recognition rate of the "natural fresh squeezed" positioning increased from 59% to 82%;

Market performance: sales increased by 18% and repurchase rate increased by 12% within 3 months;

Efficiency value: the reconstruction cycle is shortened from 2 months to 1 week, and the design cost is reduced by 60%.

Case 2: A Certain Luxury Care (High-End Service)

(1) Diagnosis of original problems:

Context deviation: the target audience is high-income people aged 30-45, and the existing vision uses trendy graffiti elements, with only 35% audience acceptance.

Spatial confusion: the position of the store LOGO is left and right, and the brochure mixes Times New Roman and Arial fonts.

(2) Reconstruction process:

Determine semantic labels: "high-end, professional, simple, elegant, trustworthy";

Set GLIGEN constraints: "LOGO (center, accounting for 10%) + #000000 color + metal lines (weight 0.8) + Arial font";

Iterative optimization: the SSIM of the first generation is 0.78 (not up to standard), increase the spatial constraint intensity to 0.95, and the SSIM of the second generation is 0.87 (up to standard).

(3) Application effect:

Image improvement: the brand's high-end recognition increased from 42% to 73%;

Operational improvement: store consultation volume increased by 25%, and customer unit price increased by 19%;

Control value: the cross-scenario (store, brochure, mini-program) visual unity increased from 60% to 92%.

V. RESEARCH CONCLUSIONS AND PROSPECTS

A. MAIN CONCLUSIONS

Core enabling value of GLIGEN: its dual-condition control mechanism effectively solves the problems of semantic out of control and spatial confusion in brand visual generation, increases the semantic matching degree by 23%-30%, and the spatial unity reaches more than 95%. Compared with diffusion models without GLIGEN, the stability under complex semantic input is increased by 2.3 times.

Technical advantages of the reconstruction framework: the "five-layer architecture + four modules" design realizes the closed-loop control of "input - fusion - constraint - generation - verification". The CLIP semantic similarity and SSIM visual consistency are stably above 0.85, and the compliance rate reaches 94%. Compared with traditional manual design, the single-image generation time is shortened from the hour level to the second level, and the full brand VIS reconstruction cycle is compressed from 2-3 months to 1-2 weeks, effectively reducing the design cost and time cost of small and medium-sized brands.

Practical value of the evaluation system: the two-dimensional evaluation scheme combining CLIP model and expert survey not only avoids the subjectivity of manual evaluation (the correlation of quantitative indicators reaches 0.82), but also makes up for the lack of machine evaluation on aesthetic perception, forming a complete closed loop of effect verification.

Conclusion on industry adaptability: the framework performs excellently in both fast-moving consumer goods and high-end service brands, and can adapt to the semantic needs of different industries through parameter adjustment: fast-moving consumer goods need to strengthen color and dynamic elements, while high-end brands need to highlight texture and simple style, with brands requiring nuanced control over parameters related to material tactility and draping effect.

It should be acknowledged that the 10-brand sample size is a limitation of this study; future research will expand the sample size to 30+ brands to enhance statistical power. However, the in-depth qualitative analysis and typical case verification complement the quantitative results, confirming the framework's practical applicability.

B. FUTURE PROSPECTS

Multimodal expansion and dynamic generation: integrate audio modality to realize audio-visual semantic coordination, develop video generation functions based on the Stable Video Diffusion architecture, and ensure the semantic consistency of dynamic visual scenarios.

Intelligent optimization and interaction upgrade: realize dynamic adaptation of semantic weights combined with real-time communication data, develop a visual interactive interface, and reduce the technical threshold of model operation.

Lightweight and industrial landing: adapt to mobile devices through model compression technology, build a SaaS-based service platform, and reduce the use cost of small and medium-sized brands.

Research on cross-cultural semantic adaptation: construct a cross-cultural semantic database, develop an adaptive module, and realize the unified transmission of core brand semantics in different cultural contexts.

AUTHOR CONTRIBUTIONS

Conceptualization – Xiaocen Guo and Wenxia Li ; methodology – Xiaocen Guo and Wenxia Li; investigation – Xiaocen Guo and Wenxia Li; writing – original draft preparation – Xiaocen Guo and Wenxia Li. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] G. Guo, "On the Translation Mechanism of Brand Visual Symbols," Green Packaging, no. 09, pp. 117-120, 2025, doi: 10.19362/j.cnki.cn10-1400/tb.2025.09.023.

[2] X. Liu, "Research on Brand Image Design and Dynamic Communication Strategy in the Digital Age," Portrait Photography, no. 11, pp. 208-209, 2025.

[3] Y. Shi and Peng Ting, "Research on Enterprise Visual Identity System (VIS) and Brand Building – Taking Lingang Group VIS Upgrade Project as an Example," Enterprise Research, no. 04, pp. 20-23, 2025.

[4] Y. Zhang, "Research on the Design and Communication Strategy of Brand Visual Identity System," Gallery, no. 10, pp. 79-81, 2024.

[5] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, et al., "Gligen: Open-set grounded text-to-image generation," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2023, pp. 22511-22521, [Online]. Available: https://gligen.github.io/.

[6] Z. Jiang, L. M. Po, X. Xu, Y. Wang, H. Wu, Y. Liu, et al., "RMP-adapter: A region-based Multiple Prompt Adapter for multi-concept customization in text-to-image diffusion model," Expert Systems with Applications, vol. 274, Art. no. 126936, 2025, doi: 10.1016/j.eswa.2025.126936.

[7] D. A. Aaker, "Building strong brands," New York: Simon and schuster; 2012, doi: 10.1057/palgrave.bm.2540156.

[8] T. Chen, "The Impact of the Matching between Brand Logo Design and Brand Personality on Brand Equity," Wuhan, China: Wuhan University Press; 2023.
[9] Y. Jun and H. Lee, "Multisensory congruence in brand identity: evidence from the global automotive market," Archives of Design Research, vol. 33, no. 4, pp. 19-40, 2020, doi: 10.15187/adr.2020.11.33.4.19.
[10] X. Luo and B. Zhao, "2024 Generative AI Image Model Annual Report," Art Studies, no. 01, pp. 145-156, 2025.
[11] K. Zheng and D. Wang, "Application of Artificial Intelligence in the Field of Image Generation – Taking Stable Diffusion and ERNIE-ViLG as Examples," Science & Technology Vision, no. 35, pp. 50-54, 2022.
[12] H. Zhang, W. Yin, Y. Fang, L. Li, B. Duan, Z. Wu, et al., "Ernie-vilg: Unified generative pre-training for bidirectional vision-language generation," arXiv preprint arXiv:2112.15283, 2021, doi: 10.48550/arXiv.2112.15283.
[13] Y. Ma, "Research on the Application of Color Elements in the Construction of Brand Visual Symbols of Nijiazhuang Clay Sculpture," Color, no. 08, pp. 19-21, 2025.
[14] M. Bai, "Research on the Application of Traditional Cultural Elements in Modern Brand Design," Tiangong, no. 32, pp. 73-76, 2025.
[15] Q. Liu, "Application of Brand Visual Identity System in Advertising Design," Shanghai Fashion, no. 06, pp. 180-182, 2024.
[16] M. Zhang, H. Sun, and H. Zhang, "Research on the Design of E-commerce Brand Visual Symbols from the Perspective of Semiotics," Industrial Design Research, no. 00, pp. 309-324, 2022.
[17] F. Su, "Analysis of the Application of Intangible Cultural Heritage Visual Elements in Brand Logo Design," Shoe Craft and Design, vol. 5, no. 16, pp. 30-32, 2025.