Journal of Design Intelligence
https://www.ffspub.com/index.php/jdi
ISSN 3104-9001 (print)
ISSN 3136-3032 (online)
E-mail: jdi.office@ffspub.com



Article

A Prompt-Optimized Vision-Language Model System for Fine-Grained Fashion Style Recognition

Rui Mai, Ting Chen, Xiao Li, Xiaoqun Dai †
College of Textile and Clothing Engineering, Soochow University, Suzhou 215000, China
†*E-mail: daixqsz@suda.edu.cn*

Received: April 17, 2026 / Revised: June 11, 2026 / Accepted: June 24, 2026 / Published online: July 27, 2026

Abstract: With the accelerated digital transformation of the fashion industry, accurate clothing style recognition has become essential for intelligent e-commerce retrieval, personalized outfit recommendation, and fashion trend analysis. Vision-language models (VLMs), with their cross-modal semantic alignment and zero-shot inference capabilities, offer an efficient technical pathway for this task. However, when applied to the fashion domain, general-purpose VLMs face several challenges, including vague prompt semantics, insufficient fine-grained discrimination, difficulty in capturing composite style characteristics, and high deployment costs. To address these issues, this paper presents a comprehensive clothing style recognition system based on Qwen3-VL-Plus. A four-level structured prompt template—comprising style definitions, visual feature guidance, and confidence distribution constraints—is introduced to direct the model's attention toward fine-grained attributes such as silhouette, fabric, color, and decoration. A confidence distribution modeling mechanism based on Softmax normalization is proposed to convert cross-modal similarity scores into interpretable confidence distributions, enabling quantitative representation of composite styles. Furthermore, INT8 quantization and inference pipeline optimization are employed to achieve lightweight deployment, allowing the system to operate stably in an ordinary CPU environment. On the FashionStyle14 dataset, the proposed system achieves a Top-1 accuracy of 69.21% and a macro-average F1 score of 66.38%, outperforming CLIP by 17.98 and 18.21 percentage points, respectively. INT8 quantization yields a 36.77% reduction in inference time and a 75.12% reduction in model size. In a user experience evaluation involving 30 participants, the system attains an overall satisfaction score of 8.60 out of 10, with 96.67% of users recognizing the interpretability of the recognition results. Experimental results demonstrate that the proposed approach offers significant advantages in recognition accuracy, interpretability, and deployment practicality, providing a complete solution with both technical feasibility and application value for intelligent clothing style recognition in the fashion industry.

Keywords: Clothing Style Recognition; Vision-Language Model; Prompt Engineering; Confidence Distribution; Lightweight Deployment
https://doi.org/10.64509/jdi.14.118

1 Introduction

The digital transformation of the fashion industry is profoundly reshaping the entire chain of garment design, production, distribution, and consumption. The integration of e-commerce and social media has generated massive volumes of fashion visual data, with clothing images exhibiting rapid growth [1, 2]. Clothing style recognition—automatically determining the style category of a garment from its image—has become a critical technology connecting visual perception and fashion semantics, with applications ranging from intelligent product categorization and personalized recommendation on e-commerce platforms to outfit trend analysis and digital archiving of traditional costumes [3–6]. However, clothing style is inherently abstract and subjective, requiring the synthesis of multi-dimensional information including silhouette, fabric, color, decorative details, and wearing context. This makes automated style recognition a persistently challenging topic in the intersection of computer vision and fashion.

In recent years, vision-language models (VLMs) have introduced a new paradigm for clothing style recognition.

† Corresponding author: Xiaoqun Dai
*Academic Editor: Zhaohui Wang
© 2026 The authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Models such as CLIP [7] employ cross-modal contrastive learning to map images and texts into a unified semantic space, achieving powerful zero-shot generalization. The Qwen-VL series [8] further optimizes fine-grained visual perception and instruction-following capabilities, providing superior support for vertical domain tasks. Unlike traditional pure vision models that extract only shallow features such as texture and color, VLMs can transform abstract semantic concepts like style into computable feature vectors, thus theoretically possessing stronger fine-grained discrimination ability.

Despite these advances, existing VLMs still face three core challenges in clothing style recognition. First, generic prompts such as "This is a picture of [style name] clothing" provide insufficient guidance, enabling only shallow semantic matching and failing to activate the model's focus on key visual features–silhouette, cut, fabric texture, and decorative details–that determine stylistic differences. Second, the mainstream single-label hard-classification paradigm cannot represent the composite style characteristics prevalent in real-world outfits, and raw similarity scores lack probabilistic interpretability, undermining trust in downstream decisions. Third, the inference process of large-parameter VLMs heavily relies on high-performance GPUs, making deployment on ordinary consumer-grade devices or mobile terminals prohibitively expensive and restricting industrial application [9–11].

To address these issues, this paper constructs a high-precision, interpretable, and lightweight clothing style recognition system based on Qwen3-VL-Plus. At the methodological level, three core strategies are proposed. First, a domain-specific structured prompt system is introduced, integrating professional dimensions such as silhouette, fabric, color, decorative details, and wearing context through hierarchical semantic guidance of "style definition + visual feature guidance + multi-label output constraint." Second, a confidence distribution modeling module is designed based on Softmax normalization, converting cross-modal similarity scores into interpretable percentage probability distributions and enabling simultaneous output of Top-1 classification results and full-category composite style probabilities. Third, a lightweight inference scheme combining INT8 weight quantization, prompt pre-encoding cache, and image preprocessing optimization is adopted, significantly reducing model storage, memory consumption, and inference latency while preserving recognition accuracy, thereby enabling stable operation on ordinary CPU environments without dedicated GPUs. At the system level, a front-end interactive platform integrating image upload, style recognition, result visualization, and history management is built using the Streamlit framework.

The main contributions of this paper are:

(1) A domain-specific structured prompt optimization system that translates professional fashion design knowledge into model-comprehensible semantic instructions;

(2) A confidence distribution modeling and output mechanism that unifies academic evaluation metrics (Top-1 accuracy) with application-level multi-style confidence output;

(3) A lightweight inference scheme balancing accuracy and efficiency, enabling stable CPU-only operation and significantly lowering deployment barriers.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 details the proposed methodology. Section 4 presents experimental validation. Section 5 concludes and discusses future work.

2 Related Work

2.1 Clothing Style Recognition

Research on clothing style recognition has evolved from traditional machine learning to deep learning, with each stage exhibiting distinct characteristics and limitations.

Early studies primarily employed hand-crafted feature extraction operators combined with traditional classifiers for style determination. Wang *et al*. [3] proposed an intelligent fashion recommender system using fuzzy logic for rule-based matching of garment attributes. While this approach explicitly linked user requirements with clothing features, it relied on manually defined fixed rule sets, limiting its ability to capture abstract styles such as vintage or minimalism, and could not dynamically adapt to rapidly evolving fashion trends. Lee *et al*. [12] developed a garment pattern shape information system that constructed a silhouette feature library through image analysis. Although achieving accurate extraction of physical garment morphology, this method overemphasized shape features while neglecting other critical style determinants such as color matching and decorative elements, resulting in poor performance when distinguishing visually similar styles like Korean vs. Japanese or street vs. punk. Cheng *et al*. [13] proposed an intelligent clothing search system that integrated style features into the retrieval process for the first time, but its fixed keyword matching strategy could not effectively respond to emerging fashion trends. Overall, traditional methods share common limitations: hand-crafted features have limited expressive capacity, are sensitive to variations in lighting and pose, and struggle to capture the full connotation of abstract semantic concepts like style.

With the advancement of deep learning, convolutional neural networks (CNNs) have gradually become the mainstream approach for clothing style recognition. Hidayati *et al*. [14] proposed a full-body image-based clothing genre classification method that exploited combinations of style elements for category discrimination, though its feature engineering approach showed limited generalization in diverse outfit scenarios. Wang *et al*. [15] proposed a multi-feature fusion method for personalized clothing recommendation, combining texture, color, and silhouette features to improve recommendation diversity. However, this method relied on large-scale annotated data, and its large parameter count and slow inference speed made deployment on resource-constrained devices challenging. Moreover, pure vision deep learning models generally suffer from insufficient interpretability—their style determination process operates as a "black box," unable to explain classification rationales to users [16, 17].

2.2 Vision-Language Models and Fashion Applications

The emergence of vision-language models has brought a paradigm shift to fine-grained image recognition. The dual-encoder architecture represented by CLIP [7] employs contrastive learning on large-scale image-text pairs to map both modalities into a unified semantic space, endowing models with powerful zero-shot classification capabilities. Building on this architecture, the Qwen-VL series [8] further enhances fine-grained visual perception and instruction-following abilities, providing superior foundational performance for vertical domain applications.

In the fashion domain, VLMs have demonstrated initial application potential. Kim *et al.* [18] leveraged VLM-guided fine-grained textual attributes with cross-domain contrastive optimization to significantly improve fashion image retrieval accuracy. Wang *et al.* [19] proposed FashionVQA, introducing the visual question answering paradigm into the fashion domain and opening new pathways for clothing attribute recognition.

However, a critical bottleneck persists across existing research: VLM performance is highly dependent on prompt design quality. In general scenarios, simple category name prompts may suffice, but in highly specialized vertical domains like clothing style, generic prompts fail to effectively activate the model's domain knowledge. Liu *et al.* [20] systematically investigated prompt design guidelines in text-to-image generation tasks, demonstrating that structured and specific prompts significantly improve generation quality. Cao *et al.* [21] proposed AdaCLIP, which employs learnable hybrid prompts for zero-shot anomaly detection, achieving notable performance gains. These studies indicate that task-specific prompt optimization is a key pathway to unlocking the potential of VLMs, providing theoretical support for the domain-specific prompt design in this work.

2.3 Model Lightweight Deployment

The inference process of large-parameter deep learning models typically relies on high-performance GPUs, severely limiting their practical application in resource-constrained scenarios. Model quantization is one of the core technical approaches to address this issue, converting model weights and activations from high-precision floating-point to low-precision integer representations to reduce storage and computational overhead. Zhao *et al.* [22] proposed a distribution-adaptive INT8 quantization method that dynamically adjusts quantization parameters during training, significantly compressing model size while maintaining classification accuracy.

From an industrial application perspective, lightweight deployment represents a critical step in transitioning technical achievements from laboratory to practical scenarios. Existing clothing style recognition research has predominantly focused on algorithmic performance improvement, with insufficient attention to actual deployment conditions and hardware requirements, which has to some extent limited the industrialization of related technologies.

2.4 Our Motivation

In summary, although significant progress has been made in clothing style recognition research, three key limitations remain. At the semantic understanding level, fixed rules or keyword libraries struggle to adapt to dynamically changing fashion semantics, and general-purpose VLMs lack domain-specific prompt adaptation for clothing. At the output format level, the single-label hard-classification paradigm cannot characterize the composite style features prevalent in real-world outfit scenarios, and the interpretability of recognition results is insufficient. At the deployment practicality level, large model sizes and high hardware requirements lack end-user-oriented lightweight validation. To address these challenges, this paper proposes a clothing style recognition system based on prompt-optimized VLMs, integrating three core strategies—structured domain prompts, confidence distribution modeling, and lightweight inference—to achieve synergistic improvement in recognition accuracy, interpretability, and deployment practicality.

3 Methodology

3.1 Problem Formulation and System Overview

Figure 1 presents the overall architecture of the proposed fashion style recognition system, comprising five main processing stages: user input, data preprocessing, VLM inference, post-processing, and result output, followed by its application in fashion design support, e-commerce style classification, and fashion trend analysis.

The clothing style recognition task can be formally defined as follows: given an input clothing image, the system outputs a confidence probability distribution over a predefined set of 14 style categories (conservative, dressy, ethnic, fairy, feminine, gal, girlish, kireime-casual, lolita, mode, natural, retro, rock, and street). The probabilities of all 14 style categories sum to 100%, with each individual probability value ranging from 0% to 100%. Based on this distribution, the system outputs both a Top-1 classification result for academic evaluation and a full-category probability distribution to represent the composite style composition of the garment.

Built upon the Qwen3-VL-Plus pre-trained VLM, the system comprises three core processing pipelines: a prompt optimization module responsible for generating domain-specific structured prompt templates and performing pre-encoding cache; a multimodal inference module that extracts image features and style semantic features via the visual encoder and text encoder respectively, and computes cross-modal similarity scores; and a result parsing module that performs probability normalization on raw scores and outputs structured recognition results. Additionally, the system integrates an INT8 quantization engine and an image preprocessing optimization unit to ensure efficient inference in an ordinary CPU environment. The front-end interaction layer is built on the Streamlit framework, providing users with image upload, result visualization, and history management functionalities.

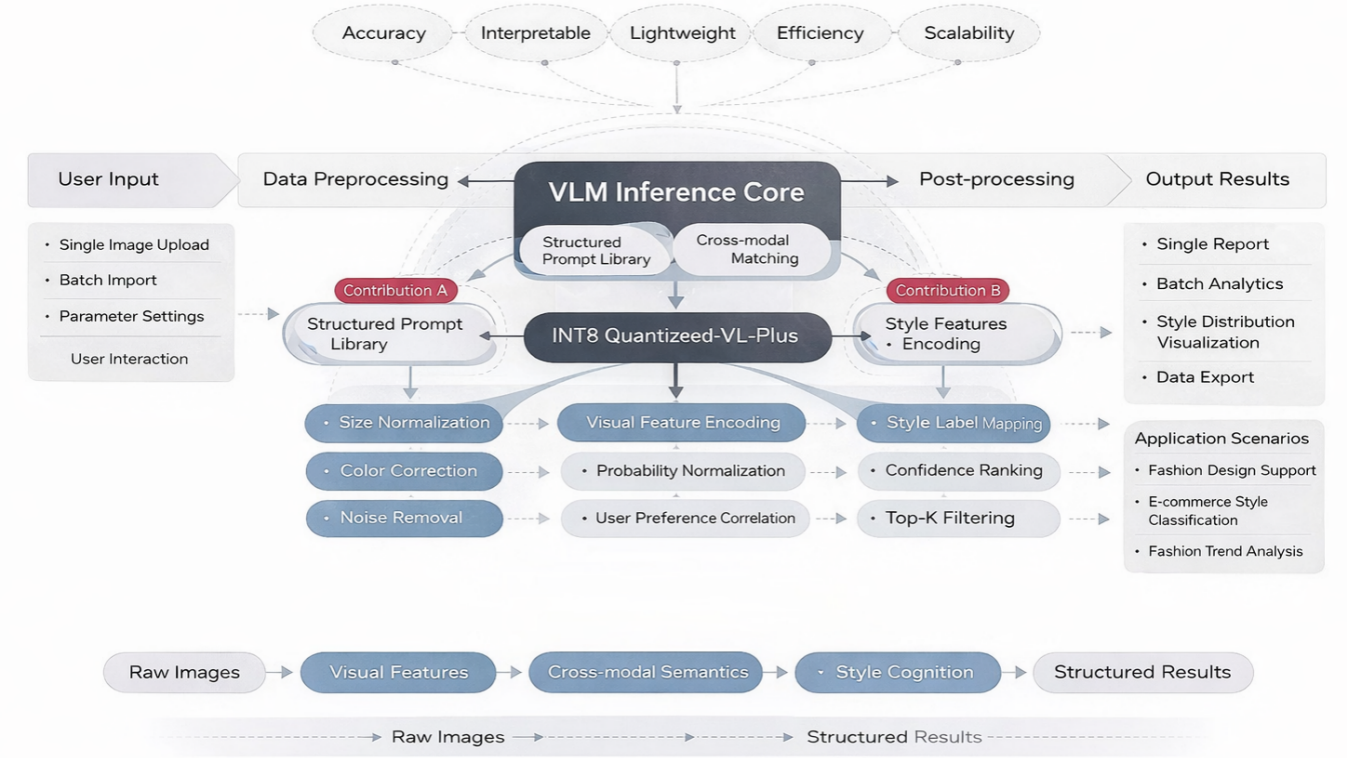


Figure 1: Overall framework of the proposed clothing style recognition system.

3.2 Structured Prompt Optimization for the Fashion Domain

Prompts serve as the core guiding medium for VLMs to perform zero-shot inference, and their semantic precision and structural integrity directly determine the effectiveness of fine-grained style recognition. In preliminary experiments, we observed that using the generic prompt template "This is a picture of [style name] clothing" yielded a Top-1 accuracy of merely 48.35% for the native Qwen3-VL-Plus model, with confusion rates exceeding 30% for styles with blurred boundaries. As illustrated in Figure 2, the input image is first processed by the four-level structured prompt, followed by multimodal semantic analysis, feature confidence estimation, weighted score calculation, and Softmax normalization to generate the final confidence distribution. Analysis revealed that generic prompts only enable shallow semantic matching based on style names, failing to guide the model toward capturing the core visual features that determine stylistic differences—particularly struggling to distinguish between visually similar categories such as conservative vs. kireime-casual, and retro vs. natural.

To address this issue, we systematically reviewed research findings in fashion stylistics and garment design engineering, establishing that clothing style determination is not single-dimensional feature matching, but rather the synergistic outcome of five core elements: silhouette, fabric, color, decorative details, and wearing context. Silhouette serves as the most intuitive visual carrier of style, with four basic types—H-line, waist-cinched, straight, and loose—covering the vast majority of everyday clothing styles. Fabric, as the material basis of style expression, directly influences stylistic tonality through its texture and drapability; for instance, crisp suiting fabric naturally points toward business casual style, while soft cotton-linen blends are strongly associated with natural casual style. In terms of color, low-saturation neutral tones are typical of mode and kireime-casual styles, while high-saturation color blocking is more common in street and girlish styles. Decorative details such as buttons, belts, prints, and logos serve as key supplementary evidence for distinguishing similar styles. Wearing context provides contextual semantic support for style judgment.

Based on the above theoretical foundation, we constructed a four-level structured prompt system. The first level comprises task instructions and role setting, explicitly positioning the model to perform recognition from a professional fashion design perspective. The second level contains professional feature definitions for the 14 style categories, with each style precisely characterized by quantifiable visual features. For example, mode style is defined as "clean tailoring, smooth lines, loose or fitted silhouette, low-saturation solid colors, no superfluous decoration, emphasis on fabric texture itself"; conservative is defined as "formal tailoring, neutral tones, fitted or slightly relaxed silhouette, functional details (belt/pockets/concealed placket), suitable for workplace office scenarios" (see Appendix A for complete definitions of all 14 styles). The third level is the visual attribute guidance layer, requiring the model to focus on the four core features of the garment itself—silhouette, fabric, color, and decorative details—while completely ignoring irrelevant information such as background, model, and accessories. The fourth level is the output format constraint layer, strictly standardizing the output order and format across eight fields—category, silhouette, fabric, pattern, decoration, style, color, and recognition rationale—ensuring structural consistency of results.

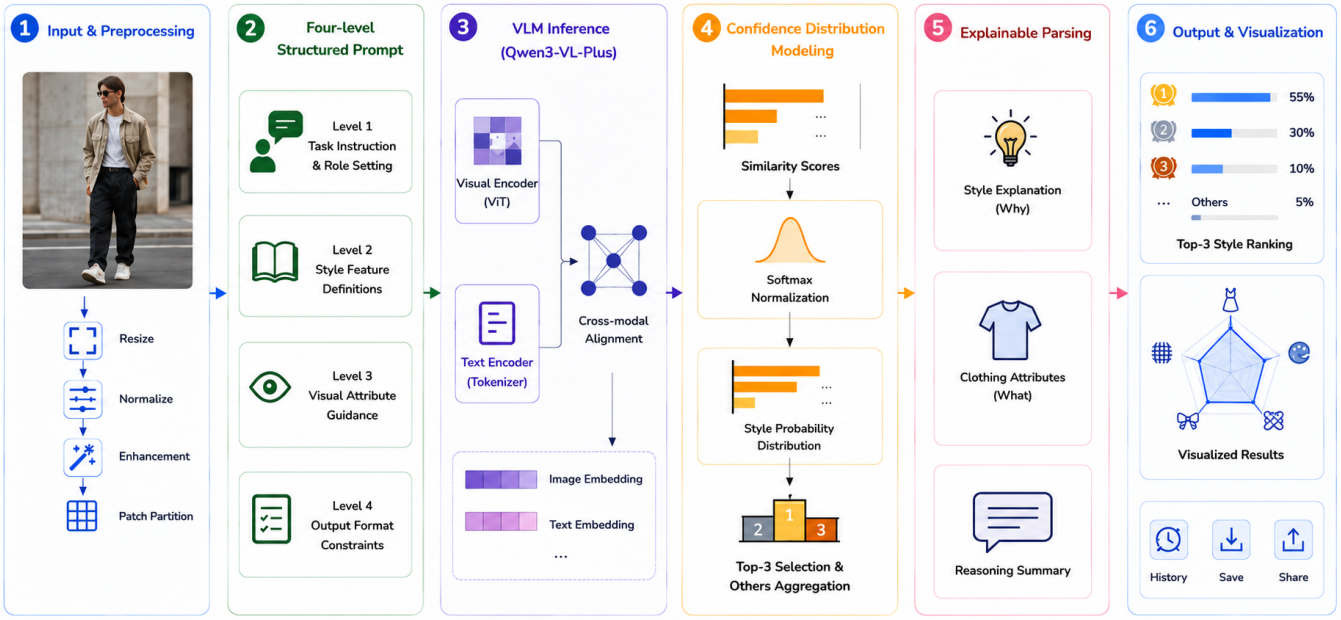


Figure 2: Overall workflow of the proposed clothing style recognition system.

To further strengthen the semantic guidance of core style features, a feature weighting mechanism is introduced, assigning differentiated weight coefficients to different dimensional attributes:

$$S_{\text{style}}(c) = \sum_{i=1}^{n} \omega_i \cdot f_i(c) \tag{1}$$

where $S_{\text{style}}(c)$ denotes the final matching score of the garment for style category c; ω_i represents the weight coefficient for the i-th feature dimension, with silhouette and fabric assigned the highest weights as core determinants of clothing style, while color matching, decorative details, and wearing context are assigned correspondingly adjusted weights as auxiliary features, with all weight coefficients summing to 1; $f_i(c)$ denotes the matching degree of the i-th feature dimension on style c, with values ranging from 0 to 1; and n is the total number of feature dimensions. This weighting mechanism guides the model to prioritize core professional features that determine style categories during inference, while attenuating the interference of non-core features and irrelevant information such as background and accessories, thereby further improving the fine-grained discrimination accuracy for styles with blurred boundaries.

The feature matching scores are not manually assigned but are automatically inferred by Qwen3-VL-Plus through the proposed four-level structured prompt framework. Given an input clothing image, the vision-language model first extracts multimodal visual representations and performs semantic matching with the predefined textual descriptions of all 14 fashion style categories. During this process, the model evaluates the semantic consistency of five key clothing attributes, including silhouette, fabric, color, decoration, and wearing context, thereby producing the corresponding feature matching scores. The feature weights in Equation (1) are predefined empirical coefficients that reflect the relative importance of each clothing attribute in fashion style recognition and remain fixed during inference. Subsequently, the weighted feature score is calculated according to Equation (1), and Softmax normalization is applied to convert the weighted scores into a confidence distribution over all predefined fashion style categories. Therefore, the final prediction is generated through prompt-guided multimodal semantic matching and confidence estimation rather than free-form text generation.

3.3 Confidence Distribution Modeling and Output Mechanism

While prompt optimization enables the model to capture multi-dimensional clothing style features semantically, the raw cross-modal semantic similarity scores output by VLMs are merely vector distance metrics in the feature space. These scores suffer from inherent deficiencies—unbounded numerical ranges, inconsistent scales across samples, and non-unit sums—making them unsuitable for direct presentation and analysis as probability results. Preliminary experimental data indicate that approximately 42% of samples in the FashionStyle14 dataset simultaneously exhibit visual features of two or more styles. The traditional single-label hard-classification approach not only loses critical style distribution information but also significantly increases the confusion rate for styles with blurred boundaries.

To address this issue, we designed an independent confidence distribution modeling module as a standardized post-processing unit between model inference and result output. This module employs the Softmax normalization function to map the 14-dimensional raw similarity vector into probability space:

$$P(c) = \frac{\exp(S_{\text{style}}(c))}{\sum_{j=1}^{M} \exp(S_{\text{style}}(j))} \times 100 \tag{2}$$

where $P(c)$ denotes the percentage confidence of the garment image corresponding to style c, ranging within $[0, 100\%]$; $S_{\text{style}}(c)$ is the matching score for style c computed from

Equation (1); and $M = 14$ is the total number of style categories. After normalization, the confidence values of all 14 styles sum strictly to 100%, with each confidence value possessing clear probabilistic interpretation, intuitively quantifying the degree of matching between the garment image and each style category.

Based on this probability distribution, the system output comprises two core components. The first is the Top-1 recognition result, outputting the style name with the highest confidence and its corresponding percentage, serving as the core classification result for quantitative evaluation in experiments. The second is the full-category probability distribution, outputting all 14 style names and their corresponding percentages in descending order of confidence, comprehensively presenting the composite style composition of the garment. This provides users with interpretable recognition results while supplying multi-dimensional style data support for downstream applications such as intelligent e-commerce retrieval and personalized outfit recommendation. Furthermore, guided by the structured prompts, the system simultaneously outputs the garment's category, silhouette, fabric, pattern, decoration, color composition, and a natural language recognition rationale, forming a complete interpretable recognition chain from visual features to style determination.

3.4 Lightweight Inference Optimization

The inference process of large-parameter VLMs heavily relies on high-performance GPU devices, representing one of the core reasons limiting the industrial application of existing technologies in clothing style recognition. This study addresses this issue through lightweight system transformation from two dimensions: model quantization and inference pipeline optimization.

For model quantization, INT8 weight quantization technology is employed to compress both the visual encoder and text encoder of the Qwen3-VL-Plus model. The core logic of quantization is to convert the model's original 32-bit floating-point (FP32) weight parameters into 8-bit integer (INT8) representations, reducing model storage footprint by over 75.12% while preserving the model's cross-modal feature alignment capability and recognition accuracy to the greatest extent possible. To minimize accuracy degradation caused by the quantization process, a calibration quantization scheme is adopted: based on a small set of representative samples from the FashionStyle14 dataset, quantization calibration is performed by statistically analyzing the activation value distributions across model layers and dynamically optimizing quantization parameter selection, ensuring that the quantized model's style recognition accuracy does not experience significant degradation. After quantization, the model no longer requires GPU parallel computation and can perform smooth inference on an ordinary consumer-grade CPU.

For inference pipeline optimization, a prompt pre-encoding cache strategy is adopted. The fixed prompts for the 14 clothing styles are pre-encoded into semantic feature vectors via the text encoder and stored in a local cache. During each inference process, only a single image feature extraction and cross-modal similarity computation need to be performed, eliminating redundant prompt encoding computations. This strategy reduces per-image inference time by over 30%. At the image preprocessing level, the standardized input image size is fixed at 448×448 pixels, the native supported resolution of the model, with bilinear interpolation for size scaling. Adaptive histogram equalization is also introduced to mitigate the interference of illumination variations on visual feature extraction, ensuring both feature extraction stability and consistency while improving inference efficiency.

4 Experiments

4.1 Experimental Settings

4.1.1 Dataset

Experiments were conducted on the FashionStyle14 fine-grained clothing style dataset [23]. This dataset contains 14 mainstream clothing style categories: conservative, dressy, ethnic, fairy, feminine, gal, girlish, kireime-casual, lolita, mode, natural, retro, rock, and street. The dataset is organized in a folder-based classification structure, with each style corresponding to an independent folder. Samples are real-world clothing images covering diverse shooting angles, lighting conditions, background environments, and wearing poses. The dataset contains 13,126 full-body clothing images covering 14 fashion styles, with approximately 1,000 images per category, resulting in a relatively balanced class distribution.

4.1.2 Experimental Environment

Aligned with the core design objective of lightweight deployment, all experiments were conducted in a CPU-only hardware and software environment. Hardware configuration: Intel Core i5-12400 hexa-core processor (2.5 GHz base frequency, 4.4 GHz maximum turbo frequency), 16 GB DDR4 3200 MHz RAM, 512 GB NVMe SSD storage, no dedicated GPU. Software environment: Windows 11 Professional 64-bit operating system, Python 3.10.14, PyTorch 2.2.0, Transformers 4.38.2, OpenCV-Python 4.9.0.80, Streamlit 1.32.0, NumPy 1.26.4, Pandas 2.2.1, Scikit-learn 1.4.1. Qwen3-VL-Plus utilized the official pre-trained weights from Tongyi Qianwen. All comparison models used officially released open-source pre-trained weights. All random seeds were fixed during experiments to ensure reproducibility.

4.1.3 Evaluation Metrics

Five core evaluation metrics were selected: Top-1 Accuracy, Macro-Average Precision, Macro-Average Recall, Macro-Average F1-score, and average per-image inference time. Top-1 Accuracy is defined as the proportion of test samples where the model's highest-probability predicted category matches the ground truth label. Macro-average metrics are computed by first calculating the corresponding metric for each category individually and then taking the arithmetic mean across all 14 categories, thereby eliminating bias introduced by imbalanced category distributions. Average inference time covers the full pipeline from image input to result output, including image preprocessing, feature extraction, cross-modal matching, and result parsing, measured in milliseconds.

4.1.4 Comparison Models

Two groups of comparison models were selected to comprehensively validate the performance advantages of the proposed system. The VLM group includes: CLIP (ViT-B/32) as the classic baseline for cross-modal zero-shot classification [7]; Qwen-VL-Max as a high-performance general-purpose VLM from the same model family [8]; and the proposed optimized system as the core experimental group. The traditional pure vision model group includes: VGG16 as a representative classic deep CNN [24]; and MobileNetV2 as a representative lightweight CNN [25]. All models used officially released pre-trained weights and were tested under identical experimental environment, dataset, and preprocessing pipeline conditions. For the CNN-based baselines (VGG16 and MobileNetV2), the pre-trained models were employed as visual feature extractors. The extracted visual features were evaluated under the same prompt-guided inference protocol as the proposed system to generate predictions over the predefined 14 FashionStyle14 categories. No task-specific fine-tuning or additional classifier training was performed.

4.2 Comparative Experiments

To validate the comprehensive performance advantages of the proposed optimized system, horizontal comparative experiments were conducted on the FashionStyle14 test set. All models performed inference directly using pre-trained weights without task-specific fine-tuning. Experimental results are presented in Figure 3.

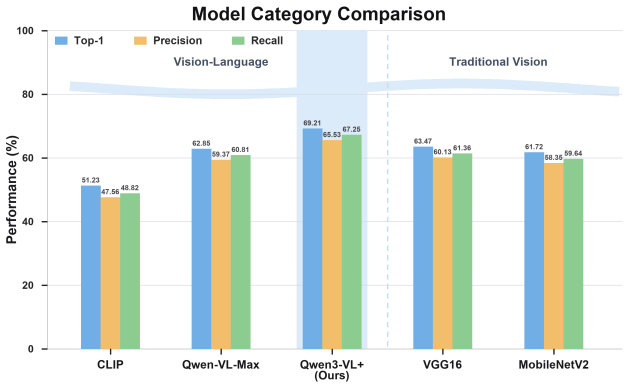


Figure 3: Comparative experimental results

Within the VLM group, the classic baseline CLIP achieved only 51.23% Top-1 accuracy and 48.18% macro-average F1-score. This result indicates that early general-purpose VLMs possess limited cross-modal representation capacity and, without domain adaptation, can only perform shallow semantic matching based on style names, failing to effectively distinguish the 14 boundary-blurred clothing styles. Qwen-VL-Max, leveraging stronger fine-grained visual perception and semantic understanding capabilities, improved Top-1 accuracy to 62.85%, representing an 11.62% point gain over CLIP. The proposed system, building upon Qwen3-VL-Plus with domain-specific prompt optimization and confidence distribution modeling, further elevated Top-1 accuracy to 69.21%—a 6.36% point improvement over Qwen-VL-Max—and macro-average F1-score by 6.30% points. This substantial gap fully validates the effectiveness of the proposed domain-specific optimization strategies, demonstrating that even for high-performing general-purpose large models, targeted domain adaptation can yield considerable performance gains.

In comparison with traditional pure vision models, VGG16 and MobileNetV2 achieved Top-1 accuracies of 63.47% and 61.72%, respectively, with macro-average F1-scores of 60.74% and 58.99%—all significantly lower than the proposed system's 69.21% and 66.38%. This performance discrepancy stems from the fundamentally different core technical approaches of the two model categories: traditional CNNs can only extract shallow visual features such as texture, contour, and color, failing to comprehend abstract high-level semantic concepts like "style," and thus struggle to distinguish style categories with highly similar visual features. In contrast, VLMs establish a unified representation space for visual features and textual semantics through cross-modal contrastive learning, transforming abstract style semantics into computable feature vectors and thereby enabling more accurate style recognition. In terms of precision and recall, the proposed system outperformed VGG16 by 5.40% and 5.89% points respectively, and MobileNetV2 by 7.18% and 7.61% points, indicating not only higher overall recognition accuracy but also stronger classification stability and generalization capability across individual style categories, with reduced cross-confusion rates between similar styles—making it more suitable for fine-grained style recognition demands in real-world fashion scenarios.

4.3 Ablation Study

To individually validate the independent effectiveness of the three core optimization strategies, three ablation experiment groups were designed with the native Qwen3-VL-Plus model as the baseline. The baseline group used the model's native inference logic with a generic single-sentence prompt template and none of the proposed optimization strategies. Ablation Group 1 introduced only the structured prompt optimization (including the feature weighting mechanism) on top of the baseline, replacing the generic prompt template with the proposed domain-specific fine-grained prompt system. Ablation Group 2 further incorporated the confidence distribution modeling and style confidence normalization output mechanism on top of Group 1. The complete system further integrated the lightweight inference optimization scheme—including INT8 weight quantization, prompt pre-encoding cache, and image preprocessing optimization—on top of Group 2. All experiments were conducted on the same test set under identical experimental conditions. Results are presented in Figure 4.

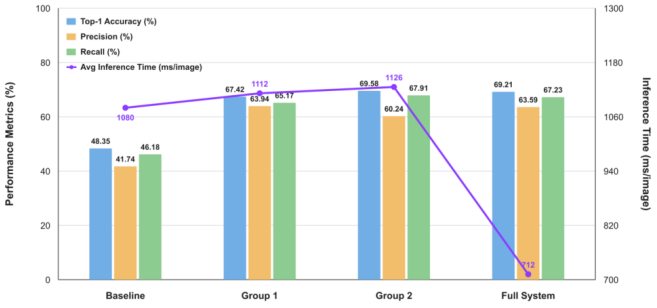


Figure 4: Ablation experiment results

The ablation results clearly reveal the independent contributions and synergistic effects of each optimization module. The baseline model, under generic prompt conditions, achieved only 48.35% Top-1 accuracy and 45.47% macro-average F1-score, with both precision and recall below 50%. This outcome aligns fully with the practical performance characteristics of fine-grained classification tasks: general-purpose VLMs, despite broad pre-training data coverage, lack vertical domain depth. When relying solely on shallow semantic matching through style names, they cannot effectively distinguish the 14 boundary-blurred clothing styles—particularly for visually similar categories such as conservative vs. kireime-casual and retro vs. natural, where confusion rates exceeded 40%.

With the introduction of only structured prompt optimization (Group 1), Top-1 accuracy surged to 67.42% and macro-average F1-score to 64.56%—improvements of 19.07 and 19.09 percentage points, respectively, over the baseline. This leap represents the largest single-factor gain in the ablation study, fully validating the core value of domain-specific prompts: by incorporating quantifiable professional visual features such as silhouette, fabric, color, and decoration into the prompts, the model is provided with explicit judgment criteria, fundamentally resolving the semantic vagueness problem of generic prompts and substantially reducing cross-confusion rates between similar styles. Inference time increased modestly from 1087 ms to 1112 ms (approximately 2.30%), attributable to the increased text encoding and contextual learning computational overhead as prompt length expanded from roughly 50 to approximately 1200 characters—an impact well within acceptable bounds.

Building upon prompt optimization with the addition of confidence distribution modeling (Group 2), Top-1 accuracy further improved to 69.58% and macro-average F1-score to 67.07%—gains of 2.16 and 2.51 percentage points over Group 1, respectively. This result demonstrates that the confidence distribution modeling, by converting raw similarity scores into standardized probability distributions through Softmax normalization, not only addresses the inability of traditional single-label outputs to represent composite styles, but also reduces the error introduced by hard classification through soft classification, further enhancing fine-grained classification stability. The additional inference time of approximately 14 ms stems from the post-processing overhead of Softmax normalization computation, result ranking, and structured parsing—fully consistent with the theoretical computational complexity.

After the complete integration of lightweight inference optimization (complete system), Top-1 accuracy experienced a marginal decrease of 0.37 percentage points, with macro-average F1-score decreasing by 0.68 percentage points, and precision and recall decreasing by 0.70 and 0.66 percentage points, respectively—all accuracy metric fluctuations controlled within an acceptable range of under 1 percentage point. Simultaneously, average per-image inference time was dramatically reduced from 1126 ms to 712 ms—a 36.77% improvement in inference speed—with model storage footprint reduced by 75.12% and runtime memory consumption reduced by 62.34%.

As shown in Table 1, INT8 quantization reduced the model storage footprint from 4.18 GB to 1.04 GB, while runtime memory consumption decreased from 6.08 GB to 2.29 GB. Meanwhile, the average inference time was reduced from 1126 ms to 712 ms, demonstrating that the proposed lightweight deployment strategy significantly improves efficiency while maintaining competitive recognition performance.

This outcome aligns perfectly with the inherent characteristics of INT8 weight quantization: the conversion of weights from 32-bit floating-point to 8-bit integer representation inevitably introduces minor numerical precision loss, which is entirely within acceptable bounds for non-safety-critical tasks such as clothing style recognition. To minimize quantization loss, the calibration quantization scheme based on the FashionStyle14 dataset was adopted, dynamically adjusting quantization thresholds through statistical analysis of activation value distributions from representative samples, thereby controlling precision degradation to the minimum level. Furthermore, the inference efficiency gains from prompt pre-encoding cache and image preprocessing optimization further amplified the engineering benefits of the lightweight scheme, fully offsetting the temporal overhead introduced by the preceding two optimization steps and ultimately achieving a substantial leap in inference speed.

Taken together, structured prompt optimization contributed approximately 90% of the overall accuracy improvement, serving as the core driving force for system performance enhancement; confidence distribution modeling provided critical value at the interpretability level while contributing approximately 10% of the accuracy gain; and lightweight inference achieved a favorable balance between accuracy and efficiency, enabling the system to genuinely possess the feasibility for large-scale deployment on ordinary consumer-grade devices. The three optimization strategies each have distinct emphases yet are functionally complementary, jointly achieving the optimal balance among recognition accuracy, interpretability, and deployment efficiency.

4.4 User Experience Evaluation

Unlike the comparative experiments conducted on the FashionStyle14 dataset, the purpose of the user study was to evaluate the practical usability of the deployed system rather

Table 1: Lightweight deployment performance before and after INT8 quantization.

Metric	Before Quantization (FP32)	After Quantization (INT8)	Improvement
Model Size	4.18 GB	1.04 GB	−75.12%
Runtime Memory Consumption	6.08 GB	2.29 GB	−62.34%
Average Inference Time	1126 ms	712 ms	−36.77%

than benchmark recognition accuracy. Therefore, real-world web images with diverse backgrounds were intentionally adopted to better simulate practical application scenarios.

While the preceding comparative and ablation experiments have provided quantitative validation of model algorithmic performance under standardized dataset conditions, such metrics only reflect performance under ideal testing conditions and cannot measure the actual user experience and comprehensive perception when the system is deployed and operated through the front-end interactive interface by ordinary users. To this end, this section conducts a user experience evaluation leveraging the system's front-end interactive platform to assess the system's engineering practicality and user acceptance.

A total of 30 volunteers were recruited for this evaluation, comprising 13 males and 17 females, aged between 18 and 35, including students, working professionals, and a small number of fashion industry practitioners. All volunteers possessed basic computer operation skills but had received no professional training in clothing style classification, thus representing the cognitive level and usage habits of ordinary users. Prior to the experiment, volunteers were only briefly informed of the basic system operation procedures, with no guidance provided regarding style determination criteria, ensuring that the test results could authentically reflect the system's performance under natural usage scenarios. Each volunteer independently completed recognition tasks for 10 randomly selected real-world clothing images sourced from publicly available internet resources, covering diverse shooting angles, lighting conditions, background complexity, and clothing categories, without any standardized preprocessing. After completing all recognition tasks, volunteers filled out a system satisfaction questionnaire, rating five dimensions—interface design aesthetics and layout rationality, system operation smoothness and convenience, image upload and recognition response speed, satisfaction with recognition results, and recognition result interpretability—on a 0 to 10 scale, and provided recommendation willingness and improvement suggestions.

All 30 questionnaires were validly returned. The statistical results (Table 2 & Figure 5) show that the system received positive evaluations across all dimensions, with overall performance meeting the expected design objectives.

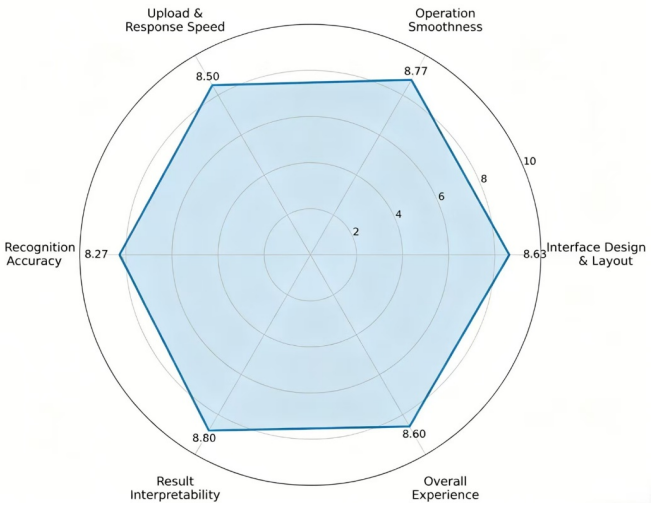


Figure 5: System evaluation chart

At the basic system experience level, the interface design aesthetics and layout rationality dimension achieved a mean score of 8.63, with 90% of users giving scores of 8 or above, indicating that the functionally partitioned layout adopted by the system aligns with ordinary users' visual cognitive habits, presenting a clean and intuitive interface. The system operation smoothness and convenience dimension achieved a mean score of 8.77, with 93.33% of users rating it as satisfactory or above, consistent with the behavioral observation during the experiment that users could independently complete the full operation workflow without additional guidance, validating the low learning barrier characteristic of the front-end interaction design. The image upload and recognition response speed dimension achieved a mean score of 8.50, with 86.67% of users considering the waiting time within an acceptable range, reflecting the practical effectiveness of the INT8 quantization and inference pipeline optimization scheme in improving system responsiveness.

Among the core functional dimensions, recognition result interpretability performed most prominently, achieving a mean score of 8.80, with 96.67% of users expressing recognition of the structured feature analysis and multi-style probability distribution output format, considering that this design effectively lowered the comprehension barrier for algorithm outputs and clearly presented the core rationale for style determination. The satisfaction with recognition results dimension achieved a mean score of 8.27, with 80% of users considering the recognition results to have high consistency with

Table 2: User satisfaction questionnaire results

Dimension	Mean Score (out of 10)	8-10 Score (%)	6-7 Score (%)	0-5 Score (%)
Interface design aesthetics and layout rationality	8.63	90	10	0
System operation smoothness and convenience	8.77	93.33	6.67	0
Image upload and recognition response speed	8.50	86.67	13.33	0
Satisfaction with recognition results	8.27	80	20	0
Recognition result interpretability	8.80	96.67	3.33	0
Overall system usage experience	8.60	90	10	0

their own subjective cognition, with only minor deviations observed on complex samples with blurred style boundaries and mixed elements—a phenomenon consistent with the general technical characteristics of fine-grained style classification tasks.

In terms of overall evaluation and recommendation willingness, the overall system usage experience achieved a mean score of 8.60, with 90% of users giving satisfactory or above ratings. Recommendation willingness statistics showed that 40% of users expressed strong willingness to recommend the system to others (Table 3), 50% expressed willingness to recommend, only 10% remained neutral, and no users expressed unwillingness to recommend. These results fully demonstrate that the complete system design—from algorithmic performance to front-end interaction—can satisfy the daily usage needs of ordinary users and possesses favorable engineering deployment value.

Table 3: User recommendation willingness statistics

Willingness	Number of Users	Percentage (%)
Very willing	12	40.00
Willing	15	50.00
Neutral	3	10.00
Unwilling	0	0.00
Very unwilling	0	0.00

Feedback from the open-ended questions further corroborates the above conclusions. The most frequently mentioned system advantages centered on three aspects: clean and clear interface, detailed and comprehensible recognition results, and relatively fast response speed. The main issues encountered by users during usage included insufficient recognition accuracy for certain niche styles and the need for improvement in batch recognition functionality. Suggested improvements mainly included adding mobile device adaptation, supporting real-time video recognition, and expanding style category coverage. The above user feedback is highly consistent with the technical strengths and existing limitations of the proposed system, providing clear user-oriented guidance for subsequent system functionality iteration and performance optimization.

4.5 System Application Demonstration

This section demonstrates the system's application effectiveness in practical scenarios through typical recognition cases. Figure 6 shows the front-end interactive interface, which adopts a functionally partitioned layout, clearly dividing the image upload area, recognition result display area, and history record viewing area. Users can submit images through conventional selection or drag-and-drop operations, with the interaction approach conforming to daily usage habits.

Figure 7 illustrates the single-image recognition result for a typical garment. Taking the suit shown in the figure as an example, the structured recognition result output by the system includes: category as "long trench coat style suit," silhouette as "H-line straight," fabric as "suiting fabric," pattern as "solid color," decoration as "double-breasted buttons, belt," style determination as "kireime-casual style (55%) + mode style (30%) + conservative (15%)," and color composition as "off-white (70%) + light gray (30%)." The system simultaneously generates a natural language recognition rationale: "The garment presents an H-line straight silhouette with clean and smooth tailoring, consistent with the silhouette characteristics of mode and kireime-casual styles; the crisp suiting fabric conveys a business casual sensibility; the off-white solid color with low-saturation palette suits kireime-casual and mode styles; the double-breasted button and belt design adds formality, pointing toward kireime-casual and conservative." This case intuitively demonstrates the system's characteristics of comprehensive recognition dimensions, structured results, and strong interpretability, validating the system's capability for precise capture of fine-grained garment features and accurate style determination.

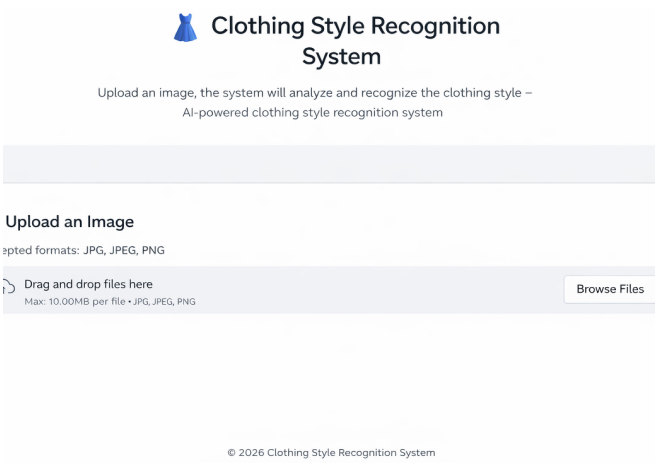


Figure 6: Front-end interactive interface

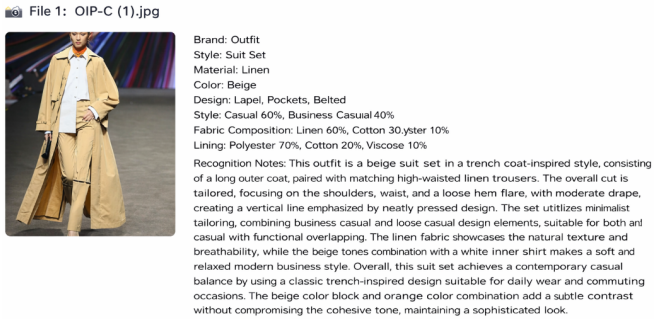


Figure 7: Example of single-image garment recognition result

5 Conclusions

This paper addresses the practical demands of fine-grained clothing style recognition, targeting core issues faced by general-purpose VLMs in the fashion domain—insufficient prompt guidance, limited fine-grained discrimination accuracy, and high deployment costs—by constructing a complete clothing style recognition system based on Qwen3-VL-Plus. The work encompasses the full research pipeline from algorithm optimization and system design to experimental validation.

At the methodological level, three core optimization strategies are proposed. First, a domain-specific four-level structured prompt system was constructed, integrating task instructions, style feature definitions, visual attribute guidance,

and output format constraints. A feature weighting mechanism was introduced to strengthen the semantic guidance of core style features, effectively resolving the problems of semantic vagueness in generic prompts and weak style discrimination capability. Second, a confidence distribution modeling mechanism was designed, employing Softmax normalization to achieve standardized mapping from cross-modal similarity scores to percentage probability distributions. This breakthrough overcomes the limitation of traditional single-label outputs that cannot represent composite style characteristics, achieving unification of academic evaluation and application-oriented output. Third, a lightweight inference scheme combining INT8 calibration quantization, prompt pre-encoding cache, and image preprocessing optimization was adopted, substantially reducing model storage footprint, runtime memory consumption, and per-image inference latency while ensuring no significant loss in recognition accuracy, enabling stable system operation on ordinary consumer-grade CPU environments. At the system level, a front-end interactive platform integrating image upload, style recognition, result visualization display, and history record management was built, forming a complete implementation solution from algorithm optimization to practical application.

Experimental results on the FashionStyle14 dataset demonstrate that the proposed system achieves a Top-1 accuracy of 69.21% and a macro-average F1-score of 66.38%, outperforming CLIP by 17.98 and 18.21 percentage points respectively, and surpassing the native Qwen3-VL-Plus by 20.86 and 20.92 percentage points respectively. All core metrics significantly exceed those of traditional pure vision models and other general-purpose VLM baselines. Ablation experiments verify that structured prompt optimization contributes approximately 90% of the overall accuracy improvement, while confidence distribution modeling provides critical value at the interpretability level and contributes approximately 10% of the accuracy gain. After INT8 quantization, inference speed improves by 36.77%, model size is reduced by 75.12%, and both inference efficiency and environmental adaptability are significantly enhanced. In user experience testing with 30 participants, the system achieves an overall satisfaction score of 8.60 out of 10, with 96.67% of users recognizing the interpretability of results and 90% expressing willingness to recommend the system, validating its usability and user acceptance in practical application scenarios.

This work still has limitations and areas for improvement. The current dataset scale and style category coverage are limited, and recognition accuracy for niche or emerging fashion styles needs enhancement. Recognition results may deviate when garments are heavily occluded or captured at extreme shooting angles. Future work will proceed along the following directions: expanding the dataset with more samples and style categories, incorporating emerging fashion styles, traditional ethnic costumes, and cross-regional outfit patterns to enhance recognition robustness for niche styles; deepening the integration of professional attributes such as garment pattern-making and craftsmanship with the prompt system to further improve fine-grained discrimination of boundary-blurred styles; extending static image recognition to dynamic visual scenarios such as short videos and live streaming for real-time style analysis; exploring more efficient on-device inference technologies to adapt the system to lightweight terminals such as mobile applications, further lowering hardware deployment barriers; and, in combination with the digital preservation needs of intangible cultural heritage costumes, constructing a dedicated style feature library for traditional ethnic clothing to apply this system to the digital classification and archiving of traditional garments, providing technical support for the preservation of clothing cultural heritage.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or non-profit sectors.

Author Contributions

Conceptualization, R.M. and X.L.; methodology, R.M. and T.C.; software, R.M.; validation, R.M.; formal analysis, R.M.; investigation, R.M.; resources, X.L.; data curation, R.M.; writing—original draft preparation, R.M.; writing—review and editing, R.M., X.L., T.C. and X.D.; visualization, R.M.; supervision, T.C., X.L. and X.D.; project administration, X.D.; funding acquisition, X.D. All authors have read and agreed to the published version of the manuscript.

Conflict of Interest

All the authors declare that they have no conflict of interest.

Data Avaliable

The source data supporting the findings of this study have not been uploaded to, and will not be made publicly available through, any open-source platform or public data repository. The data may be obtained from the corresponding author upon reasonable request, subject to applicable restrictions.

References

[1] Casciani, D., Chkanikova, O., Pal, R.: Exploring the nature of digital transformation in the fashion industry: Opportunities for supply chains, business models, and sustainability-oriented innovations. Sustainability: Science, Practice and Policy **18**(1), 773–795 (2022). https://doi.org/10.1080/15487733.2022.2125640

[2] Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1096–1104 (2016). https://doi.org/10.1109/CVPR.2016.124

[3] Wang, L.C., Zeng, X.Y., Koehl, L., Chen, Y.: Intelligent fashion recommender system: Fuzzy logic in personalized garment design. IEEE Transactions on Human-Machine Systems, **45**(1), 95–109 (2015). https://doi.org/10.1109/THMS.2014.2364398

[4] Yue, X., Zhang, C., Fujita, H., Liu, J., Li, Y.: Clothing fashion style recognition with design issue graph [J]. Applied Intelligence, **51**(6), 3548–3560 (2021). https://doi.org/10.1007/s10489-020-01950-7

[5] Ortega Ballesteros, G.: Personalized style-specific outfit generation from garment wardrobes. Master's Thesis, Universitat Politècnica de Catalunya, Spain (2025)

[6] Paulauskaite-Taraseviciene, A., Noreika, E., Purtokas, R., Lagzdinyte-Budnike, I., Daniulaitis, V., Salickaite-Zukauskiene, R.: An intelligent solution for automatic garment measurement using image recognition technologies. Applied Sciences, **12**(9), 4470 (2022). https://doi.org/10.3390/app12094470

[7] Hafner, M., Katsantoni, M., Köster, T., Marks, J., Mukherjee, J., Staiger, D., Ule, J., Zavolan, M.: CLIP and complementary methods. Nature Reviews Methods Primers, **1**(1), 20 (2021). https://doi.org/10.1038/s43586-021-00018-1

[8] Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., *et al.:* Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025). https://doi.org/10.48550/arXiv.2511.21631

[9] Shen, S., Li, L.H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.W., Yao, Z., Keutzer, K.: How much can CLIP benefit vision-and-language tasks? arXiv preprint arXiv:2107.06383 (2021). https://doi.org/10.48550/arXiv.2107.06383

[10] Ghosh, A., Acharya, A., Saha, S., Jain, V., Chadha, A.: Exploring the frontier of vision-language models: A survey of current methodologies and future directions. arXiv preprint arXiv:2404.07214 (2024). https://doi.org/10.48550/arXiv.2404.07214

[11] Wang, C., Wang, Z., Xu, X., Tang, Y., Zhou, J., Lu, J.: Q-VLM: Post-training Quantization for Large Vision-Language Models. In Proceedings of the 38th International Conference on Neural Information Processing Systems, pp.114553–114573 (2024). https://doi.org/10.52202/079017-3638

[12] Lee, J., Sul, I.H.: Construction of garment pattern shape information system using image analysis and shape recognition techniques. International Journal of Clothing Science and Technology, **28**(4), 543–555 (2016). https://doi.org/10.1108/IJCST-10-2015-0114

[13] Cheng, C., Liu, D.S.-M.: An Intelligent clothing search system based on fashion style. In 2008 International Conference on Machine Learning and Cybernetics, pp. 1592–1597 (2008). https://doi.org/10.1109/ICMLC.2008.4620660

[14] Hidayati, S.C., You, C.W., Cheng, W.H., Hua, K.L.: Learning and recognition of clothing genres from full-body images. IEEE Transactions on Cybernetics, **48**(5), 1647–1659 (2017). https://doi.org/10.1109/TCYB.2017.2712634

[15] Wang, Z., Wang, X.Y., Xu, J.P.: Personalized clothing recommendation via multi-feature fusion. Wool Textile Journal, **53**(12), 101–108 (2025). https://doi.org/10.19333/j.mfkj.20240707401

[16] Hua, Y.Y., Zhang, D.C., Ge, S.M.: Research progress on interpretability of deep learning models [J]. Journal of Information Security, **5**(3), 1–13 (2020). https://doi.org/10.19363/J.cnki.cn10-1380/tn.2020.05.01

[17] Ji, S.L., Li, J.F., Du, T.Y., Li, B.: Survey on Techniques, Applications and Security of Machine Learning Interpretability. Journal of Computer Research and Development, **56**(10), 2071–2096 (2019). https://doi.org/10.7544/issn1000-1239.2019.20190540

[18] Kim, E., Kim, S., Jung, C., Hahm, Y., Cho, S.: Fashion image retrieval with vision-language model guided fine-grained textual attributes and cross-domain contrastive optimization. IEEE Access, **13**, 192403–192415 (2025). https://doi.org/10.1109/ACCESS.2025.3630632

[19] Wang, M., Mahjoubfar, A., Joshi, A.: FashionVQA: A domain-specific visual question answering system. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 3514–3519 (2023). https://doi.org/10.1109/CVPRW59228.2023.00356

[20] Liu, V., Chilton, L.B.: Design guidelines for prompt engineering text-to-image generative models. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, pp. 1–23 (2022). https://doi.org/10.1145/3491102.3501825

[21] Cao, Y., Zhang, J., Frittoli, L., Cheng, Y., Shen, W., Boracchi, G.: AdaCLIP: Adapting CLIP with hybrid learnable prompts for zero-shot anomaly detection. In European Conference on Computer Vision, pp. 55–72 (2024). https://doi.org/10.1007/978-3-031-72761-0_4

[22] Zhao, K., Huang, S., Pan, P., Li, Y., Zhang, Y., Gu, Z., Xu, Y.: Distribution adaptive INT8 quantization for training CNNs. In Proceedings of the AAAI Conference on Artificial Intelligence. Virtual Event, pp. 3483–3491 (2021). https://doi.org/10.1609/aaai.v35i4.16462

[23] Takagi, M., Simo-Serra, E., Iizuka S., Ishikawa, H.: What Makes a Style: Experimental Analysis of Fashion Prediction. In 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 2247–2253 (2017). https://doi.org/10.1109/ICCVW.2017.263

[24] Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In International Conference on Learning Representations (ICLR), pp. 1–14 (2015)

[25] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018). https://doi.org/10.1109/CVPR.2018.00474

Appendix A Complete Definitions of All 14 Styles

You are a clothing style analysis expert with a professional background in fashion design. Please conduct a comprehensive and multi-dimensional analysis of the input clothing image.

This recognition task covers 14 mainstream clothing styles. Please strictly match the styles according to the following professional characteristics:

1. Conservative Style: Formal tailoring, neutral tones, a fitted or slightly relaxed silhouette, and restrained functional details such as belts, pockets, or concealed plackets; typically associated with workplace and office settings.
2. Dressy Style: Elegant and refined tailoring, a fitted or waist-defined silhouette, smooth or lustrous fabrics, sophisticated colors, and polished decorative details suitable for formal occasions, ceremonies, or social events.
3. Ethnic Style: Traditional or regional silhouettes, culturally distinctive motifs, embroidery, weaving, tassels, or other handcrafted details, often combined with natural or heritage fabrics and colors that convey a clear regional identity.
4. Fairy Style: Lightweight and flowing fabrics such as chiffon or tulle, pale or translucent colors, an A-line or softly draped silhouette, and romantic details such as lace, ruffles, gathers, or delicate floral decoration.
5. Feminine Style: Soft, graceful, or waist-emphasizing silhouettes, delicate fabrics, harmonious or pastel colors, and refined details such as pleats, lace, bows, or subtle embellishment, creating an elegant and womanly impression.
6. Gal Style: A body-conscious, cropped, or short silhouette, bold or high-saturation colors, conspicuous prints and fashionable materials, and eye-catching details that convey a glamorous, youthful, and trend-driven street or nightlife aesthetic.
7. Girlish Style: Youthful and sweet styling, soft or bright colors, fitted or flared silhouettes, and playful details such as bows, frills, lace, puff sleeves, or small prints, creating a cute and lively visual impression.
8. Kireime-Casual Style: A polished balance between formal and casual elements, clean and moderately structured tailoring, an H-line or gently fitted silhouette, neutral or light colors, and restrained practical details suitable for commuting and refined everyday wear.
9. Lolita Style: A bell-shaped or voluminous skirt silhouette supported by petticoat-like structure, modest proportions, and elaborate Victorian- or Rococo-inspired details such as lace, ruffles, ribbons, bows, decorative buttons, and ornate prints.
10. Mode Style: Clean tailoring, smooth lines, a loose or fitted silhouette, low-saturation solid colors, no superfluous decoration, and an emphasis on the texture and structural quality of the fabric itself.
11. Natural Style: Relaxed or loose silhouettes, soft cotton, linen, knit, or other natural-texture fabrics, earthy or muted colors, and simple, understated details that create a comfortable and unforced everyday appearance.
12. Retro Style: Classic period-inspired silhouettes, nostalgic color schemes, vintage prints, and characteristic details such as traditional collars, buttons, pleats, or distressed textures, producing a clear historical or old-fashioned aesthetic.
13. Rock Style: Leather, denim, or other firm-texture fabrics, black or dark colors, fitted, asymmetric, or distressed silhouettes, and rebellious details such as metal hardware, studs, zippers, chains, or ripped elements.
14. Street Style: Loose, oversized, or O-shaped silhouettes, casual fabrics, graphic prints, logos, color blocking, and details associated with hoodies, cargo garments, denim, or sportswear, creating a free, expressive, and urban aesthetic.

Please focus on the following four core features of the clothing itself: silhouette, fabric, color, and decorative details. Completely ignore irrelevant information such as background, models, and accessories.

The output must strictly follow the field order and format below without omitting any field:

- Category: [Specific clothing category]
- Silhouette: [Core silhouette feature, such as X-shape, A-shape, O-shape, H-shape, etc.]
- Fabric: [Main fabric material, such as linen, combed cotton, suiting fabric, knitwear, etc.]
- Pattern: [Pattern type, such as solid color, print, stripes, plaid, etc.]
- Decoration: [Main decorative details, such as double-breasted buttons, large pockets, belts, lace, etc.]
- Style: [All 14 style names and their confidence percentages, arranged from highest to lowest confidence, with the percentages summing to 100%]
- Color: [Primary color percentage%, secondary color percentage%, arranged from highest to lowest proportion, with all color percentages summing to 100%]
- Recognition Basis: [Provide a detailed explanation of the style judgment based on the four core features: silhouette, fabric, color, and decoration, using professional and logically clear language]