Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 20 No. 21 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijim.v20i21.63711

PAPER

Product Artistic Style Transfer via Integrated Mobile Visual Search and Generative Models

Chen Kang (✉), Zheng Qi

Changzhou Vocational Institute of Mechatronic Technology, Changzhou, China

kc1644@czimt.edu.cn

ABSTRACT

Mobile product artistic style transfer is increasingly required for e-commerce visualization, augmented reality try-on, and personalized design. Existing approaches typically rely on predefined styles or manually specified reference images, while mobile visual search and generative modeling are generally implemented as separate processes. Moreover, conventional diffusion models are difficult to deploy under the sub-second interaction constraints imposed by mobile devices. To address these limitations, a search-guided style-prior conditioning mechanism was proposed, in which mobile visual search was reformulated as a prior-supply module for generative tasks. Retrieval results were encoded into a global style-tone vector and local style-texture features, which were aggregated through temperature-based soft weighting to construct dual-granularity style priors. Differentiability was preserved during soft weighting within the candidate set, allowing the generative loss to be back-propagated to the candidate re-ranking and style encoding modules, thereby aligning the retrieval representation with the generative conditioning space. During generations, a lightweight diffusion network was employed, in which depthwise separable convolutions and inverted residual bottlenecks were incorporated to reduce computational complexity. The dual-granularity style priors were injected through global affine modulation and local attention mechanisms. During training, denoising, style consistency, content fidelity, and retrieval ranking losses were jointly optimized; for deployment, consistency distillation and 8-bit integer quantization were adopted. Experimental results demonstrated automatic style-prior recommendation and sub-second style transfer were achieved on mobile devices, providing a feasible pathway toward an integrated search-and-generation architecture.

KEYWORDS

mobile visual search, generative models, product artistic style transfer, lightweight diffusion models, style-prior conditioning, mobile inference

1 INTRODUCTION

Rapid on-device preview of artistic product effects has become desired by consumers and designers as a result of the rapid development of mobile e-commerce,

Kang, C., Qi, Z. (2026). Product Artistic Style Transfer via Integrated Mobile Visual Search and Generative Models. *International Journal of Interactive Mobile Technologies (iJIM)*, 20(21), pp. 173–188. https://doi.org/10.3991/ijim.v20i21.63711

Article submitted 2026-07-30. Revision uploaded 2026-09-26. Final acceptance 2026-10-10.

© 2026 by the authors of this article. Published under CC-BY.

augmented reality try-on, personalized product display, and mobile design assistance [1]. Mobile augmented reality has also increasingly supported interactive visual experiences by integrating virtual content with real-world environments through mobile devices [2]. In product artistic style transfer, a specified artistic style is required to be transferred while product content structure is preserved, such that generated results retain product recognizability and target artistic expressiveness [3]. Mobile visual search has been widely used for image-based retrieval of products and similar designs [4]. Lightweight on-device retrieval has been enabled by methods such as deep learning hashing [5], and retrieval can be completed within seconds on ordinary mobile phones [6]. If retrieval results are transformed into style priors compatible with product content, the need for one-by-one selection of reference images by users can be reduced, and style transfer can be made more natural and more immediate [7]. With improvements in mobile-device computational power and advances in generative models, on-device style transfer has been rendered feasible [8]. Meanwhile, lightweight deep neural networks such as MobileNet, ShuffleNet, and EfficientNet have provided practical architectural foundations for visual processing on mobile platforms, although computational cost remains an important deployment constraint [9]. Mobile visual search is therefore extended in this study from an information retrieval tool into a prior-supply module for generative tasks; by this extension, integrated mobile search and generation is promoted, and a new technical pathway for product artistic style transfer is provided.

However, significant gaps are still present in existing research at three levels: style source, retrieval-generation linkage, and mobile deployment. With respect to style source, style specification in style transfer methods is largely dependent on predefined style sets or reference images manually provided by users [10]. Although stylization quality has been continuously improved by Gram matrix-based neural style transfer and by subsequent methods based on convolutional neural networks, attention modules, and Transformers [11], product visual content has not been used to automatically recommend style priors compatible with the product, and adaptation to real-time mobile interaction and stylistic diversity remains difficult. With respect to retrieval-generation linkage, mobile visual search research has long been focused on retrieval stages such as keypoint detection, feature extraction, and index matching [12]. Retrieval results are directly presented to users, retrieval ranking objectives are inconsistent with generation quality objectives, and style priors are not aligned with the generative conditioning space. With respect to mobile deployment, although excellent generation quality has been achieved by diffusion models [13], and content-style balance has been improved by diffusion-based methods through controlled injection of semantic and spatial information [14], their multistep denoising, large parameter counts, and memory footprint are in pronounced conflict with mobile constraints on computation, storage, and energy consumption. Recent work on adaptive generative visual systems has similarly shown that generation latency and edge deployment remain important considerations when generative content is incorporated into interactive visual environments [15]. Studies of edge-based intelligent systems further indicate that local processing can substantially reduce latency and energy consumption, but computational resources at edge devices remain limited [16]. Attempts have therefore been made to construct lightweight neural style transfer models using depthwise separable convolutions and inverted residual bottlenecks, enabling rapid on-device inference [17]. Related lightweight real-time vision frameworks also demonstrate that low computational complexity and efficient inference are essential for deployment on resource-constrained platforms [18]. However, style specification in existing mobile style transfer models

remains dominated by fixed styles or manual selection, retrieval and generation are not jointly optimized, and a mobile-oriented joint distillation and quantization pipeline is lacking.

To address the aforementioned gaps, a search-guided style-prior conditioning mechanism is proposed. In this mechanism, candidate reference images obtained through mobile visual search are encoded into a global style-tone vector and local style-texture features. These are aggregated into dual-granularity priors through candidate re-ranking and temperature-based soft weighting, such that style sourcing is shifted from one-by-one user specification to automatic recommendation by search. In the generation stage, a lightweight diffusion network is adopted; computational cost is reduced by depthwise separable convolutions and inverted residual bottlenecks, and dual-granularity priors are injected through global affine modulation and local attention. In the training stage, denoising, style consistency, content fidelity, and retrieval ranking losses are jointly optimized. In the deployment stage, consistency distillation and 8-bit integer quantization are adopted, enabling sub-second on-device inference. The principal contributions are threefold: a search-guided style-prior conditioning mechanism is proposed; a mobile-oriented lightweight diffusion style transfer network is designed; and a pipeline for joint optimization of search and generation and for mobile deployment is constructed.

2 METHODOLOGY

2.1 Problem definition and overall framework

A product content image captured or uploaded by a user through a mobile device is denoted as I_c, and the style retrieval library is denoted as:

$$D = \{(I_i^s, e_i)\}_{i=1}^{N} \tag{1}$$

where, N is the retrieval library size, I_i^s is the i-th artistic product reference image, $e_i \in \mathrm{R}^d$ is its global style vector used for approximate nearest neighbor retrieval, and d is the vector dimension. In the million-scale index, only compressed global vectors are stored; local style-texture features are not written into the inverted file with product quantization (IVF-PQ) primary index, but are instead computed on the candidate set or read from a candidate cache. A stylized product image I_{gen} is to be generated such that the low-frequency layout and semantic structure of I_c are preserved in content structure, and the dual-granularity style prior formed by the retrieved reference set is approached in style. No unique correct artistic style is assumed to exist for the same product; the retrieval module is used to recommend, from the candidate library, a style reference set compatible with the query content. Model size, peak memory, single-inference latency, and energy consumption are included among mobile constraints; therefore, automatic recommendation of style priors and conditioned generation are realized within mobile computational constraints, without reliance on server-side ultra-large generative models.

The overall pipeline is composed of mobile retrieval, candidate local encoding, soft-weighted aggregation, adaptive injection, lightweight diffusion denoising, and distillation-quantization deployment. The overall architecture is shown in Figure 1. In this architecture, dual-granularity style priors are extracted on the left through mobile visual search and are then adaptively injected into the lightweight diffusion

denoising backbone in the center, while sub-second on-device inference is ensured by the joint optimization and distillation-quantization module on the right.

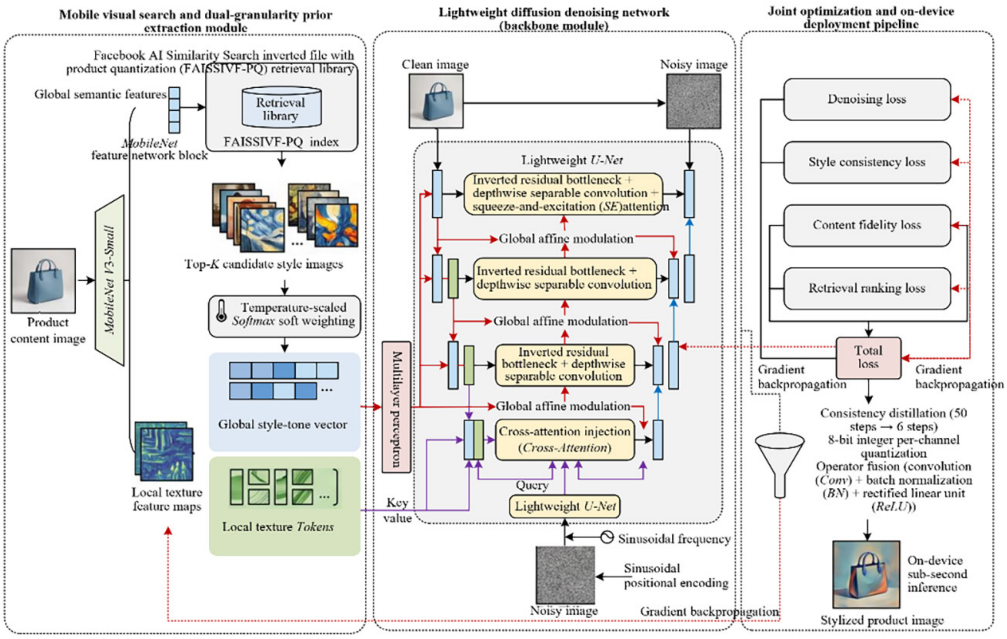


Fig. 1. Overall architecture of the search-guided style-prior conditioning mechanism

After an image is uploaded by a user, query features are extracted by a mobile lightweight encoder, and approximate nearest neighbor search is performed in a local or edge-side vector library. A candidate set is first obtained according to global vectors, candidate local texture features are then computed and re-ranked, and Top-K style reference images are selected therefrom. The reference images are not directly used as input to the generation network; instead, they are encoded into two structured priors, namely a global style-tone vector c_g and a local style-texture representation c_l. Color distribution, contrast, and overall style tendency are described by c_g, whereas local brushstrokes, materials, and texture patterns are described by c_l. During denoising, noisy image x_t, time step t, and c_g and c_l are simultaneously accepted as conditions by the lightweight diffusion U-Net. The statistical distribution of denoising features is adjusted by the global prior through affine modulation, and texture information is injected by the local prior through attention so that fine-grained control over the retrieved style is achieved while the product content structure is preserved during generation. During training, approximate nearest neighbor candidate generation, as a discrete retrieval step, is not included in backpropagation, whereas the re-ranking, soft weighting, and style encoding modules within the candidate set are jointly optimized with the denoising network. During deployment, the model is compressed using consistency distillation and 8-bit integer quantization. During retrieval, only low-dimensional query features are uploaded or computed locally, and frequent uploading of original product images is avoided; candidate local features are computed only for a small number of candidates. Style priors can be cached and reused, and repeated encoding is not required when the same retrieval result is continuously previewed by a user. If retrieval times out or fails, switching is performed to a default style prior or the most recent cached prior so that interaction is not interrupted.

2.2 Search-guided dual-granularity style-prior extraction

MobileNetV3-Small is adopted as the mobile lightweight encoder E_v so that a balance is achieved between receptive field and computational cost. After an input image is preprocessed to 256×256, global semantic features and local texture features are simultaneously output by the encoder:

$$f_g = \text{Pool}(E_v(I_c)), \quad F_l = E_v^l(I_c) \tag{2}$$

where, $f_g \in R^{512}$ is the global feature, and $F_l \in R^{28\times28\times48}$ is the local texture feature map. Product category, overall color, and semantic layout are captured by the global feature, whereas edges, brushstrokes, and material responses are preserved by the local features. Reference images in the retrieval library are pre-encoded into global style vectors e_i and stored in a Facebook AI Similarity Search (FAISS) IVF-PQ index. A candidate set is first obtained by retrieval according to global cosine similarity; subsequently, local features F_i are computed only for the candidate reference images, and re-ranking is performed using fused global-local similarity:

$$\text{sim}_i = \cos(f_g, e_i) + \lambda\cos(\text{Pool}(F_l), \text{Pool}(F_i)) \tag{3}$$

where, $\lambda = 0.4$ is used to balance semantic similarity and texture similarity. After re-ranking, Top-K candidates are selected, and soft weighting coefficients are generated using temperature-scaled softmax:

$$\alpha_i = \frac{\exp(\text{sim}_i/\tau)}{\sum_{j\in\text{TopK}} \exp(\text{sim}_j/\tau)} \tag{4}$$

In experiments, $K = 5$ and $\tau = 0.07$ were taken. Since no pixel-level spatial correspondence is present among different reference images, local features F_i are first converted, by adaptive pooling and a linear projection $P_l()$, into local texture tokens that are independent of absolute position and are then aggregated. The dual-granularity style priors are defined as:

$$c_g = \sum_{i\in\text{TopK}} \alpha_i e_i, \quad c_l = \sum_{i\in\text{TopK}} \alpha_i P_l(F_i) \tag{5}$$

where, color distribution, contrast, and overall style tendency are described by c_g, whereas local brushstrokes, materials, and texture patterns are described by c_l. Approximate nearest neighbor candidate generation and Top-K truncation are discrete operations; no gradients are computed with respect to them themselves. After the candidate set is determined, sim_i, α_i, the local projection, and the style encoding modules are differentiable. Generation losses are back-propagated through c_g and c_l to the candidate re-ranking and style-encoding modules so that style representations within the candidate set are constrained by both retrieval objectives and generation quality. During training, the retrieval vector index is updated on an epoch basis, so that index representations remain consistent with the current encoder. Under the premise that approximate nearest neighbor search itself is not treated as a differentiable operator, joint alignment between candidate retrieval representations and the generative conditioning space is achieved.

2.3 Lightweight diffusion generation and joint optimization

A pixel-space lightweight diffusion U-Net is adopted as the generation backbone, and an independent variational autoencoder encoder or decoder is not introduced. The clean image is denoted by x_0, and the diffusion forward process is given by:

$$q(x_t \mid x_0) = \mathrm{N}\left(x_t; \sqrt{\bar{\alpha}_t} x_0, (1-\bar{\alpha}_t) I\right) \tag{6}$$

where, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. The denoising network ε_θ is conditioned on the noisy image x_t, time step t, and dual-granularity style priors c_g and c_l:

$$\varepsilon_\theta = \varepsilon_\theta(x_t, t, c_g, c_l) \tag{7}$$

To adapt to mobile computational constraints, a four-stage downsampling and upsampling structure is employed by the U-Net, the channel numbers are set to 32, 64, 128, and 256, and two residual blocks are contained in each stage. Standard convolutions are replaced by depthwise separable convolutions, and the computational cost is approximately $1/C_{\text{out}} + 1/k^2$ times that of standard convolution, where C_{out} is the number of output channels and k is the convolution kernel size. An inverted residual bottleneck structure is adopted inside the residual blocks, in which dimensionality is first expanded, then depthwise convolution is performed, and dimensionality is finally reduced; the expansion ratio is set to 4, and lightweight squeeze-and-excitation attention is incorporated to enhance channel selection capability. The time step is mapped to a 128-dimensional embedding by sinusoidal positional encoding and a two-layer multilayer perceptron, and the embedding is injected layer by layer. By this design, the parameter count and floating-point operations are reduced while denoising expressiveness is maintained so that the diffusion model can be operated under mobile computational constraints.

Condition injection is performed through a joint scheme of global modulation and local attention. For the global style prior c_g, affine parameters are generated by a multilayer perceptron:

$$\gamma, \beta = \mathrm{MLP}(c_g), \quad \mathrm{h} = \gamma \odot \mathrm{GN}(\mathrm{h}) + \beta \tag{8}$$

where, GN() denotes group normalization and $\odot$ denotes channel-wise multiplication. This operation is applied after the normalization layers in residual blocks, such that the mean and variance of denoising features are adaptively adjusted by the global style tone. For the local style prior c_l, spatial positions of denoising features are used as queries, and local texture tokens are used as keys and values:

$$Q = W_q \mathrm{h}, \quad K = W_k c_l, \quad V = W_v c_l \tag{9}$$

$$A = \mathrm{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_a}}\right), \quad \mathrm{h}' = \mathrm{h} + W_o(AV) \tag{10}$$

where, d_a is the attention channel dimension. Absolute spatial alignment among different reference images is not required by this design; therefore, overall style direction is controlled by c_g, and local style information is injected by c_l into different

spatial positions in the form of texture tokens. During training, 10% condition dropout is applied to the conditions to support classifier-free guidance at inference, and the guidance scale is set to 1.5. Compared with direct concatenation of style images and noisy images along the channel dimension, significant expansion of backbone channels is avoided by this joint injection scheme, making it more suitable for mobile devices.

A two-stage strategy is adopted for training. In the first stage, the basic denoising capability of the diffusion model is pre-trained on a large-scale style transfer dataset so that the network is endowed with the ability to recover image structure from noise. In the second stage, joint search-generation fine-tuning is performed, in which the candidate re-ranking and soft weighting modules are connected to the denoising network through a differentiable interface. The total loss is defined as:

$$L_{\text{total}} = L_{\text{diff}} + \lambda_1 L_{\text{style}} + \lambda_2 L_{\text{content}} + \lambda_3 L_{\text{ret}} \tag{11}$$

The diffusion denoising loss is given by:

$$L_{\text{diff}} = E_{x_0,\varepsilon,t}\left[\left\|\varepsilon - \varepsilon_\theta(x_t, t, c_g, c_l)\right\|_2^2\right] \tag{12}$$

The style consistency loss constrains the generated image to approach the retrieved style prior at both global and local levels. For local texture, a Gram matrix is first computed for each reference image at the same style encoding layer and is then aggregated according to α_i so that dimensional inconsistency across different layers is avoided. The following is defined:

$$S_m = \sum_{i \in \text{TopK}} \alpha_i G(E_s^m(I_i^s)) \tag{13}$$

The style consistency loss is therefore defined as:

$$L_{\text{style}} = \left\|E_s^g(I_{\text{gen}}) - c_g\right\|_2^2 + \lambda_g \sum_{m \in M} \left\|G(E_s^m(I_{\text{gen}})) - S_m\right\|_F^2 \tag{14}$$

where, E_s^g is the global branch of the pre-trained style encoder, E_s^m is the m-th local feature layer, $G()$ is the Gram matrix, and $M = \{\text{relu1_2, relu2_2, relu3_4}\}$. The content fidelity loss is implemented using frequency-domain progressive regularization:

$$L_{\text{content}} = \sum_t \omega_t \left\|\text{LPF}(\hat{x}_0(x_t, t)) - \text{LPF}(I_c)\right\|_1 \tag{15}$$

where, LPF() is a 5×5 Gaussian low-pass filter, $\omega_t = 1 - t/T$ indicates that the content structure constraint becomes stronger as the clean image is approached, and $\hat{x}_0(x_t, t)$ is estimated from the current noise prediction. A retrieval ranking loss is used to stabilize the candidate re-ranking module and is formulated in a cross-entropy form over Top-K candidates. During joint training, gradients of L_{style} are back-propagated through c_g, c_l, and α_i to the candidate re-ranking and style encoding modules. In experiments, λ_1 = 1.0, λ_2 = 0.5, λ_3 = 0.1 and λ_g = 0.8 were taken, and the joint fine-tuning learning rate was set to 1×10^{-5}.

2.4 Mobile deployment and complexity analysis

To achieve sub-second inference on real mobile devices, three deployment optimizations are performed after joint training: consistency distillation, 8-bit integer quantization, and operator fusion. Consistency distillation is used to compress the number of diffusion steps. Let $f_\theta(x_t, t, c_g, c_l)$ denote the clean sample predicted by the student network from a noisy state, and the one-step deterministic update of the teacher model from t to $t-\Delta$ is denoted as:

$$\hat{x}_{t-\Delta} = \Phi_{\theta^-}(x_t, t \to t-\Delta; c_g, c_l) \tag{16}$$

The consistency distillation loss is defined as:

$$L_{\text{cd}} = E_{t,\Delta}\left[\left\| f_\theta(x_t, t, c_g, c_l) - \text{sg}\left(f_{\theta^-}(\hat{x}_{t-\Delta}, t-\Delta, c_g, c_l)\right)\right\|_2^2\right] \tag{17}$$

where, θ^- denotes the teacher network parameters and sg() denotes stop-gradient. In experiments, a standard 50-step teacher model is distilled into a 6-step student model. The denoising network is then subjected to 8-bit integer quantization-aware training, in which per-channel symmetric quantization is adopted, and the quantization calibration set is derived from a subset of the product artistic style training set. In operator fusion, convolution, batch normalization, and rectified linear unit operations are fused into a single operator, and depthwise separable convolutions and pointwise convolutions are mapped to efficient implementations of CoreML or android neural networks application programming interface (NNAPI), respectively. At inference, query encoding, approximate nearest neighbor search, candidate local encoding, and dual-granularity prior aggregation are first performed, followed by conditional diffusion denoising. Style priors can be cached and reused so that repeated encoding is avoided. If retrieval fails or times out, the system is degraded to a default style prior. The goal is to achieve a sub-second generation of 512×512 stylized images on mainstream smartphones, with model size controlled within 20 MB.

The computational cost of a depthwise separable convolution is taken as $1/C_{\text{out}} + 1/k^2$ of that of a standard convolution, where C_{out} is the number of output channels and k is the convolution kernel size. Because pixel-space diffusion is adopted and no independent image decoder is included, the total mobile latency is modeled as:

$$T_{\text{total}} = T_{\text{search}} + T_{\text{local}} + T_{\text{denoise}} + T_{\text{post}} \tag{18}$$

where, T_{search} is the latency of query encoding and global retrieval, T_{local} is the latency of candidate local feature computation, re-ranking, and prior aggregation, T_{denoise} is the diffusion denoising latency, and T_{post} is the post-processing latency of output cropping, format conversion, and related operations. Through distillation, the number of denoising steps is reduced from T to T', and the theoretical speedup of the denoising portion is T/T'. After quantization, weight storage is reduced from 16-bit floating-point to 8-bit integer, and the model weight footprint is approximately halved.

3 EXPERIMENTAL RESULTS AND ANALYSIS

3.1 Experimental setup and datasets

A public style transfer dataset and a self-constructed product artistic style dataset were adopted. Content images were sourced from e-commerce product images,

covering five categories: shoes, bags, furniture, electronics, and cosmetics. Style reference images used for training were obtained from WikiArt, artistically stylized product images, and public style transfer datasets. An additional 1,000,000 artistically stylized product reference images, covering 20 artistic styles, were included in the retrieval library and were used to construct a large-scale search candidate space. The training, validation, and test sets of content images were split at the image level with mutual exclusivity based on original product instances so that leakage of the same instance across sets was avoided. Cross-category and cross-style capabilities were evaluated under separate leave-one-out test settings. Mobile testing was performed on an iPhone 15 Pro and a Snapdragon 8 Gen 3 Android flagship device, which were deployed with Core ML and NNAPI, respectively. Evaluation metrics included Fréchet inception distance, learned perceptual image patch similarity, art Fréchet inception distance, style loss, content loss, and Contrastive Language–Image Pre-training (CLIP) score, as well as model size, parameter count, floating-point operations per single forward pass, latency, peak memory, energy consumption, and temperature.

The self-constructed product artistic style dataset was used, in which the total number of content images was 12,000, the total number of style reference images used for training was 8,000, and the retrieval library size was 1,000,000, covering five product categories and 20 artistic styles. The regular training, validation, and test sets were divided as 8,000, 2,000, and 2,000 images, respectively, and mutual exclusivity was maintained at the original product-instance level. Cross-category and cross-style generalization results were obtained under independent leave-one-out settings. A FAISS IVF-PQ index was adopted for the retrieval library and only compressed 512-dimensional global style vectors were stored in the primary index. Candidate local feature maps had a size of 28×28×48 and were computed or read only on the candidate set; therefore, they were not fully resident in mobile memory at the million scale. The retrieval Top-K was set to 5, and the temperature τ was set to 0.07 for style aggregation within the candidate set. The million-scale retrieval library was used to construct a large-scale candidate space, and the 8,000 style reference images used for training were used for generative model training and style supervision.

3.2 Quantitative comparison with existing methods

Table 1. Quantitative comparison with existing methods

Method	Fréchet Inception Distance ↓	Learned Perceptual Image Patch Similarity ↓	Art Fréchet Inception Distance ↓	Contrastive Language–Image Pre-Training Score ↑
Adaptive Instance Normalization (AdaIN)	42.7	0.412	38.5	0.71
Whitening and Coloring Transform (WCT)	39.4	0.389	35.7	0.73
Content-Adaptive Style Transfer (CAST)	33.2	0.352	31.4	0.75
StyTr^2	28.4	0.321	26.8	0.76
Inversion-based Style Transfer (InST)	22.1	0.276	20.4	0.79

(Continued)

Table 1. Quantitative comparison with existing methods *(Continued)*

Method	Fréchet Inception Distance ↓	Learned Perceptual Image Patch Similarity ↓	Art Fréchet Inception Distance ↓	Contrastive Language–Image Pre-Training Score ↑
StyleDiffusion	19.8	0.254	18.7	0.81
DiffStyler	18.9	0.241	17.9	0.82
MobileStyler	24.6	0.298	23.1	0.77
Proposed method	16.4	0.228	15.2	0.85
Method	**Style Loss ↓**	**Content Loss ↓**	**Parameters**	**Latency**
Adaptive Instance Normalization (AdaIN)	2.84	1.92	12 M	120 ms
Whitening and Coloring Transform (WCT)	2.61	1.85	15 M	180 ms
CAST	2.33	1.71	22 M	260 ms
StyTr^2	2.05	1.58	65 M	950 ms
InST	1.72	1.41	180 M	3.2 s
StyleDiffusion	1.54	1.32	210 M	4.1 s
DiffStyler	1.47	1.29	195 M	3.8 s
MobileStyler	1.88	1.49	18 M	620 ms
Proposed method	1.31	1.18	19.5 M	720 ms

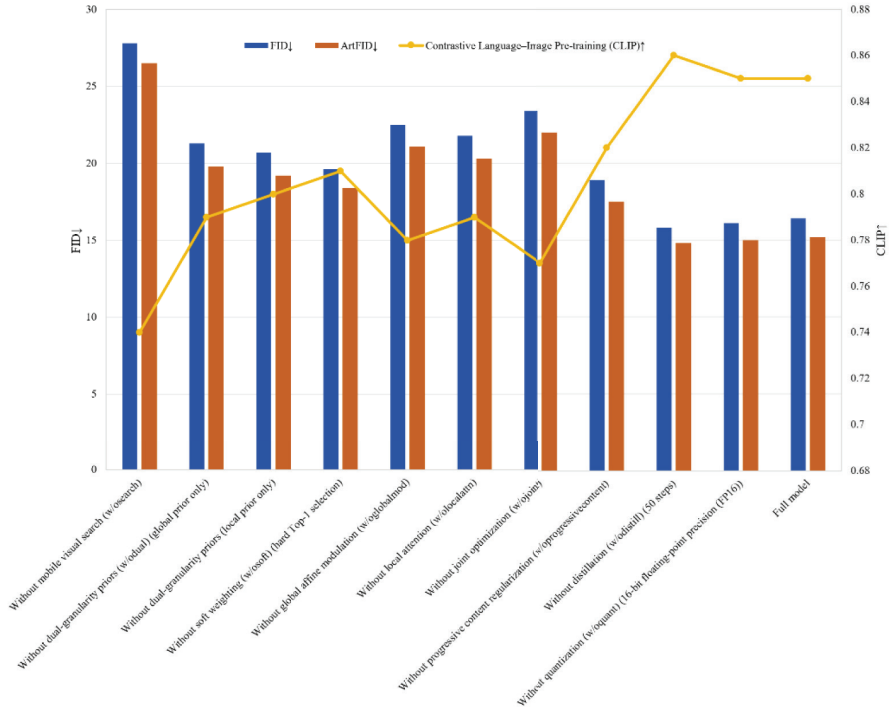
A comparison between the proposed method and eight baselines in terms of generation quality and efficiency is summarized in Table 1. The same test set, an output resolution of 512×512, and a batch size of 1 were used for all methods; latency was re-measured under the same Snapdragon 8 Gen 3 test environment, and parameter counts were counted according to the corresponding implementations. Optimal results were obtained by the proposed method across all six metrics: Fréchet inception distance, learned perceptual image patch similarity, art Fréchet inception distance, CLIP score, style loss, and content loss. Compared with MobileStyler, Fréchet inception distance was reduced from 24.6 to 16.4, art Fréchet inception distance was reduced from 23.1 to 15.2, and CLIP score was increased from 0.77 to 0.85; this gap indicates that style consistency and generation quality are improved by the search-guided dual-granularity priors. For the diffusion methods inversion-based style transfer (InST), StyleDiffusion, and DiffStyler, Fréchet inception distance values of 22.1, 19.8, and 18.9 were obtained, respectively, and art Fréchet inception distance values of 20.4, 18.7, and 17.9 were obtained, respectively; reductions of at least 2.5 and 2.7 in these two metrics were achieved by the proposed method. Meanwhile, a parameter count of approximately 19.5 M and an end-to-end latency of 720 ms were obtained by the proposed method, which is lower than the 3.2 s to 4.1 s inference time of the above diffusion baselines. A slightly lower latency of 620 ms was achieved by MobileStyler, but its Fréchet inception distance was higher by 8.2, its art Fréchet inception distance was higher by 7.9, and its CLIP score was lower by 0.08, indicating that a favorable balance between mobile quality and efficiency was achieved by the proposed method.

3.3 Ablation experiments on key components

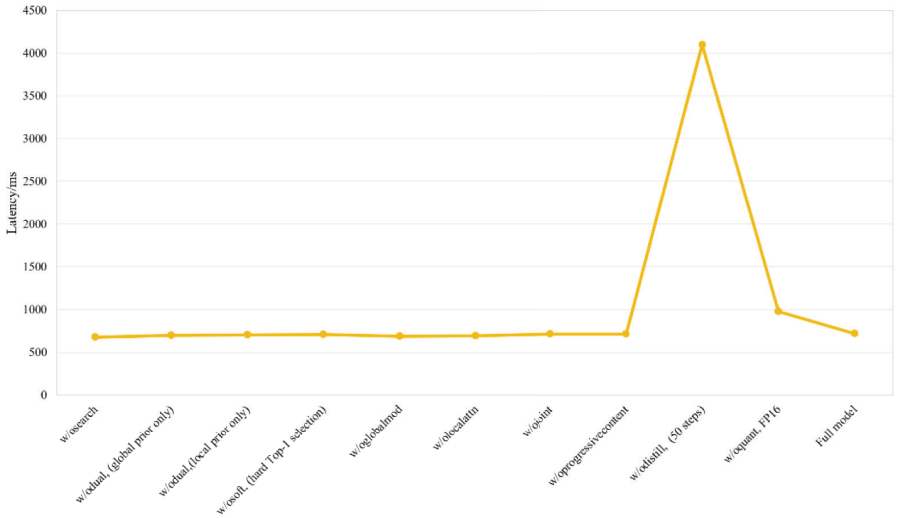
The contribution of each component to final performance is revealed by the ablation results in Figure 2. When mobile visual search was removed, Fréchet inception distance deteriorated from 16.4 to 27.8 and art Fréchet inception distance deteriorated from 15.2 to 26.5, indicating that the search-guided style prior is the primary source of performance improvement. When only global priors or only local priors were retained, Fréchet inception distance values of 21.3 and 20.7 were obtained, respectively, both inferior to those of the dual-granularity priors, reflecting the complementarity between global style tone and local texture. When soft weighting was replaced by hard Top-1, Fréchet inception distance increased to 19.6, indicating that style bias introduced by a single reference image can be alleviated by soft weighting. When global modulation or local attention was removed, Fréchet inception distance increased to 22.5 and 21.8, respectively, demonstrating that both injection schemes contribute. When joint training was removed, Fréchet inception distance increased to 23.4, verifying the necessity of joint optimization of differentiable re-ranking within the candidate set and generation. When progressive content regularization was removed, Fréchet inception distance increased to 18.9, indicating that the frequency-domain progressive constraint is beneficial for preserving product structure. Slight quality loss is introduced by distillation and quantization: when distillation was omitted, Fréchet inception distance was 15.8, but latency was 4100 ms; when quantization was omitted, Fréchet inception distance was 16.1, latency was 980 ms, and model size was larger. A Fréchet inception distance of 16.4 at a latency of 720 ms was achieved by the complete model, realizing an acceptable balance between quality and efficiency for mobile deployment.

3.4 Mobile efficiency and resource evaluation

Resource usage and inference performance of four methods on mobile devices are given in Table 2. For the proposed method, model size was 18.7 MB, parameter count was 19.5 M, and floating-point operations per denoising forward pass were 2.8 G, which falls within the deployable range for mobile devices. The lowest latency of 310 ms was achieved by Fast Neural Style, but its generation quality metrics were clearly lower than those of the proposed method, and automatic style search was not supported. Compared with MobileStyler, slightly higher model size and parameter count were exhibited by the proposed method; latency increased from 620 ms to 720 ms, peak memory increased from 380 MB to 420 MB, and energy consumption increased from 1.6 J to 1.9 J. Temperature was used to reflect device thermal state under unified test conditions and was not intended as a basis for human comfort assessment. For Quantized Diffusion, latency was 1150 ms, model size was 22.4 MB, and energy consumption was 2.8 J; reductions of approximately 37%, 16%, and 32% in these three aspects were achieved by the proposed method, respectively. When Tables 1 and 2 are considered together, 512×512 stylized generation was completed within 720 ms by the proposed method, and sub-second mobile interaction requirements were met.



a) Fréchet inception distance (FID) and art Fréchet inception distance (ArtFID) ↓



b) Latency

Fig. 2. Results of ablation experiments on key components

Table 2. Comparison of mobile efficiency and resources

Method	Model Size	Parameters	Floating-Point Operations Per Forward Pass	Latency	Peak Memory	Energy	Temperature
Fast Neural Style	8.6 MB	9.2 M	1.1 G	310 ms	210 MB	0.9 J	35.1°C
MobileStyler	17.2 MB	18.1 M	2.4 G	620 ms	380 MB	1.6 J	37.2°C
Quantized Diffusion	22.4 MB	24.3 M	3.6 G	1150 ms	520 MB	2.8 J	41.0°C
Proposed method	18.7 MB	19.5 M	2.8 G	720 ms	420 MB	1.9 J	38.5°C

3.5 User subjective evaluation and interaction experience

Subjective ratings of four methods from 30 participants are presented in Table 3. The highest scores were achieved by the proposed method in style consistency, content fidelity, aesthetic preference, operational convenience, and response satisfaction. Compared with the manual style specification, operational convenience was increased from 3.12 to 4.61 by the proposed method, indicating that user operational burden is reduced by automatic recommendation of style priors through search. A response satisfaction score of 2.41 was obtained by InST, whereas a score of 4.27 was achieved by the proposed method, indicating that interaction experience is improved by sub-second inference. Compared with MobileStyler, style consistency was increased from 3.78 to 4.52 and aesthetic preference was increased from 3.62 to 4.38 by the proposed method, indicating that perceptible quality improvement is brought by dual-granularity priors and joint optimization. Since all methods were evaluated by the same participants, a paired t-test was adopted; differences between the proposed method and other methods in style consistency and response satisfaction both reached $p < 0.01$.

Table 3. User subjective evaluation (five-point scale)

Method	Style Consistency	Content Fidelity	Aesthetic Preference	Operational Convenience	Response Satisfaction
Manual style specification	4.05 ± 0.48	4.18 ± 0.45	3.92 ± 0.51	3.12 ± 0.62	3.25 ± 0.58
Inversion-based Style Transfer (InST)	4.12 ± 0.46	3.88 ± 0.49	4.05 ± 0.47	2.95 ± 0.66	2.41 ± 0.71
MobileStyler	3.78 ± 0.52	3.95 ± 0.48	3.62 ± 0.55	3.71 ± 0.59	3.88 ± 0.54
Proposed method	4.52 ± 0.41	4.31 ± 0.47	4.38 ± 0.44	4.61 ± 0.38	4.27 ± 0.52

3.6 Generalization and robustness experiments

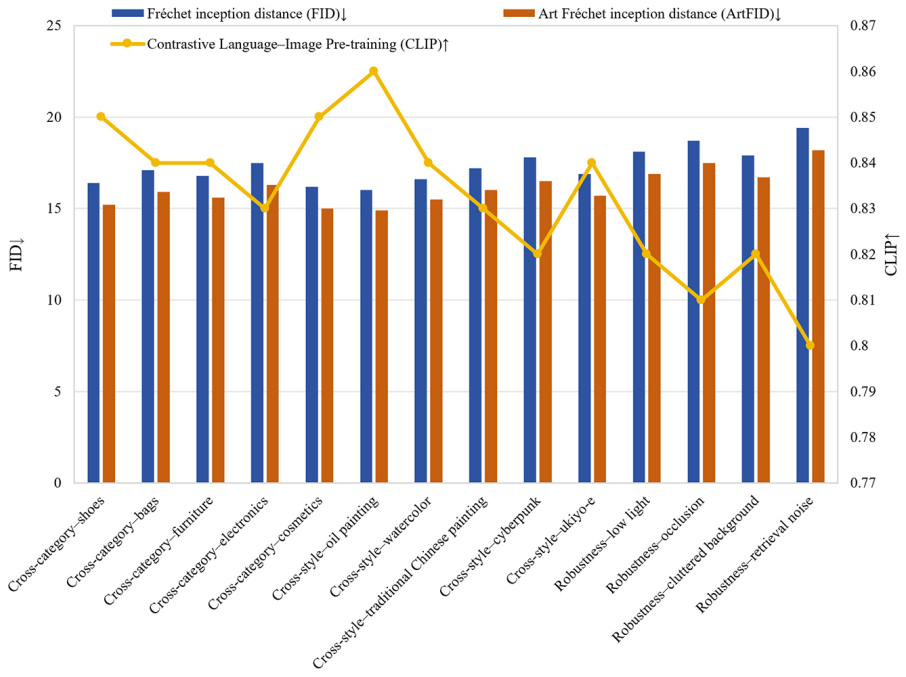


Fig. 3. Generalization and robustness

Performance variations under leave-one-category-out, leave-one-style-out, and real-world interference settings are further presented in Figure 3. When each of the five product categories was used as a leave-one-out test category excluded from the corresponding training split, Fréchet inception distance ranged from 16.2 to 17.5 and art Fréchet inception distance ranged from 15.0 to 16.3, indicating that a certain adaptability to product category variation is provided by the dual-granularity priors. Among five leave-one-out artistic styles, Fréchet inception distance values of 16.0 and 16.9 were obtained for oil painting and ukiyo-e, respectively; a value of 17.8 was obtained for cyberpunk, in which textures are complex, and color contrast is strong, so higher demands are placed on local priors. Under robustness settings, Fréchet inception distance was increased to 17.9–18.7 by low light, occlusion, and cluttered background, and was increased to 19.4 by retrieval noise, whereas the CLIP score remained above 0.80, indicating that acceptable stability is maintained in real mobile scenarios. Additional texture constraints are provided by local priors under occlusion and cluttered background conditions so that effects caused by local visual degradation of the query image are mitigated.

4 CONCLUSION

To address the problems of manual style specification, mutual separation between mobile visual search and generative models, and difficulty in rapid on-device deployment of diffusion models in mobile product artistic style transfer, a search-guided style-prior conditioning mechanism was proposed. In this mechanism, mobile visual search was reconstructed as a prior-supply module for generative tasks, retrieval candidates were encoded into a global style-tone vector and local style-texture features, and dual-granularity style priors were aggregated through candidate re-ranking and temperature-based soft weighting. Through differentiable soft weighting within the candidate set, generation losses were backpropagated to the re-ranking and style encoding modules, and alignment between retrieval representations and the generative conditioning space was achieved. On the generation side, a pixel-space lightweight diffusion U-Net was adopted, computational cost was reduced by depthwise separable convolutions and inverted residual bottlenecks, and dual-granularity priors were injected through global affine modulation and local attention. During training, denoised, style consistency, content fidelity, and retrieval ranking losses were jointly optimized. During deployment, consistency distillation and 8-bit integer quantization were adopted so that automatic style-prior recommendation and conditioned generation were achieved under mobile computational constraints. On the self-constructed product artistic style dataset, a Fréchet inception distance of 16.4, an art Fréchet inception distance of 15.2, and a CLIP score of 0.85 were obtained; model size was 18.7 MB, and the generation latency at 512×512 resolution was 720 ms. The effectiveness of the mechanism in generation quality, style consistency, on-device efficiency, and interaction experience was jointly verified by comparative experiments, ablation experiments, mobile efficiency evaluation, user subjective evaluation, and generalization and robustness experiments. A feasible solution is thereby provided for automatic style-prior recommendation and sub-second style transfer on mobile devices, and a reference is provided for the transformation of mobile visual search from a retrieval tool into a generative prior supply module.

5 REFERENCES

[1] H. Qin, D. A. Peak, and V. Prybutok, "A virtual market in your pocket: How does mobile augmented reality (MAR) influence consumer decision making?" *Journal of Retailing and Consumer Services*, vol. 58, Art. no. 102337, 2021. https://doi.org/10.1016/j.jretconser.2020.102337

[2] F. Wang, "Interactive cinematic experience design and user behavior analysis enabled by mobile augmented reality," *International Journal of Interactive Mobile Technologies*, vol. 19, no. 14, pp. 151–165, 2025. https://doi.org/10.3991/ijim.v19i14.57007

[3] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song, "Neural style transfer: A review," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 11, pp. 3365–3385, 2019. https://doi.org/10.1109/TVCG.2019.2921336

[4] A. Rahman, E. Winarko, and K. Mustofa, "Product image retrieval using category-aware siamese convolutional neural network feature," *Journal of King Saud University – Computer and Information Sciences*, vol. 34, no. 6, pp. 2680–2687, 2022. https://doi.org/10.1016/j.jksuci.2022.03.005

[5] X. S. Wei, Y. Shen, X. Sun, P. Wang, and Y. Peng, "Attribute-aware deep hashing with self-consistency for large-scale fine-grained image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 13904–13920, 2023. https://doi.org/10.1109/TPAMI.2023.3299563

[6] X. Zhang, C. Bai, and K. Kpalma, "OMCBIR: Offline mobile content-based image retrieval with lightweight CNN optimization," *Displays*, vol. 76, Art. no. 102355, 2022. https://doi.org/10.1016/j.displa.2022.102355

[7] H. Peng, W. Qian, J. Cao, and S. Tang, "Arbitrary style transfer based on attention and covariance-matching," *Computers & Graphics*, vol. 116, pp. 298–307, 2023. https://doi.org/10.1016/j.cag.2023.08.029

[8] A. Singh, V. Jaiswal, G. Joshi, A. Sanjeeve, S. Gite, and K. Kotecha, "Neural style transfer: A critical review," *IEEE Access*, vol. 9, pp. 131583–131613, 2021. https://doi.org/10.1109/ACCESS.2021.3112996

[9] J. Chi, C. K. On, H. Zhang, and S. S. Chai, "A review of deep convolutional neural networks in mobile face recognition," *International Journal of Interactive Mobile Technologies*, vol. 17, no. 23, pp. 4–19, 2023. https://doi.org/10.3991/ijim.v17i23.40867

[10] X. Zhou, Y. Zheng, and J. Yang, "Bridging the metrics gap in image style transfer: A comprehensive survey of models and criteria," *Neurocomputing*, vol. 624, Art. no. 129430, 2025. https://doi.org/10.1016/j.neucom.2025.129430

[11] Z. Zhang *et al.*, "Rethink arbitrary style transfer with transformer and contrastive learning," *Computer Vision and Image Understanding*, vol. 241, Art. no. 103951, 2024. https://doi.org/10.1016/j.cviu.2024.103951

[12] G. Jang, J. W. Lee, J. G. Lee, and Y. Liu, "Distributed fine-tuning of CNNs for image retrieval on multiple mobile devices," *Pervasive and Mobile Computing*, vol. 64, Art. no. 101134, 2020. https://doi.org/10.1016/j.pmcj.2020.101134

[13] W. Song, W. Ma, M. Zhang, Y. Zhang, and X. Zhao, "Lightweight diffusion models: A survey," *Artificial Intelligence Review*, vol. 57, no. 6, Art. no. 161, 2024. https://doi.org/10.1007/s10462-024-10800-8

[14] K. Wang, X. Liu, Y. Chang, D. Zhao, T. Xian, and X. Geng, "Semantic guidance for precise style control in diffusion image generation," *Scientific Reports*, vol. 15, no. 1, Art. no. 45581, 2025. https://doi.org/10.1038/s41598-025-28715-x

[15] I. Kyianytsia, D. Yatsiuk, H. Aldankova, O. Horobets, V. Slipchenko, and V. Dobrovolskyi, "A cyber-physical system approach to adaptive visual branding in cultural institutions," *Journal of Engineering Management and Systems Engineering*, vol. 5, no. 3, pp. 340–353, 2026. https://doi.org/10.56578/jemse050304

[16] A. K. Pallikonda, V. K. Bandarapalli, and V. Aruna, "Enhancing performance and reducing latency in autonomous systems through edge computing for real-time data processing," *Mechatronics and Intelligent Transportation Systems*, vol. 4, no. 3, pp. 154–165, 2025. https://doi.org/10.56578/mits040305

[17] I. Hwang and T. Oh, "Design and experimental research of on device style transfer models for mobile environments," *Scientific Reports*, vol. 15, no. 1, Art. no. 13724, 2025. https://doi.org/10.1038/s41598-025-98545-4

[18] Q. Chen, L. Zhang, J. Ma, and A. Khadka, "A lightweight real-time vision framework for road infrastructure monitoring in intelligent transportation systems," *Mechatronics and Intelligent Transportation Systems*, vol. 5, no. 1, pp. 1–10, 2026. https://doi.org/10.56578/mits050101

6 AUTHORS

Chen Kang is an Associate Professor and Dean of the School of Digital Creative Design at College for Digital Creative Studies, Changzhou Vocational Institute of Mechatronic Technology, and a member of the Jiangsu Industrial Design Association (JSIDA). His research interests cover product design methodology, studies on product design styles and related fields (E-mail: kc1644@czimt.edu.cn).

Zheng Qi holds a master's degree from Jiangnan University, China. She is an Associate Professor and Deputy Director of the Department of Product and Space Design at the School of Digital Creative Design, College for Digital Creative Studies, Changzhou Vocational Institute of Mechatronic Technology. Her research focuses on spatial design and household product design (E-mail: qz1937@czimt.edu.cn).