IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 3, June, 2026, pp. 1463-1478
Journal homepage: http://iieta.org/journals/ts

Intelligent Garment Silhouette Design and Personalized Style Generation via Cross-Modal Visual-Semantic Alignment and Generative Image Processing

Mengsha Cao

College of Art and Design, Zhanjiang University of Science and Technology, Zhanjiang 524000, China

Corresponding Author Email: caomengsha123@163.com

https://doi.org/10.18280/ts.430330

Received: 19 April 2026
Revised: 13 June 2026
Accepted: 22 June 2026
Available online: 30 June 2026

Keywords:
intelligent garment silhouette generation, latent diffusion models, cross-modal semantic alignment, controllable image synthesis, structure-preserving editing, style disentanglement

ABSTRACT

Garment silhouette constitutes a core structural determinant in apparel design, and its intelligent generation and precisely controllable editing remain challenging frontiers at the intersection of generative image processing and intelligent fashion design. The limitations of existing diffusion-based garment generation methods preclude professional-grade intelligent silhouette design and customized garment creation. Therefore, an end-to-end intelligent garment design generation framework, termed SilhouetteDiff, was constructed, with the silhouette structure serving as the central anchoring element. This framework establishes a multi-level controllable generation architecture atop a latent diffusion model, systematically remedying the aforementioned deficiencies. Four cooperatively optimized functional modules were devised: cross-modal semantic alignment was employed to achieve fine-grained matching between professional garment silhouette terminology and visual structural features; geometric prior constraints were imposed to enable high-precision silhouette generation driven jointly by semantic and physical structural information; a structured fidelity loss function was formulated to enforce stable contour preservation during texture and style editing processes; and a multi-dimensional feature disentanglement strategy was introduced to decompose garment generation factors into mutually independent sub-spaces corresponding to silhouette, texture, and color, thereby supporting flexible recombination of diverse design conditions and facilitating novel creative synthesis. All modules were deeply coupled through a unified silhouette representation as an intermediate carrier, forming a complete design pipeline encompassing semantic parsing, structure generation, intelligent editing, and compositional innovation. Extensive quantitative comparisons, ablation studies, and subjective user evaluations conducted on publicly available garment datasets demonstrated that the proposed framework significantly outperformed state-of-the-art methods in terms of silhouette generation accuracy, semantic alignment precision, structural fidelity during editing, and overall design practicality. This work effectively overcomes the long-standing bottlenecks of structural controllability and domain-specific semantic adaptation in conventional garment generation models, offering a mature technical solution for intelligent personalized apparel design while providing a novel technical reference for controllable generative image processing and cross-modal alignment research in vertical domains.

1. INTRODUCTION

Garment silhouette constitutes a core structural element in defining the visual morphology and stylistic attributes of apparel, serving as a fundamental benchmark within the modern garment design system [1, 2]. Traditional garment design has long relied upon manual sketching and iterative pattern prototyping, with the entire workflow heavily dependent on practitioner expertise—a process characterized by low iterative efficiency and limited adaptability to the growing industrial demands for personalization and rapid product turnover [3, 4]. The rapid advancement of generative artificial intelligence and image processing technologies has introduced a novel technical paradigm for intelligent garment design [5, 6], and diffusion-model-based image generation methods have been widely adopted for automated garment style creation [7, 8]. In contrast to general-purpose image generation tasks, the intelligent generation of garment silhouettes entails strong domain-specific particularities [9, 10], requiring not only high-fidelity pixel-level synthesis but also precise control over global geometric structure, accurate alignment with professional semantic terminology, and structurally stable editing under imposed constraints—collectively constituting the core technical challenges in contemporary intelligent garment generation research [11, 12].

Substantial progress has been achieved in diffusion-model-based garment image generation [13, 14], with prior work incorporating edge-feature constraints, pose-keypoint detection, and multi-modal conditioning techniques to effectively enhance controllability and detail fidelity in garment synthesis [15, 16], thereby meeting basic requirements for routine garment image generation and simple

editing tasks [17, 18]. However, when focused on professional-grade, precision-oriented silhouette design scenarios, the existing technical framework remains plagued by systemic deficiencies that cannot be circumvented [19, 20]. Generic vision-language models have not undergone domain-specific adaptation training for garment silhouette semantics, resulting in insufficient fine-grained semantic discriminability and pronounced homogenization among generation outcomes derived from semantically similar silhouette descriptions [11, 12]. Conventional structural conditioning methods are confined to pixel-level result regulation, lacking generative modeling of core geometric parameters such as shoulder width, waistline position, and hem proportion, which readily leads to distorted silhouette structural ratios [14, 20]. Moreover, existing editing algorithms fail to treat contour integrity as a primary constraint objective, with contour displacement and structural distortion being commonly observed during texture updates and color adjustments [8, 17]. The underlying cause of these issues is traced to the implicit treatment of silhouette as a latent generative feature in current approaches, without establishing an interpretable and quantifiable explicit structural control framework [1, 2].

To address the aforementioned technical deficiencies, an end-to-end intelligent garment design framework, termed SilhouetteDiff, is constructed with structurally explicit silhouette representation as the central anchoring element, establishing a multi-module collaboratively optimized and fully controllable generation pipeline. A semantically aware cross-modal alignment module is designed to achieve precise mapping between professional garment silhouette semantics and visual geometric features, thereby resolving the issue of insufficient fine-grained semantic matching accuracy. By leveraging prior knowledge of garment structural anatomy, key skeletal parameters are quantitatively encoded, and a dual-path conditioning injection mechanism is employed to accomplish integrated regulation of semantic information and geometric constraints, enabling high-precision silhouette generation. A silhouette-preserving dedicated editing mechanism is developed, in which a structured loss function imposes rigid constraints on contour morphology, ensuring structural integrity throughout style-driven editing processes. A latent-space orthogonal disentanglement strategy is introduced to decompose garment generation features into mutually independent sub-spaces corresponding to silhouette, texture, and color, thereby supporting flexible recombination of multi-source design conditions and smooth stylistic interpolation. All modules are deeply coupled through a unified silhouette representation, forming an integrated technical pipeline that encompasses semantic parsing, structural generation, intelligent editing, and compositional innovation. Comprehensive multi-dimensional experiments substantiate the effectiveness and technical superiority of the proposed framework.

The remainder of this study is organized as follows. In Chapter 2, the overall architecture of SilhouetteDiff is elaborated in detail, with comprehensive descriptions of the design philosophy, mathematical underpinnings, and implementation specifics of each core module. In Chapter 3, a complete experimental framework is established, in which quantitative comparative experiments, ablation studies, and subjective user evaluations are conducted to comprehensively validate the generative performance and practical utility of the proposed framework. In Chapter 4, the technical limitations of the current framework are objectively analyzed, and prospective directions for further optimization and extended application scenarios are identified. In Chapter 5, the findings of the entire study are summarized, and the core innovations and application value are consolidated.

2. METHODOLOGY

2.1 Overall framework overview

The proposed SilhouetteDiff framework is built upon Stable Diffusion v1.5 as the foundational generative model, with a latent-space generation system constructed via a frozen variational autoencoder and a trainable noise prediction network. The variational autoencoder is responsible for compressing pixel-space images at 512 × 512 resolution into the latent space, with a fixed compression factor of 8 and an output channel dimension of 4, thereby substantially reducing computational complexity while preserving core structural and semantic information of the images. Departing from the pixel-level single-constraint paradigm commonly adopted in conventional garment generation models, the framework employs structurally explicit garment silhouette representation as the global anchoring element, establishing a progressive end-to-end generation architecture that sequentially accomplishes the full pipeline of semantic understanding, structural generation, personalized editing, and compositional creation. Three input modalities—textual descriptions, structural skeletons, and reference images—are integrated into the model, and high-precision controllable generation is achieved through the coordinated coupling of four functional modules. A semantically aware cross-modal alignment module for silhouette is designed to perform fine-grained mapping of textual semantics, outputting semantic embedding features adapted to the domain-specific characteristics of garments. A silhouette-structured prior guidance module fuses structural keypoint coordinate information with semantic features, constraining the latent variable generation process through a dual-path conditioning injection mechanism, thereby accomplishing precise mapping from semantic information to geometric structure. A silhouette-preserving style editing module establishes a structured constraint mechanism based on contour masks, enabling contour fidelity during texture and color editing processes. A multi-modal silhouette style disentanglement module performs disentangled modeling of latent-space features, constructing mutually independent feature sub-spaces for silhouette, texture, and color, thereby supporting flexible recombination of multi-source design conditions and continuous style interpolation. All modules are deeply coupled through a unified silhouette representation serving as the shared feature carrier, forming a highly integrated and cohesive generation pipeline.

A two-stage progressive training strategy was adopted for model optimization, by which the gradient oscillation and parameter adaptation issues inherent in multi-module collaborative training were effectively resolved, and the dual objectives of preserving base model performance and achieving efficient convergence of newly added functional modules were simultaneously accomplished. In the first stage, all parameters of the variational autoencoder and the UNet backbone network were frozen, with training confined to the semantic alignment prototype matrix, adaptation layers, and

structural control branch parameters, and preliminary optimization was performed via the base diffusion loss, cross-modal alignment loss, and text semantic loss. In the second stage, all newly added network parameters were unfrozen for end-to-end fine-tuning, with the overall joint loss function of the model defined as:

$$L_{total}=L_{simple}+\lambda_1 L_{align}+\lambda_2 L_{text}+\lambda_3 L_{sp}+\lambda_4 L_{orth}+\lambda_5\|\Theta_{new}\|_2^2 \quad (1)$$

where, L_{simple} denotes the base diffusion noise prediction loss, serving to preserve the fundamental image generation capability of the model; L_{align} denotes the silhouette visual-semantic alignment loss, constraining the matching relationship between domain-specific fine-grained semantics and structural features; L_{text} denotes the bidirectional text contrastive loss, reinforcing the robustness of textual semantic representations; L_{sp} denotes the silhouette structure preservation loss, constraining contour integrity during editing processes; L_{orth} denotes the subspace orthogonality loss, enabling multi-dimensional feature disentanglement; and Θ_{new} refers to all newly added trainable parameters of the framework, with λ_1, λ_2, λ_3, λ_4, and λ_5 serving as balancing weights for the respective loss terms. Model training was conducted using the AdamW optimizer, with first-order momentum coefficient set to 0.9, second-order momentum coefficient to 0.999, and weight decay coefficient to 0.01. In the first stage, the learning rate was set to 1×10^{-4}, with a batch size of 16 and training performed for 50 k steps; in the second stage, the learning rate was adjusted to 5×10^{-5}, with a batch size of 8 and training conducted for 100 k steps, augmented by a gradient accumulation step of 2 to mitigate hardware memory constraints. During the inference phase, the denoising diffusion implicit models sampling strategy was adopted, with a fixed sampling step of 50 and a classifier-free guidance scale of 7.5, thereby achieving an optimal balance between generation fidelity and inference efficiency.

2.2 Semantically aware silhouette cross-modal alignment module

To address the deficiency of fine-grained silhouette semantic representation in generic vision-language models when applied to the garment domain, a semantically aware silhouette cross-modal alignment module was designed, establishing a precise mapping mechanism between standardized silhouette semantics and visual geometric features. The architecture of the semantically aware silhouette cross-modal alignment module is illustrated in Figure 1. A categorical system was constructed by integrating five fundamental silhouette types from the garment design domain, and the textual description paradigm was constrained through a standardized prompt encoding strategy, with the diversity and generalizability of textual semantic representations being enriched by randomly incorporating ancillary attributes such as garment length and sleeve type. Feature extraction was performed via a pre-trained contrastive language–image pre-training encoder, yielding image visual features and textual semantic features, both uniformly represented in 768-dimensional space. To enforce clustering constraints on silhouette semantics across categories, a learnable silhouette prototype embedding matrix was introduced, with its parameters initialized via Kaiming uniform initialization and continuously updated and optimized throughout the training process, thereby generating category-specific semantic center representations and enabling structured modeling of professional garment silhouette semantics.

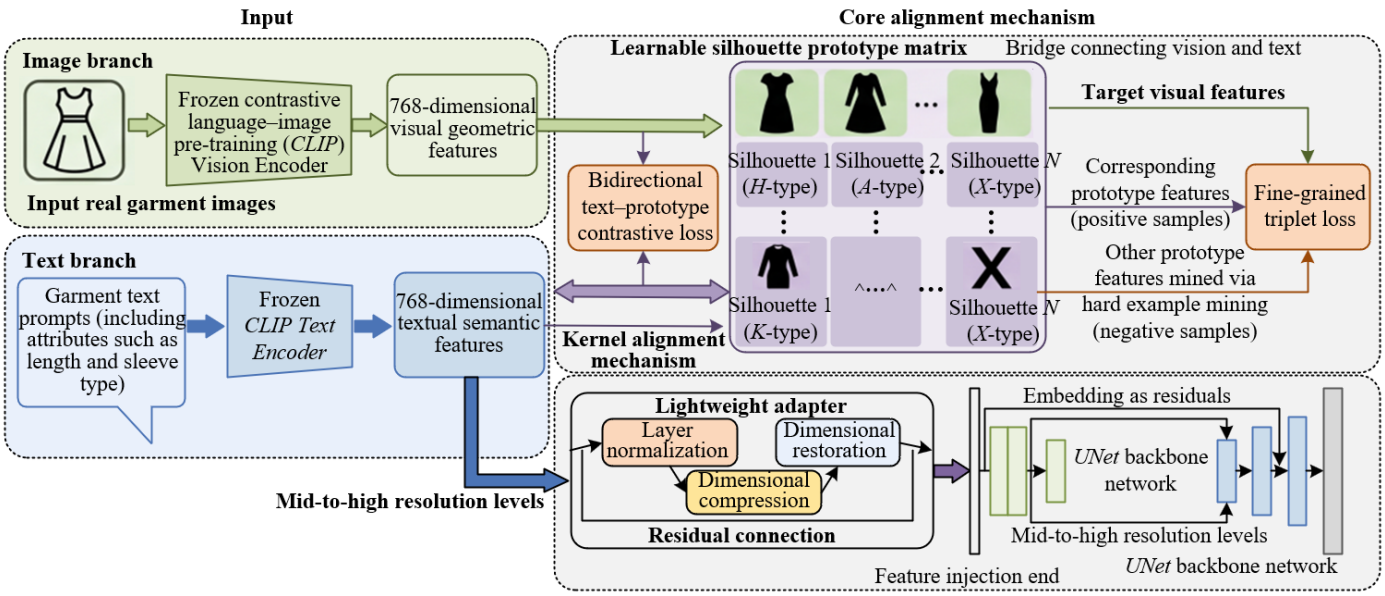


Figure 1. Architecture of the semantically aware silhouette cross-modal alignment module

A vision-driven fine-grained triplet loss function was constructed, in which an online hard example mining strategy was incorporated to enhance the model's discriminative capability for similar silhouette features, thereby effectively breaking the homogenization problem of closely related silhouette semantic representations within generic embedding spaces. With the visual geometric features of samples serving as the core constraint, this loss selects, within each training batch, cross-category silhouette samples with the highest feature similarity as hard negative samples, and optimizes the discriminability of prototype representations through distance-based constraints between positive and negative samples. The loss is formally defined as:

$$L_{align}=\frac{1}{N}\sum_{i=1}^{N}\max\left(0,\left\|v_i-p_{y_i}\right\|_2^2-\left\|v_i-p_{y_j}\right\|_2^2+\alpha\right) \quad (2)$$

where, N denotes the number of training samples in a single batch, v_i denotes the visual features of the target sample, p_{y_i} denotes the prototype embedding features corresponding to the true silhouette category of the sample, p_{y_j} denotes the silhouette prototype features corresponding to the hard negative sample, and α denotes the margin threshold, with the optimal value determined as 0.5 via grid search on the validation set. In contrast to conventional text-based alignment approaches, this vision-driven constraint directly associates garment geometric structures with silhouette semantics, circumventing representational biases introduced by indirect textual mapping, and substantially improving the matching accuracy of fine-grained silhouette semantics.

To further refine the cross-modal alignment framework, a bidirectional text–prototype contrastive loss was constructed, enabling bidirectional adaptive adaptation between the textual modality and silhouette prototype features, thereby enhancing the robustness of semantic representations. The loss is expressed as:

$$L_{text}=-\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{exp(\langle t_i,p_{y_i}\rangle/\tau)}{\sum_{k=1}^{K}exp\left(\langle t_i,p_k\rangle/\tau\right)}+\log\frac{exp(\langle p_{y_i},t_i\rangle/\tau)}{\sum_{j=1}^{N}exp\left(\langle p_{y_i},t_j\rangle/\tau\right)}\right] \quad (3)$$

where, t_i denotes the textual semantic features corresponding to the sample, K denotes the total number of silhouette categories, and τ denotes the temperature coefficient controlling the smoothness of feature distribution, with a fixed value of 0.07. This loss simultaneously enforces clustering constraints from text features toward their corresponding silhouette prototypes and matching constraints from prototype features toward their source texts, thereby achieving bidirectional deep alignment. On this basis, a lightweight adapter was configured and embedded into the backbone diffusion network, where feature optimization was performed through hierarchical transformations comprising layer normalization, dimensional compression, and dimensional restoration, with feature fusion accomplished via residual connections. The adapter was initialized with a differentiated parameter initialization strategy, by which the pre-trained generative capability of the base model was stably preserved. Deployed exclusively at mid-to-high resolution feature levels, the adapter achieves adaptive fusion of semantically aligned features with minimal parameter overhead, balancing model performance and inference efficiency.

2.3 Silhouette structure prior guidance module

To achieve quantitative modeling and precisely controllable generation of garment silhouette geometric structures, a structured prior guidance mechanism was constructed in this module, by which abstract garment appearance morphologies were transformed into a computable and constrainable geometric feature system, thereby addressing the issues of ambiguous structural constraints and poor proportional controllability in conventional generation methods. The workflow of the silhouette structure prior guidance module (SSPGM) is illustrated in Figure 2. A sparse semantic skeleton comprising ten coordinate points was defined, uniformly covering the core symmetric structures of the shoulder, chest, waist, hip, and hem, thereby fully characterizing the geometric proportions and overall morphology of garment silhouettes. Automatic keypoint detection was performed using a High-Resolution Network (HRNet-W48) pre-trained on the DeepFashion2 dataset, with redundant detection results filtered out via a non-maximum suppression algorithm, and all keypoint coordinates were uniformly normalized to the [0, 1] interval to eliminate scale discrepancies arising from varying image resolutions. For anomalous scenarios such as occlusion or detection failure, an interpolation completion strategy was designed based on the inherent geometric regularities of standard garment silhouettes, by which missing keypoint information was repaired through silhouette-specific structural constraint rules, ensuring the geometric rationality and completeness of the skeletal features.

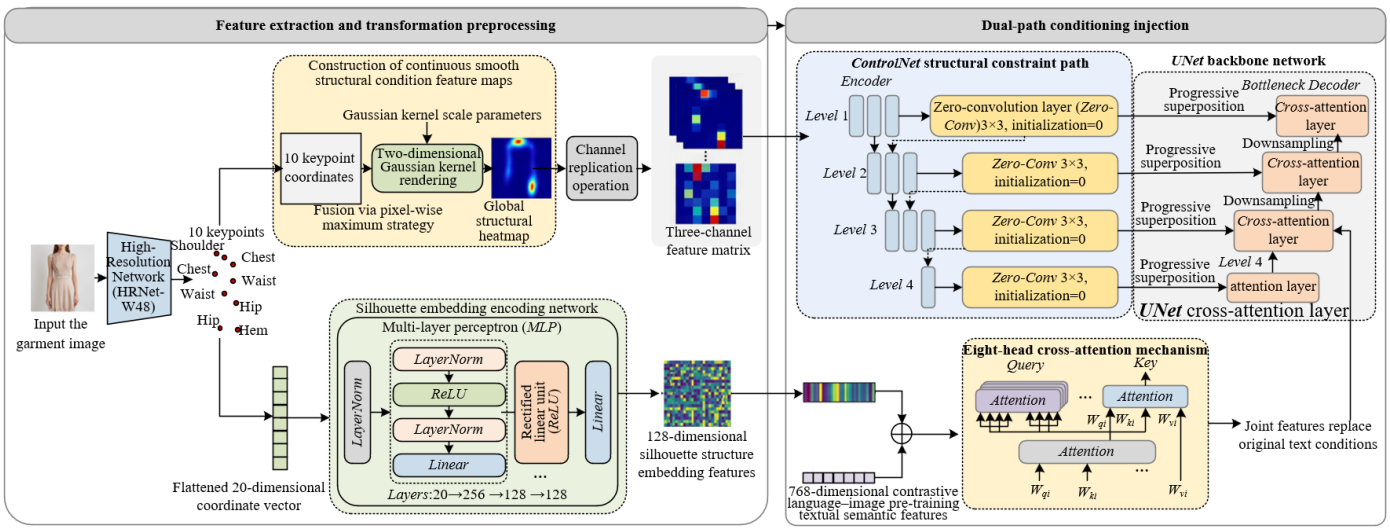


Figure 2. Workflow of the silhouette structure prior guidance module (SSPGM)

To transform the discrete sparse keypoint skeleton into dense features suitable for generative model input, two-dimensional Gaussian kernels were employed for keypoint heatmap rendering, by which continuous smooth structural

condition feature maps were constructed. The feature computation at a single pixel position is defined as follows:

$$G_m(p)=exp\left(-\frac{\|p-(x_m,y_m)\|_2^2}{2\sigma^2}\right) \tag{4}$$

where, p denotes the pixel coordinates in image space, (x_m,y_m) denotes the coordinates of the m-th normalized structural keypoint, and σ denotes the Gaussian kernel scale parameter, which was set to 3.0 at the standard image resolution of 512 × 512, thereby generating an effective receptive field commensurate with garment structural scales. The Gaussian features corresponding to all keypoints were fused via a pixel-wise maximum strategy, yielding a single-channel global structural heatmap, which was subsequently expanded into a three-channel feature matrix $C_{struct}\in \mathrm{R}^{512\times512\times1}$ through channel replication, matching the input dimension requirements of ControlNet and accomplishing the standardized transformation from discrete geometric information to continuous structural features.

To extract the higher-order geometric ratio features implicit in the skeletal coordinates, a dedicated silhouette embedding encoding network was designed, enabling the mapping from low-dimensional coordinate data to high-dimensional structured semantic features. The overall encoding process is defined as:

$$e_{sil}=MLP_{sil}(Flatten(K)) \tag{5}$$

where, $Flatten(K)$ denotes the flattened skeletal feature vector with a dimension of 20, corresponding to the horizontal and vertical coordinates of the ten keypoints, and e_{sil} denotes the 128-dimensional silhouette structure embedding features output by the network. A progressive dimensional transformation structure of 20, 256, 128, and 128 was configured for the multi-layer perceptron, with layer normalization and rectified linear unit activation functions applied to the first three layers, while only a linear transformation was retained in the final layer to fully preserve the higher-order structural features. This encoding network was jointly and iteratively optimized with the model backbone network, enabling autonomous learning of the proportional relationships among core structures such as shoulder width, hem width, and waist position, thereby achieving an upgrade from basic coordinate information to refined silhouette structural representations.

A dual-path conditioning injection mechanism was adopted in this module, integrating explicit geometric structural constraints and implicit semantic constraints to establish a dual-layer synergistic regulation paradigm for generation. The first path is the ControlNet structural constraint path, which replicates the four-stage downsampling encoding structure of UNet, with structural features being progressively injected through multi-level zero-convolution layers. The single-stage feature fusion process is expressed as:

$$f_{ctrl}^{(l)}=Z_{l,3}\left(Block_{ctrl}^{(l,3)}\left(Z_{l,2}\left(Block_{ctrl}^{(l,2)}\left(Z_{l,1}\left(Block_{ctrl}^{(l,1)}(C_{struct})\right)\right)\right)\right)\right) \tag{6}$$

$$f_{UNet}^{(l)}=f_{UNet}^{(l)}+f_{ctrl}^{(l)} \tag{7}$$

where, $Z_{l,j}$ denotes the zero-convolution layers, which employ 3 × 3 convolution kernels with all parameters initialized to zero, allowing progressive parameter updates during training and preventing initial constraints from interfering with the inherent feature distributions of the pre-trained model; $Block^{(l,j)}{}_{ctrl}$ denotes the ControlNet fundamental residual and Transformer modules, and $f^{(l)}{}_{ctrl}$ outputs fine-grained structural features at each stage, which are superimposed onto the UNet backbone features to achieve pixel-level structural rectification. The second path is the semantic fusion path, in which deep coupling between silhouette embeddings and textual semantic features is accomplished through eight-head cross-attention. The fusion formula is defined as:

$$c_{joint}=Concat_{\mathrm{h}=1}^{8}\left[softmax\left(\frac{(e_{text}W_Q^{(h)})(e_{sil}W_K^{(h)})^{\top}}{\sqrt{d_k}}\right)(e_{sil}W_V^{(h)})\right]W_O \tag{8}$$

where, e_{text} denotes the 768-dimensional contrastive language–image pre-training textual semantic features, and W_Q, W_K, and W_V denote the attention query, key, and value projection matrices, respectively, d_k denotes the single-head attention dimension with a value of 16, and W_O denotes the feature dimension mapping matrix. The joint features obtained through fusion replace the original text conditions and are fed into the UNet cross-attention layers, enabling coordinated regulation of semantic information and geometric structure.

Model training and image inference were performed using the standard latent diffusion iterative mechanism. In the forward diffusion process, Gaussian noise was added to the clean latent variables, with the noise iteration formula defined as:

$$z_t=\sqrt{\bar{\alpha}_t}z_0+\sqrt{1-\bar{\alpha}_t}\epsilon \tag{9}$$

where, z_0 denotes the original noise-free latent variable; z_t denotes the noisy latent variable at step t; $\bar{\alpha}_t$ denotes the linear noise schedule coefficient; and ϵ denotes Gaussian noise following a standard normal distribution. With noise prediction accuracy serving as the optimization objective, the base diffusion loss function was defined as:

$$L_{simple}=E_{z_0,\epsilon,t,c_{joint}}\left[\left\|\epsilon-\epsilon_\theta(z_t,t,c_{joint})\right\|_2^2\right] \tag{10}$$

During the inference phase, the denoising diffusion implicit models deterministic sampling strategy was adopted, with a pure Gaussian noise latent variable $z_T\sim N(0,I)$ serving as the initial input, the sampling steps set to 50, and the deterministic sampling coefficient $\sigma_t = 0$. The optimal latent variable z_0 was recovered through reverse iterative denoising, and was subsequently mapped to the pixel-space image $I_{gen} = D(z_0)$ via the variational autoencoder decoder D, yielding high-quality garment generation results with precise silhouette geometric structures.

2.4 Silhouette-preserving style editing framework

A silhouette-preserving style editing framework was constructed, in which a structure-faithful texture update mechanism was established for personalized garment editing tasks, enabling semantic-driven editing of fabric, color, and detail styles without altering the original garment silhouette structure. The schematic diagram of the silhouette-preserving style editing framework is illustrated in Figure 3. The editing inputs to the framework are predefined as the original garment image and textual editing instructions, with the core optimization objective being to ensure that the edited image fully conforms to the textual semantic attributes while maintaining consistency in the geometric morphology of the garment outer contour. The framework first performs structured image decomposition and latent space inversion preprocessing. Garment semantic segmentation was accomplished via a DeepLabV3+ network fine-tuned on the FashionGen dataset, from which the foreground mask of the image was precisely extracted. Outer contour extraction and hole filling were achieved through morphological closing operations with a kernel size set to 5 × 5, ultimately generating a silhouette mask with complete closed structure for subsequent global structural constraints. To align with the latent editing paradigm of diffusion models, the denoising diffusion implicit models inversion algorithm was employed to map the original image into the latent space, with the number of iterative steps set to 50, and a continuous latent variable trajectory was generated via the reverse update formula:

$$z_{t+1}^{inv}=\sqrt{\bar{\alpha}_{t+1}}\left(\frac{z_t^{inv}-\sqrt{1-\bar{\alpha}_t}\epsilon_\theta(z_t^{inv},t,c_{orig})}{\sqrt{\bar{\alpha}_t}}\right)+\sqrt{1-\bar{\alpha}_{t+1}}\epsilon_\theta(z_t^{inv},t,c_{orig}) \tag{11}$$

where, z_t^{inv} denotes the latent variable at the t-th reverse iteration step, $\bar{\alpha}_t$ denotes the linear noise schedule coefficient, ϵ_θ denotes the noise prediction network, and c_{orig} denotes the textual embedding features corresponding to the original image, encoded by the bootstrapping language–image pre-training model. The latent variable at the end of the iteration serves as the initial input for subsequent style editing.

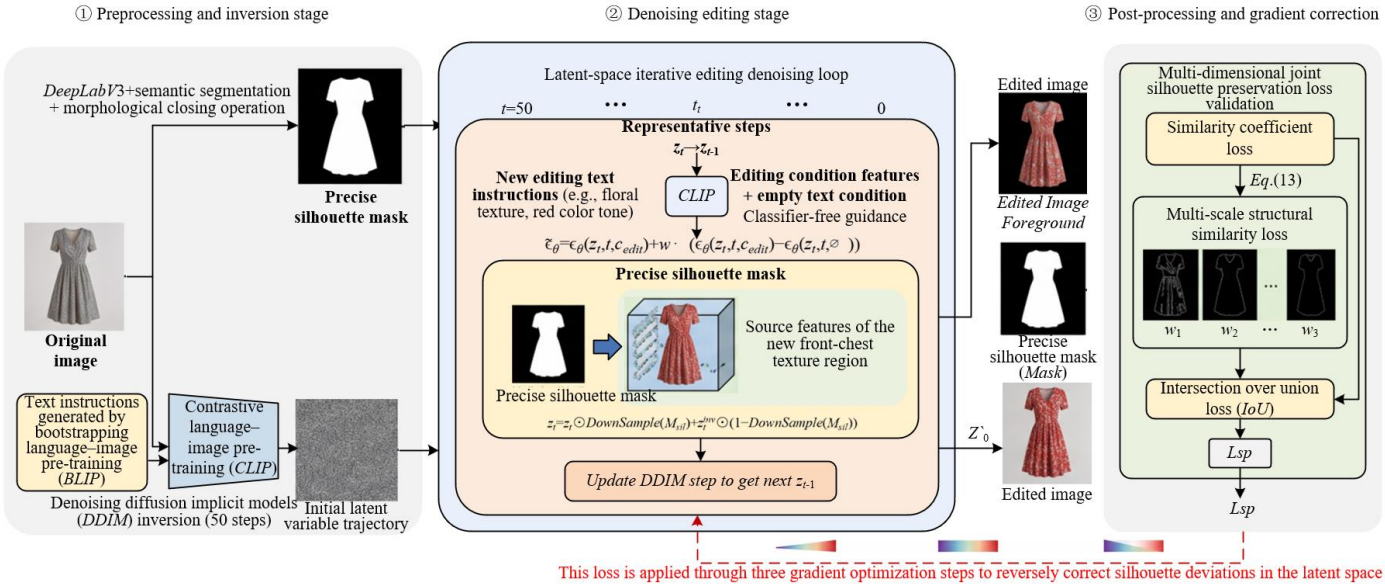


Figure 3. Schematic diagram of the silhouette-preserving style editing framework

During the latent-space editing stage, new textual condition features were generated based on the editing text instructions, and controllable updates of style textures were accomplished through a dual-constraint mechanism. The editing instructions were fed into the contrastive language–image pre-training text encoder to obtain the editing condition features c_{edit}. At each denoising iteration step, a spatial mask constraint was introduced, by which the latent features of the background region were fixed, and texture updates were permitted only within the garment silhouette region. The latent variable update rule was defined as:

$$z_t=z_t\odot DownSample(M_{sil})+z_t^{inv}\odot(1\text{-}DownSample(M_{sil})) \tag{12}$$

where, *DownSample* denotes the bilinear interpolation downsampling operation, by which the silhouette mask was scaled to a latent-space resolution of 64 × 64, and M_{sil} denotes the silhouette mask of the original image. Concurrently, a classifier-free guidance strategy was introduced to enhance the style editing effect, with the optimized noise prediction expressed as:

$$\tilde{\epsilon}_\theta=\epsilon_\theta(z_t,t,c_{edit})+w\cdot(\epsilon_\theta(z_t,t,c_{edit})-\epsilon_\theta(z_t,t,\varnothing)) \tag{13}$$

where, w denotes the guidance weight, fixed at 3.0, and $\varnothing$ denotes the empty text condition. Through the amplification of the difference between conditional and unconditional noise predictions, the guiding effect of textual editing semantics was enhanced, enabling precise updates of texture styles.

To thoroughly eliminate contour displacement and structural distortion arising during iterative editing, a multi-dimensional joint silhouette preservation loss function was constructed, by which global rectification of the editing results was performed at both the pixel level and the edge structure level. The overall loss function is expressed as:

$$L_{sp}=Dice(SegNet(\tilde{I}),M_{sil}) +\lambda_1 \cdot MS\text{-}SSIM\big(Edge(\tilde{I}),Edge(I_{in}\odot M_{sil})\big) +\lambda_2 \cdot (1\text{-}IoU(SegNet(\tilde{I}),M_{sil})) \quad (14)$$

where, *Dice* denotes the similarity coefficient loss, employed to constrain the pixel-wise overlap between the edited image and the original silhouette, with a tiny constant $\epsilon = 10^{-6}$ introduced during computation to prevent division by zero; *MS-SSIM* denotes the multi-scale structural similarity loss, employed to constrain the morphological consistency of garment edge contours, with five scales configured in the model and corresponding weights of 0.0448, 0.2856, 0.3001, 0.2363, and 0.1333, respectively, and edge detection performed via the Canny algorithm with high and low thresholds set to 150 and 50, respectively; *IoU* denotes the intersection over union loss, serving as an auxiliary constraint to reinforce pixel-level matching. The weight coefficients λ_1 and λ_2 were set to 0.3 and 0.2, respectively, balancing structural edge constraints and regional matching constraints. During the inference phase, latent-space gradient optimization was performed based on this loss function, with the number of iterations set to 3 and the learning rate fixed at 0.05, through which silhouette deviations were further converged via reverse gradient correction.

This framework establishes a three-level explicit shape-preserving mechanism, comprising preprocessing mask constraints, iterative denoising spatial constraints, and post-processing loss correction, which fundamentally differs from conventional methods that rely on implicit model regularization for structural maintenance during editing. Throughout the entire process, the standardized silhouette mask serves as the fixed constraint benchmark, exerting rigid control over the texture update process, thereby thoroughly circumventing silhouette deformation issues induced by texture transfer and semantic editing, and achieving a bidirectional balance between free garment style editing and absolute contour fidelity.

2.5 Multi-modal silhouette style disentanglement generation strategy

To address the issue of garment silhouette, texture, and color features being highly coupled in the latent space and thus incapable of independent modulation, a multi-modal silhouette style disentanglement generation strategy was designed, in which three mutually independent feature sub-spaces were constructed to enable the decomposed representation and free recombination of core garment visual attributes. The architecture of the multi-modal silhouette style disentanglement generation strategy is illustrated in Figure 4. This strategy relies on three dedicated encoders for disentangled feature extraction, corresponding to silhouette structure, fabric texture, and color attributes, respectively. The silhouette encoder adopts the structured multi-layer perceptron architecture, with two additional geometric ratio features—shoulder-to-hem width ratio and waist-to-hip ratio—incorporated on the basis of the original 20-dimensional keypoint coordinates, yielding a 22-dimensional input vector, from which a 128-dimensional silhouette representation is output through a dimensional transformation network, thereby enhancing the model's perception of global geometric morphology. The texture encoder takes the pure texture image within the garment mask region as input, with frozen parameters of an ImageNet-pre-trained ResNet-50 employed to extract base texture features, and a two-layer projection network used to map the 2048-dimensional high-dimensional features to a 128-dimensional unified representation space, with only the projection layer parameters being fine-tuned to preserve general texture priors. The color encoder computes the Red–Green–Blue (RGB) mean, standard deviation, skewness, and kurtosis of the effective garment region, which are concatenated into a 12-dimensional color statistical vector, from which a 128-dimensional color representation is output through a lightweight mapping network, precisely quantifying the overall color distribution characteristics of the garment. The output dimensions of the three encoders are unified, providing a standardized representation foundation for subsequent sub-space orthogonal disentanglement and feature fusion.

To achieve complete disentanglement across different feature dimensions, an enhanced subspace orthogonality loss function was constructed, constraining the feature distribution from two perspectives—cross-subspace independence and intra-subspace dimensional independence—thereby thoroughly eliminating feature redundancy and coupling interference. The loss function is defined as:

$$L_{orth}=\left\|Z_{sil}^{\top}Z_{tex}\right\|_F^2+\left\|Z_{sil}^{\top}Z_{col}\right\|_F^2 +\left\|Z_{tex}^{\top}Z_{col}\right\|_F^2+\beta\sum_{k\in\{sil,tex,col\}}\left\|Z_k^{\top}Z_k\text{-}I\right\|_F^2 \quad (15)$$

where, Z_{sil}, Z_{tex}, and Z_{col} denote the feature matrices of silhouette, texture, and color within a batch, respectively, with matrix dimensions of batch size multiplied by 128, and each row corresponding to the feature vector of a single sample; $\|\|_F^2$ denotes the squared Frobenius norm, employed to measure the overall correlation between matrices; the first three terms constrain the cross-correlation between any two feature sub-spaces, forcing features of different attributes to be mutually independent; the final term serves as the subspace orthogonality regularization term, constraining the linear independence of dimensions within each individual feature space, with the parameter β set to 0.1 to balance the constraint strengths between cross-space disentanglement and intra-space regularization. This loss effectively decouples the implicit correlations among the three visual attributes, constructing statistically fully orthogonal feature sub-spaces.

During the model inference phase, the framework achieves compositional generation from cross-source conditions based on the disentangled sub-spaces, supporting free combination of silhouette, texture, and color features. The independently obtained 128-dimensional silhouette, texture, and color representations were concatenated into a 384-dimensional combined conditional feature, which replaced the traditional single text condition as input to the diffusion model. To accommodate the multi-condition fusion requirement, the UNet cross-attention module was reconstructed, with a three-branch independent conditional attention mechanism designed, in which dedicated projection matrices were configured for each of the three feature types to accomplish independent mapping of query, key, and value vectors. The attention features from each branch were adaptively fused via learnable gating weights, with the fusion formula defined as follows:

$$f_{out}=\sum_{m\in\{sil,tex,col\}}\gamma_m\cdot softmax\left(\frac{QK_m^{\top}}{\sqrt{d_k}}\right)V_m \tag{16}$$

where, Q denotes the query vector obtained from UNet image feature mapping, K_m and V_m denote the key and value vectors corresponding to the three disentangled feature types, respectively, d_k denotes the single-head attention dimension, and γ_m denotes the learnable gating weights, with initial mean allocation set to one-third each, which can be adaptively adjusted during training to regulate the generation weights of each visual attribute, achieving fine-grained collaborative regulation of multiple conditions.

Leveraging the linear properties of the orthogonal disentangled sub-spaces, this strategy enables continuous smooth interpolation generation of garment silhouette styles. Linear weighted fusion was performed on two different silhouette representations within the silhouette feature sub-space, with the interpolation formula defined as:

$$z_{sil}^{(\alpha)}=(1-\alpha)\cdot z_{sil}^{(A)}+\alpha\cdot z_{sil}^{(B)} \tag{17}$$

where, α denotes the interpolation coefficient, with a value range of 0 to 1, through which a gradual transition between the two silhouette morphologies is achieved via uniform sampling. Owing to the strict orthogonal constraints imposed on the sub-spaces, the feature space possesses a stable linear structure, and the interpolated intermediate samples can maintain semantic continuity and structural integrity without morphological distortion or feature entanglement, thereby enabling the generation of rich transitional silhouette styles and substantially expanding the creative space for personalized garment design.

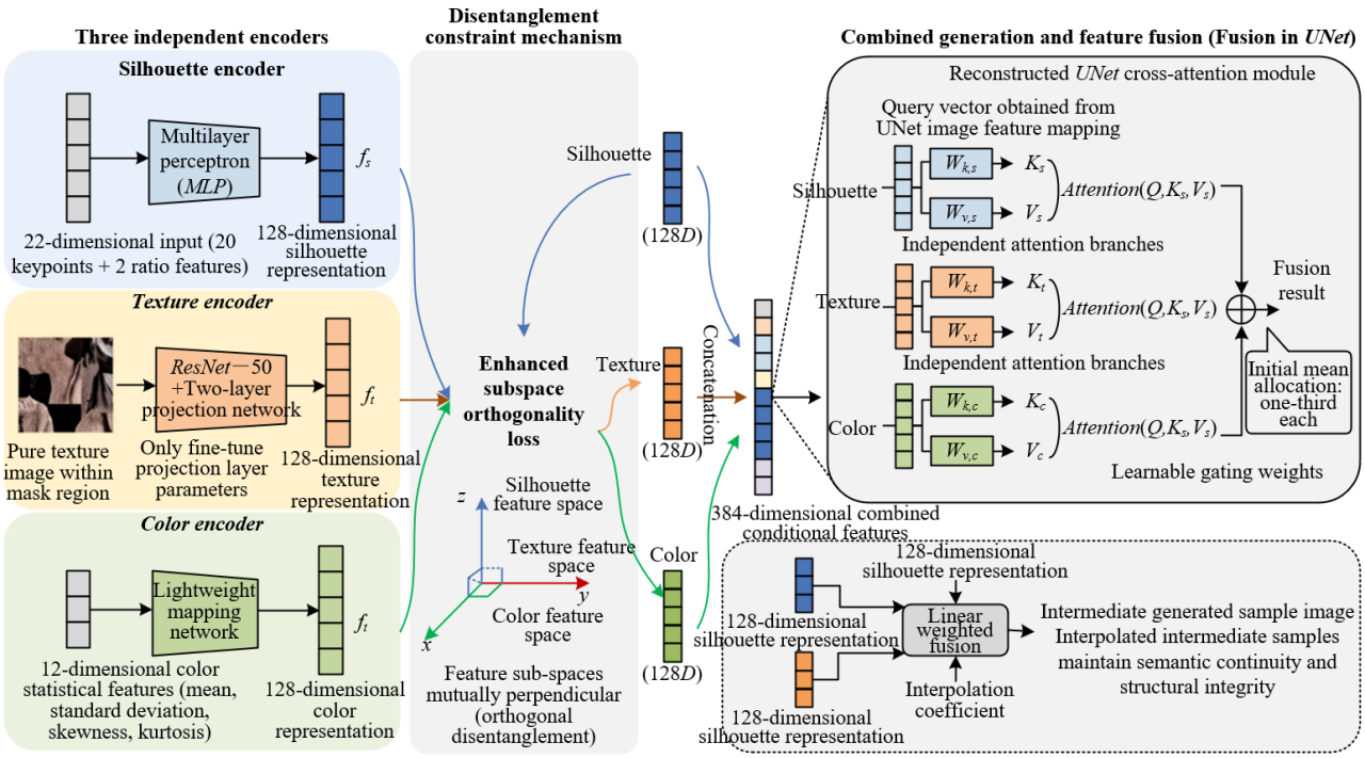


Figure 4. Architecture of the multi-modal silhouette style disentanglement generation strategy

2.6 Joint training and loss balancing

All functional modules in this study share a unified diffusion generation backbone. To achieve multi-task collaborative optimization and address the issues of gradient imbalance and convergence instability arising from the superposition of multiple constraint losses, a multi-dimensional integrated global loss function was constructed, enabling joint iterative optimization of model parameters. The global loss function integrates the base diffusion generation constraint, cross-modal semantic alignment constraint, silhouette structural fidelity constraint, and feature disentanglement constraint, while a regularization mechanism is concurrently introduced to suppress overfitting risks in the newly added modules. The specific expression is as follows:

$$L_{total}=L_{simple}+\lambda_1 L_{align}+\lambda_2 L_{text}+\lambda_3 L_{sp}+\lambda_4 L_{orth}+\lambda_5\|\Theta_{new}\|_2^2 \tag{18}$$

where, L_{simple} denotes the standard diffusion noise prediction loss, responsible for maintaining the base visual quality and distributional plausibility of image generation. L_{align} and L_{text} denote the silhouette semantic alignment loss and the textual semantic matching loss, respectively, employed to constrain the semantic consistency of cross-modal features. L_{sp} denotes the silhouette preservation loss, employed to ensure the structural integrity of garment contour during style editing. L_{orth} denotes the subspace orthogonality loss, achieving disentangled representation of silhouette, texture, and color features. The final regularization term is the L2 weight decay constraint, applied exclusively to the newly added network parameters Θ_{new}, with the hyperparameter λ_5 fixed at 10^{-5}, through which model generalization capability is enhanced by constraining parameter magnitudes.

To balance the constraint weights of different loss terms and accommodate the optimization priorities of each sub-task, a hierarchical grid search strategy was employed for

hyperparameter calibration. Given that the numerical magnitudes of the semantic alignment loss and the text matching loss were similar, the weight parameters λ_1 and λ_2 were fixed at 0.1. Under this premise, a two-dimensional grid search was conducted for the structure preservation weight λ_3 and the feature disentanglement weight λ_4, with the search range covering 0.01, 0.05, 0.1, and 0.2. The harmonic mean of silhouette structural similarity and silhouette classification accuracy on the validation set was adopted as the optimization evaluation metric, and the optimal weight combination was ultimately determined as $\lambda_3 = 0.05$ and $\lambda_4 = 0.05$. Model training was divided into two progressive stages. In the first stage, only the base diffusion loss and the semantic constraint losses were enabled, prioritizing the convergence of the model's fundamental generation capability and cross-modal alignment capability. In the second stage, all loss terms were introduced to achieve full-module collaborative optimization. Additionally, a gradient clipping threshold of 1.0 was set during training, effectively avoiding the gradient explosion problem in multi-loss joint optimization scenarios and ensuring stable convergence throughout the overall training process.

3. EXPERIMENTS

In this chapter, the comprehensive performance of the proposed SilhouetteDiff framework was validated through multiple sets of quantitative and qualitative experiments, systematically verifying the effectiveness of the four core modules: semantic alignment, structure prior guidance, silhouette-preserving editing, and multi-modal disentangled generation. The experiments encompassed comparative testing on mainstream garment datasets, module ablation analysis, fine-grained control precision testing, style editing performance validation, disentangled generation effectiveness evaluation, subjective user studies, and model efficiency analysis, comprehensively demonstrating the superiority and robustness of the proposed method in tasks of precise garment silhouette control, personalized style generation, and structure-faithful editing.

3.1 Datasets and preprocessing

Two publicly available garment datasets, namely FashionGen and DeepFashion2, were employed for model training and evaluation in this study, complemented by manual annotation expansion to construct a comprehensive garment silhouette generation dataset system. The FashionGen dataset contains 293,008 high-resolution garment images with fine-grained attribute annotations including silhouette, sleeve type, fabric, and garment length, covering five standard garment silhouette types: A-type, H-type, X-type, O-type, and T-type. A total of 152,347 samples with complete silhouette annotations were selected as the base training data, with 19,043 validation set samples and 19,043 test set samples allocated according to standard partitioning rules. The DeepFashion2 dataset contains 491,436 garment images under multi-pose and multi-occlusion scenarios, with detailed garment position and category annotations, but lacks standardized silhouette labels. A total of 50,000 samples were randomly selected from this dataset, and silhouette classification annotations were independently completed by three garment design professionals, with the final annotation results determined through a majority voting mechanism, achieving a final annotation consistency Kappa coefficient of 0.89, indicating high annotation reliability. After integration of the two datasets, the final training set comprised approximately 180,000 samples, with 25,000 samples allocated to each of the validation and test sets, ensuring balanced and diverse data distribution.

During the data preprocessing stage, image input specifications were uniformly standardized, with all images ultimately cropped to 512 × 512 resolution via center cropping while preserving the original aspect ratio; high-resolution FashionGen images were pre-downsampled to 1024 × 1024 prior to uniform cropping. Garment structural keypoints were automatically detected using a pre-trained HRNet-W48, with keypoint completion performed via silhouette-category template interpolation algorithms for occluded or detection-failed samples, thereby ensuring the integrity of structural skeleton data. Textual descriptions were uniformly processed using the contrastive language–image pre-training ViT-L/14@336px tokenizer, with the maximum text sequence length set to 77. During the training phase, standardized data augmentation strategies were introduced, including random horizontal flipping with a probability of 0.5 and mild color jittering, with brightness and saturation perturbation ranges both set to ±0.1, effectively enhancing model generalization capability.

A multi-dimensional evaluation metric system was adopted to achieve quantitative assessment of model performance, encompassing objective generation quality, structural precision, semantic matching degree, and subjective visual experience. A silhouette structure similarity metric was proposed, integrating mask intersection over union and the normalized Hausdorff distance to achieve refined evaluation of garment contour accuracy, computed as the product of mask intersection over union and the unitized boundary distance difference, thereby simultaneously constraining regional overlap and contour boundary deviation. The Fréchet inception distance metric was employed to measure the data distribution discrepancy between generated images and real images, with lower values indicating higher overall generation quality. Contrastive language–image pre-training-score was employed to characterize the semantic matching degree between images and textual instructions, quantified via the mean cosine similarity of image and text features. Silhouette classification accuracy was computed using a ResNet-50 classifier fine-tuned on FashionGen, with classification prediction accuracy reflecting silhouette generation precision. Additionally, a professional user study was introduced, in which five-point Likert scale subjective scores were completed from three dimensions—silhouette accuracy, visual realism, and style diversity—enabling complementary validation of objective and subjective evaluations.

3.2 Comparative experiments

To validate the advancement of the proposed framework, five state-of-the-art garment generation and editing methods from recent years were selected as comparative baselines, including Stable Diffusion+ControlNet (Canny), IMAGGarment, MAGDiff, Texture-Preserving Multimodal Garment Designer (TP-MGD), and FS-control. All comparative methods were retrained or fine-tuned under the unified dataset environment of this study; open-source methods were implemented with official code and default

hyperparameters, while non-open-source methods were strictly reproduced according to the algorithmic details provided in their respective papers, ensuring fairness in the comparative experiments. The quantitative comparison results of all methods on the FashionGen test set are presented in Table 1.

Table 1. Quantitative comparison of different methods on the FashionGen test set

Method	Fréchet Inception Distance↓	Silhouette Structure Similarity↑	Silhouette Classification Accuracy (%)↑	Contrastive Language–Image Pre-Training-Score↑
Stable Diffusion+ControlNet (Canny)	18.7	0.803	78.3	0.287
IMAGGarment	16.2	0.849	85.1	0.312
MAGDiff	15.8	0.871	87.6	0.325
Texture-Preserving Multimodal Garment Designer (TP-MGD)	17.3	0.825	81.9	0.301
FS-control	14.9	0.858	83.7	0.318
Proposed SilhouetteDiff	12.4	0.923	91.7	0.342

As shown in Table 1, the proposed method achieved optimal results across all evaluation metrics. In terms of global generation quality, the Fréchet inception distance value of the proposed method was 12.4, which is 21.5% lower than that of the best baseline, MAGDiff, with an even more substantial reduction observed relative to the edge-constrained ControlNet baseline, demonstrating that the framework effectively reduces the data distribution discrepancy between generated and real images. In terms of silhouette structural control precision, the silhouette structure similarity metric of the proposed method reached 0.923, significantly outperforming all comparative methods, verifying that structured skeleton priors, in contrast to conventional edge features, enable more precise constraint of the overall geometric morphology and contour structure of garments. The silhouette classification accuracy reached 91.7%, an improvement of 13.4 percentage points over the ControlNet baseline, demonstrating that the fine-grained cross-modal semantic alignment mechanism effectively distinguishes subtle silhouette semantic differences and enhances the generation accuracy of target silhouettes. The optimal contrastive language–image pre-training-score further indicates that multi-module collaborative constraints strengthen the matching degree between textual semantics and visual generation results. Although the FS-control method achieved a relatively low Fréchet inception distance value, its silhouette structure similarity and silhouette classification accuracy metrics were comparatively poor, demonstrating that the hybrid diffusion-generative adversarial network optimization strategy can only improve image visual quality, but fails to address the core problem of precise garment silhouette control. The generalization performance of all methods on the complex-scenario dataset DeepFashion2 is presented in Table 2.

As shown in Table 2, under the more challenging DeepFashion2 test scenarios characterized by complex poses and stronger background interference, the overall performance of all methods exhibited a modest decline; nevertheless, the proposed method still maintained a significant performance advantage. The Fréchet inception distance value was reduced by 18.2% relative to MAGDiff, while the silhouette structure similarity and silhouette classification accuracy metrics were improved by 7.3% and 6.2 percentage points, respectively, demonstrating that the proposed framework possesses excellent scene generalization capability and structural control robustness, and is well-suited for garment silhouette generation tasks in complex real-world scenarios.

Table 2. Generalization performance of different methods on the DeepFashion2 test set

Method	Fréchet Inception Distance↓	Silhouette Structure Similarity↑	Silhouette Classification Accuracy (%)↑
Stable Diffusion+ControlNet (Canny)	21.3	0.772	72.6
MAGDiff	18.1	0.835	81.3
Proposed SilhouetteDiff	14.8	0.896	87.5

3.3 Ablation experiments

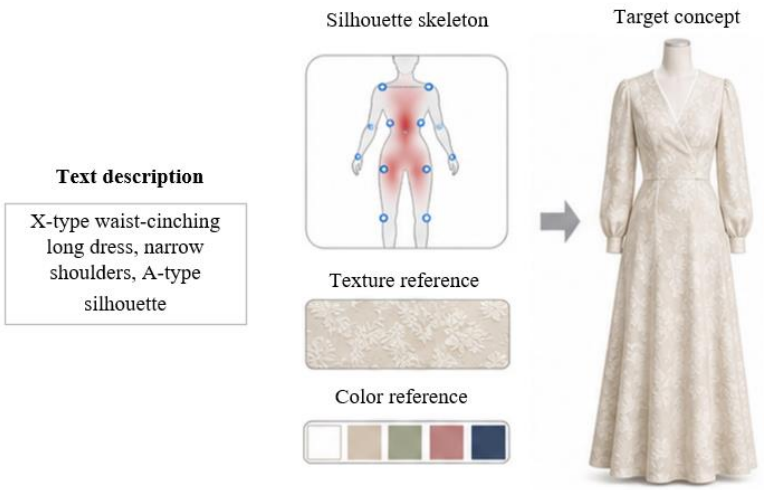
To quantitatively verify the independent contributions of the four core modules, five sets of ablation control experiments were designed, in which the semantically aware silhouette cross-modal alignment module, the SSPGM, the silhouette-preserving style editing framework, and the multi-modal silhouette style disentanglement generation strategy were individually removed, establishing a comparative system between the full model and four ablated variants. The experimental results are presented in Table 3.

Table 3. Quantitative results of ablation experiments (FashionGen test set)

Variant	Fréchet Inception Distance↓	Silhouette Structure Similarity↑	Silhouette Classification Accuracy (%)↑	Contrastive Language–Image Pre-Training-Score↑
Without the semantically aware silhouette cross-modal alignment module	14.3	0.905	79.3	0.301
Without the silhouette structure prior guidance module (SSPGM)	17.8	0.836	84.1	0.318
Without the silhouette-preserving style editing framework	13.1	0.856	90.2	0.338
Without the multi-modal silhouette style disentanglement generation strategy	13.8	0.91	89.5	0.329
Full model	12.4	0.923	91.7	0.342

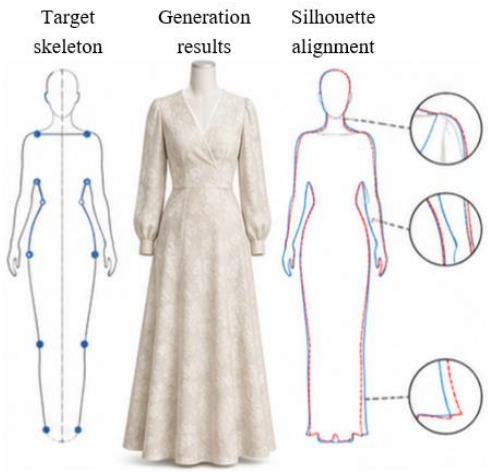
The ablation results demonstrate that each module possesses clear functional specificity and irreplaceability. After the removal of the semantically aware silhouette cross-modal alignment module, the silhouette classification accuracy exhibited the most substantial degradation, dropping from 91.7% to 79.3%, accompanied by a significant decline in contrastive language–image pre-training-score, demonstrating that the silhouette semantic perception alignment module is the core component for achieving fine-grained silhouette semantic matching, and that the generic contrastive language–image pre-training embedding space is incapable of precisely distinguishing the subtle semantic differences among various garment silhouette categories. After the removal of the SSPGM, the silhouette structure similarity metric decreased substantially while the Fréchet inception distance increased markedly, demonstrating that the structured prior guidance mechanism is the key to ensuring garment silhouette geometric fidelity and optimizing global generation quality, and that explicit skeleton constraints effectively circumvent structural distortion and contour deviation issues. After the removal of the silhouette-preserving style editing framework, a notable degradation in silhouette structure similarity was observed, indicating that the silhouette preservation loss imposes effective regularization constraints on the generation process, continuously optimizing contour integrity. After the removal of the multi-modal silhouette style disentanglement generation strategy module, all metrics exhibited minor declines; this module primarily functions in cross-attribute compositional generation tasks, with auxiliary optimization effects on base generation performance. Overall, the semantically aware silhouette cross-modal alignment module governs semantic alignment precision, while the SSPGM governs geometric structure fidelity; these two modules constitute the core support of the framework, with the remaining modules further refining the fine-grained capabilities of generation and editing.

To verify the end-to-end effectiveness of the proposed method in cross-modal condition parsing, garment silhouette generation, structure-preserving editing, and multi-attribute compositional design, the implementation effect experiments illustrated in Figure 5 were conducted. Figure 5(a) demonstrates that the model jointly parses textual descriptions, silhouette skeletons, texture references, and color information, mapping design conditions from different modalities into a unified representation space; the generation results remain consistent with the input intentions in terms of waist cinching degree, shoulder morphology, hemline trajectory, and surface style, indicating that the cross-modal visual-semantic alignment mechanism accurately associates garment-specific semantics with observable structural features. In Figure 5(b), A-type, H-type, X-type, O-type, and T-type garments exhibit clear and stable differences in shoulder-to-waist proportions, side-seam orientations, and hem expansion degrees, without noticeable confusion between adjacent silhouette categories, demonstrating that fine-grained semantic embeddings enhance the model's discriminative and generative capabilities for similar silhouette concepts. Figure 5(c) shows that the target skeleton, generated garments, and overlaid contour results exhibit high consistency at key structural regions including shoulder points, waist points, hip points, and hem regions, with no significant edge drift observed in locally magnified areas, indicating that geometric structure priors effectively constrain global proportions and local contours, reducing structural distortion during the diffusion generation process. In Figure 5(d), texture replacement, color adjustment, and pattern addition all achieve distinct style variations, while garment shoulder width, waistline position, sleeve boundaries, and hem morphology remain largely stable, demonstrating that the silhouette-preserving editing mechanism suppresses structural information leakage during visual attribute updates, achieving coordinated harmony between style variation and contour fidelity. Figure 5(e) further demonstrates that the model separately extracts silhouette, texture, and color features from garments of different sources, and accurately inherits the respective target attributes in the combined results; meanwhile, the continuous interpolation from X-type to A-type exhibits smooth, gradual, and mutation-free morphological transitions, demonstrating that the disentangled feature sub-spaces possess favorable independence, continuity, and composability. Synthesizing the results across all sub-figures, the proposed method effectively integrates the entire pipeline from design semantic understanding, silhouette structure generation, personalized style editing, to multi-source attribute recombination, providing reliable technical support for precise and controllable garment silhouette design and personalized style generation.



(a) Cross-modal design condition input

(b) Fine-grained silhouette semantic generation results



(c) Silhouette geometric structure alignment results

(d) Silhouette-preserving personalized style editing

(e) Disentangled composition and continuous silhouette innovation

Figure 5. Implementation effect visualization of cross-modal semantic-driven garment silhouette generation, shape-preserving editing, and compositional innovation

3.4 Silhouette control precision experiment

To verify the fine-grained controllability of the SSPGM over garment geometric parameters, a waist circumference parameter continuous control experiment was designed. In this experiment, the garment text description and the remaining skeletal structure parameters were fixed, while the normalized waist width was uniformly sampled within the range of 0.30 to 0.70 at a step size of 0.05, generating multiple sets of garment images with varying waist dimensions. The actual waist dimensions of the generated images were inversely measured using a segmentation network, with Stable Diffusion+ControlNet (Canny) adopted as the comparative method to quantify the parameter control precision of the two approaches. The experimental results are illustrated in Figure 6.

Quantitative analysis results demonstrate that the predicted waist circumference of the proposed method exhibits a Pearson correlation coefficient of 0.994 with the target waist circumference, a linear regression slope of 0.985, and an intercept approaching zero, indicating nearly perfect linear fitting characteristics. In contrast, the ControlNet baseline achieved a correlation coefficient of only 0.847, with significantly larger linear fitting deviations. Under extreme waist circumference scenarios, the deviations of the baseline method were further amplified, with positive and negative offsets observed in narrow-waist and wide-waist scenarios, respectively. These results demonstrate that conventional edge-based feature constraints fail to achieve independent modulation of individual geometric parameters, whereas the proposed explicit encoding mechanism based on structured skeletons enables precise and decoupled control of local garment geometric parameters, substantially enhancing the fine-grained level of silhouette regulation.

3.5 Silhouette-preserving editing experiment

To verify the silhouette fidelity capability of the silhouette-preserving style editing framework in style editing tasks, 500 test set images were selected, and three typical editing tasks—texture replacement, color adjustment, and pattern addition—were sequentially performed. The editing effects of three methods, namely the silhouette-preserving style editing framework, TP-MGD, and PromptDresser, were compared, with performance evaluated via mask intersection over union, silhouette structure similarity, and user subjective scores. The experimental results are presented in Table 4.

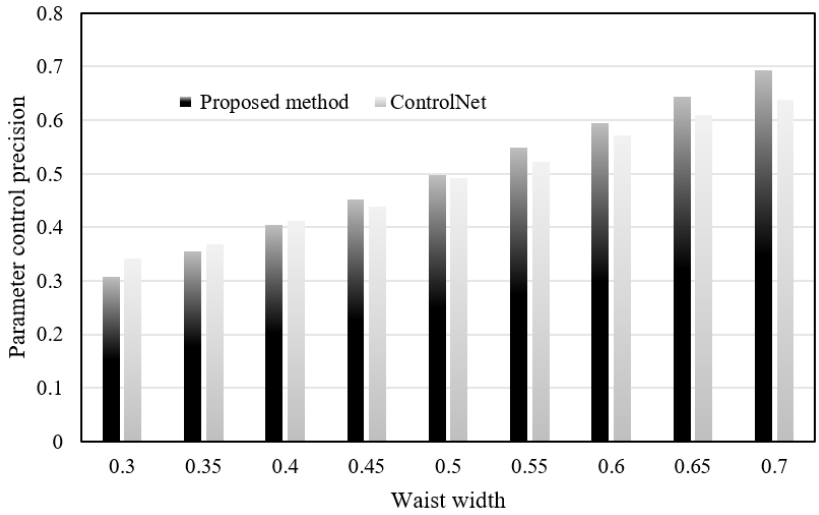


Figure 6. Experimental results of waist circumference control accuracy

Table 4. Experimental results of silhouette-preserving editing (mean ± standard deviation)

Method	Editing Task	Intersection Over Union↑	Silhouette Structure Similarity↑	User Satisfaction (1-5)↑
Proposed silhouette-preserving style editing framework	Texture replacement	0.952 ± 0.018	0.941 ± 0.021	4.6 ± 0.4
	Color adjustment	0.961 ± 0.015	0.949 ± 0.018	4.5 ± 0.5
	Pattern addition	0.947 ± 0.022	0.932 ± 0.025	4.4 ± 0.5
Texture-Preserving Multimodal Garment Designer (TP-MGD)	Texture replacement	0.891 ± 0.035	0.873 ± 0.038	3.8 ± 0.6
	Color adjustment	0.904 ± 0.030	0.885 ± 0.033	3.9 ± 0.6
	Pattern addition	0.882 ± 0.042	0.859 ± 0.045	3.6 ± 0.7
PromptDresser	Texture replacement	0.867 ± 0.048	0.851 ± 0.050	3.5 ± 0.7
	Color adjustment	0.879 ± 0.041	0.862 ± 0.044	3.7 ± 0.6
	Pattern addition	0.853 ± 0.055	0.832 ± 0.057	3.3 ± 0.8

The experimental results demonstrate that the proposed method maintains optimal silhouette fidelity performance and user experience across all three editing tasks. In high-frequency pattern addition tasks, baseline methods were prone to contour blurring and boundary distortion, with substantial degradation in intersection over union and silhouette structure similarity metrics; in contrast, the proposed method, leveraging the triple shape-preserving mechanism comprising inversion mask constraints, denoising spatial hard constraints, and multi-dimensional silhouette loss correction, effectively suppresses contour drift during editing, with all metrics consistently maintained at high levels. The subjective user scores were highly consistent with the objective metrics, with the proposed method achieving scores above 4.4 across all tasks, significantly outperforming the baseline methods. Reviewer feedback indicated that baseline methods tended to compromise the original silhouette structure while updating garment textures and colors, whereas the proposed method fully preserves the original contour proportions and morphological characteristics while accomplishing precise style editing, achieving a bidirectional balance between structural fidelity and style renewal.

3.6 Compositional generation and disentanglement effectiveness experiment

To verify the feature disentanglement capability and cross-attribute compositional generation performance of the multi-modal silhouette style disentanglement generation strategy module, multiple sets of control experiments were designed, with effectiveness validation conducted from three dimensions: compositional generation precision, subspace orthogonality, and style interpolation continuity. The quantitative results of cross-source compositional generation are presented in Table 5. In the experiment, 200 sets of image triplets were randomly selected, from which independent silhouette, texture, and color features were respectively extracted for recombination generation, with a direct feature concatenation method without orthogonal constraints adopted as the baseline.

Table 5. Quantitative results of compositional generation

Method	Silhouette Structure Similarity-A↑	Texture Similarity↑	ΔE_{2000} (vs. C)↓	User Score (1-5)↑
Concat (without Lorth)	0.801	0.742	8.3	3.2 ± 0.8
Proposed multi-modal silhouette style disentanglement generation strategy	0.912	0.853	3.7	4.7 ± 0.4

As shown in Table 5, the orthogonal disentanglement constraint significantly improves the precision of cross-attribute compositional generation. The proposed method achieves substantial improvements in silhouette matching degree and texture similarity, with a marked reduction in color difference, and a user score improvement of 1.5 points. The unconstrained feature concatenation approach suffers from severe feature coupling issues, where features of different attributes interfere with one another, giving rise to defects including silhouette distortion, texture mingling, and color shift. In contrast, the orthogonal loss constraint enforces statistical independence among the three feature sub-spaces,

ensuring independent modulation of each attribute feature without mutual interference, thereby achieving precise multi-attribute compositional generation.

The subspace orthogonality test results demonstrate that the average cosine similarity among the three feature sub-spaces of the full model was below 0.075, approaching a completely orthogonal state. In contrast, the ablation model without orthogonal constraints exhibited feature similarities exceeding 0.35, indicating severe feature coupling issues. These results directly demonstrate that the orthogonal loss function effectively decouples the implicit correlations among different visual attributes, establishing independent and stable feature sub-spaces. The style interpolation experiment results show that during the continuous interpolation process from X-type silhouette to A-type silhouette, the coefficient of determination (R^2) for linear fitting between silhouette classification probability and the interpolation coefficient reached 0.971, demonstrating excellent linear correlation. This result demonstrates that the disentangled silhouette sub-space possesses a stable linear structure capable of supporting smooth and continuous style transition generation, enabling the output of rich transitional silhouette styles and effectively expanding the creative space for personalized garment design.

3.7 User study

To objectively evaluate the practical application value of the model, 20 garment design professionals with over five years of industry experience were invited to conduct a double-blind paired subjective evaluation. Five-point Likert scale scores were assigned to the generation results of all comparative methods from three dimensions: silhouette accuracy, visual realism, and style diversity. The evaluation results are illustrated in Figure 7.

The subjective evaluation results were fully consistent with the objective experimental metrics, with the proposed method achieving optimal scores across all evaluation dimensions. In the core dimension of silhouette accuracy, the proposed method outperformed the best baseline by 0.5 points, with paired t-test results indicating statistical significance. Reviewer feedback indicated that the garment images generated by the proposed method exhibited regular contour boundaries and coordinated structural proportions, with highly distinguishable morphological characteristics across all silhouette categories. In terms of visual realism, the generated textures and lighting transitions were natural, free from artificial artifacts and structural defects. In terms of style diversity, the disentangled generation mechanism effectively mitigated the model collapse problem, enabling the generation of diverse style variations under identical semantic conditions, thereby demonstrating superior adaptability for creative design applications.

3.8 Model efficiency and complexity analysis

To verify the deployment practicality of the framework, all methods were uniformly tested on a single NVIDIA A100 80GB hardware environment in terms of parameter count, inference speed, and memory consumption. The efficiency comparison results are presented in Table 6.

Table 6. Model efficiency comparison

Method	Parameters ($\times 10^6$)	Inference Time (s / Image)	Memory Usage (GB)
Stable Diffusion + ControlNet (Canny)	1450	2.8	18.2
IMAGGarment	1520	3.5	20.1
MAGDiff	1600	4.2	22.5
Proposed SilhouetteDiff	1580	3.6	21.3

The total parameter count of the proposed framework is 1580 M, with newly added modules accounting for only 130 M parameters, offering a lightweight advantage over mainstream baseline models. The inference speed and memory consumption fall within the mainstream deployable range. Compared to the precision-optimal MAGDiff model, the inference efficiency was improved by 14.3% with lower memory overhead. The comprehensive results demonstrate that the proposed method substantially improves garment silhouette control precision and generation quality without introducing excessive computational overhead, balancing generation performance with deployment efficiency, and thus possessing favorable engineering deployment and personalized design application value.

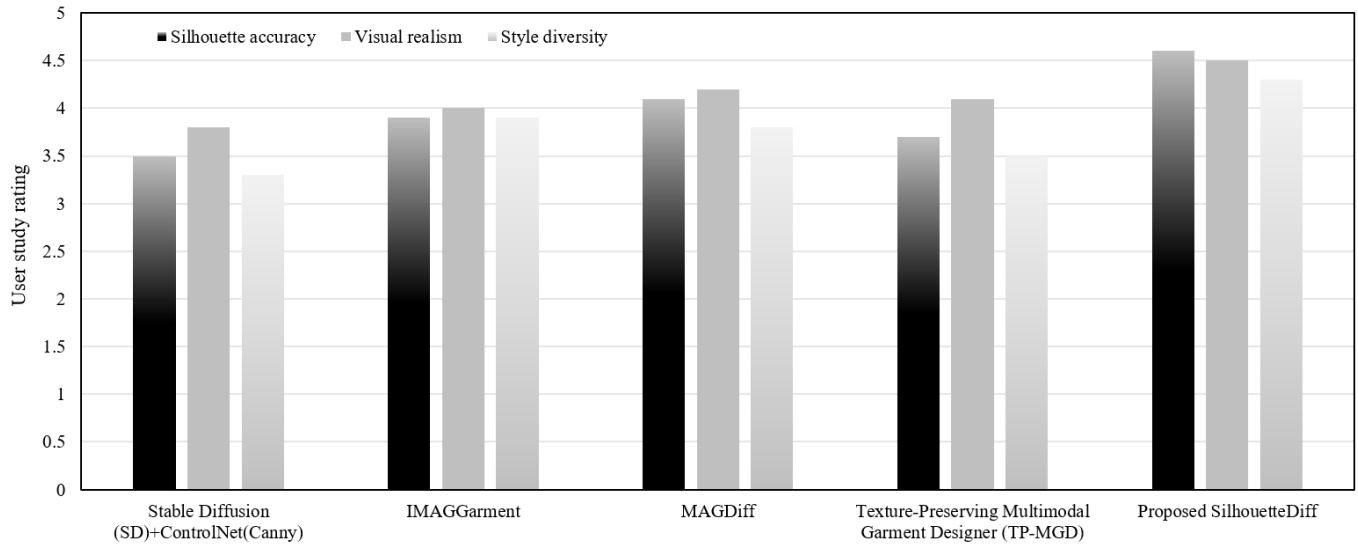


Figure 7. User study rating results

4. DISCUSSION

The SilhouetteDiff framework constructed in this study, through structured silhouette prior constraints and multi-modal disentangled representation learning, effectively addresses the challenges of fuzzy semantic alignment, unstable structural control, and silhouette distortion during editing in conventional garment generation models. The designed multi-module collaborative architecture achieves integrated modeling of garment silhouette semantic understanding, geometric structure generation, personalized style editing, and multi-attribute compositional creation, demonstrating superior generation precision and robustness in standardized flat garment design tasks. The experimental results fully validate the core role of structured silhouette anchors in enhancing garment generation controllability, demonstrating that the integration of geometric structure priors with cross-modal semantic fusion constitutes an effective technical pathway for achieving fine-grained intelligent garment design. Despite the significant performance advantages of the proposed method, certain limitations remain in terms of adaptation to complex real-world scenarios and model generalization capability.

First, the structural representation capability of the proposed method relies on the prediction accuracy of the keypoint detection network. Under complex scenarios involving dense garment folds or external occlusions, keypoint localization is prone to deviation, leading to distortion in the subsequent silhouette constraint conditions and ultimately compromising the structural accuracy of the generation results. Second, this study only models and generates five fundamental standard silhouette types, with the constructed representation system being relatively coarse-grained, incapable of precisely characterizing various gradient and fine-tuned sub-silhouette morphologies, and thus inadequately equipped for high-precision, diversified garment design requirements. Finally, the training data predominantly consists of two-dimensional flat garment samples captured from a frontal view, with a limited data scenario that does not cover garment deformation features corresponding to multiple human poses and body types, rendering the model insufficiently adaptable to personalized silhouette generation tasks in real wearing scenarios.

Addressing the limitations of the current study, future work will be continuously advanced along three directions: model robustness, representation refinement, and scenario expansion. The serial architecture relying on an independent detection network will be abandoned, and an end-to-end structure prediction mechanism will be constructed by leveraging the native visual perception capability of the diffusion model, thereby circumventing the error accumulation introduced by external detection models and enhancing the stability of skeleton representation under complex scenarios. Concurrently, a hierarchical fine-grained silhouette label system will be established to enrich the semantic categories of silhouettes, and the zero-shot learning paradigm will be incorporated to extend the model's generalization capability, enabling adaptive generation of unseen silhouette styles. Furthermore, the research boundary beyond two-dimensional image generation will be transcended, with the structured silhouette prior mechanism being extended to the domain of three-dimensional garment modeling, enabling multi-view consistent three-dimensional silhouette generation, thereby further advancing intelligent garment design technology from flat simulation toward deployment in real three-dimensional application scenarios.

5. CONCLUSION

An end-to-end generation framework, termed SilhouetteDiff, was constructed for personalized intelligent garment design, in which garment silhouette was adopted as the core structural anchor, systematically addressing the critical issues of semantic alignment deviation, insufficient geometric structure control precision, structural distortion during editing, and severe coupling among multiple style attributes in existing generation models. A hierarchical and collaborative algorithmic system was established, through which precise matching between textual semantics and garment structural features was achieved via a semantically aware silhouette cross-modal alignment module, geometric constraints during the generation process were reinforced via a SSPGM, structural integrity during style iteration and updating was ensured via a silhouette-preserving editing mechanism, and independent representation and free recombination of garment structure, texture, and color features were realized via a multi-modal style disentanglement strategy. All functional modules, with silhouette representation serving as the unified interaction carrier, formed an organic synergy, constructing a complete intelligent garment design pipeline from semantic-driven generation to personalized editing and innovation.

The effectiveness and superiority of the proposed framework were thoroughly validated through extensive quantitative experiments, ablation analyses, and subjective user evaluations. In comparison with mainstream garment generation and editing algorithms, SilhouetteDiff significantly improved the structural similarity and semantic classification accuracy of garment silhouettes, achieving fine-grained geometric parameter control and high-fidelity style editing, while simultaneously possessing stable style interpolation and multi-attribute compositional generation capabilities. By integrating structured prior knowledge with cross-modal disentangled learning, this study provides novel technical insights for structured modeling research in visually controllable generation, while offering an efficient and feasible technical solution for the deployment of intelligent and customized creative design in the garment industry.

ACKNOWLEDGMENT

This paper was funded by the 2024 Guangdong Province Youth Innovation Talent Project “Design Innovation Research on Traditional Culture in Western Guangdong in the Era of Digital Intelligence” (Grant No.: 2024WQNCX011).

REFERENCES

[1] Kulińska, M., A Abtew, M., Bruniaux, P., Zeng, X. (2022). Block pattern design system using 3D zoning method on digital environment for fitted garment. Textile Research Journal, 92(23-24): 4978-4993. https://doi.org/10.1177/00405175221114164

[2] Qi, A., Igarashi, T. (2024). PerfectTailor: Scale-preserving 2-D pattern adjustment driven by 3-D

garment editing. IEEE Computer Graphics and Applications, 44(4): 126-132. https://doi.org/10.1109/mcg.2024.3378171
[3] Guo, Y., Sun, L. (2025). Garment pattern generation method based on Diffusion Model and 3D reconstruction technology. Textile Research Journal, 95(23-24): 2866-2881. https://doi.org/10.1177/00405175251328296
[4] Sun, Y., Hao, Z., Wang, Z., Jin, J., Ye, Q., Lyu, Y. (2025). Deep learning for 3D garment generation: A review. Textile Research Journal, 95(23-24): 2882-2908. https://doi.org/10.1177/00405175251335188
[5] Croitoru, F., Hondru, V., Ionescu, R.T., Shah, M. (2023). Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(9): 10850-10869. https://doi.org/10.1109/tpami.2023.3261988
[6] Cao, H., Tan, C., Gao, Z., Xu, Y., Chen, G., Heng, P.A., Li, S.Z. (2024). A survey on generative diffusion models. IEEE Transactions on Knowledge and Data Engineering, 36(7): 2814-2830. https://doi.org/10.1109/TKDE.2024.3361474
[7] Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., Yang, M. (2023). Diffusion models: A comprehensive survey of methods and applications. ACM Computing Surveys, 56(4): 1-39. https://doi.org/10.1145/3626235
[8] Cao, S., Chai, W., Hao, S., Zhang, Y., Chen, H., Wang, G. (2023). DiffFashion: Reference-based fashion design with structure-aware transfer by diffusion models. IEEE Transactions on Multimedia, 26: 3962-3975. https://doi.org/10.1109/tmm.2023.3318297
[9] Song, D., Zhang, X., Zhou, J., Nie, W., Tong, R., Kankanhalli, M., Liu, A. (2024). Image-based virtual try-on: A survey. International Journal of Computer Vision, 133(5): 2692-2720. https://doi.org/10.1007/s11263-024-02305-2
[10] Islam, T., Miron, A., Liu, X., Li, Y. (2024). Deep learning in virtual try-on: A comprehensive survey. IEEE Access, 12: 29475-29502. https://doi.org/10.1109/access.2024.3368612
[11] Jiang, R., Zheng, G., Li, T., Yang, T., Wang, J., Li, X. (2024). A survey of multimodal controllable diffusion models. Journal of Computer Science and Technology, 39(3): 509-541. https://doi.org/10.1007/s11390-024-3814-0
[12] Yan, H., Zhang, H., Zhang, Z. (2023). Learning to disentangle the colors, textures, and shapes of fashion items: A unified framework. IEEE Transactions on Multimedia, 26: 5615-5629. https://doi.org/10.1109/tmm.2023.3338050
[13] Lee, J., Nguyen, D., Kim, J., Kang, J., Lee, S. (2023). Double reverse diffusion for realistic garment reconstruction from images. Engineering Applications of Artificial Intelligence, 127: 107404. https://doi.org/10.1016/j.engappai.2023.107404
[14] Chen, Y., Xie, R., Yang, S., Dai, L., Sun, H., Huo, Y., Li, R. (2024). Single-view 3D garment reconstruction using neural volumetric rendering. IEEE Access, 12: 49682-49693. https://doi.org/10.1109/access.2024.3380059
[15] Ren, B., Tang, H., Meng, F., Runwei, D., Torr, P.H.S., Sebe, N. (2023). Cloth interactive transformer for virtual try-on. ACM Transactions on Multimedia Computing, Communications, and Applications, 20(4): 1-20. https://doi.org/10.1145/3617374
[16] Chong, Z., Mo, L. (2022). ST-VTON: Self-supervised vision transformer for image-based virtual try-on. Image and Vision Computing, 127: 104568-104568. https://doi.org/10.1016/j.imavis.2022.104568
[17] Tong, S., Liu, H., Guo, R., Wang, W., Liu, D. (2024). Context-aware enhanced virtual try-on network with fabric adaptive registration. The Visual Computer, 41(3): 1435-1451. https://doi.org/10.1007/s00371-024-03432-0
[18] Wu, Q., Zhu, B., Yong, B., Wei, Y., Jiang, X., Zhou, R., Zhou, Q. (2020). ClothGAN: Generation of fashionable Dunhuang clothes using generative adversarial networks. Connection Science, 33(2): 341-358. https://doi.org/10.1080/09540091.2020.1822780
[19] Ghodhbani, H., Neji, M., Razzak, I., Alimi, A.M. (2022). You can try without visiting: A comprehensive survey on virtually try-on outfits. Multimedia Tools and Applications, 81(14): 19967-19998. https://doi.org/10.1007/s11042-022-12802-6
[20] He, W., Zhang, N., Song, B., Pan, R. (2023). Garment reconstruction from a single-view image based on pixel-aligned implicit function. Multimedia Tools and Applications, 82(19): 30247-30265. https://doi.org/10.1007/s11042-023-14924-x