Structure-Aware and Frequency-Guided Diffusion Framework for Multimodal Fashion Image Editing

Xin Chen[1,2], Jie Zhang[1,2*], Peng Zhang[1,2], Shasha Meng[1,2]

[1] Engineering Research Center of Artificial Intelligence for Textile Industry Ministry of Education, Institute of Artificial Intelligence, Donghua University, Shanghai, China
[2] Shanghai Engineering Research Center of Industrial Big Data and Intelligent System, Institute of Artificial Intelligence, Donghua University, Shanghai, China

Abstract. Fashion image editing aims to modify target garment attributes under textual and reference guidance while preserving non-target contents. However, existing methods often suffer from inaccurate garment localization, insufficient preservation of high-frequency textures, and unnatural transitions near edited boundaries. To address these issues, we propose a structure-aware and frequency-guided framework for multimodal fashion image editing. Specifically, we design a Structure-Guided Textual Mask Network to predict geometry-aware editing regions by leveraging refined textual structural cues and human-centric priors, where a structural prior reweighting mechanism is introduced to improve localization accuracy. We further develop an adaptive frequency-domain texture enhancement module to inject high-frequency fabric details from a reference image during late denoising, and employ a boundary-band soft fusion strategy to ensure smooth visual transitions. In addition, we construct a new dataset, DFEdit, for fine-grained multimodal fashion image editing. Experimental results show that the proposed method achieves competitive performance in terms of editing fidelity, texture consistency, and visual quality, showing its effectiveness for intelligent fashion image editing applications.

Keywords: Fashion Image Editing, Diffusion Models, Multimodal processing, Frequency-domain Texture Enhancement

1 Introduction

Fashion image editing focuses on altering garments or accessories in an image to match user-specified appearances, serving as a key problem in computer vision-driven intelligent computing. High-quality editing supports realistic visualization of personalized clothing, benefiting applications such as fashion design [1] and e-commerce [2]. Existing methods are generally divided into image-guided approaches [3][4] and text-guided approaches [5][6].

In recent years, increasing attention has been paid to local image editing, which restricts modifications to user-specified regions in order to avoid unnecessary interference with irrelevant areas [7] [8]. Although existing methods have made progress in enforcing local editing constraints, most of them still operate at a coarse-

grained region level and do not explicitly model the functional structures of garments. In fashion images, clothing is not a generic visual region, but a composition of semantically and geometrically meaningful parts, such as skirts and bodices [9] [10]. These garment substructures exhibit strong spatial dependencies and texture continuity. Consequently, coarse-grained editing masks often fail to accurately localize editable areas, leading to structural inconsistency and misaligned local modifications. Another major challenge lies in preserving fine-grained garment texture during the editing process. Fashion images often contain highly structured fabric patterns, such as stripes, plaid, knitted textures, and repetitive decorative details, which are dominated by high-frequency visual information. However, existing editing methods often struggle to maintain such texture details after local manipulation, producing artifacts such as blurring, texture discontinuity, and pattern drift within edited regions [11] [8], as illustrated in Fig. 1. Moreover, many early fashion editing methods are based on GAN architectures [12] [13], which are often difficult to optimize and less effective in generating high-quality local details. More recently, diffusion models [14] [15] [16] have demonstrated strong capability in controllable image generation and editing, providing a more powerful foundation for fine-grained fashion image manipulation.
Nevertheless, directly applying diffusion-based editing methods to fashion scenarios remains challenging, because they still lack explicit modeling of garment structure, and the preservation of high-frequency texture-specific.

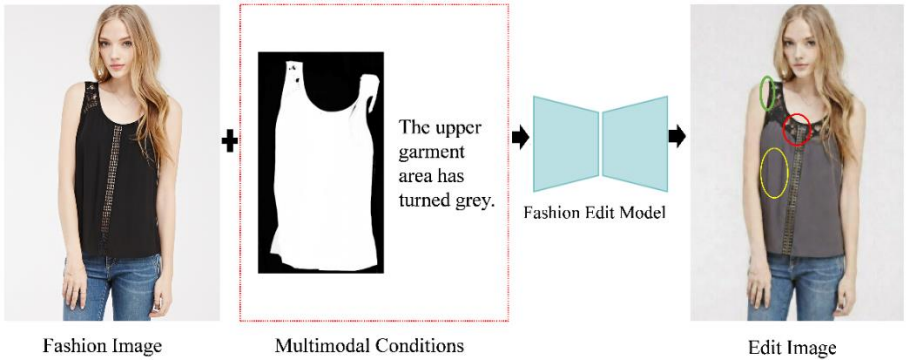

Fig. 1. Limitations of existing fashion editing methods, particularly in accurately identifying the editing area (red area), maintaining the integrity of the details of the high-frequency fabric textures (yellow area), and ensuring the continuity between the editing area and the non-edited area in terms of structure (green area).

We propose a structure-aware and frequency-guided multimodal framework for fashion image editing. Specifically, we first introduce a Structure-Guided Textual Mask Network (SGTMNet) to predict target editing regions by jointly exploiting fine-grained textual cues, human-centric structural priors, and boundary-aware mask refinement. In particular, to better adapt region localization to different editing instructions, we further design a text-driven structural prior reweighting mechanism, by enabling text-adaptive fusion of foreground cues, DensePose representations, and semantic parsing maps,

SGTMNet can more accurately localize the garment regions relevant to the target editing intent. We then adopt an adaptive frequency-domain texture enhancement mechanism to preserve realistic fabric details during the denoising process. We disentangle latent features into low-frequency structure and high-frequency texture, and progressively inject reference fabric details in the late diffusion stage to better preserve fine-grained textures while keeping overall semantics consistent. A boundary-band soft fusion strategy is further introduced to smooth transitions between edited and unedited regions, reducing visible artifacts. In addition, we construct DFEdit, a multimodal fashion editing dataset that differs from existing datasets by providing explicit alignment among region-level prompts, garment masks, and texture references, enabling more precise localized supervision. Our contributions are summarized as follows:

- We propose a structure-aware and frequency-guided diffusion framework for multimodal fashion image editing.
- We design a SGTMNet to localize editable garment regions under fine-grained textual guidance.
- We introduce an adaptive frequency-domain texture enhancement to preserve high-frequency fabric details, together with a soft fusion strategy to improve transition quality near edited region boundaries.
- We construct a DFEdit dataset for fine-grained multimodal fashion image editing.

2 Related Work

2.1 Text-guided image synthesis

Text-guided image synthesis focuses on generating or modifying visual content according to natural language descriptions, and has become a representative research topic in intelligent computing and computer vision. Earlier studies in this area were mainly built upon generative adversarial networks (GANs) [17] [18] [19], whereas more recent progress has been largely driven by diffusion models [20] [21] [22], which provide stronger generative capacity and higher image fidelity. Fashion-Gen [23] introduced a large-scale dataset for fashion synthesis and applied StackGAN [17] to generate clothing images from textual descriptions. FashionG [24] further improved the realism of generated fashion images by combining global and local style constraints. AttnGAN [25] enhanced the correspondence between text and image content through attention mechanisms, enabling finer semantic alignment. With the emergence of latent diffusion models such as Stable Diffusion [20] , text-driven generation has achieved substantial gains in visual quality and semantic consistency. In addition, classifier-free guidance [26] [22] has become an effective strategy for improving controllability in text-conditioned synthesis. In fashion-related scenarios, classifier-guided diffusion has also been introduced for attribute manipulation [27], while SGDiff [28] extends diffusion generation by incorporating style-aware guidance. Beyond direct text-to-image generation, diffusion models have also shown strong potential in text-guided image editing. SDEdit [11] edits an existing image by adding noise and reconstructing

the result with diffusion priors. DiffEdit [8] automatically estimates editable regions by comparing conditional prompts, thereby enabling semantic modification while keeping irrelevant content unchanged. MasaCtrl [29] improves local controllability by redesigning self-attention interactions and introducing mask-guided editing. Despite their promising performance in general image synthesis and editing, these methods are not specifically tailored for fashion images, where garment region precision and texture fidelity are especially critical.

2.2 Fashion image editing

Fashion image editing aims to modify garment-related attributes, such as color, texture, or style, based on user-provided conditions while maintaining person identity, pose, and background consistency. Early studies in this area mainly adopted GAN-based frameworks [12] [13] [6] [30]. Although these methods demonstrated the feasibility of semantic garment editing, they often suffered from limited detail generation and weak performance on high-resolution images. More recently, diffusion models have become a more effective solution for fashion image editing. MGD [5] is among the earlier diffusion-based methods that uses text and sketch inputs as dual conditions for garment generation and manipulation. Other approaches, including SDG [31], Hybrid Diffusion [32], and DiffusionCLIP [33] further exploit CLIP-based multimodal alignment [34] to improve the semantic accuracy of text-driven editing. In local editing settings, many methods [32] [35][22] rely on manually annotated masks to restrict the editable area. While effective, this strategy requires expensive human annotation and is difficult to scale to large datasets or practical applications. To reduce this dependency, DiffEdit [8] and Prompt-to-Prompt [36] attempt to infer editable regions automatically from textual conditions. TextFit [7] further introduces an editing region localization module to support text-only fashion editing. However, most of these approaches still operate at a relatively coarse spatial level, which makes them less effective for precise garment localization. Recent efforts have begun to address the structural and textural requirements of fashion editing more explicitly. DPDEdit[37] employs Grounded-SAM to obtain more accurate editing regions and utilizes a reference branch to extract high-frequency texture information from exemplar images for texture-preserving editing. MGD [5] incorporates multimodal conditions through text inversion, and Ti-MGD [38] extends this framework by introducing fabric texture as an additional modality for pattern control. Although these methods improve editing flexibility and controllability, fine-grained multimodal fashion editing still remains challenging. In particular, two issues are not yet fully resolved: accurate localization of garment structures and faithful preservation of high-frequency fabric details. These challenges motivate the proposed structure-aware and frequency-guided framework.

3 Method

We propose a method for fine-grained local garment editing conditioned on text and texture guidance. Given an input fashion image x_0, a text prompt P, and a texture reference x_t, the model generates an edited image x_G. The generated result is expected

to preserve the original structural information, such as pose and identity, while enabling localized fashion modification. The overall framework is illustrated in Fig. 2

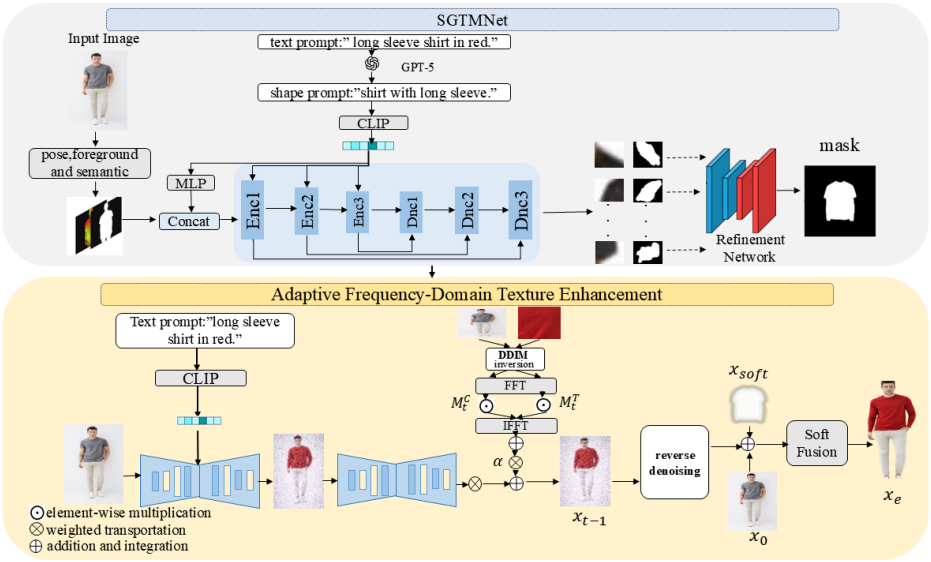

Fig. 2. Overview of our method.

3.1 Preliminaries

Our approach is based on the Stable Diffusion framework [20], integrates an autoencoder with a diffusion model. The autoencoder first compresses the input image x_0 into a compact latent code z_0, enabling more efficient diffusion learning.:

$$L_{AE}=\|x_0-D(E(x_0))\|^2, \tag{1}$$

Where E and D denote the encoder and decoder.

After encoding into latent z_0. diffusion is applied by progressively injecting Gaussian noise over T steps:

$$q(z_t|z_{t-1})=N(z_t;\sqrt{1-\beta_t}z_{t-1},\beta_t I), \tag{2}$$

Where β_t controls the noise magnitude. The reverse process is learned via a noise predictor ϵ_θ using the denoising loss.

$$L_{DM}=E_{t,z_0,\epsilon}[||\epsilon_\theta(z_t)-\epsilon||^2], \tag{3}$$

Starting from noisy latents, the model iteratively removes noise to recover z_0, which is then decoded into the final image.

In this work, we adopt DDIM-based inversion and sampling [39] to support image editing. Compared with the stochastic DDPM process, DDIM provides a deterministic

mapping between an image and its latent trajectory, which is more suitable for reconstruction-based editing tasks.

3.2 Structure-Guided Textual Mask Network

Accurate localization of editable garment regions is a key prerequisite for fine-grained fashion image editing. To this end, we propose the SGTMNet, a region prediction module that identifies editable garment areas under fine-grained textual guidance. SGTMNet adopts a lightweight U-Net as the backbone and injects textual conditions into the encoding process in a layer-wise manner, enabling effective semantic-spatial feature fusion for mask prediction.

To improve spatial awareness, SGTMNet incorporates multiple human- and garment-related structural priors. Specifically, Graphonomy [40] and DensePose [41] are used to extract foreground human regions and body structure cues from the input image, while an additional semantic parsing map is obtained using a pretrained segmentation model [42]. These structural representations are concatenated and used together with the input image to provide explicit guidance on human anatomy and garment layout. Since editable regions are mainly determined by structural attributes described in the text prompt, we further introduce a shape-oriented prompt refinement strategy. GPT-5 is used to parse the original textual input and generate a geometry-focused mask prompt by retaining structural descriptors while removing attributes irrelevant to region localization, such as color and texture. The refined prompt T is then encoded by a pretrained CLIP text encoder to obtain the textual embedding e_t. Based on the input image and the textual embedding, the mask prediction network produces an initial coarse editing mask.

We introduce a text-driven structural prior reweighting mechanism to further improve mask localization. Although multiple structural priors are beneficial, different editing instructions rely on them to different extents. For example, sleeve-related editing depends more on arm and upper-body cues, whereas skirt or trouser-related editing relies more heavily on lower-body parsing information. Direct concatenation of all priors treats them equally and may introduce irrelevant structural noise. To address this issue, we use the refined textual embedding e_t to dynamically predict the importance of each structural prior. Let S_f, S_d, and S_p denote the foreground map, densepose map, and semantic parsing map, respectively. A lightweight multi-layer perceptron network (MLP) is applied to e_t to produce normalized prior weights.

$$[w_f, w_d, w_p] = Softmax(MLP(e_t)), \tag{4}$$

Where w_f, w_d, and w_p represent the relative importance of the three structural priors under the current editing instruction, and the reweighted structural representation is then formulated as:

$$\tilde{S} = \mathrm{Concat}(w_f S_f, w_d S_d, w_p S_p), \tag{5}$$

In this way, the refined prompt not only provides semantic guidance for mask prediction, but also explicitly modulates which structural cues should be emphasized for the current target garment transformation. Based on the input image I, the

reweighted structural prior S, and the textual embedding e_t, the backbone network predicts an initial coarse editing mask:

$$M_c = G_\theta(I, \tilde{S}, e_t), \tag{6}$$

Where G_θ denotes the U-Net-based mask prediction network. This text-adaptive prior fusion strategy improves the sensitivity of SGTMNet to target garment attributes and reduces interference from structurally irrelevant regions.

However, directly predicted masks often suffer from inaccurate boundary delineation due to the imbalance between boundary and interior pixels. To alleviate this issue, Inspired by BPR [43], only local high-resolution patches within the boundary neighborhood are extracted and analyzed locally. Subsequently, instead of replacing the refined results as the output, the BASNet [44] predict-refine idea is referred to. Two complementary pieces of information are learned for each boundary patch: one is whether the local area truly belongs to the boundary position that needs to be corrected with priority, and the other is how to perform local residual correction for this area relative to the coarse mask. Further, the local prediction results of each boundary patch are re-mapped back to the original image space and fused with the coarse mask output by the backbone network to obtain the final refined mask. The fusion is not a simple overlay, but is modulated by the boundary confidence information predicted by the boundary branch.

$$M = M_c + R(C_b \odot R_b), \tag{7}$$

Where C_b denotes the boundary confidence map, R_b denotes the residual correction map predicted by the boundary branch, $\odot$ represents element-wise multiplication, and $R(\cdot)$ denotes the remapping operation from local boundary patches to the full-resolution image space. Boundary refinement is guided by both confidence-aware selection and local residual correction, leading to more accurate mask boundaries.

Unlike existing text-based segmentation models, SGTMNet predicts editable regions according to the target editing intent rather than merely segmenting the garment category already present in the image. Benefiting from the proposed structural prior reweighting and the boundary-aware refinement mechanism, SGTMNet achieves more precise garment localization.

3.3 Adaptive Frequency-Domain Texture Enhancement

Text prompts mainly control high-level semantics, but they struggle to describe fine-grained fabric patterns such as stripes or other material-level textures. These high-frequency details are crucial for realistic fashion editing; relying only on text often yields correct semantics but weak or blurred texture reconstruction. To overcome this, we integrate text-guided diffusion editing with adaptive frequency-domain texture enhancement. The diffusion model first performs semantic clothing modification according to the prompt, and then a frequency-based module injects texture cues from a reference fabric image into the edited region, enhancing high-frequency details and improving realism.

Given a content image x_0, a text prompt P, and a texture reference image x_t, we first encode the text prompt using a text encoder:

$$E_p = CLIP(P) \tag{8}$$

Where E_p denotes the text embedding used to guide semantic editing through cross-attention. Meanwhile, the editing mask M predicted by the localization module constrains subsequent modifications to the target garment region.

To retain the source image structure while supporting controllable editing, we apply DDIM inversion to map both the input image and texture reference into their corresponding latent trajectories, formulated as follows:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} f_\theta(x_t, t) + \sqrt{1 - \bar{\alpha}_{t-1}} \epsilon_\theta(x_t, t), \tag{9}$$

$$f_\theta(x_t, t) = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}, \tag{10}$$

Where ϵ_θ represents the noise estimation network. Using DDIM inversion, the content image I^C and texture image I^T are mapped into latent sequences.

The model conducts text-guided editing, where latent features are iteratively updated at each denoising step based on the text embedding E_p:

$$\epsilon_t = \epsilon_\theta\left(x_t^C, t, E_p\right), \tag{11}$$

$$\hat{x}_{t-1}^C = Denoise(x_t^C, \epsilon_t), \tag{12}$$

Where $\hat{x}_{t-1}^C$ denotes the text edited latent feature. To preserve non-edited regions, we apply mask-constrained updating:

$$\tilde{x}_{t-1}^C = M \odot \hat{x}_{t-1}^C + (1 - M) \odot x_{t-1}^C, \tag{13}$$

This stage determines the target garment semantics and structure. However, although the edited result becomes semantically aligned with the text prompt, it may still lack realistic fabric textures.

After semantic editing, we enhance the edited garment by transferring texture cues from the reference image in the frequency domain. We apply Fast Fourier Transform (FFT) to the latent feature x_t^T of the texture image to extract its high-frequency components:

$$F_t^T = FFt(x_t^T) \odot M_t^T, \tag{14}$$

$$x'^T_t = IFFT(F_t^T), \tag{15}$$

Where $FFT(\cdot)$ and $IFFT(\cdot)$ denote the Fast Fourier Transform and its inverse, and M_t^{T} is the texture mask that preserves the high-frequency bands. Similarly, the text edited latent feature is transformed into the frequency domain to retain its low-frequency structural information:

$$F_t^C = FFt(x_t^C) \odot M_t^C, \tag{16}$$

$$x'^{C}_{t} = IFFT(F^{C}_{t}), \tag{17}$$

Where M^{C}_{t} is the complementary low-frequency mask used to preserve structural and color information. In practice, M^{C}_{t} and M^{T}_{t} are designed as complementary masks, such that the content branch maintains garment structure while the texture branch contributes fine-grained fabric details.

After frequency decomposition, the extracted texture and content components are fused during denoising as:

$$x'_{t} = \tilde{x}^{C}_{t} + M \odot \alpha_{t,l}\left(x'^{C}_{t} + x'^{T}_{t} - \tilde{x}^{C}_{t}\right), \tag{18}$$

Where M is the editing mask and α is an adaptive fusion coefficient that controls the texture injection strength according to the current diffusion timestep t and feature level l. Specifically, we define

$$\alpha_{t,l} = \alpha_{max} \cdot \left(1 - \frac{t}{T}\right)^{\gamma} \cdot \frac{l}{L}, \tag{19}$$

Where T is denoising steps, L is the number of injected feature levels, γ controls the timestep scheduling curve, and α_{max} denotes the maximum texture fusion strength. This formulation suppresses texture injection at early timesteps and low-resolution layers, while progressively strengthening it at later timesteps and high-resolution layers. The fused latent feature xt is then fed into the reverse denoising process:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} f_{\theta}(x'_{t}, t) + \sqrt{1 - \bar{\alpha}_{t-1}} \epsilon_{\theta}(x'_{t}, t). \tag{20}$$

Through this mechanism, the generated image achieves both semantic consistency with the text prompt and improved texture fidelity with respect to the reference fabric, resulting in more realistic garment materials.

To preserve person identity and maintain the integrity of non-edited regions, we adopt a boundary-band constrained soft blending strategy. Instead of applying smoothing over the entire editing region, the proposed method restricts the transition to a narrow band around the predicted editing mask. This design avoids unnecessary blur inside edited garment regions while ensuring smooth appearance transitions near region boundaries. We first construct a boundary band region B using morphological operations:

$$B = Dilate(M, r) - Erode(M, r), \tag{21}$$

Where r controls the band width. The inner edited region and outer preserved region are defined as:

$$R_{in} = M - B, R_{out} = (1 - M) - B, \tag{22}$$

To achieve smooth blending only within the boundary band, we generate a soft mask by applying Gaussian smoothing to the binary mask:

$$\tilde{M} = G_\sigma * M, \tag{23}$$

The final synthesized image $\tilde{x}_{soft}$ are fefined as :

$$\tilde{x}_{soft}(p) = \begin{cases} x_e(p), & p \in R_{in} \\ x_0(p), & p \in R_{out} \\ \tilde{M}(p)x_e(p) + (1 - \tilde{M}(p))x_0(p), & p \in B \end{cases} \tag{24}$$

Where x_e is the edited image. In this way, the edited region remains sharp, while only the boundary area is softly blended to reduce visible discontinuities.

3.4 The DFEdit Dataset

To support fine-grained localized fashion image editing, we construct a new dataset To support fine-grained localized fashion editing, we build DFEdit on DeepFashion-MultiModal[9] and Fashion-Gen[23], providing images, parsing maps, attributes, and text with explicit region–text–texture alignment. Parsing annotations from DeepFashion-MultiModal are converted into binary masks for key garment regions, then combined with attributes to form region-level prompts. Compared with global-description datasets, DFEdit enables finer localized supervision. To preserve realistic details, texture patches are extracted within valid mask areas via a sliding-window strategy, providing explicit cues for texture-aware editing. DFEdit dataset is illustrated in Fig. 3.

Fig. 3. Comparison of DFEdit with other fashion datasets

4 Experiments

4.1 Experimental Settings

Experiments are conducted on the DFEdit dataset with an extended test set of 12,000 image–text pairs for evaluation. We compare our method against three diffusion-based approaches, including SDEDit [11], TextFit [7], and DiffEdit [8], all retrained on DFEdit under identical settings to ensure fairness. Training is performed on a single

NVIDIA A40 GPU with images resized to 512×256, SGTMNet is trained for 200 epochs. For editing, Stable Diffusion v1.4 is adopted as the backbone and fine-tuned on DFEdit for 200k iterations with AdamW (learning rate 1×10^{-5}), using gradient accumulation due to memory limits, and the same initialization is applied to all baselines. Evaluation includes FID [45] and LPIPS [46] for visual quality, along with CLIP score [47] to measure text–image semantic consistency

4.2 Comparison with State-Of-The-Art Methods

We evaluate our approach against representative diffusion-based baselines on the DFEdit test set. As reported in Table 1, our method attains the lowest FID and LPIPS scores, and also performs competitively on CLIP-S, demonstrating improved visual quality and more accurate localized editing results.

Table 1. Quantitative results on the DFEdit dataset.

Method	FID	LPIPS	CLIP-S
SDEdit	16.10	0.172	28.08
DiffEdit	16.72	0.101	26.52
TextFit	10.68	0.050	27. 85
Ours	**9. 85**	**0.047**	**28.04**

As shown in Fig. 4. SDEdit often causes over-editing and affects surrounding regions. DiffEdit roughly localizes editable areas but tends to produce blurred boundaries and broken textures. TextFit offers better region control, yet still suffers from texture distortion and detail loss in complex garment scenarios. In contrast, our method preserves person identity and pose more effectively, while generating clearer structures and more consistent textures.

Fig. 4. Visual comparison between results of our method and baseline approaches.

Fig. 5 further illustrates representative results under diverse editing scenarios, including color modification, category replacement, and structural adjustment. The results demonstrate that our method produces visually coherent edits while maintaining local texture integrity.

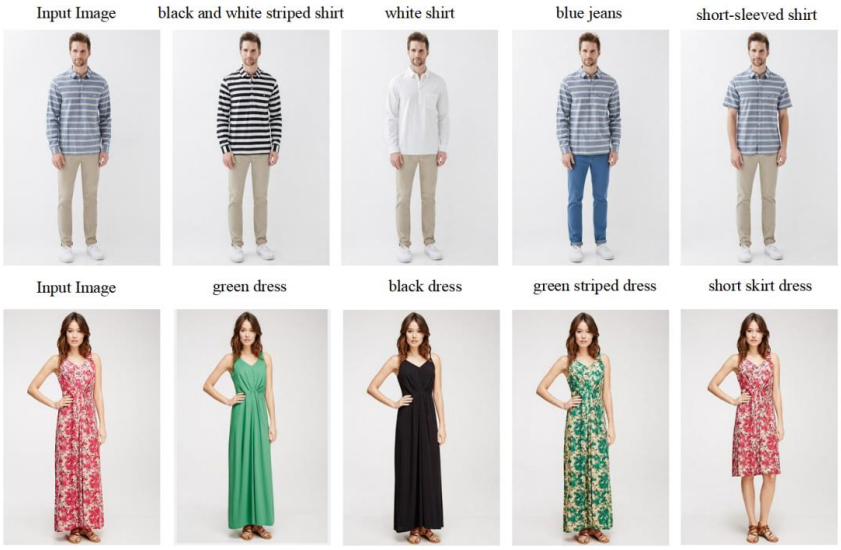

Fig. 5. Qualitative research on different text editing methods

To evaluate perceptual quality in practical scenarios, we conduct a user study with 20 participants on 2000 samples. Participants select the best results among different methods based on realism, garment structure and texture consistency, and text alignment. As shown in Table 2, our method achieves the highest average preference across all criteria, demonstrating superior human-perceived quality.

Table 2. User study results comparing baseline methods and our method

Method	Image realism	Texture consistency	Text Matching Degree
SDEdit	196	227	163
TextFit	427	294	276
DiffEdit	391	268	384
Ours	**986**	**1211**	**1177**

Overall, both qualitative and quantitative results demonstrate that, our method achieves a better trade-off between editing accuracy and image realism compared with SDEdit, DiffEdit, and TextFit.

4.3 Ablation study

We conduct ablation studies on SGTMNet, the adaptive frequency-domain texture enhancement module (AF-TE), and the flexible fusion strategy to evaluate their contributions to region localization, texture preservation, and boundary quality. As shown in Fig. 6, SGTMNet produces more accurate and stable editing masks than the baseline TextFit. Without structural priors and prompt refinement, the baseline often yields coarse or incomplete garment regions. In contrast, SGTMNet generates masks with clearer boundaries and better alignment with garment geometry.

Fig. 6. Visual comparison of editing region masks generated by our method and TextFit.

In addition, as reported in Table 3. compared with the baseline setting without SGTMNet and frequency-domain texture enhancement module, incorporating both modules yields significant improvements in FID, LPIPS and CLIP-S.

Table 3. Quantitative ablation results of our method on DFEdit dataset

SGTMNet	AF-TE	FID	LPIPS	CLIP-S
		14.32	0.324	26.35
√		13.21	0.296	27.05
	√	12.03	0.217	27.74
√	√	**9. 85**	**0.047**	**28.04**

As shown in Fig. 7, after introducing the flexible fusion strategy, significant improvements have been achieved in the clarity of high-definition textures, the sharpness of the edges of the clothing area, and the rationality.

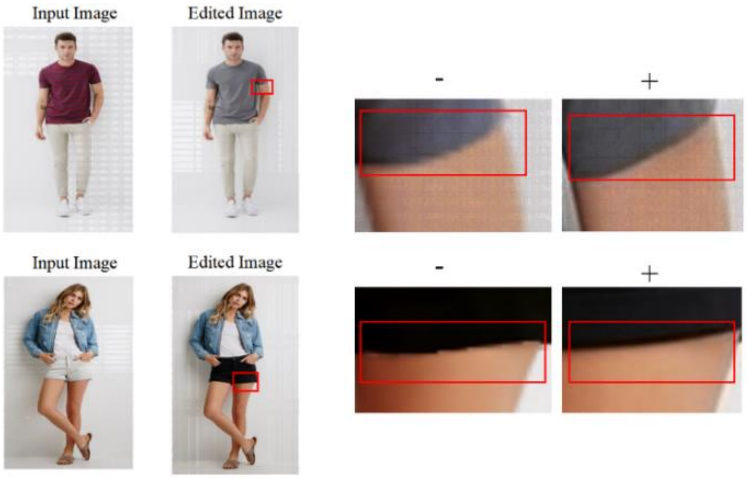

Fig. 7. Ablation study of the soft fusion strategy

These results confirm that SGTMNet improves editing region localization, AF-TE enhances texture fidelity, and the soft fusion strategy yields more natural boundary transitions.

5 Conclusion

This paper presents a diffusion-based fashion image editing framework with adaptive frequency-domain texture enhancement, integrating fine-grained garment localization and a soft fusion mechanism to achieve more accurate and visually consistent local edits. In addition, we introduce the DFEdit dataset to support detailed fashion editing studies. Experimental results show clear improvements over representative methods in editing accuracy, texture retention, and overall visual quality, demonstrating the effectiveness of the proposed approach.

Acknowledgements. This work is supported by the Shanghai Soft Science Research Funds [grant number 25692107300], National Natural Science Foundation of China under Grant No. 52005099, the National Key R \& D Program of China under Grant No.2019YFB1706300, Shanghai Frontier Science Research Center for Modern Textiles, Donghua University , and the Fundamental Research Funds for the Central Universities.

Disclosure of Interests. the authors have no competing interests.

References

[1] Sorger R, Udale J. The fundamentals of fashion design[M]. Bloomsbury Publishing, 2020.
[2] Jain V, Malviya B, Arya S. An overview of electronic commerce (e-Commerce)

[J]. Journal of Contemporary Issues in Business and Government, 2021, 27(3): 666.
[3] Han X, Wu Z, Wu Z, et al. Viton: An image-based virtual try-on network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7543-7552.
[4] Choi S, Park S, Lee M, et al. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 14131-14140.
[5] Baldrati A, Morelli D, Cartella G, et al. Multimodal garment designer: Human-centric latent diffusion models for fashion image editing[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 23393-23402.
[6] Patashnik O, Wu Z, Shechtman E, et al. Styleclip: Text-driven manipulation of stylegan imagery[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 2085-2094.
[7] Wang T, Ye M. Texfit: Text-driven fashion image editing with diffusion models[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(9): 10198-10206.
[8] Couairon G, Verbeek J, Schwenk H, et al. Diffedit: Diffusion-based semantic image editing with mask guidance[J]. arXiv preprint arXiv:2210.11427, 2022.
[9] Jiang Y, Yang S, Qiu H, et al. Text2human: Text-driven controllable human image generation[J]. ACM Transactions on Graphics (TOG), 2022, 41(4): 1-11.
[10] Liu Z, Luo P, Qiu S, et al. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1096-1104.
[11] Meng C, He Y, Song Y, et al. Sdedit: Guided image synthesis and editing with stochastic differential equations[J]. arXiv preprint arXiv:2108.01073, 2021.
[12] Nam S, Kim Y, Kim S J. Text-adaptive generative adversarial networks: manipulating images with natural language[J]. Advances in neural information processing systems, 2018, 31.
[13] Li B, Qi X, Lukasiewicz T, et al. Manigan: Text-guided image manipulation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 7880-7889.
[14] Zhu L, Yang D, Zhu T, et al. Tryondiffusion: A tale of two unets[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 4606-4615.
[15] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[C]//Proceedings of Advances in Neural Information Processing Systems. 2020: 6840-6851.
[16] Dhariwal P, Nichol A. Diffusion models beat gans on image synthesis[C]//Proceedings of Advances in Neural Information Processing Systems. 2021: 8780-8794.
[17] Zhang H, Xu T, Li H, et al. Stackgan: Text to photorealistic image synthesis with stacked generative adversarial networks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 5907-5915.
[18] Zhang H, Koh J Y, Baldridge J, et al. Cross-modal contrastive learning for text-to-image generation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2021: 833-842.
[19] Tao M, Tang H, Wu F, et al. Df-gan: A simple and effective baseline for text-to-

image synthesis[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2022: 16515-16525.
[20] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2022: 10674-10685.
[21] Saharia C, Chan W, Saxena S, et al. Photorealistic text-to-image diffusion models with deep language understanding[C]//Proceedings of Advances in Neural Information Processing Systems. 2022: 36479-36494.
[22] Nichol A Q, Dhariwal P, Ramesh A, et al. Glide: Towards photorealistic image generation and editing with text-guided diffusion models[C]//Proceedings of the International Conference on Machine Learning. 2022: 16784-16804.
[23] Rostamzadeh N, Hosseini S, Boquet T, et al. Fashion-gen: The generative fashion dataset and challenge[J]. arXiv preprint arXiv:1806.08317, 2018.
[24] Jiang S, Li J, Fu Y. Deep learning for fashion style generation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(9): 4538-4550.
[25] Xu T, Zhang P, Huang Q, et al. Attngan: Fine-grained text to image generation with attentional generative adversarial networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1316-1324.
[26] Ho J, Salimans T. Classifier-free diffusion guidance[C]//Proceedings of NeurIPS Workshop on Deep Generative Models. 2021.
[27] Kong C, Jeon D, Kwon O, et al. Leveraging off-the-shelf diffusion model for multi-attribute fashion image manipulation[C]//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. 2023: 848-857.
[28] Sun Z, Zhou Y, He H, et al. Sgdiff: A style guided diffusion model for fashion synthesis[C]//Proceedings of the ACM International Conference on Multimedia. 2023: 8433-8442.
[29] Cao M, Wang X, Qi Z, et al. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing[C]//Proceedings of the IEEE International Conference on Computer Vision. 2023: 22560-22570.
[30] Xia W, Yang Y, Xue J H, et al. Tedigan: Text-guided diverse face image generation and manipulation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2021: 2256-2265.
[31] Liu X, Park D H, Azadi S, et al. More control for free! image synthesis with semantic diffusion guidance[C]//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. 2023: 289-299.
[32] Avrahami O, Lischinski D, Fried O. Blended diffusion for text-driven editing of natural images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2022: 18208-18218.
[33] Kim G, Kwon T, Ye J C. Diffusionclip: Text-guided diffusion models for robust image manipulation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2022: 2426-2435.
[34] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of the International Conference on Machine Learning. 2021: 8748-8763.
[35] Avrahami O, Fried O, Lischinski D. Blended latent diffusion[J]. ACM Transactions on Graphics, 2023, 42(4): 1-11.
[36] Hertz A, Mokady R, Tenenbaum J, et al. Prompt-to-prompt image editing with

cross-attention control[C]//Proceedings of the International Conference on Learning Representations. 2022.
[37] Wang X, Cheng Z, Wang J, et al. Dpdedit: Detail-preserved diffusion models for multimodal fashion image editing[C]//Proceedings of the IEEE International Conference on Multimedia and Expo. 2025: 1-6.
[38] Baldrati A, Morelli D, Cornia M, et al. Multimodal-conditioned latent diffusion models for fashion image editing[J]. arXiv preprint arXiv:2403.14828, 2024.
[39] Song J, Meng C, Ermon S. Denoising diffusion implicit models[J]. arXiv preprint arXiv:2010.02502, 2020.
[40] Gong K, Gao Y, Liang X, et al. Graphonomy: Universal human parsing via graph transfer learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 7450-7459.
[41] Guler R A, Neverova N, Kokkinos I. Densepose: Dense human pose estimation in the wild[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7297-7306.
[42] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
[43] Tang C, Chen H, Li X, et al. Look closer to segment better: Boundary patch refinement for instance segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13926-13935.
[44] Qin X, Zhang Z, Huang C, et al. Basnet: Boundary-aware salient object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 7479-7489.
[45] Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[C]//Proceedings of Advances in Neural Information Processing Systems. 2017.
[46] Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 586-595.
[47] Hessel J, Holtzman A, Forbes M, et al. Clipscore: A reference-free evaluation metric for image captioning[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2021: 7514-7528.