GLYPHFASHION: FINE-GRAINED TEXT-AWARE FASHION IMAGE EDITING VIA DIFFUSION MODELS

Yanting Zhang[1]*, **Jingyi Guo**[1], **Huanwen Zheng**[1], **Cairong Yan**[1], **Yonggang Qi**[2], **Gaoang Wang**[3]*
[1]Donghua University, [2]Beijing University of Posts and Telecommunications, [3]Zhejiang University
{ytzhang, cryan}@dhu.edu.cn, {2232945, 2222760}@mail.dhu.edu.cn
qiyg@bupt.edu.cn, gaoangwang@intl.zju.edu.cn

ABSTRACT

With the rapid advancement of diffusion models in image generation and editing, multimodal garment image editing has emerged as an important research direction in intelligent fashion design. However, visual text rendering on garments often suffers from issues such as illegible characters, blurred boundaries, and spelling errors. To address these challenges, we propose GlyphFashion, a diffusion-based multimodal framework for fine-grained, text-aware garment image editing. The proposed framework introduces a unified text-aware conditioning module and integrates sketch priors, color cues, and region mask contexts through ControlNet-based conditional branches, enabling precise geometric constraints and context-aware text generation throughout the denoising process. Furthermore, to alleviate the lack of high-quality text annotations in existing garment editing datasets, we construct an OCR-aware fine-grained editing dataset based on IGPair. Experimental results demonstrate that, compared with existing methods, our approach significantly improves the consistency between generated text and visual appearance, while maintaining high stability in challenging scenarios such as complex textures and small-scale printed texts.

1 INTRODUCTION

In fashion design and the apparel industry, there has been a long-standing and pressing demand for rapid generation and localized modification of garment visual content. However, traditional approaches such as manual illustration or editing via graphic design software are often labor-intensive, time-consuming, and highly dependent on expert experience, making it difficult to balance efficiency and quality. In recent years, the rapid progress of diffusion models (Ho et al., 2020) in image generation and editing has opened new technological pathways for intelligent garment image synthesis and editing, providing important support for interactive fashion design and AI-driven creative workflows.

With the rapid development of denoising diffusion probabilistic models (Ho et al., 2020), diffusion-based methods have demonstrated superior performance over traditional generative approaches in image quality, diversity, and controllability. Stable Diffusion (Rombach et al., 2022) enables high-resolution image synthesis via latent diffusion modeling. Subsequently, multi-modal conditional frameworks such as T2I-Adapter (Mou et al., 2024), ControlNet (Zhang et al., 2023), InstructPix2Pix (Brooks et al., 2023), and MagicQuill (Liu et al., 2025) further improve controllability, allowing users to generate or edit images using text, sketches, or reference inputs. More recently, models including DALL·E 3 (Betker et al., 2023) and Stable Diffusion 3 (Esser et al., 2024) have achieved significant gains in realism and semantic alignment, highlighting the potential of diffusion models for practical creative applications.

To address the challenges of imprecise character rendering and blurry graphical patterns that commonly appear in general T2I models, recent works have explored incorporating text encoders, glyph-aware modeling, or OCR-guided supervision such as GlyphControl (Yang et al., 2023), TextDiffuser (Chen et al., 2023), TextDiffuser-2 (Chen et al., 2024), and ControlText (Jiang et al., 2025)

[*] Corresponding authors.

to enhance text readability and semantic consistency. These approaches have achieved promising results in tasks such as scene text generation, document layout synthesis, and 2D graphic design, demonstrating the potential of text-aware supervision in improving visual text fidelity within diffusion-based generation frameworks.

However, when applied to localized garment editing, especially in scenarios involving small-scale text and complex fabric textures, existing diffusion-based methods still struggle to simultaneously ensure text readability, content accuracy, and natural integration with garment style. More concretely, current approaches exhibit the following limitations: (1) In image-based text rendering, textual conditions are often entangled with other visual attributes, leading to cross-condition interference in the feature space and limiting precise control over text appearance. (2) In many existing systems, text is only used as a high-level semantic cue or implicit condition during generation, without explicitly encoding the target character content. As a result, generated outputs may appear text-like but deviate from the user-specified content at the character or semantic level, leading to spelling errors or mismatches. (3) Lack of high-quality datasets and systematic evaluation protocols for garment text editing. Existing public garment datasets mainly focus on garment appearance or virtual try-on, and rarely provide fine-grained annotations for printed text.

To address the aforementioned limitations in garment-oriented text editing, we introduce **GlyphFashion**, a diffusion-based fine-grained garment image editing framework that takes glyph-level visual detail generation as its core modeling objective. The framework incorporates a Text-Aware Conditioning Module (TACM) that explicitly encodes glyph appearance and spatial layout information, and injects these representations into the diffusion denoising process. This design enables the model to perceive character stroke structures and text layouts on non-planar garment surfaces. By decoupling glyph conditions from garment-related visual controls, cross-condition interference is effectively reduced. To further enhance controllability in garment text editing, we explicitly integrate visual conditions such as sketch priors, color cues, and region masks, and inject them into the diffusion backbone via a ControlNet-based multi-modal control branch to maintain consistency between the edited regions and the overall garment structure. Meanwhile, text prompts and image prompts are separately encoded and injected into the model, and an OCR-aware feature loss is incorporated to improve the correctness and visual fidelity of the generated characters. In addition, we construct an OCR-aware fine-grained garment editing dataset based on the IGPair (Shen et al., 2025) benchmark. For each image, we augment dense textual captions, precise OCR annotations, region-level masks, color conditions, and corresponding sketch priors, thus providing systematic support for training and evaluation. During inference, we adopt a pixel-space region-preserving blending strategy that retains model-generated outcomes only within target editing regions, while directly inheriting the original pixels elsewhere, ensuring that non-edited areas remain visually untouched while localized editing effects are preserved. The main contributions of this work are summarized as follows:

- We propose GlyphFashion, a diffusion-based fine-grained garment editing framework that leverages ControlNet to integrate sketch, color, and mask conditions, enabling collaborative control over content, placement, and shape on non-planar garment surfaces while alleviating texture discontinuity and unnatural artifacts.
- We introduce a unified Text-Aware Conditioning Module (TACM) that explicitly models glyph strokes and spatial layouts, incorporating an OCR-aware perceptual loss to enhance character readability and semantic correctness, thereby achieving fine-grained and controllable text generation on garments.
- We construct an OCR-aware garment editing benchmark based on IGPair, enriched with dense captions, OCR annotations, region masks, and sketch priors. Extensive experiments demonstrate that our approach significantly improves text fidelity, spatial alignment, and texture continuity compared with existing diffusion-based editing methods.

2 Related work

2.1 Controllable Image Editing

The rapid development of diffusion models has significantly advanced the field of image editing. Building upon this progress, text-guided editing methods based on text-to-image (T2I) generation models (Mou et al., 2024; Ye et al., 2023; Ruiz et al., 2023; Lee et al., 2025) allow users to modify images through natural language instructions, greatly improving the accessibility and convenience of image editing. However, relying solely on textual conditions suffers from inherent limitations in expressing complex spatial relationships, fine-grained structures, and cross-modal semantics, making it difficult to achieve precise and highly targeted editing control. To overcome these limitations, recent studies have explored interactive image editing methods that integrate multi-modal conditions. By introducing cross-modal alignment and fusion mechanisms, these approaches are able to better interpret user intent and enable multi-dimensional control over both content and style, demonstrating stronger expressive power and broader application potential.

SmartBrush (Xie et al., 2023) pioneered a differentiable mask precision adjustment mechanism and innovatively incorporated the BLIP (Li et al., 2022b) vision-language model into a multi-task training framework, enabling the system to automatically infer the optimal editing granularity based on user-provided sketch contours. Building upon this idea, BrushNet (Ju et al., 2024) further proposed a decomposed dual-branch diffusion architecture that processes masks through a VAE (Kingma & Welling, 2013) encoder and performs hierarchical UNet (Ronneberger et al., 2015) feature fusion, together with a blurred blending strategy. This design transforms mask-based inpainting into a plug-in module that can be seamlessly integrated with arbitrary pretrained diffusion models. FramePainter (Zhang et al., 2025) reformulated static image editing as a video generation task. Built upon Stable Video Diffusion (Blattmann et al., 2023), it introduces a cross-frame matching attention mechanism that effectively addresses the limitations of traditional temporal modeling in long-range motion understanding. MagicQuill (Liu et al., 2025) presents a new generation of natural interaction paradigms, where a sparse control encoder maps discrete user inputs into continuous latent representations. Combined with a CoTracker-based attention alignment module and a multimodal large language model–driven “draw-and-guess” mechanism, MagicQuill significantly lowers the operational barrier of professional editing tools, allowing users to perform complex edits through intuitive interactions such as doodling and dragging.

2.2 Text Generation

With the advancement of diffusion models, substantial progress has been made in synthesizing text within images. Nevertheless, integrating clear, accurate, and structurally consistent text into images remains a challenging problem. In diffusion-based text generation research, innovations in conditional control mechanisms have become a key pathway toward improving visual text quality. AnyText (Tuo et al., 2023) introduces auxiliary latent modules and text embedding modules combined with OCR models to support multilingual text generation. Diff-Text (Zhang et al., 2024) leverages sketch images as priors and employs local attention constraints and contrastive image-level prompts to enable scene text generation in arbitrary languages. FontDiffuser (Yang et al., 2024) adopts multi-scale content aggregation blocks and a style contrastive refinement module to generate complex font styles through the denoising diffusion process. TextCrafter (Du et al., 2025) addresses confusion, omission, and blurring in complex visual text generation via a three-stage framework consisting of instance fusion, region isolation, and text-focused generation. Notably, AnyText2 (Tuo et al., 2024) achieves improved inference efficiency through a novel WriteNet + AttnX architecture. Its text embedding module integrates four encoders corresponding to position, font, color, and glyph, enabling independent attribute control for each text line. This design not only improves text accuracy for both Chinese and English characters but also significantly enhances the overall photorealism of generated images. Although existing methods have achieved notable progress in scene text generation and planar graphic design, their performance remains limited in fine-grained garment text editing due to non-planar surfaces and complex textures. Our approach aims to enhance the stability and controllability of garment text editing by explicitly modeling glyph appearance and spatial layout.

3 Method

3.1 Architecture Overview

GlyphFashion is a diffusion-based multimodal fine-grained garment image editing framework designed to enhance the generation quality and controllability of printed text on clothing. Built upon a pretrained diffusion backbone, the proposed method introduces a Text-aware Conditioning Module together with multimodal control branches to jointly model character-level glyph details, spatial lay-

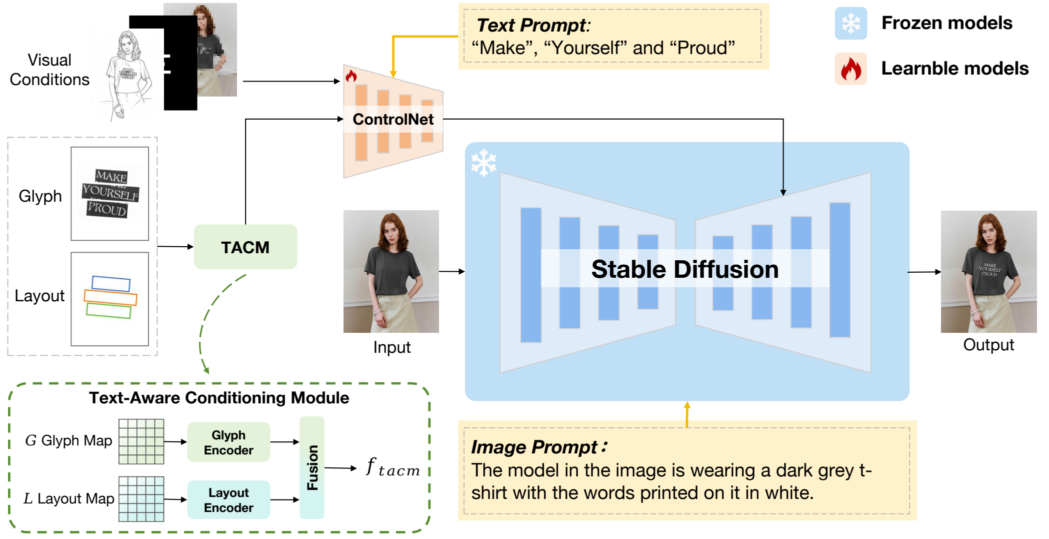

Figure 1: **Model overview.** Our model generates the target text on garment regions of the input image. Specifically, glyph appearance and spatial layout cues are injected into the generation process through the Text-Aware Conditioning Module (TACM), while additional visual conditions are integrated via ControlNet, including sketch priors, region masks, and color cues. The final output is guided jointly by text prompts and image prompts. During inference, the glyph condition is obtained by rendering the target text into a bitmap image using system fonts.

outs, and texture consistency in garment regions. An overview of the entire framework is provided in Figure 1.

Our approach is instantiated upon the Latent Diffusion Model (LDM) (Rombach et al., 2022) paradigm. LDM performs diffusion and denoising in a compressed latent space rather than the full pixel space, thus significantly reducing computational overhead while preserving high-quality visual fidelity. This property has made LDM widely adopted in high-resolution image synthesis and editing scenarios. In a standard diffusion model, the forward diffusion process gradually adds Gaussian noise to the data, transforming the original data distribution into an isotropic Gaussian distribution. Given a time step t, the forward process can be formulated as:

$$q(\mathbf{z}_t \mid \mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t; \sqrt{\alpha_t}\mathbf{z}_0, (1-\alpha_t)\mathbf{I}), \tag{1}$$

where z_0 denotes the original latent representation, z_t represents the noisy latent variable at time step t, and α_t is a predefined noise scheduling coefficient.

The reverse denoising process is learned by a parameterized neural network ϵ_θ, whose objective is to predict the added noise given the noisy latent variable and the conditioning information.

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{z}_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathcal{C})\|_2^2\right], \tag{2}$$

where C denotes the conditioning information, which is typically implemented as text embeddings in standard text-to-image generation tasks.

As illustrated in Figure 1, GlyphFashion is built upon a latent diffusion backbone and performs controllable local editing of garment text through three categories of input signals: text-aware conditioning, ControlNet-based multimodal controls, and textual prompt encodings. Given an input garment image and user-specified multimodal editing conditions, the goal of GlyphFashion is to modify only the specified region while preserving structural alignment, texture continuity, and textual semantic fidelity. The model takes as input the original image I, editing mask M, target textual content P_{text}, global caption P_{cap}, glyph and layout maps G, L, and sketch/color-based control conditions S, C. The model generates new printed text only within regions where $M = 1$, while ensuring that the non-edited regions of the output image $\hat{I}$ remain identical to the input.

To achieve the above objective, we adopt a Latent Diffusion Model (LDM) as the base generator and integrate text-aware features, ControlNet guidance branches, and global textual prompts to form the following conditional noise prediction formulation:

$$\epsilon_\theta(z_t, E_{txt}(P_{cap}, P_{text}), E_{glyph}(G, L), E_{ctrl}(S, C, M)) \tag{3}$$

where E_{text} denotes the text embeddings produced by a CLIP (Radford et al., 2021) text encoder, E_{glyph} represents the glyph-and-layout features generated by the proposed text-aware conditioning module, and E_{ctrl} corresponds to the sketch/mask/color features extracted by the ControlNet branch.

The final result is obtained by performing a region-level blending between the model-generated output and the original image, ensuring that the editing operation remains strictly localized in space and does not induce unintended modifications to the global garment structure or background regions.

$$\mathbf{I}_{\text{out}} = \mathbf{M} \odot \hat{\mathbf{I}} + (1 - \mathbf{M}) \odot \mathbf{I}. \tag{4}$$

3.2 Text-aware Conditioning Module

To enhance the diffusion model's awareness of garment text content, we propose a Text-Aware Conditioning Module (TACM), which explicitly models the glyph appearance and spatial layout of printed text. In contrast to approaches that rely solely on semantic text embeddings, our method incorporates spatially aligned visual conditions to directly encode character structures, thereby providing more fine-grained textual constraints to the diffusion process.

In implementation, to avoid entanglement between character content and geometric layout, TACM takes as input a glyph map $G \in R^{1 \times H \times W}$, and a layout map $L \in R^{1 \times H \times W}$, which respectively encode character appearance and geometric placement of textual regions. To avoid entanglement between character shape and layout structure, TACM employs two independent convolutional encoding branches for feature extraction:

$$F_G = \mathcal{E}_G(G), F_L = \mathcal{E}_L(L), \tag{5}$$

where $\mathcal{E}_G(\cdot)$ and $\mathcal{E}_L(\cdot)$ consist of a series of convolutional layers followed by nonlinear activations. Both outputs maintain identical spatial resolution to ensure alignment during subsequent feature fusion.

Both feature maps retain the same spatial resolution to preserve alignment. The encoded glyph and layout features are then fused via a convolutional fusion layer f to produce the final TACM features:

$$F_{tacm} = f(F_G + F_L). \tag{6}$$

The resulting F_{tacm} is spatially aligned and can be injected into the diffusion network, where it is integrated with ControlNet features at multiple scales to provide persistent, fine-grained structural guidance during denoising.

By providing explicit structural priors, TACM enables conditional injection of both textual content and geometric layout, allowing the diffusion model to separately reason about character appearance and typographic configuration. This effectively mitigates issues such as stroke discontinuities and deformations, while achieving precise layout control, and demonstrates particularly strong performance in challenging scenarios, such as curved garments or complex fabric textures.

3.3 Training Objectives and Inference Strategy

In garment-focused text generation and editing tasks, relying solely on the diffusion model's denoising objective is insufficient to guarantee text readability and structural stability. To address this issue, inspired by prior work in text-aware image generation (e.g., RepText (Wang et al., 2025)), we introduce an OCR-aware auxiliary supervision signal during training to enhance the model's ability to capture textual visual characteristics.

Specifically, during diffusion training, the model first predicts the noise at timestep t (denoted as ϵ_t), from which an estimate of the noise-free latent variable $\hat{z_0}$ is obtained. This latent is subsequently decoded through the VAE decoder into pixel space to produce the reconstructed image $\hat{x_0}$. Since the training data provides precise annotations of text regions, we localize the corresponding text patches from both the original image x_0 and the reconstructed image $\hat{x_0}$, and feed these cropped regions into a pretrained OCR network for feature extraction.

Within the OCR model, we extract the intermediate feature representations prior to the final fully connected layer as high-level semantic descriptors of the text regions. By constraining the discrepancy between the generated and ground-truth text features within the OCR feature space, the model is encouraged to produce text that is structurally coherent and more easily recognizable. The OCR-aware loss is defined as follows:

$$\mathcal{L}_{\text{ocr}} = \sum_{p} \frac{1}{hw} \sum_{h,w} \left\| \mathbf{m}_p - \mathbf{m}'_p \right\|_2^2 , \tag{7}$$

where m_p and m'_p denote the OCR-extracted feature representations at location p for the ground-truth and generated text regions, respectively.

Finally, the overall training objective of the model is formulated as a combination of the diffusion denoising loss and the OCR-aware supervision loss:

$$\mathcal{L} = \mathcal{L}_{denoise} + \lambda \mathcal{L}_{ocr}, \tag{8}$$

where λ denotes the weighting factor used to balance the denoising loss and the OCR-based auxiliary supervision, which is set to a small constant in our experiments. This auxiliary loss is applied only during training and introduces no additional computational overhead at inference time.

4 Experiment

4.1 Experimental Setup

Datasets. The dataset used in this work is constructed by filtering and restructuring the publicly available IGPair dataset (Shen et al., 2025). We first apply PP-OCRv3 (Li et al., 2022a) to detect and recognize printed text in IGPair clothing images, obtaining both the spatial locations and corresponding textual content. To ensure data reliability and consistency, a series of filtering rules is designed to constrain OCR confidence scores, text-region scales, and occlusion conditions, thereby removing low-quality samples that are unsuitable for text editing tasks. For annotation and conditioning construction, we employ the LLaVA (Liu et al., 2023) model to regenerate dense image-level captions for each sample, providing more complete and consistent textual descriptions. In addition, sketch priors are generated for each detected text region using an unsupervised method generating automatic line drawings (Chan et al., 2022), which serve as structural constraints for local editing. In addition, simplified color condition maps are constructed by downsampling the original images to a low resolution and subsequently upsampling them back to the original size, providing color guidance during the text editing process. During test set construction, to assess the model's actual ability in text generation and replacement, we remove the detected text regions using a diffusion-based inpainting model (Yu et al., 2023) to obtain input images without text. The model outputs are then compared against the original text-containing images to quantitatively and qualitatively evaluate text generation quality, spatial alignment, and texture continuity. In total, we construct an OCR-aware clothing image editing dataset consisting of 10,258 training samples and 2,000 test samples, providing a reliable basis for training and evaluation of the proposed method. The dataset covers diverse clothing categories, texture types, and text layout styles, offering varied and challenging testing conditions for assessing text generation and editing performance under realistic fashion scenarios.

Metrics. Following AnyText (Tuo et al., 2023), we adopt two key metrics to evaluate the accuracy of text rendering: Sentence Accuracy (Sen.ACC) and Normalized Edit Distance (NED). The evaluation procedure is as follows: for each generated image, text lines are cropped based on the annotated bounding boxes and then recognized by an OCR model. Let s_{gt} denote the ground-truth text content and s_{pred} the recognized prediction. If s_{pred} exactly matches s_{gt}, the text line is considered correct, and this criterion is used to compute Sen.ACC. During evaluation, all methods use the same OCR recognizer to ensure fairness and consistency. In addition, we compute the NED between s_{pred} and s_{gt} to measure character-level similarity. To further assess overall visual fidelity,

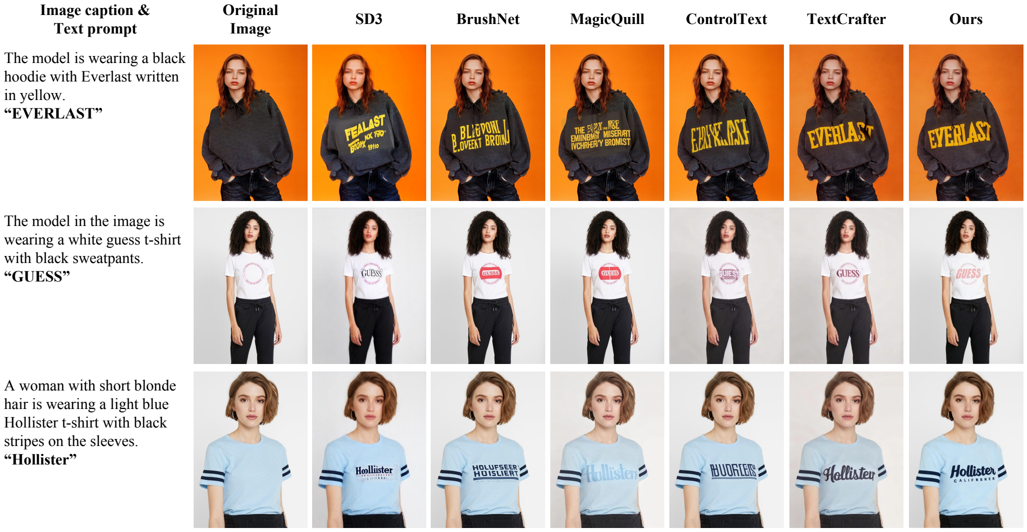
Image caption & Text prompt
Original Image
SD3
BrushNet
MagicQuill
ControlText
TextCrafter
Ours
The model is wearing a black hoodie with Everlast written in yellow.
"EVERLAST"
The model in the image is wearing a white guess t-shirt with black sweatpants.
"GUESS"
A woman with short blonde hair is wearing a light blue Hollister t-shirt with black stripes on the sleeves.
"Hollister"

Figure 2: Qualitative comparison between GlyphFashion and other baselines on our dataset. GlyphFashion preserves the overall image quality while producing text that is both accurate and readable.

we adopt Structural Similarity (SSIM) (Wang et al., 2004) and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) as perceptual quality metrics, and employ CLIPscore (Hessel et al., 2021) to evaluate cross-modal alignment between textual semantics and the generated image. Together, these metrics assess the outputs from multiple perspectives, including text readability, character-level consistency, and holistic visual quality, providing a comprehensive evaluation of model performance on clothing text editing tasks.

Training. Our method is built upon the pretrained latent diffusion model Stable Diffusion (SD 1.5) and incorporates a ControlNet architecture to enable multimodal condition injection. During both training and inference, all images are resized to a fixed resolution of 256×256. The associated conditioning inputs (including glyph maps, position maps, sketches, and masks) are resized accordingly to ensure spatial alignment. During training, the backbone network of the pretrained diffusion model is kept frozen, while the ControlNet module and the proposed Text-aware Conditioning Module are optimized. We adopt the AdamW optimizer with a learning rate of 1×10^{-5}. The model is trained for 10,000 steps with a batch size of 1. The loss weights are determined empirically, with the OCR-aware loss set to 0.05 to balance character-level accuracy and overall visual quality.

4.2 Quantitative Comparison

Table 1 summarizes the quantitative comparison results across different methods on the clothing text editing task. We compare our approach against several representative diffusion-based image generation and editing models. All methods are evaluated on the same test set using identical text prompts, mask configurations, and OCR-based evaluation protocols to ensure fair comparison. With respect to text rendering metrics, our method achieves the best performance on both Sen.ACC and NED. This demonstrates that the proposed text-aware conditioning and OCR-guided supervision are effective in mitigating common issues observed in diffusion-based clothing text generation, such as missing characters and spelling errors, thereby improving character-level accuracy and readability. In contrast, approaches that do not explicitly model textual structure tend to produce degraded text when dealing with small-scale fonts or complex backgrounds. Regarding visual quality and cross-modal consistency, our method maintains competitive performance on SSIM and LPIPS, while achieving a relatively high CLIP score. This indicates that enhancing text readability does not compromise global garment structure, texture continuity, or semantic alignment between text and image. In summary, the quantitative results show that our method achieves a more favorable balance among text readability, character-level consistency, and holistic visual quality. This advantage can be attributed

Table 1: Quantitative results on our dataset. Bold numbers indicate the best performance for each metric, while underlined values denote the second-best performance.

Method	SSIM ↑	LPIPS ↓	Sen.ACC ↑	NED ↑	ClipScore ↑
SD3 (Esser et al., 2024)	0.8516	0.2198	0.4912	0.6055	26.29
BrushNet (Ju et al., 2024)	0.8999	0.0826	0.5842	0.7124	27.33
MagicQuill (Liu et al., 2025)	0.9099	0.1508	0.5938	0.7284	27.97
ControlText (Jiang et al., 2025)	0.9035	0.1921	0.6938	0.7827	28.89
TextCrafter (Du et al., 2025)	0.9146	0.1635	0.7491	0.8028	30.56
Ours	**0.9392**	**0.0691**	**0.7676**	**0.8146**	**30.63**

to the proposed text-aware conditioning module, which explicitly models glyph appearance and spatial layout, and the ControlNet-based multimodal injection mechanism that enables fine-grained and stable text editing in complex garment scenarios.

4.3 Qualitative Comparison

To more intuitively analyze the differences among methods in the fashion text editing task, we further present qualitative comparisons across various approaches, as shown in Figure 2. From the perspective of text generation quality, existing methods often struggle to maintain stable character structures in scenarios involving complex garment textures or small-scale text, frequently resulting in blurry characters, broken strokes, or spelling errors. In contrast, our method produces clearer and structurally intact character shapes with sharper contours and well-preserved stroke details, while maintaining high consistency with the user-specified target text. This observation suggests that explicitly modeling glyph appearance and incorporating OCR-aware supervision improves character-level generation reliability in diffusion models. Regarding spatial layout and local consistency, several comparison methods tend to introduce position shifts or unintentionally affect textures outside the editing region, leading to visually unnatural results. Our approach maintains accurate spatial alignment between the generated text and the underlying garment structure, while preserving the original appearance in non-edited areas, thereby achieving a favorable balance between local modification and global consistency. We attribute this to the persistent structural constraints provided by the text-aware conditioning during the denoising process and the region-preserving blending strategy adopted during inference. Overall, the qualitative results further validate the advantages of the proposed method in the fashion text editing task. With multimodal text-aware conditioning and structured supervision, our model achieves clearer, more stable, and more controllable text generation under challenging garment scenarios, consistent with the quantitative findings.

4.4 Ablation Study

To assess the contribution of each key component in our framework, we further conduct ablation studies focusing on the impact of the OCR-aware loss term and the proposed text-aware conditioning module on clothing text editing performance. Specifically, we investigate how these components influence both text generation accuracy and overall visual quality. We design two ablation variants as follows: (1) w/o OCR loss, where the OCR-aware character-level supervision is removed during training, and optimization relies solely on the diffusion reconstruction loss; (2) w/o Text-aware Conditioning Module, where the proposed text-aware conditioning module is discarded, leaving only the basic conditional diffusion process with ControlNet for editing. All ablated models are trained under the same configurations and evaluated on the same test set to ensure fairness. The ablation results are summarized in Table 2. We observe that removing the OCR-aware loss leads to noticeable degradation in text-related metrics such as Sen.ACC and NED, with generated text exhibiting more frequent character omissions and spelling errors. This demonstrates that character-level supervision plays an important role in improving textual semantic fidelity. A more severe performance drop is observed when removing the text-aware conditioning module, where the model struggles to maintain text readability and spatial alignment, particularly in cases involving complex garment textures or small-scale text. In such settings, the generated text often degenerates into blurred or unstable patterns. Overall, the ablation results show that the text-aware conditioning

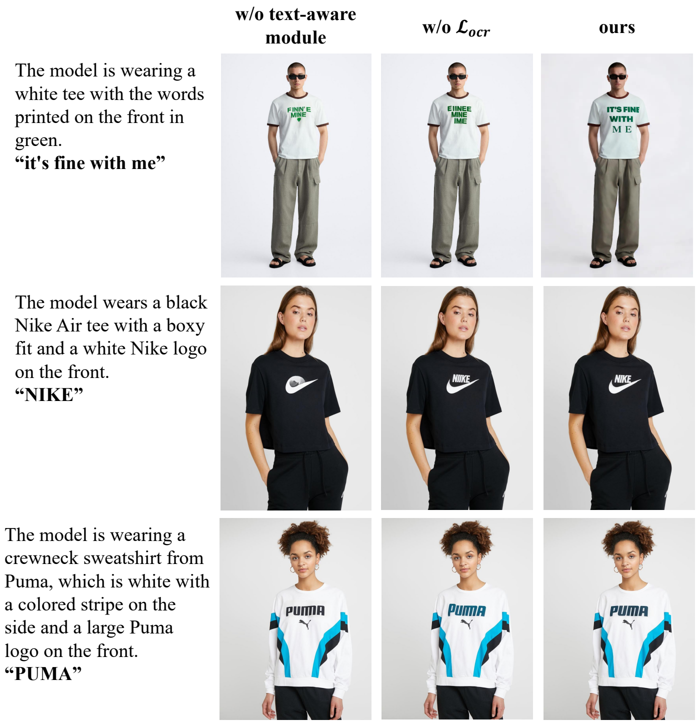
w/o text-aware module
w/o $\mathcal{L}_{ocr}$
ours
The model is wearing a white tee with the words printed on the front in green.
"it's fine with me"
The model wears a black Nike Air tee with a boxy fit and a white Nike logo on the front.
"NIKE"
The model is wearing a crewneck sweatshirt from Puma, which is white with a colored stripe on the side and a large Puma logo on the front.
"PUMA"

Figure 3: Qualitative results of the ablation study.

Table 2: Quantitative results of the ablation study.

Method	SSIM ↑	LPIPS ↓	Sen.ACC ↑	NED ↑	ClipScore ↑
w/o TACM	0.9127	0.0717	0.7262	0.6816	29.49
w/o $\mathcal{L}_{ocr}$	0.9288	0.0729	0.7455	0.7153	29.93
ours	**0.9392**	**0.0691**	**0.7676**	**0.8146**	**30.63**

module and the OCR-aware loss provide complementary structural and character-level supervision, leading to more reliable text generation during localized garment editing.

5 Conclusion

This work addresses the challenge of generating printed text during localized garment image editing by proposing a multimodal, fine-grained diffusion-based framework for fashion text editing. By introducing text-aware conditional modeling and OCR-guided supervision, the method improves the controllability and stability of text generation on garments. Explicit modeling of glyph appearance and spatial layout further alleviates issues such as blurry or inaccurate characters. Despite its effectiveness, the framework still has limitations, including handling multilingual text and higher-resolution editing. Future work will explore stronger text structure modeling and more efficient multimodal interaction to further advance diffusion-based intelligent fashion design.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62206046 and No. 62477006) and Donghua University 2024 Cultivation Project of Discipline Innovation.

REFERENCES

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Caroline Chan, Frédo Durand, and Phillip Isola. Learning to generate line drawings that convey geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7915–7925, 2022.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *Advances in Neural Information Processing Systems*, 36:9353–9387, 2023.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. In *European Conference on Computer Vision*, pp. 386–402. Springer, 2024.

Nikai Du, Zhennan Chen, Shan Gao, Zhizhou Chen, Xi Chen, Zhengkai Jiang, Jian Yang, and Ying Tai. Textcrafter: Accurately rendering multiple texts in complex visual scenes. *arXiv preprint arXiv:2503.23461*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Bowen Jiang, Yuan Yuan, Xinyi Bai, Zhuoqun Hao, Alyson Yin, Yaojie Hu, Wenyu Liao, Lyle Ungar, and Camillo J Taylor. Controltext: Unlocking controllable fonts in multilingual text rendering without font annotations. *arXiv preprint arXiv:2502.10999*, 2025.

Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European Conference on Computer Vision*, pp. 150–168. Springer, 2024.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Taekyung Lee, Donggyu Lee, and Myungjoo Kang. Pointt2i: Llm-based text-to-image generation via keypoints. *arXiv preprint arXiv:2506.01370*, 2025.

Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoting Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, et al. Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system. *arXiv preprint arXiv:2206.03001*, 2022a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Zichen Liu, Yue Yu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Wen Wang, Zhiheng Liu, Qifeng Chen, and Yujun Shen. Magicquill: An intelligent interactive image editing system. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13072–13082, 2025.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4296–4304, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023.

Fei Shen, Xin Jiang, Xin He, Hu Ye, Cong Wang, Xiaoyu Du, Zechao Li, and Jinhui Tang. Imagdressing-v1: Customizable virtual dressing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6795–6804, 2025.

Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. *arXiv preprint arXiv:2311.03054*, 2023.

Yuxiang Tuo, Yifeng Geng, and Liefeng Bo. Anytext2: Visual text generation and editing with customizable attributes. *arXiv preprint arXiv:2411.15245*, 2024.

Haofan Wang, Yujia Xu, Yimeng Li, Junchen Li, Chaowei Zhang, Jing Wang, Kejia Yang, and Zhibo Chen. Reptext: Rendering visual text via replicating. *arXiv preprint arXiv:2504.19724*, 2025.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22428–22437, 2023.

Yukang Yang, Dongnan Gui, Yuhui Yuan, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. Glyphcontrol: Glyph conditional control for visual text generation. *Advances in Neural Information Processing Systems*, 36:44050–44066, 2023.

Zhenhua Yang, Dezhi Peng, Yuxin Kong, Yuyi Zhang, Cong Yao, and Lianwen Jin. Fontdiffuser: One-shot font generation via denoising diffusion with multi-scale content aggregation and style contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 6603–6611, 2024.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023.

Lingjun Zhang, Xinyuan Chen, Yaohui Wang, Yue Lu, and Yu Qiao. Brush your text: Synthesize any scene text on images via diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7215–7223, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Yabo Zhang, Xinpeng Zhou, Yihan Zeng, Hang Xu, Hui Li, and Wangmeng Zuo. Framepainter: Endowing interactive image editing with video diffusion priors. *arXiv preprint arXiv:2501.08225*, 2025.