TexFit: Text-Driven Fashion Image Editing with Diffusion Models

Tongxin Wang, Mang Ye*

National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence
Hubei Key Laboratory of Multimedia and Network Communication Engineering
School of Computer Science, Hubei Luojia Laboratory, Wuhan University, Wuhan, China
{wangtx, yemang}@whu.edu.cn

Abstract

Fashion image editing aims to edit an input image to obtain richer or distinct visual clothing matching effects. Existing global fashion image editing methods are difficult to achieve rich outfit combination effects while local fashion image editing is more in line with the needs of diverse and personalized outfit matching. The local editing techniques typically depend on text and auxiliary modalities (*e.g.*, human poses, human keypoints, garment sketches, etc.) for image manipulation, where the auxiliary modalities essentially assist in locating the editing region. Since these auxiliary modalities usually involve additional efforts in practical application scenarios, text-driven fashion image editing shows high flexibility. In this paper, we propose *TexFit*, a **Tex**t-driven **F**ashion **i**mage Edi**t**ing method using diffusion models, which performs the local image editing only with the easily accessible text. Our approach employs a text-based editing region location module to predict precise editing region in the fashion image. Then, we take the predicted region as the generation condition of diffusion models together with the text prompt to achieve precise local editing of fashion images while keeping the rest part intact. In addition, previous fashion datasets usually focus on global description, lacking local descriptive information that can guide the precise local editing. Therefore, we develop a new DFMM-Spotlight dataset by using region extraction and attribute combination strategies. It focuses locally on clothes and accessories, enabling local editing with text input. Experimental results on the DFMM-Spotlight dataset demonstrate the effectiveness of our model. Code and Datasets are available at https://texfit.github.io/.

Introduction

The purpose of fashion image editing is to manipulate the clothes and accessories in the given fashion image to show the desired outfit combination for users. Successful applications of fashion image editing algorithms enable users and designers to visualize a personalized clothing in a realistic and visually convincing manner (Jiang and Fu 2017; Pernuš et al. 2023; Baldrati et al. 2023; Lin et al. 2023). This topic has inspiring potential different fields, such as online apparel sales and social media. With the development of generative models, significant research efforts have been devoted towards fashion image manipulation (Jiang and Fu 2017; Patashnik et al. 2021; Xia et al. 2021; Huang et al. 2022; Pernuš et al. 2023; Baldrati et al. 2023; Lin et al. 2023).

Most of existing fashion image editing methods are designed to perform global editing, which refers to the overall style and semantic manipulation of fashion images. They generally follow a pipeline by manipulating the latent code mapped to a latent space, generating an edited image. Due to the holistic character of image manipulation, these methods unable to meet the personalized demand for precise local editing of fashion images. Other than global editing, local editing methods can apply local constraints to restrict the execution region of the editing actions in the fashion image to acquire a fashion image that is modified in the specified region. Compared to global editing methods, local fashion image editing is superior in obtaining diverse and personalized outfit-matching effects. Thus in this paper, we focuses on local editing for fashion images.

For local fashion image editing, existing methods are usually based on GAN designs (Kim, Kim, and Lee 2019; Dong et al. 2020), and generally involve multimodal data (*e.g.*, text, human poses, cloth segmentation, human keypoints, garment sketches, etc.) for local guided editing (Kim, Kim, and Lee 2019; Dong et al. 2020; Baldrati et al. 2023). These local editing methods suffer from two limitations. First, they depend on auxiliary modalities other than text. Compared with text modality, other modalities such as human poses and garment sketches involve additional efforts in practical application scenarios. Secondly, most of the current methods are designed based on GAN. GAN-based methods are not easily trained and struggle to produce high-quality generated images with abundant details. More recently, diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Dhariwal and Nichol 2021; Rombach et al. 2022) have demonstrated their dominance in image generation and editing. Additionally, diffusion models are easier to train than GAN-based methods. Hence we tend to develop a **text-only** local fashion image editing method based on the diffusion models, as illustrated in Figure 1.

In view of the problem of how to use text only for local fashion image editing, it is not difficult to find that existing approach Multimodal Garment Designer (Baldrati et al. 2023) completes the precise positioning of the editing region in the fashion image because of the usage of additional

*Corresponding author.

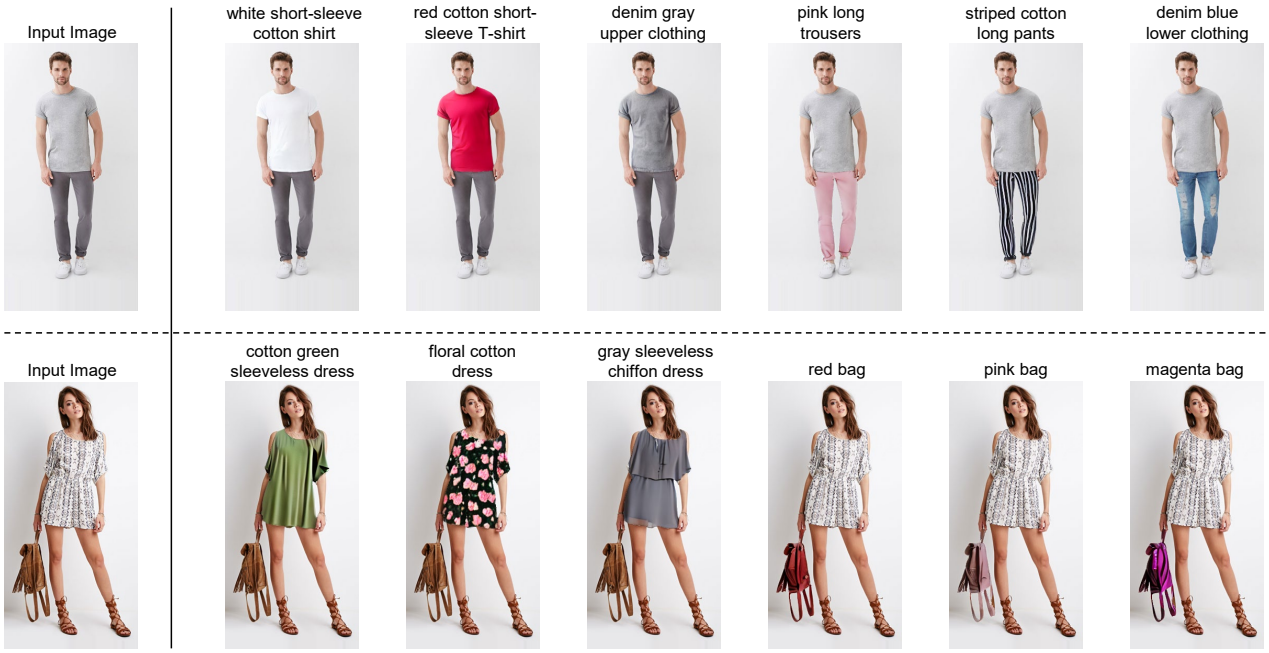

Figure 1: Given an input fashion image and a text prompt, our proposed method *TexFit* can perform precise local image editing and obtain rich outfit-matching effects.

data (*e.g.*, garment sketches and human keypoints modalities), which cannot be completed if only text modality data is used. We propose a local fashion image editing method *TexFit* using diffusion models, which employs a text-based editing region location module (ERLM) to explicitly provide the regions to be edited. The rationale of the location module lies in the fact that the prompt text already contains region information implicitly. For example, for a text prompt like "denim short-sleeve blue shirt", the editing definitely focuses on the "shirt" region of the upper part of the body. Therefore, we introduce the ERLM to explore the hidden region mask modality information to assist the editing of fashion images. With the assistance of the ERLM, we can perform local editing of fashion images by only using the text prompt, without considering auxiliary modalities such as human poses, human keypoints, garment sketches, etc. Figure 2 demonstrates the difference in input modalities between our proposed *TexFit* and other local fashion image editing methods, and it is observed that *TexFit* is more concise and accessible. We employ the diffusion model architecture against GAN for local fashion image editing, which brings quality assurance to our image editing results.

In addition, the current datasets Fashion-Gen (Rostamzadeh et al. 2018) and DeepFashion-MultiModal (Jiang et al. 2022) lack local descriptive sentences that can guide the precise editing of fashion images. To address this issue and also satisfy the testing requirement of our proposed method, we used the region extraction and attribute combination method on the basis of the DeepFashion-MultiModal dataset to create a new fashion image-region-text pair dataset, termed as DFMM-Spotlight dataset. It is expected to facilitate the development of text-driven fashion

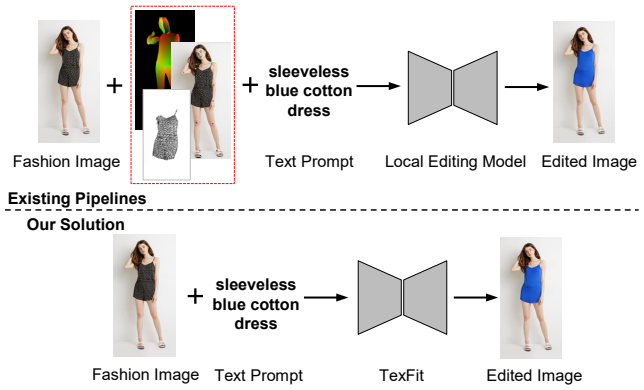

Figure 2: Comparison in input modalities between *TexFit* and other local fashion image editing methods.

image editing task. Our contributions are as follows:

- We propose a text-driven fashion image editing method using diffusion models, which only using the text as the initial generation condition, and could achieve close to the real effect of fashion image generation.
- We propose an editing region location model based on the text prompt to explicitly locate the editing region.
- We also create a new DFMM-Spotlight dataset, which is an image-region-text pair dataset that enables fine-grained text-guided local image editing.
- Experimental results on the DFMM-Spotlight dataset indicate that *TexFit* outperforms other comparative methods in terms of image fidelity and consistency between the editing region and the text prompt.

Related Works

Text-to-Image Generation Text-to-image generation is an important and challenging task, which aims to generate realistic images from natural language descriptions. Most of the early works are based on GANs (Reed et al. 2016; Zhang et al. 2017, 2018a; Xu et al. 2018). (Reed et al. 2016) firstly proposed the GAN-based method by combining text embedding vectors with a generator, it can effectively capture the semantic information in the description and generate realistic images. StackGAN (Zhang et al. 2017) and Stackgan++ (Zhang et al. 2018a) used a multi-stage progressive generative network structure to gradually increase the image resolution during the generation process. AttnGAN (Xu et al. 2018) applied the attention mechanism to the text-to-image generation process in order to align the text description and image content more accurately. Recently, diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Dhariwal and Nichol 2021; Rombach et al. 2022) have received much attention as an effective generative method, and have achieved state-of-the-art results in text-to-image generation. The diffusion model uses a stepwise diffusion process to gradually generate high-quality images by updating the noisy signal several times.

Text-Driven Image Editing Text-driven image editing is designed to enable precise editing of an image from a given text description. Text-driven image editing with GANs has been investigated extensively (Dong et al. 2017; Nam, Kim, and Kim 2018; Li et al. 2020; Patashnik et al. 2021; Xia et al. 2021). As the most powerful competitor of GANs recently, diffusion model shows its extraordinary performance in image editing. SDG (Liu et al. 2023), Blended Diffusion (Avrahami, Lischinski, and Fried 2022), and DiffusionCLIP (Kim, Kwon, and Ye 2022) leveraged the image-text feature alignment capability of CLIP (Radford et al. 2021) to perform text-driven editing of the image. Many works (Avrahami, Lischinski, and Fried 2022; Avrahami, Fried, and Lischinski 2023; Nichol et al. 2022) explored the possibility of using a manual mask to edit the specified region of an image while leaving the rest unchanged. However, providing manual mask is still an enormous work. DiffEdit (Couairon et al. 2022) and Prompt-to-Prompt (Hertz et al. 2022) achieved the goal of text-only editing by automatically predicting a mask before image editing. In the fashion domain, FICE (Pernuš et al. 2023) employed latent code regularization to enhance the GAN inversion process by leveraging CLIP textual embeddings to guide the fashion image editing process. Multimodal Garment Designer (Baldrati et al. 2023) proposed a multimodal-conditioned fashion image editing solution based on latent diffusion models.

Method

In this section, we propose a fashion image editing method using only text. Specifically, given a fashion image $\mathbf{x}_0$, and the editing text prompt P. We expect to obtain a new fashion image $\tilde{\mathbf{x}}$ edited according to the text prompt P based on the original fashion image. The characteristics of the new fashion image, such as human body pose and identity, have to be consistent with the original image, and the manipulation in the image should be in accordance with the text prompt P. Since we only take the text prompt P and original fashion image $\mathbf{x}_0$ as our initial input during editing, different from (Baldrati et al. 2023), we do not use auxiliary data such as human poses, human keypoints and garment sketches that can provide editing region location information. In order to address this issue, we propose an text-based editing region location module to explicitly locate the editing region of the fashion image. Then, the editing region extracted by the location module is employed as the mask conditional input of diffusion models to complete fashion image editing. An overview of our method is shown in Figure 3.

Preliminaries

Diffusion Models In simple terms, diffusion models are a class of probabilistic generative models that turn noise to a representative data sample. The diffusion model consists of two processes: a forward diffusion and a reverse denoising process. Given data $\mathbf{x}_0 \sim q(\mathbf{x})$, the forward diffusion process adds Gaussian noise to it over T steps:

$$\begin{aligned} q(\mathbf{x}_{1:T}|\mathbf{x}_0) &= \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \\ q(\mathbf{x}_t|\mathbf{x}_{t-1}) &= \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \end{aligned} \tag{1}$$

where β_t denotes the noise variance schedule, *i.e.*, $\{\beta_t \in (0,1)\}_{t=1}^{T}$. This forword process brings $\mathbf{x}_t$ progressively closer to the standard Gaussian noise as t increases. By reversing the forward diffusion process, we can obtain the reverse denoising process:

$$\begin{aligned} p_\theta(\mathbf{x}_{0:T}) &= p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \\ p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) &= \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)), \end{aligned} \tag{2}$$

where $\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$ is learnable while $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)$ is fixed as contants (Ho, Jain, and Abbeel 2020). In practice, in order to train the diffusion model, we have the objective function:

$$\mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}, t} \left\| \epsilon - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right\|^2, \tag{3}$$

where $\mathbf{x}_0$ denotes the input data, $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ represents the Gaussian noise added to the input data while t is the denoising timestep. $\boldsymbol{\epsilon}_\theta$ is the noise predictor that is used to estimate the added noise, which is typically implemented using U-Net (Ronneberger, Fischer, and Brox 2015). Once $\boldsymbol{\epsilon}_\theta$ is trained it can be employed to generate an image from a completely random noise image after T steps of denoising.

Latent Diffusion Models Differently from earlier diffusion models that operate at the image pixel level (Dhariwal and Nichol 2021; Nichol et al. 2022), latent diffusion models (LDMs) (Rombach et al. 2022) employ a pre-trained autoencoder to compress the image into a low-dimensional latent space for diffusion. The pre-trained autoencoder is composed of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. To be specific, given an image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$, the encoder $\mathcal{E}$ maps $\mathbf{x}$ into a latent representation $\mathbf{z} = \mathcal{E}(x)$, and the decoder $\mathcal{D}$ is adopted to reconstruct the image from the latent, *i.e.*,

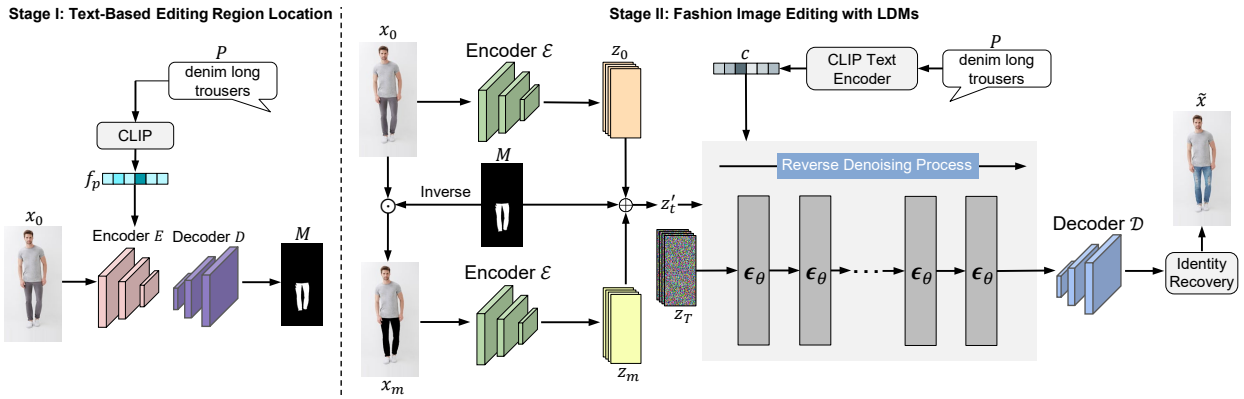

Figure 3: Overview of our *TexFit*. We divide the entire editing process into two stages. In the first stage, we locate the editing region in the fashion image based on the text prompt, and then in the second stage we employ LDMs to precisely edit the visual content within the editing region of the fashion image.

$\tilde{\mathbf{x}} = \mathcal{D}(\mathbf{z}) = \mathcal{D}(\mathcal{E}(\mathbf{x}))$, where $\mathbf{z} \in \mathbb{R}^{h \times w \times 4}$, h, w are downsampled from H, W. By substituting the data point $\mathbf{x}$ in Eq. (3) with the encoded latent $\mathbf{z}$, the training objective function of LDMs can be derived as:

$$\mathbb{E}_{\mathbf{z}_0, \boldsymbol{\epsilon}, t} \left\| \epsilon - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t) \right\|^2. \tag{4}$$

When it comes to conditional generation, this can be implemented with a extended conditional denoising autoencoder $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c})$, where $\mathbf{c}$ denotes the conditional embedding. For a text-guided diffusion model, $\mathbf{c}$ could be the conditional embedding vector of a text prompt.

Considering the advantage that LDMs iteratively denoise data in a low-dimensional latent representation space to generate an image, this manner significantly reduces the computational resources required for image generation. Therefore, we develop our *TexFit* on top of the Stable Diffusion(Rombach et al. 2022) model.

Text-Driven Fashion Image Editing

Text-driven fashion image editing consists of two stages. Since text prompts implicitly contain region information needed for image editing, we use the Editing Region Location Module (ERLM) to locate and uncover hidden editing region information in the first stage. After obtaining the predicted editing region, we take it together with the text prompt as the generating condition for LDMs to perform the fashion image editing in the second stage.

Stage I: Text-Based Editing Region Location Inspired by the pose-to-parsing module in (Jiang et al. 2022), we design the ERLM. Given a text prompt P describing a local outfit in a fashion image $\mathbf{x}_0$, we expect to obtain a region mask $M \in \{0,1\}^{H \times W \times 1}$ corresponding to this description. The text prompt P is first fed into the pre-trained CLIP model to obtain a text embedding f_p. And then the ERLM which is composed of an encoder E and a decoder D takes the fashion image $\mathbf{x}_0$ and f_p as input. The function of layer i of the encoder E can be described as follows:

$$f_{x_i} = E_i([f_{x_{i-1}}, \mathcal{B}(f_p)]), \tag{5}$$

where f_p is broadcasted to have the same spatial size with $f_{x_{i-1}}$ by the spatial broadcast operation $\mathcal{B}(\cdot)$, and f_{x_0} is set to $\mathbf{x}_0$. The function of layer i of the decoder D is:

$$\tilde{f}_{x_i} = D_i([f_{x_i}, \tilde{f}_{x_{i-1}}]). \tag{6}$$

We fed the output $\tilde{f}_x$ of the last layer from the decoder D into fully convolutional layers to predict the editing region mask M. The cross-entropy loss is employed to train the ERLM on the DFMM-Spotlight dataset.

Stage II: Fashion Image Editing with LDMs Given the predicted editing region M obtained in stage I, the masked image can be represented by $\mathbf{x}_m = (1 - M) \odot \mathbf{x}_0$, where $\odot$ denotes the element-wise multiplication operator. In order to preform the generation under the masked condition, we extend $\mathbf{z}_t$ defined in $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c})$ with $\mathbf{z}'_t = [\mathbf{z}_t, m, \mathbf{z}_m]$, where $\mathbf{z}_t$ is concatenated along the channel dimension by m and $\mathbf{z}_m$, $m \in \{0,1\}^{h \times w \times 1}$ is downsampled from M, and $\mathbf{z}_m = \mathcal{E}(\mathbf{x}_m)$ is the latent embedding of masked image $\mathbf{x}_m$. As a result, we get the final training objective function:

$$\mathbb{E}_{\mathbf{z}_0, m, \mathbf{z}_m, \boldsymbol{\epsilon}, t, P} \left\| \epsilon - \boldsymbol{\epsilon}_\theta(\mathbf{z}'_t, t, \mathbf{c}) \right\|^2, \tag{7}$$

where $\mathbf{c} = \tau_\theta(P)$, τ_θ denotes the pre-trained CLIP text encoder taking the text prompt P as input.

During the inference process, we employ the classifier-free guidance technique introduced in (Ho and Salimans 2021), where the noise prediction at each step is weighted by the combination of unconditional and conditional predictions. let $\mathbf{c}_\varnothing = \tau_\theta(\text{“”})$ as the unconditional embedding, the noise prediction at each inference step can be computed by:

$$\tilde{\boldsymbol{\epsilon}}_t = \boldsymbol{\epsilon}_\theta(\mathbf{z}'_t, t, \mathbf{c}_\varnothing) + w \cdot (\boldsymbol{\epsilon}_\theta(\mathbf{z}'_t, t, \mathbf{c}) - \boldsymbol{\epsilon}_\theta(\mathbf{z}'_t, t, \mathbf{c}_\varnothing)), \tag{8}$$

where w represents the guidance scale, higher guidance scale encourages to generate the image that is closely linked to the text prompt P.

To keep the identity of the person in the fashion image and the rest of the image except the editing region unchanged, we combine the edited fashion image $\mathbf{x}_e$ generated by the decoder $\mathcal{D}$ after the inference process with the original fashion image $\mathbf{x}_0$. The final image $\tilde{\mathbf{x}}$ can be obtained by:

$$\tilde{\mathbf{x}} = M \odot \mathbf{x}_e + (1 - M) \odot \mathbf{x}_0, \tag{9}$$

where M denotes the editing region predicted in stage I.

The DFMM-Spotlight Dataset

The text in the current fashion image-text pair dataset is mostly the description of the whole fashion image, lacking local descriptive information that can guide the precise editing of fashion images. In order to address this problem, we collect a new fashion image-region-text pair dataset called **DFMM-Spotlight**, highlighting local cloth.

Data Collection

Data Source We use the DeepFashion-MultiModal dataset (Jiang et al. 2022) as our data source. It contains 11,484 full-body images with human parsing labels of 24 classes. For each image, the dataset provides human parsing annotations including 24 semantic labels of clothes (*top, outer, skirt, dress, pants, rompers*), body components (*hair, face, skin*), and accessories (*eyeglasses, belt, bag, etc.*). Meanwhile, each image is also annotated with clothes shape, texture attributes and a textual description.

Region Prompt Extraction The human parsing labels provided by the DeepFashion-MultiModal dataset can be used to extract **Region Prompt**. We selected five types of semantic labels from them, which are upper clothes (*top*), lower clothes (*pants*), outer clothes (*outer*), dresses (*dress, rompers*), and accessories (*eyeglass, belt, bag*). The pixels in the part of the image that matches the selected semantic label will be set to one, and the rest will be set to zero, resulting in a region prompt image.

Attributes Combination for Text Prompt The clothes shape attributes in the DeepFashion-MultiModal dataset include the length of upper clothes and lower clothes. The length of upper clothes is described as follows: *sleeveless, short-sleeve, medium-sleeve, long-sleeve,* and *not long-sleeve*; The length categories for lower clothes are *three-point, medium short, three-quarter,* and *long*. They can be called **length attribute**. Texture attributes mainly include clothes colors and clothes fabrics. Clothes colors fall into *floral, graphic, striped, pure color, lattice,* and *color block*. As for *pure color*, we classify them into specific colors (*black, gray, red, blue, etc.*) by identifying the HSV color space of the corresponding region. Clothes fabrics consist of *denim, cotton, leather, furry, knitted,* and *chiffon*. Similarly, they are referred to as **color attribute** and **fabric attribute** respectively. For each region prompt extracted in the previous step, we look for the **cloth text** (*e.g.*, *tank top, T-shirt, shorts, trousers*) in the textual description annotation. Finally, we combine the **length attribute**, **color attribute**, **fabric attribute**, and **cloth text** as the **Text Prompt**.

Comparison with Other Datasets

We split the DFMM-Spotlight dataset into a training set with 21377 image-region-text pairs and a test set with 2379 pairs following the original split setting in the DeepFashion-MultiModal dataset. We will make this dataset publicly available and hope that it can aid in the investigation of techniques for the task of local fashion image editing.

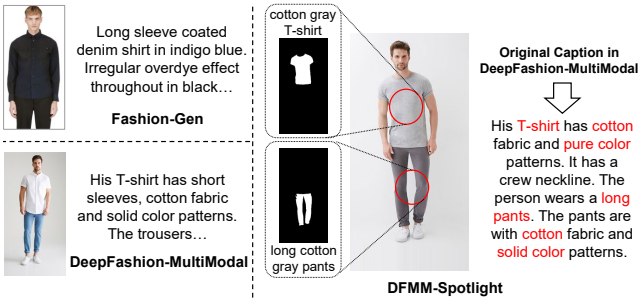

Figure 4: Comparison of DFMM-Spotlight with other fashion datasets.

We compare the samples in the DFMM-Spotlight dataset with those in the current fashion dataset Fashion-Gen and DeepFashion-MultiModal, as illustrated in Figure 4. The text in Fashion-Gen describes a single clothing in the fashion image, although the fashion image also contains other clothes, which is prone to ambiguity. The textual description in the DeepFashion-MultiModal dataset is an overview of all the outfits worn in the fashion image. As a result, neither of the current datasets Fashion-Gen nor DeepFashion-MultiModal describes the fine-grained correspondence between text and garment regions, and thus cannot be applied to the local fashion image editing task. Compared with above datasets, our newly collected DFMM-Spotlight dataset acts like a spotlight, which can illuminate local clothing regions and associate them with brief text prompts to facilitate the local fashion image editing.

Experiments

Experimental Settings

Datasets The experiments are performed on the DFMM-Spotlight dataset. Since the number of test set pairs for DFMM-Spotlight is only 2379, we extends the test set to evaluate fashion image editing model introduced in the second stage. Specifically, we search for several textual descriptions describing the same cloth category (*e.g.*, *tank top, T-shirt, shorts, trousers*) for each text prompt in the dataset. After the extension, we finally get the expanded test set with 10845 image-region-text pairs.

Baselines We choose three Stable Diffusion based image editing methods SDEdit (Meng et al. 2021), SD-Inpaint[1], and DiffEdit (Couairon et al. 2022) as our comparable baselines. SDEdit partially adds noise to the input image and then denoise it for editing. We employ the SDEdit editing technique in the Img2Img function of Stable Diffusion. We set the strength parameter to 0.8, consistent with the original paper. SD-Inpaint is developed on the basis of Stable Diffusion with the extra capability of inpainting the pictures by using a mask. DiffEdit is an editing method that does not require a manual mask, the same as our proposed *TexFit* method. DiffEdit can produce an automatically computed mask by

[1]https://huggingface.co/runwayml/stable-diffusion-inpainting

Figure 5: Qualitative comparison between the competing methods and our proposed *TexFit*.

comparing the prediction noise guided by the source text prompt and the editing text prompt.

Implementation Details All experiments are performed on a single NVIDIA RTX 3090. We downsample all images to 512×256 resolution in the experiments. ERLM is trained on DFMM-Spotlight for 100 epochs with a batch size of 8, adopting the Adam optimizer (Kingma and Ba 2015) and the learning rate is set as 1×10^{-4}. We employ Stable Diffusion v1.4 as the pre-trained model for our second-stage fashion image editing module and initialize additional channel weights after restoring the non-inpainting checkpoint. We finetune it for 140k steps on the DFMM-Spotlight dataset, using the AdamW optimizer (Loshchilov and Hutter 2018) and setting the learning rate to 1×10^{-5}. To save memory, we adopt the strategy of mixed precision (Micikevicius et al. 2018) and gradient accumulation, where the steps for gradient accumulation is set to 4 and the batch size is set to 1. For inference, we employ the PNDM scheduler (Liu et al. 2021) with 50 steps of iteration and set the classifier-free guidance scale w to 7.5. In order to make a fair comparison, we employ Stable Diffusion v1.4 backbone finetuned on DeepFashion-MultiModal for SDEdit and DiffEdit. The finetuning hyperparameters of the backbone are consistent with our second-stage model.

Evaluation Metrics We adopt Fréchet Inception Distance (FID) (Heusel et al. 2017) and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018b) to quantitatively assess the sample fidelity of the generated fashion images. Furthermore, to estimate whether the edited fashion images match the input text prompts, we employ the CLIP Score (CLIP-S) (Hessel et al. 2021). CLIP Score can be used to evaluate the correlation between a generated caption for an image and the actual content of the image. It has been found to be heavily associated with human judgement. We compute CLIP-S by filling the rest of the fashion image with white pixels except for the ground truth editing region.

Method	**FID** (↓)	**LPIPS** (↓)	**CLIP-S** (↑)
SDEdit†	16.34	0.188	**28.18**
SD-Inpaint	12.90	0.063	26.87
DiffEdit†	17.53	0.105	26.28
TexFit (Ours)	**10.77**	**0.052**	28.10

Table 1: Quantitative comparison between baselines and our proposed method on the expanded DFMM-Spotlight test dataset. †: employ Stable Diffusion backbone finetuned on the DeepFashion-MultiModal dataset.

Comparison with Baselines

We report the quantitative results of our *TexFit* and it competing methods on the DFMM-Spotlight test dataset in Table 1. In terms of the results of the FID and LPIPS metrics, our proposed *TexFit* method is superior regarding the fidelity of the fashion image after editing. By observing the results of CLIP-S, we find that *TexFit* has the competitive alignment result between the relevant region of the edited fashion image and the text prompt compared to other methods.

We show the qualitative comparison between the competing methods and our proposed method in Figure 5. It can be seen that *TexFit* can precisely locate the region of the fashion image to be edited according to the text prompt and present the semantic modifications in line with the text prompt. In

Editing Region Location			Fashion Image Editing		Metrics		
DiffEdit	ERLM (Ours)	GT Mask	SD-Inpaint	TexFit (Ours)	FID (↓)	LPIPS (↓)	CLIP-S (↑)
✓			✓		12.92	0.113	27.77
	✓		✓		12.90	0.063	26.87
		✓	✓		13.61	0.073	27.35
✓				✓	58.19	0.179	28.48
	✓			✓	10.77	**0.052**	28.10
		✓		✓	**10.51**	0.060	**28.80**

Table 2: Ablation study of editing region location and fashion image editing modules.

Method	Image Fidelity	Text Matching	ID Preservation
SDEdit	125	74	38
SD-Inpaint	51	133	69
DiffEdit	261	186	52
TexFit (Ours)	**563**	**607**	**841**

Table 3: Results of the human-subject evaluation of the compared methods and our proposed method.

contrast, DiffEdit, another method that can automatically generate the mask of the image editing region shows less precision, which is manifested by some deviations of the located editing image regions. We display the visualization results of region generation in Figure 6, through which it is obvious that our proposed ERLM can focus on key editing regions compared to DiffEdit, so as to obtain more precise edited fashion images.

We conduct the human-subject study to evaluate our method based on the human judgment. We organize 36 users to evaluate on 1000 groups of study cases. The invited users are requested to single out the images generated by different methods with the best performance in terms of image fidelity, text matching and id preservation. We report the detailed image selection results in Table 3. Our *TexFit* outperforms the other methods in each evaluation term.

Ablation Study

We conduct ablation studies on the editing region location (**Stage I**) and fashion image editing (**Stage II**) parts of our proposed method. The results are shown in Table 2. ERLM refers to our proposed editing region location module in **Stage I**. GT Mask denotes the ground truth region mask in DFMM-Spotlight dataset. According to Table 2, both our fashion image editing module and ERLM show the best performance when the editing region location and fashion image editing module techniques are fixed respectively, which proves the effectiveness of these two modules. It is worth noting that when we combine the editing region mask generated by DiffEdit with our second-stage fashion image editing model, the fidelity of the image is greatly decreased, which indicates that the editing region located by DiffEdit lacks accuracy and causes considerable interference to our editing process. This further corroborates the validity of our proposed ERLM.

Figure 6: Visual comparison of editing region masks generated by *TexFit* and DiffEdit.

Conclusion

In this paper, we propose a text-driven fashion image editing method based on diffusion models, which allows local editing of fashion images using readily available text in real application scenarios. The key design of our model lies in the application of the ERLM, which explicitly mines out the hidden editing region information in the text prompt. Furthermore, we collect a DFMM-Spotlight dataset based on the existing fashion DeepFashion-MultiModal dataset, which can provide fine-grained correspondence between the text prompt and the editing region for local fashion image editing. We conduct experiments on the newly collected DFMM-Spotlight dataset to demonstrate the effectiveness of our proposed method.

Acknowledgements

This work is partially supported by National Natural Science Foundation of China under Grant (62176188, 62066021) and the Special Fund of Hubei Luojia Laboratory (220100015).

References

Avrahami, O.; Fried, O.; and Lischinski, D. 2023. Blended latent diffusion. *TOG*, 42(4): 1–11.

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images. In *CVPR*, 18208–18218.

Baldrati, A.; Morelli, D.; Cartella, G.; Cornia, M.; Bertini, M.; and Cucchiara, R. 2023. Multimodal Garment Designer: Human-Centric Latent Diffusion Models for Fashion Image Editing. In *ICCV*.

Couairon, G.; Verbeek, J.; Schwenk, H.; and Cord, M. 2022. DiffEdit: Diffusion-based semantic image editing with mask guidance. In *ICLR*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. In *NeurIPS*, 8780–8794.

Dong, H.; Liang, X.; Zhang, Y.; Zhang, X.; Shen, X.; Xie, Z.; Wu, B.; and Yin, J. 2020. Fashion editing with adversarial parsing learning. In *CVPR*, 8120–8128.

Dong, H.; Yu, S.; Wu, C.; and Guo, Y. 2017. Semantic image synthesis via adversarial learning. In *ICCV*, 5706–5714.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-or, D. 2022. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *ICLR*.

Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*, 7514–7528.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, volume 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *NeurIPS*, 6840–6851.

Ho, J.; and Salimans, T. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Huang, L.; Liu, Y.; Wang, B.; Pan, P.; and Jin, R. 2022. A Trend-Driven Fashion Design System for Rapid Response Marketing in E-commerce. In *AAAI*, 13179–13181.

Jiang, S.; and Fu, Y. 2017. Fashion style generator. In *IJCAI*, 3721–3727.

Jiang, Y.; Yang, S.; Qiu, H.; Wu, W.; Loy, C. C.; and Liu, Z. 2022. Text2human: Text-driven controllable human image generation. *TOG*, 41(4): 1–11.

Kim, B.-K.; Kim, G.; and Lee, S.-Y. 2019. Style-controlled synthesis of clothing segments for fashion image manipulation. *TMM*, 22(2): 298–310.

Kim, G.; Kwon, T.; and Ye, J. C. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *CVPR*, 2426–2435.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Li, B.; Qi, X.; Lukasiewicz, T.; and Torr, P. H. 2020. Manigan: Text-guided image manipulation. In *CVPR*, 7880–7889.

Lin, A.; Zhao, N.; Ning, S.; Qiu, Y.; Wang, B.; and Han, X. 2023. FashionTex: Controllable Virtual Try-on with Text and Texture. In *SIGGRAPH*.

Liu, L.; Ren, Y.; Lin, Z.; and Zhao, Z. 2021. Pseudo Numerical Methods for Diffusion Models on Manifolds. In *ICLR*.

Liu, X.; Park, D. H.; Azadi, S.; Zhang, G.; Chopikyan, A.; Hu, Y.; Shi, H.; Rohrbach, A.; and Darrell, T. 2023. More control for free! image synthesis with semantic diffusion guidance. In *WACV*, 289–299.

Loshchilov, I.; and Hutter, F. 2018. Decoupled Weight Decay Regularization. In *ICLR*.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *ICLR*.

Micikevicius, P.; Narang, S.; Alben, J.; Diamos, G.; Elsen, E.; Garcia, D.; Ginsburg, B.; Houston, M.; Kuchaiev, O.; Venkatesh, G.; et al. 2018. Mixed Precision Training. In *ICLR*.

Nam, S.; Kim, Y.; and Kim, S. J. 2018. Text-adaptive generative adversarial networks: manipulating images with natural language. In *NeurIPS*.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; Mcgrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *ICML*, 16784–16804. PMLR.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2085–2094.

Pernuš, M.; Fookes, C.; Štruc, V.; and Dobrišek, S. 2023. FICE: Text-Conditioned Fashion Image Editing With Guided GAN Inversion. arXiv:2301.02110.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative adversarial text to image synthesis. In *ICML*, 1060–1069. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241. Springer.

Rostamzadeh, N.; Hosseini, S.; Boquet, T.; Stokowiec, W.; Zhang, Y.; Jauvin, C.; and Pal, C. 2018. Fashion-Gen: The Generative Fashion Dataset and Challenge. arXiv:1806.08317.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. In *ICLR*.

Xia, W.; Yang, Y.; Xue, J.-H.; and Wu, B. 2021. Tedigan: Text-guided diverse face image generation and manipulation. In *CVPR*, 2256–2265.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 1316–1324.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 5907–5915.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2018a. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *TPAMI*, 41(8): 1947–1962.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018b. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 586–595.