DiffCloth: Diffusion Based Garment Synthesis and Manipulation via Structural Cross-modal Semantic Alignment

Xujie Zhang[1*] Binbin Yang[2*] Michael C. Kampffmeyer[4] Wenqing Zhang[1]
Shiyue Zhang[1] Guansong Lu[3] Liang Lin[2] Hang Xu[3] Xiaodan Liang[1,5†]
[1]Shenzhen Campus of Sun Yat-Sen University [2]Sun Yat-Sen University
[3]Huawei Noah's Ark Lab [4]UiT The Arctic University of Norway [5]MBZUAI
{zhangxj59,yangbb3,zhangwq76,zhangshy223}@mail2.sysu.edu.cn,michael.c.kampffmeyer@uit.no
luguansong@huawei.com,linliang@ieee.org,chromexbjxh@gmail.com,xdliang328@gmail.com

(a) Garment Synthesis **(b) Garment Manipulation**

Figure 1. Results of our proposed DiffCloth. DiffCloth is able to produce garments with part-level semantics well-aligned to the prompt and allows for precise manipulation of the generated results by simply modifying the text description.

Abstract

Cross-modal garment synthesis and manipulation will significantly benefit the way fashion designers generate garments and modify their designs via flexible linguistic interfaces. However, despite the significant progress that has been made in generic image synthesis using diffusion models, producing garment images with garment part level semantics that are well aligned with input text prompts and then flexibly manipulating the generated results still remains a problem. Current approaches follow the general text-to-image paradigm and mine cross-modal relations via simple cross-attention modules, neglecting the structural correspondence between visual and textual representations in the fashion design domain. In this work, we instead introduce DiffCloth, a diffusion-based pipeline for cross-modal garment synthesis and manipulation, which empowers diffusion models with flexible compositionality in the fashion domain by structurally aligning the cross-modal semantics. Specifically, we formulate the part-level cross-modal alignment as a bipartite matching problem between the linguistic Attribute-Phrases (AP) and the visual garment parts which are obtained via constituency parsing and semantic segmentation, respectively. To mitigate the issue of attribute confusion, we further propose a semantic-bundled cross-attention to preserve the spatial structure similarities between the attention maps of attribute adjectives and part nouns in each AP. Moreover, DiffCloth allows for manipulation of the generated results by simply replacing APs in the text prompts. The manipulation-irrelevant regions are recognized by blended masks obtained from the bundled attention maps of the APs and kept unchanged. Extensive experiments on the CM-Fashion benchmark demonstrate that

*Equal contribution. †Corresponding author.

DiffCloth both yields state-of-the-art garment synthesis results by leveraging the inherent structural information and supports flexible manipulation with region consistency.

1. Introduction

Leveraging artificial intelligence to generate and alter garment images based on control signals from a variety of modalities has the potential to revolutionize the fashion design process. Particularly, cross-modal garment synthesis [6, 13, 14, 18, 19, 28, 39] and manipulation by linguistic interfaces have gradually attracted increasing attention from the academic community. Unfortunately, the visual semantics in the fashion domain are different from those in generic image generation tasks due to its inherent structural property, *e.g.* each type of garment has a distinct shape and can be partitioned into several garment parts. However, existing work [6, 13, 14, 18, 19, 28, 39] on cross-modal garment synthesis are primarily built on two-stage pipelines of generic generative transformers and ignore the structural correspondences between the garment images and the input text prompts. This leads to imprecise cross-modal semantic alignment and poor semantic compositionality.

Given the recent success of diffusion models [23, 26, 29, 31], which provide flexible control of the generative process through guidance mechanisms, departing from prior approaches and leveraging diffusion models appears a natural approach. However, we observe the following two semantic issues when applying state-of-the-art text-based image generation models to the fashion domain: 1) Garment Part Leakage, where one or more of the garment parts described in the prompt are not actually generated in the image; and 2) Attribute Confusion, where the attributes and the garment parts are wrongly paired or some attributes are ignored in the generated image. Examples of the aforementioned issues are provided in Fig. 2. In Fig. 2(a), examples of garment part leakage are provided, where the model fails to generate the pockets in the dusty rose jacket and the button fastening in the blue shirt. In Fig. 2(b), examples of attribute confusion are provided, where the color attributes 'blue' and 'brown' bind to the incorrect garment parts and the 'plain white' attribute is missing in the striped shirt.

To solve the above issues, we propose **DiffCloth**, a diffusion model with structural semantic consensus guidance to achieve accurate fine-grained part-level semantic alignment. To be specific, a semantic segmentor is trained to explore the visual structure and divide the visual garment into part-level images, *e.g.*, sleeves, body piece, hood, etc. Additionally, a constituency parsing tree is leveraged as a linguistic structural parser to extract the collection of Attribute-Phrases. By formulating the cross-modal semantic alignment as a bipartite matching problem between these two sets of semantic components, we

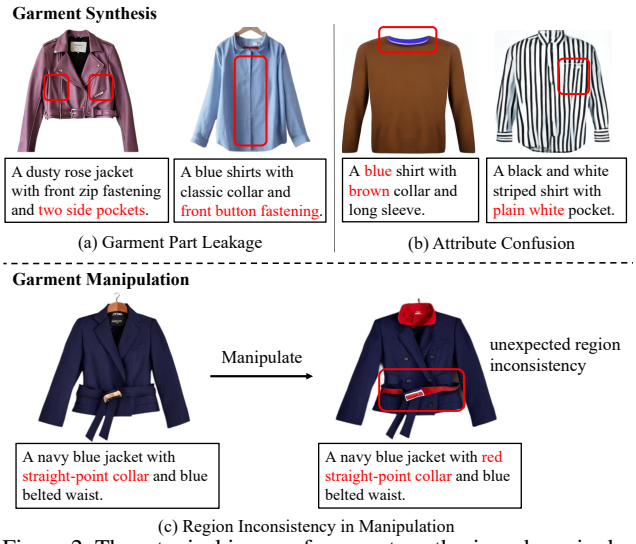

Figure 2. Three typical issues of garment synthesis and manipulation. (a) Garment Part Leakage: one or more of the garment parts described in the prompt are not accurately generated; (b) Attribute Confusion: the attributes and garment parts are wrongly paired or some attributes are ignored; (c) Region Inconsistency in Manipulation: the manipulation-irrelevant regions are carelessly modified.

introduce a Hungarian matching loss as the summation of CLIP-similarities [24] between the part-level images and the Attribute-Phrases. This Hungarian matching loss can be used to guide the diffusion model to achieve structural consensus across images and text. Furthermore, we propose a semantic-bundled cross-attention module to avoid the aforementioned attribute confusion issue. Specifically, we observe that the attention maps of the attribute adjective and the garment part nouns are different when attribute confusion occurs while they share similar spatial structures when attributes are matched to the correct garment parts. Hence, we propose to preserve the spatial structure similarity between the attribute adjective and the garment part subject in the cross-attention module by a semantic-bundled loss, which aims to minimize the Jensen-Shannon divergence [8] between these two maps. This semantic-bundle loss is also utilized to guide the sampling process of DiffCloth.

In order to further allow easy manipulation of the generated images, DiffCloth introduces a mechanism to manipulate input images based solely on changes in the input text prompt and, unlike prior approaches [1, 21], does not require explicit masking of the areas that should be changed. By injecting the cross-attention maps during the diffusion steps, DiffCloth can automatically find which pixels should be attended to and should be modified. When for instance changing the attribute, *e.g.* "long sleeve" $\rightarrow$ "short sleeve", only the cross-attention maps of the bundled Attribute-Phrase need to be changed and the attention maps of other textual tokens can be frozen. Moreover, we propose a consistency loss to prevent irrelevant content from

being carelessly edited. An example of unexpected region inconsistency is given in Fig.2 (c), where the blue belt is wrongly modified to a red one. The consistency loss is further designed to preserve the pixel-level consistency of the exclusive area indicated by the attention map of the changed tokens. Comprehensive experiments on the CM-Fashion benchmark demonstrate that DiffCloth yields state-of-the-art generation results in garment synthesis and further supports flexible manipulation by editing the text prompt in a user-friendly manner.

Our main contributions are summarized as follows:

- We propose a structural semantic consensus guidance to address the structural semantic alignment across visual garments and linguistic attribute-phrases as a bipartite matching problem via the Hungarian algorithm.
- We propose a new semantic-bundled cross-attention, which encourages spatial structure similarity between the cross-attention maps of attributes and part subjects, to alleviate attribute confusion issues.
- We introduce a region consistency mechanism to prevent irrelevant content from being modified during garment manipulation.
- Extensive experiments on the CM-Fashion benchmark verify the superiority of DiffCloth, particularly in terms of accurate text-image alignment for both garment synthesis and manipulation.

2. Related work

Text-guided image synthesis. Early works explored text-guided image synthesis in the context of GANs [32, 36, 37, 40, 42] or VQVAE [33]. However, more recent research has demonstrated impressive results using large-scale auto-regressive models [27, 38] and diffusion models [23, 26, 29, 31]. Stable Diffusion [29] proposes to encode an image with an autoencoder and then leverage a diffusion model to generate continuous feature maps in the latent space. Imagen [31] addresses the importance of language understanding by using a frozen T5 [25] encoder, a dedicated large language model. However, generating images that faithfully align with the input prompt remains challenging. To enforce heavier reliance on the text, classifier-free guidance [12, 23, 31] allows extrapolating text-driven gradients to better guide the generation by strengthening the reliance on the text. Despite this, the semantic flaws of text-to-image models still exist. Recent work has begun to address this issue, such as Composable Diffusion models [20], which compose multiple outputs of a pre-trained diffusion model. Each output is tasked with capturing different image components which are then joined using compositional operators to attain a unified image. StructureDiffusion [7] and Attend-and-Excite [3] optimize the attention map calculation for better image generation. However, these attempts still fall short of generating garment images with fine-grained compliance with the input prompts as the structural correspondences between garment representations of the two modalities are often ignored. In this work, we strive to achieve part-level cross-modal semantic alignment by aligning those visual and linguistic structured representations in a fine-grained manner.

Image Manipulation with Generative Models. A number of techniques [1, 15, 30, 34] have been developed based on diffusion models to enable editing, personalization and inversion to token space. Dreambooth [30] and Imagic [15] involve fine-tuning of the generative models. ImagenEditor [34] frames editing as text-guided image inpainting, and involves user specified masks. Blended diffusion [1] provides a clip-guided mask-based editing method. However, the mask provided by the user is often not accurate enough, and there will be disharmony in the editing boundary. More recently, Prompt-to-Prompt [9] explored mask-free image editing through the interaction of attention maps. However, the manipulation results often affect content that is irrelevant to the modification, leading to unsatisfactory results. In this paper, we explore an attention-based garment manipulation method by injecting the attention maps of the target Attribute-Phrase (AP) while keeping other regions unchanged using a mask that blends the attention maps.

3. Methodology

Our proposed DiffCloth is built on Stable Diffusion [29], which we briefly review in Sec. 3.1. We then introduce our structural semantic consensus guidance in Sec. 3.2, which addresses the problem of garment part leakage. Our semantic-bundled cross-attention mechanism is then presented in Sec. 3.3 in order to avoid the confusion between attributes before we present our garment manipulation in Sec. 3.4. An overview of DiffCloth is provided in Fig. 3.

3.1. Preparatory

Stable Diffusion. Our proposed DiffCloth is built on Stable Diffusion [29], which consists of an autoencoder model and a diffusion model. The autoencoder is trained to encode an image x_0 as lower-resolution latent maps z_0 for efficient diffusion training:

$$L_{AE} = \|x_0 - \text{Dec}(\text{Enc}(x_0))\|^2, \tag{1}$$

where L_{AE} is the reconstruction loss for training the encoder Enc and decoder Dec. $z_0 = \text{Enc}(x_0)$ and x_0 can be approximately reconstructed by $\text{Dec}(z_0)$. The diffusion model contains two stages: a diffusion and a denoising stage. In the diffusion stage, z_0 is gradually transformed into a normal distribution by gradually adding noise for T

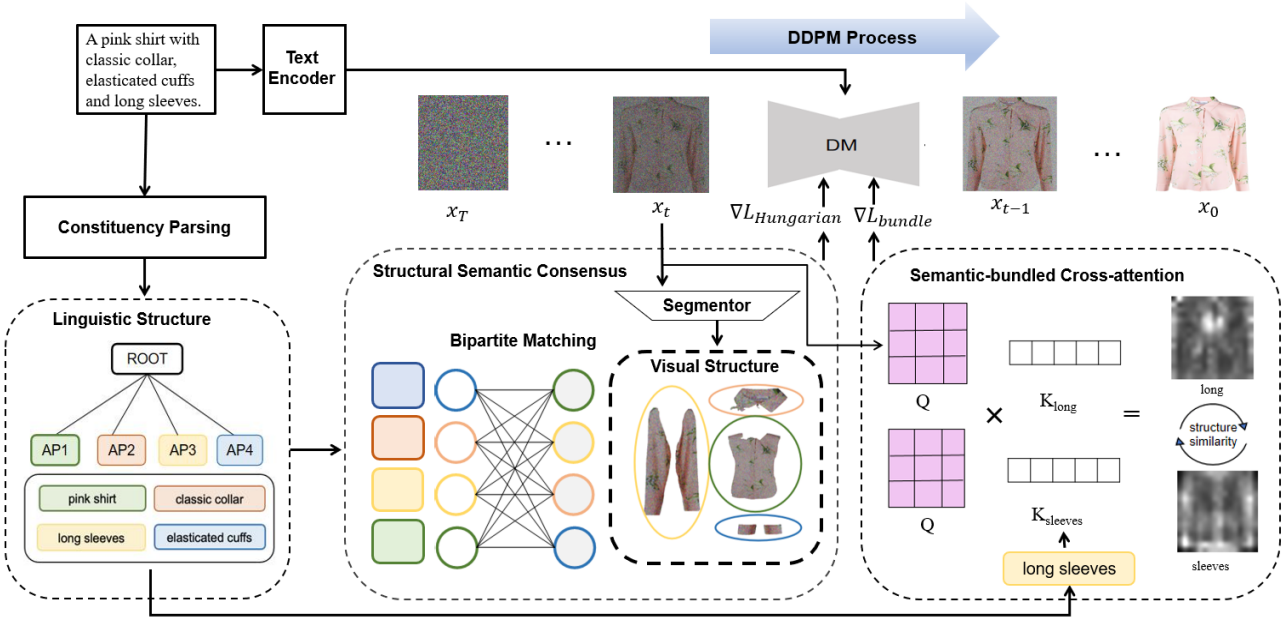

Figure 3. Overview of DiffCloth. During the diffusion step, we leverage constituency parsing to extract the text structure and obtain a tree of all attribute- phrases (APs). Given this structure information, the structural semantic consensus partitions the garment images using a segmentor into multiple visual parts, which are then matched with the APs using a bipartite matching to get structural semantic alignment. This generates the $L_{Hungarian}$ loss. Similarly, to preserve structure similarity between the attention maps of the attribute adjectives and the corresponding garment part subjects we introduce semantic-bundled cross-attention, which addresses the attribute confusion issue via the L_{bundle} loss. More specifically, query Q is obtained from the visual representation X_t, while keys K are computed for each word. L_{bundle} then aims to encouraging similar attention maps for each AP. Finally, the losses are used to refine the feature representation of the diffusion model at each step.

steps following the Gaussian transition $q(z_t|z_{t-1})$:

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

where β denotes the noise scale, $\mathbf{I}$ is the identity matrix, and z_t is the latent of the timestep t .

By optimizing a noise estimator ϵ_θ, the model is trained to reverse the diffusion process and generate images from random noise by optimizing the loss L_{DM}:

$$L_{\text{DM}} = \mathbb{E}_{t,z_0,\epsilon}\left[\|\epsilon_\theta(z_t) - \epsilon\|^2\right]. \tag{3}$$

A synthesized image x_0^* is obtained by denosing noise x_T for T steps and decoding it using the decoder $x_0^* = \text{Dec}(z_0^*)$.

DiffCloth is trained on the garment images by optimizing Eq. (3) and sampled using the guidance from our proposed structure semantic consensus and semantic-bundled losses, which will be detailed in the following sections.

3.2. Structural Semantic Consensus Guidance

Our structural semantic consensus guidance is based on the intuition that there are structural similarities between visual and textual representations in cross-modal garment synthesis. As shown in Fig. 3, a segmentor trained on noisy inputs can be used to partition garment images into multiple visual parts that adhere to the standard structural patterns used by humans in garment design.[1] The visual structured components can be denoted as:

$$\mathbf{V} = [V_{full}, V_1, V_2, ..., V_m], \tag{4}$$

where V_{full} denotes the full garment image and V_i is the i^{th} part image of V_{full} indicated by the mask M_i, *e.g.*, sleeves, body piece, hood.

Similarly, we can obtain the text structure by leveraging constituency parsing to extract a tree of all Attribute-Phrases (APs), which are crucial for depicting the semantic components of a garment image:

$$\mathbf{W} = [W_{full}, W_1, W_2, ..., W_m], \tag{5}$$

where W_{full} denotes the full prompt and W_i is the i^{th} linguistic AP in the tree structure *e.g.*, 'blue sweater', 'classic hood', 'long sleeves', where meaningless conjunctions, *e.g.*, 'and', 'with' are omitted.

Bipartite Matching. In order to generate garment images with part-level consensus between these two collections of visual and linguistic components, we formulate the cross-modal semantic alignment as a set-to-set bipartite matching problem. Our objective is to find a permutation,

[1] More details are provided in Sec. 4.

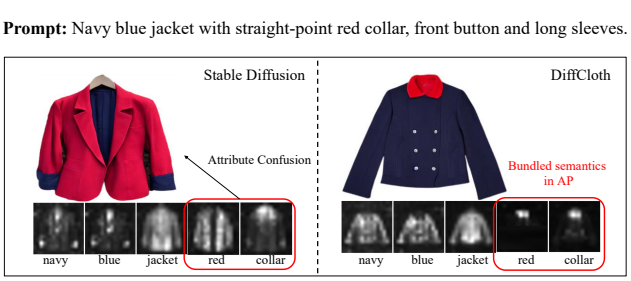

Figure 4. Visualization of the cross-attention of Stable Diffusion and our DiffCloth.

σ, of the set of m semantic components, which minimizes the pair-wise semantic matching loss $L_{match}(V_i, W_{\hat{\sigma}(i)})$:

$$\hat{\sigma} = \arg\min_{\sigma} \sum_{i}^{m} L_{match}(V_i, W_{\hat{\sigma}(i)}), \tag{6}$$

where $L_{match}(V_i, W_{\hat{\sigma}(i)}) = \text{CLIP}(V_i, W_{\hat{\sigma}(i)})$ and $\text{CLIP}(\cdot,\cdot)$ denotes the CLIP similarity [24]. The optimal matching is obtained by using the Hungarian algorithm [17]. Further, we define our Hungarian matching loss to compute the hierarchical structure alignment score on $\mathbf{V}$ and $\mathbf{W}$ by calculating the part-level and image-level alignment scores:

$$L_{\text{Hungarian}}(\mathbf{V}, \mathbf{W}) = \sum_{i}^{m} \text{CLIP}(V_i, W_{\hat{\sigma}(i)}) + \text{CLIP}(V_{full}, W_{full}). \tag{7}$$

The latent code z_t in the t^{th} denoising step of the diffusion model is then refined via the structural semantic consensus guidance:

$$\hat{z}_t \leftarrow z_t + \alpha \cdot \nabla_{z_t} L_{Hungarian}(\phi(z_t), \mathbf{W}), \tag{8}$$

where $\phi(z_t)$ denotes the collection of visual structured components for the decoded image corresponding to z_t.

3.3. Semantic-bundled Cross-attention

Current text-to-image diffusion models, such as Stable Diffusion, have demonstrated that cross-attention between prompt tokens and visual feature maps results in coarse semantic alignment. However, for complex garment descriptions, a phenomenon of 'attribute confusion' arises, which can severely impact the reliability of fashion generators. Specifically, attributes and garment parts may be wrongly paired and some attributes may be ignored in the generated image, resulting in imprecise and unsatisfactory generated garments. An example of this is provided in Fig. 4, where the output image is a 'Red jacket with blue cuffs.' while the input prompt is 'Navy blue jacket with red collar.' To reveal the underlying reason for the incorrect attribution of 'red', we visualize the cross-attention maps between the visual tokens and the linguistic tokens in Fig. 4.

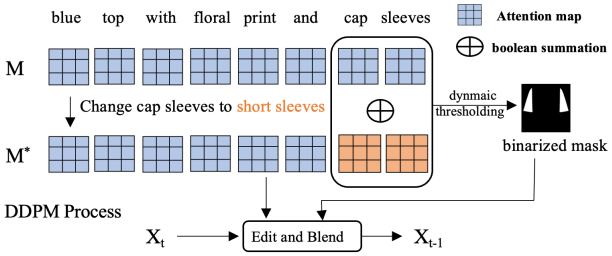

Figure 5. Overview of the garment manipulation pipeline.

It can be observed from the results given by Stable Diffusion that the attention map of 'red' is spatially similar to that of 'jacket' rather than 'collar', which leads to the unexpected mismatched attention regions for the Attribute-Phrase pair 'red collar'. To address this issue, we propose a semantic-bundled cross-attention mechanism that leverages a semantic-bundled loss L_{bundle} to preserve the spatial structure similarity between the attention maps of the attribute adjectives and the garment part subject. Formally, given an input prompt W_{full}, we first obtain the collection of attribute-phrases $\{W_1, W_2, ..., W_m\}$ using the aforementioned linguistic parsing tree. Our goal is to make the attention maps of the N_i attribute adjectives for the AP W_i, $\{W_i^j\}_{j=1}^{N_i}$, and the part noun $W_i^{N_i+1}$, *i.e.*, $\{M_i^j\}_{j=1}^{N_i+1}$ share similar spatial structures. We therefore regard an attention map M_i^j as a multi-dimensional probability distribution and define the internal structural similarity for W_i as:

$$d_{IS}(V_{full}, W_i) = \sum_{(j,k)\in\binom{N_i+1}{2}} d_{JS}(M_i^j, M_i^k), \tag{9}$$

where $\binom{N_i+1}{2}$ denotes the 2-combination set of the $N_i + 1$ indexes, d_{JS} is the Jensen-Shannon Divergence [8], and the attention mask M_i^j is obtained from the cross-attention between the text token W_i^j and image V_{full}. We then define the semantic-bundled loss for $\{W_i\}_{i=1}^m$ as

$$L_{bundle}(V_{full}, \mathbf{W}) = \sum_{i=1}^{m} d_{IS}(V_{full}, W_i). \tag{10}$$

Similarly to Eq. (8), we again shift the latent code $\hat{z}_t$ to bundle the semantics of attribute adjectives and the part noun in the APs in the denosing stage:

$$z_t' \leftarrow \hat{z}_t - \beta \cdot \nabla_{z_t} L_{bundle}(z_t, \mathbf{W}). \tag{11}$$

3.4. Region Consistency for Garment Manipulation

DiffCloth is inspired by Prompt-to-Prompt [9] and allows manipulation of the generated images by simply modifying the input text prompt. Formally speaking, given an original prompt input and its W, we can locally manipulate

Figure 6. Results of DiffCloth on the garment synthesis task for some difficult examples that require the precise generation of fine-grained details. DiffCloth outperforms existing SOTA methods and is capable of generating semantically-correct results. The boxes are used to highlight specific areas that should contain the elements highlighted in the text.

an output image I that is generated from W by simply modifying W to W^* which will result in the updated image I^*. For example, we can change a text token W_i^j to $W_i^{j,*}$ and replace its attention map M_i^j with a new one $M_i^{j,*}$ in each diffusion step. However, we find that this simple application of Prompt-to-Prompt [9] degrades our bundled semantics for APs that were introduced in Sec. 3.3 and may lead to attribute confusion problems in the editing phase.

To preserve the bundled semantics for attribute-phrases during manipulation, as shown in Fig. 5, we propose to replace the attention maps of all tokens $\{W_i^j\}_{j=1}^{N_i+1}$ in an Attribute-Phrase W_i rather than solely handling the token we need to change. For example, if we want to change the attribute of the sleeves, *e.g.*, "long sleeves" $\rightarrow$ "short sleeves", we need to inject the attention maps of both "long" and "sleeves". Following Prompt-to-Prompt [9], we need to run the diffusion step again by merging the new attention maps $\{M_i^{j,*}\}_{j=1}^{N_i+1}$ with the fixed ones. In the t^{th} denoising step, we can then use the semantic-bundled guidance in Eq. (11) again to preserve the internal structural similarity for $\{M_i^{j,*}\}_{j=1}^{N_i+1}$.

Another issue with garment manipulation is how to avoid editing regions that are not relevant to the Attribute-Phrase W_i^* that is being modified. To address this, we select a dynamic threshold p as the first quartile of the pixel activations in the attention map M_i^j and use it to binarize M_i^j to a mask B_i^j by thresholding. In this way, we obtain binarized masks $\{B_i^j\}_{j=1}^{N_i+1}$ and $\{B_i^{j,*}\}_{j=1}^{N_i+1}$ according to W_i and W_i^*, respectively. The irrelevant region is then indicated by the blended mask B_i:

$$B_i = (\bigoplus_{j=1}^{N_i+1} B_i^j) \bigoplus (\bigoplus_{j=1}^{N_i+1} B_i^{j,*}), \tag{12}$$

where $\bigoplus$ denotes boolean summation. Similarly, when modifying multiple APs, *e.g.*, $\{W_i\}_{i\in\Gamma}$, we can compute a global mask B across $\{W_i\}_{i\in\Gamma}$ as $B = \bigoplus_{i\in\Gamma} B_i$, where Γ denotes the indexes of the APs that are being manipulated.

The region consistency is encouraged in each denoising step by blending the two latent representations z_t and z_t^* using B:

$$z_{t-1}^* \leftarrow \text{Denoise}(B \cdot (z_t - z_t^*) + z_t^*) \tag{13}$$

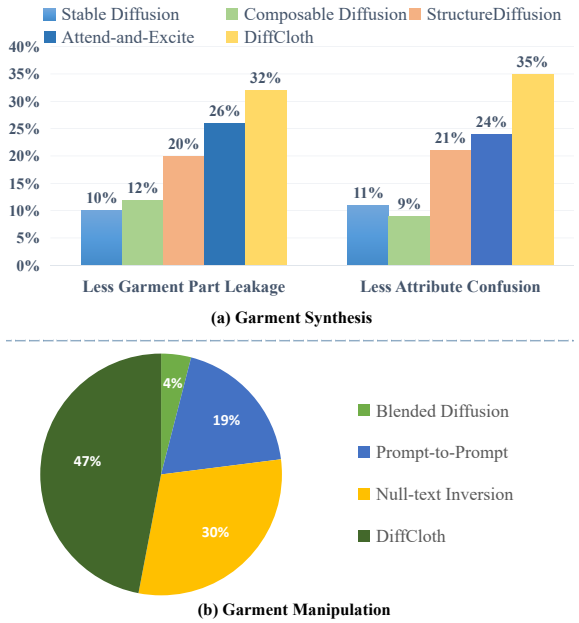

Figure 7. Human evaluation results for the garment synthesis and garment manipulation tasks.

where Denoise$(\cdot)$ denotes a DiffCloth denoising process.

4. Experiments

Datasets:

Experiments are conducted on the CM-Fashion dataset [41], which consists of garment images and their mask at resolution 512×512. This high-resolution fashion dataset contains 509,482 image-text pairs from various garment categories and is split into 409,482/100,000 training/testing pairs. In addition, we used 100,000 image-mask pairs from the training set to train the segmentor for segmenting noisy garment images into parts.

Implementation Details: The implementation closely follows Stable Diffusion [29]. However, we finetuned the model on the CM-Fashion dataset as the pre-trained Stable Diffusion did not produce garments on a homogeneous white background. Our models are trained on 8 Tesla V100 GPUs with a batch size of 32. During the generator training phase, the model is trained for 80 epochs with learning rate 1e-6.

Our segmentor is Pointrend [16], which was trained using input images with added noise. The model was trained for 150 epochs with a learning rate of 4e-5. Further details are provided in the supplementary material.

Baselines and Evaluation Metrics. For the generation step, we compare DiffCloth to the state-of-the-art methods TediGAN [35], Cogview [5], VQGAN [4], ARMANI [41], Stable Diffusion [29], Composable Diffusion [20], StructureDiffusion [7] and Attend-and-Excite [3]. To ensure fair comparisons, all models use our generator that has been trained on the CM-fashion dataset and we use the official inference code provided by the authors. For the manipulation step, we leverage Blended Diffusion [1], Prompt-to-Prompt [9], and Null-text Inversion [22] as our primary points of comparison, as this allows us to use the same diffusion model as for DiffCloth.[2] We employ three widely used metrics, namely the Fréchet Inception Distance (FID) [11], the Inception Score (IS) [2] and the CLIPScore [10] to evaluate the quality of the generation results. Furthermore, we conduct an Human Evaluation to evaluate different methods according to the text-image similarity of their results as well as their overall generation and manipulation quality. More specifically, for the garment synthesis task, we requested that participants assess the generated images based on two criteria: the extent of garment part leakage and the amount of attribute confusion. For the garment manipulation task, we instructed them to evaluate the performance based on whether a model preserves the consistency of the content in regions that are not relevant to the manipulation.

Method	FID ↓	IS ↑	CLIPScore ↑
TediGAN [35]	27.37	18.46	0.5587
Cogview [5]	12.198	23.99	0.6572
VQGAN [4]	13.249	20.33	0.6423
ARMANI [41]	12.336	24.32	0.6988
Stable Diffusion [29]	9.475	24.59	0.8169
Composable Diffusion [20]	9.499	25.91	0.8306
StructureDiffusion [7]	9.238	25.36	0.8459
Attend-and-Excite [3]	9.351	26.87	0.8241
DiffCloth(Ours)	**9.201**	**26.95**	**0.8974**

Comparison of DiffCloth to prior state-of-the-art approaches on the CM-Fashion dataset. ↓ means the lower the better, while ↑ means the opposite.

Comparison of DiffCloth to prior state-of-the-art approaches on the CM-Fashion dataset. ↓ means the lower the better, while ↑ means the opposite. Table 1.

Comparison of DiffCloth to prior state-of-the-art approaches on the CM-Fashion dataset. ↓ means the lower the better, while ↑ means the opposite.

4.1. Comparison With State-Of-The-Art Methods

Qualitative Result We provide a qualitative comparison of DiffCloth's garment generation ability compared to state-of-the-art approaches [4, 5, 26, 29]. DiffCloth is able to synthesize realistic fashion images that comply with the textual description, while prior approaches generate garment images that match the overall content of the textual description, but tends to neglect fine-grained information in the input text (red box in Fig. 6). In contrast, DiffCloth is capable of generating semantically bound parts by utilizing our proposed semantic-bundled cross-attention module. Specifically, words located within an AP generate separate

[2]Note, as Blended Diffusion [1] is not a mask-free approach, we provide it with a manually drawn mask that reflects the text description.

Figure 8. Results of DiffCloth for garment manipulation. The boxes are used to highlight specific areas that should contain the elements highlighted in the text.

attributes, which enhance DiffCloth's ability to generate semantically coherent images.

For the garment manipulation, the results in Fig. 8 demonstrate the superiority of our proposed approach. We can locally manipulate an image and maintain the consistency of the content of the manipulation-irrelevant regions by using our region consistency strategy.

Quantitative Result We apply FID [11] and IS [2] to measure the quality of the synthesized images. Further, we use the CLIPScore [10] to measure the relevance of the text to a given image. A higher CLIPScore indicates that the text is more relevant to the image. As reported in Tab. 4, our proposed DiffCloth outperforms the baselines Stable Diffusion [29], Composable diffusion [20], StructureDiffusion [7] and Attend-and-Excite [3] in all cases by a large margin, obtaining the lowest FID score and the highest IS and CLIPScore for the garment synthesis. In addition, we designed two human evaluation studies to quantitatively compare the generation and manipulation results with the baselines. For generation, we ask participants to select the generated results that exhibit minimal attribute confusion and Garment Part Leakage. For the manipulation task, we evaluate the effectiveness of the method by asking participants to select the results that best preserves the area that is irrelevant to the text modification. Aggregating the scores per model in Fig. 7, we observe that DiffCloth's results are preferred for both the garment synthesis or manipulation tasks. Furthermore, it is also noticeable that the human-based evaluation indicates a larger difference among the models compared to the machine evaluation.

Method	L1	L2	FID↓	IS ↑	CLIPScore ↑
DiffCloth†	✗	✗	9.475	24.59	0.8169
DiffCloth⋆	✓	✗	9.381	25.45	0.8821
DiffCloth∗	✗	✓	9.221	26.69	0.8423
DiffCloth	✓	✓	**9.201**	**26.95**	**0.8974**

Table 2. Quantitative results of our ablation studies. L1 and L2 denote the structural semantic consensus guidance and the semantic-bundled cross-attention, respectively.

4.2. Ablation study

In the garment synthesis task, to validate the effectiveness of the structural semantic consensus guidance and the semantic-bundled cross-attention, we design three variants of our proposed method and evaluate the performance of the different variants according to their metric scores. We denote Stable Diffusion [29] as DiffCloth†, DiffCloth without structural semantic consensus guidance as DiffCloth⋆, and denote DiffCloth without semantic-bundled cross-attention as DiffCloth∗. For the garment manipulation task, we consider DiffCloth without region consistency as our ablated model and denote it as DiffCloth$\mathcal{L}$.

As reported in Tab. 2, incorporating either the structural semantic consensus guidance or the semantic-bundled cross-attention (or both) leads to significant improvements in FID, IS and CLIPScore. These results indicate that our proposed mechanisms can produce more realistic and semantically accurate results. Additionally, as illustrated in Fig. 9, the incorporation of structural semantic consensus guidance (as DiffCloth⋆) leads to the generation of more accurate parts, whereas the exclusion of the semantic-bundled cross-attention increases attribute confusion. Finally, removing the region consistency strategy in garment manipulation causes the model to affect parts that should not be modified, as demonstrated in Fig. 9.

5. Conlusion

In this work, we propose DiffCloth, a diffusion-based pipeline for garment synthesis and manipulation, which aligns the structural cross-modal semantics between input

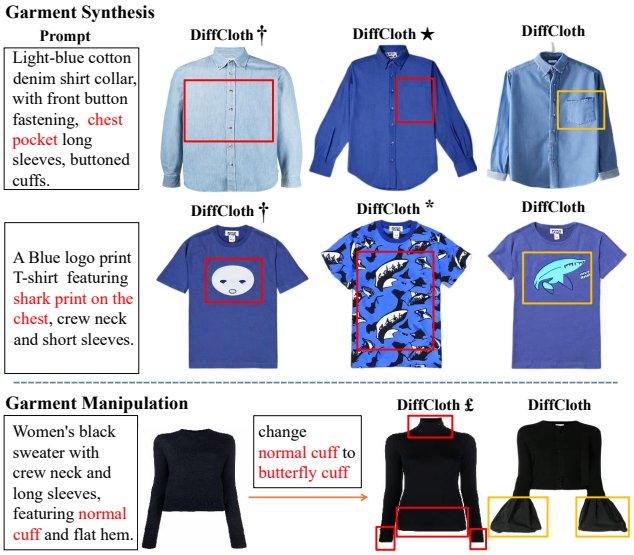

Figure 9. Qualitative results of our ablation studies for garment synthesis (top) and manipulation (bottom).

prompts and garment images to address the problem of garment part leakage and attribute confusion. Moreover, DiffCloth provides a convenient way to manipulate its generated garments by replacing the Attribute-Phrase in the text prompt, while ensuring that the content in regions unrelated to the modification is preserved using a consistency loss. Experiments on the CM-Fashion demonstrate DiffCloth's superior effectiveness compared to existing methods.

Limitation and future work: A limitation of our approach is the sensitivity to noisy text, which may make accurate correspondance matching more challenging. To address this limitation, we aim to explore how the text information can be leveraged to further strengthen the model's robustness.

6. Acknowledgement

This work was supported in part by National Key R&D Program of China under Grant No.2020AAA0109700, Guangdong Outstanding Youth Fund(Grant No.2021B1515020061), Shenzhen Science and Technology Program(Grant No.RCYX20200714114642083), Shenzhen Fundamental Research Program(Grant No.JCYJ20190807154211365), Nansha Key RD Program under Grant No.2022ZD014 and Sun Yat-sen University under Grant No.22lgqb38 and 76160-12220011. We thank MindSpore for the partial support of this work, which is a new deep learning computing framwork.[3].

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, pages 18187–18197. IEEE, 2022.

[2] Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

[3] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *arXiv preprint arXiv:2301.13826*, 2023.

[4] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. VQGAN-CLIP: open domain image generation and editing with natural language guidance. In *ECCV*, volume 13697 of *Lecture Notes in Computer Science*, pages 88–105. Springer, 2022.

[5] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021.

[6] Hao Dong, Jingqing Zhang, Douglas McIlwraith, and Yike Guo. I2t2i: Learning text to image synthesis with textual data augmentation. In *ICIP*, pages 2015–2019. IEEE, 2017.

[7] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022.

[8] Bent Fuglede and Flemming Topsøe. Jensen-shannon divergence and hilbert space embedding. In *ISIT*, page 31. IEEE, 2004.

[9] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[13] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *CVPR*, pages 7986–7994, 2018.

[14] He Huang, Philip S Yu, and Changhu Wang. An introduction to image synthesis with generative adversarial nets. *arXiv preprint arXiv:1803.04469*, 2018.

[15] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022.

[16] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross B. Girshick. Pointrend: Image segmentation as rendering. In *CVPR*, pages 9796–9805. Computer Vision Foundation / IEEE, 2020.

[3] https://www.mindspore.cn

[17] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 1955.

[18] Qicheng Lao, Mohammad Havaei, Ahmad Pesaranghader, Francis Dutil, Lisa Di Jorio, and Thomas Fevens. Dual adversarial inference for text-to-image synthesis. In *ICCV*, pages 7567–7576, 2019.

[19] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *CVPR*, pages 12174–12182, 2019.

[20] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *ECCV*, volume 13677, pages 423–439. Springer, 2022.

[21] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv: Computer Vision and Pattern Recognition*, 2021.

[22] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023.

[23] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 16784–16804. PMLR, 2022.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[25] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[26] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[27] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 2021.

[28] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, pages 1060–1069. PMLR, 2016.

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685. IEEE, 2022.

[30] 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. *arXiv preprint arXiv:2208.12242*, 2022.

[31] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[32] Ming Tao, Hao Tang, Fei Wu, Xiaoyuan Jing, Bing-Kun Bao, and Changsheng Xu. DF-GAN: A simple and effective baseline for text-to-image synthesis. In *CVPR*, pages 16494–16504. IEEE, 2022.

[33] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NIPS*, pages 6306–6315, 2017.

[34] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. *arXiv preprint arXiv:2212.06909*, 2022.

[35] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *CVPR*, pages 2256–2265. Computer Vision Foundation / IEEE, 2021.

[36] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, pages 1316–1324. Computer Vision Foundation / IEEE Computer Society, 2018.

[37] Hui Ye, Xiulong Yang, Martin Takác, Rajshekhar Sunderraman, and Shihao Ji. Improving text-to-image synthesis using contrastive learning. In *BMVC*, page 154. BMVA Press, 2021.

[38] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

[39] Mingkuan Yuan and Yuxin Peng. Text-to-image synthesis via symmetrical distillation networks. In *ACMMM*, pages 1407–1415, 2018.

[40] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *CVPR*, pages 833–842. Computer Vision Foundation / IEEE, 2021.

[41] Xujie Zhang, Yu Sha, Michael C. Kampffmeyer, Zhenyu Xie, Zequn Jie, Chengwen Huang, Jianqing Peng, and Xiaodan Liang. ARMANI: part-level garment-text alignment for unified cross-modal fashion design. In *ACMMM*, pages 4525–4535. ACMMM, 2022.

[42] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. DM-GAN: dynamic memory generative adversarial networks for text-to-image synthesis. In *CVPR*, pages 5802–5810. Computer Vision Foundation / IEEE, 2019.