FashionMAC: Deformation-Free Fashion Image Generation with Fine-Grained Model Appearance Customization

Rong Zhang[1], Jinxiao Li[1], Jingnan Wang[1], Zhiwen Zuo[1*], Jianfeng Dong[1], Wei Li[2], Chi Wang[3], Weiwei Xu[3], Xun Wang[1*]

[1]Zhejiang Gongshang University
[2]Nanjing University
[3]Zhejiang University

zhangrong@zjgsu.edu.cn, {module8627, wangjingnan751}@gmail.com, zzw@zjgsu.edu.cn, {dongjf24, liweimcc}@gmail.com, wangchi1995@zju.edu.cn, xww@cad.zju.edu.cn, wx@zjgsu.edu.cn

Abstract

Garment-centric fashion image generation aims to synthesize realistic and controllable human models dressing a given garment, which has attracted growing interest due to its practical applications in e-commerce. The key challenges of the task lie in two aspects: (1) faithfully preserving the garment details, and (2) gaining fine-grained controllability over the model's appearance. Existing methods typically require performing garment deformation in the generation process, which often leads to garment texture distortions. Also, they fail to control the fine-grained attributes of the generated models, due to the lack of specifically designed mechanisms. To address these issues, we propose FashionMAC, a novel diffusion-based deformation-free framework that achieves high-quality and controllable fashion showcase image generation. The core idea of our framework is to eliminate the need for performing garment deformation and directly outpaint the garment segmented from a dressed person, which enables faithful preservation of the intricate garment details. Moreover, we propose a novel region-adaptive decoupled attention (RADA) mechanism along with a chained mask injection strategy to achieve fine-grained appearance controllability over the synthesized human models. Specifically, RADA adaptively predicts the generated regions for each fine-grained text attribute and enforces the text attribute to focus on the predicted regions by a chained mask injection strategy, significantly enhancing the visual fidelity and the controllability. Extensive experiments validate the superior performance of our framework compared to existing state-of-the-art methods.

Project — https://github.com/module8627/FashionMAC
Supp. — https://arxiv.org/abs/2511.14031

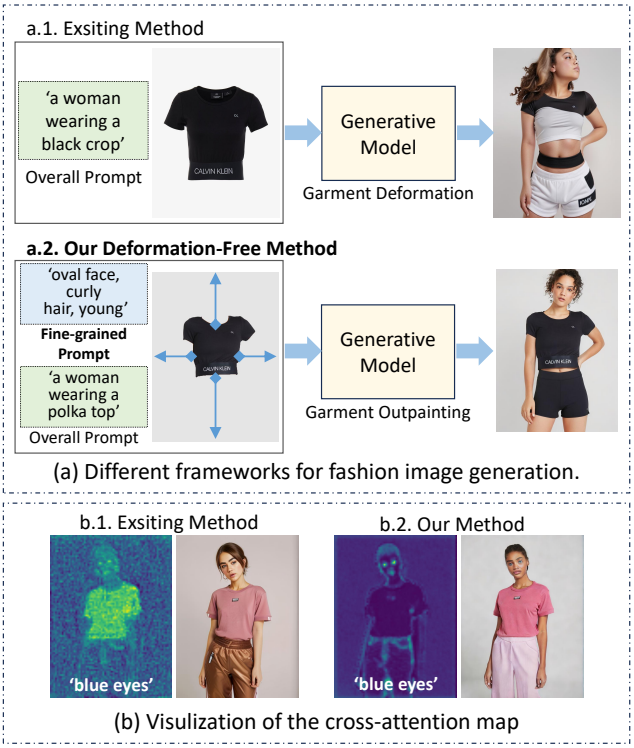

Figure 1: (a): Framework comparison of fashion image generation methods. (b): Visualization of the cross attention-map. Note that our method can preserve details better and enable accurate fine-grained text customization.

Introduction

The rapid advancement of AI technology is revolutionizing numerous industries including fashion e-commerce (Dong et al. 2021; Zhu et al. 2023; Yang et al. 2024; Dong et al. 2025). Recently, a new task called garment-centric fashion image generation (Chen et al. 2024; Lin et al. 2025; Shen et al. 2025) has gained increasing attention. Different from conventional fashion image synthesis tasks like virtual try-on, this line of research focuses on synthesizing realistic and controllable human models wearing a given garment (see Fig. 1 (a)), enabling automated fashion showcase creation for online fashion retailers. The key challenges of the task lie in two aspects: 1) faithful preservation of garment details, and 2) fine-grained controllability over the generated model's appearance.

Many approaches have been proposed to tackle this task by leveraging the strong generative capability of pre-trained

*Corresponding author.

Figure 2: Given a garment image segmented from a dressed mannequin or a person, our method can generate fashion showcase images via garment-centric outpainting under the guidance of face images or fine-grained text attributes. From top to bottom: (1) The results conditioned on the automatically generated diverse pose maps. (2) The results conditioned on different face images. (3) The results conditioned on fine-grained text prompts.

text-to-image diffusion models, *e.g.*, latent diffusion models (LDMs) (Rombach et al. 2022). MagicClothing (MC) (Chen et al. 2024) injects garment features into the LDM and conditions the generation process with textual descriptions to customize the model's appearance. Parts2Whole (Huang et al. 2024) extends this paradigm by encoding hierarchical features from images of different human parts (*e.g.*, face, hair, and clothes) and then integrating them into the generation network for region-aware synthesis. DreamFit (Lin et al. 2025) incorporates a specifically tailored lightweight encoder for efficient generation.

Despite the advances, these methods fall short in preserving the intricate garment details and freely controlling the fine-grained appearance attributes of the synthesized human models. On the one hand, the above methods typically involve deforming the reference garments to fit the synthesized model's poses in their generation process. However, as shown in Fig. 1 (a), garment deformation may lead to texture distortions or detail losses in the generated images, which can seriously limit these methods' practical usage in real-world applications. On the other hand, existing methods struggle to manipulate the fine-grained appearance attributes of the synthesized models (*e.g.*, the eye color as shown in Fig. 1 (b)). As pointed out by previous literature (Zhang et al. 2024), the cross-attention layers in LDMs have a propensity to disproportionately focus on certain tokens while ignoring others during the generation process. By examining the cross-attention maps of the fine-grained appearance text tokens in an existing method as shown in Fig. 1(b), we observe that these tokens exhibit low activations toward their corresponding generation regions, while most of their energy is diverted to irrelevant areas. Therefore, we hypothesize that the problem stems from existing methods' failure to attend to fine-grained attribute tokens during the generation process.

To address these limitations, we propose FashionMAC, a novel deformation-free garment-centric **Fashion** image generation framework with fine-grained **M**odel **A**ppearance **C**ustomization. Instead of deforming the garment to fit a target pose, FashionMAC directly outpaints a dressed garment image to synthesize a realistic human model with the guidance of text descriptions or face images, as illustrated in Fig. 1 (a). Such dressed garment images — can be easily obtained by dressing a mannequin or having a person wear the clothes — already contain natural deformations, allowing our method to bypass the complex deformation process. Our design avoids visual distortions caused by inaccurate warping and ensure faithful preservation of intricate garment details in the synthesized results. Thanks to the abundance of dressed-person images available online, our framework can be trained without requiring paired data of garments and corresponding human models, significantly improving data accessibility and real-world applicability.

Furthermore, we empower FashionMAC with the capability to control the fine-grained characteristics of the synthesized models by proposing a novel region-adaptive decou-

pled attention (RADA) mechanism. Different from existing methods that treat the input text prompt as a global one, we introduce RADA to process the overall text descriptions and fine-grained text attributes separately. Specifically, a fine-grained mask prediction module is proposed to adaptively localize the influence regions of the detailed text attributes and then these attributes are encouraged to attend to the predicted regions. In addition, we further propose a chained mask injection (CMI) strategy that utilizes the masks predicted from the previous timestep to steer RADA in the current generation timestep, providing effective mask guidance for RADA in the denoising process. These carefully crafted designs enable the proposed framework to accurately align the influenced feature map regions with the fine-grained text attributes, significantly enhancing the proposed framework's controllability over the fine-grained text attributes.

To summarize, our key contributions are as follows:

- We propose FashionMAC, a novel deformation-free garment-centric framework for fashion image generation, eliminating the need for garment deformation and ensuring faithful preservation of garment details.
- We introduce a novel region-adaptive decoupled attention mechanism along with a chained mask injection strategy to enable fine-grained model appearance customization. To the best of our knowledge, we are the first to investigate fine-grained model appearance customization in fashion image generation.
- Extensive experiments demonstrate that our method outperforms existing approaches in terms of both visual quality and appearance controllability.

Related Works

Image-Based Virtual Try-on

Image-based virtual try-on (VITON) task is designed to generate realistic try-on images based on the given garment image and the reference person image. Most prior VITON methods consist of two main phases: the cloth warping phase and the try-on generation phase. In the first phase, thin plate spline (TPS) (Duchon 1977) and appearance flow (Li, Huang, and Loy 2019) are commonly adopted for cloth warping. In the second phase, GANs (Goodfellow et al. 2014) often play a pivotal role in refining try-on results. Recently, many researchers have developed methods based on LDMs (Rombach et al. 2022) due to their superior image quality. Some approaches (Gou et al. 2023; Li et al. 2023; Wang et al. 2024a) still follow the previous two-phase framework, substituting GANs with LDMs in the second phase for synthesizing more realistic try-on results, while others (Morelli et al. 2023; Zhu et al. 2023; Yang et al. 2024; Choi et al. 2024; Kim et al. 2024; Zeng et al. 2024) build end-to-end frameworks by introducing garment encoders into LDMs and then utilizing attention modules to perform implicit garment warping. Nonetheless, few works have paid attention to apparel showcase image generation. Magic Clothing (MC) (Chen et al. 2024) is the most related work to ours in this area, which customizes characters wearing the target garment with diverse text prompts. However, compared to our method, MC falls short in cloth details and lacks precise control over the fine-grained attributes of the generated model's appearance.

Human Image Generation

Human image generation(HIG) task aims at synthesizing realistic and diverse human images. We roughly categorize HIG into transfer-based and synthesis-based methods. Given source human images and target pose conditions, transfer-based algorithms (Ma et al. 2017; Han et al. 2023; Shen et al. 2023; Lu et al. 2024) are expected to output photorealistic images with source appearance and target poses. In contrast, synthesis-based HIG methods (Frühstück et al. 2022; Fu et al. 2022; Yang et al. 2023; Ju et al. 2023; Liu et al. 2023; Wang et al. 2024b; Zhu et al. 2024; Huang et al. 2024) concentrate on synthesizing high-quality human images conditioned on poses, text prompts, or faces. Early works on synthesis-based HIG are mainly based on GANs (Frühstück et al. 2022; Fu et al. 2022; Yang et al. 2023), while many recent approaches have delved into designing specialized frameworks based on LDMs for better quality and controllability, such as incorporating more annotations (Liu et al. 2023), designing losses based on human-centric priors (Ju et al. 2023; Wang et al. 2024b), or utilizing multiple human-part images (Zhu et al. 2024). Aside from these methods, a lot of personalized text-to-image methods (Ye et al. 2023; Li et al. 2024; Shi et al. 2024; Peng et al. 2024) have been proposed to customize portraits conditioned on facial images.

Garment-Centric Fashion Image Generation

Garment-centric fashion image generation is a task to synthesize realistic fashion showcase images with specified garments under the guidance of images or text prompts. MagicClothing (Chen et al. 2024) is the first work focused on this area. It injects garment features into diffusion models to preserve details through self-attention feature integration. Parts2Whole (Huang et al. 2024) proposes a reference-based framework that separately encodes multiple human appearance aspects (e.g., face, clothes, hair) from distinct images and employs multi-image conditioning and shared attention to compose a full-body output. IMAGDressing (Shen et al. 2025) generates fashion images using a hybrid architecture that combines a garment-specific UNet with a frozen diffusion backbone. DreamFit (Lin et al. 2025) features a lightweight encoder, which significantly reduces trainable parameters. However, these methods typically require garment deformation in the generation, which often leads to garment texture distortions. Besides, they struggle to precisely control the fine-grained attributes of the generated models, due to the lack of specifically designed mechanisms.

Approach

In this section, we present FashionMAC, a deformation-free garment-centric fashion image generation framework for fine-grained model appearance customization. The proposed framework consists of two stages. In the first stage, given a garment image segmented from a dressed mannequin or a person, a garment-centric pose predictor is introduced to generate the corresponding pose that fits it. Please

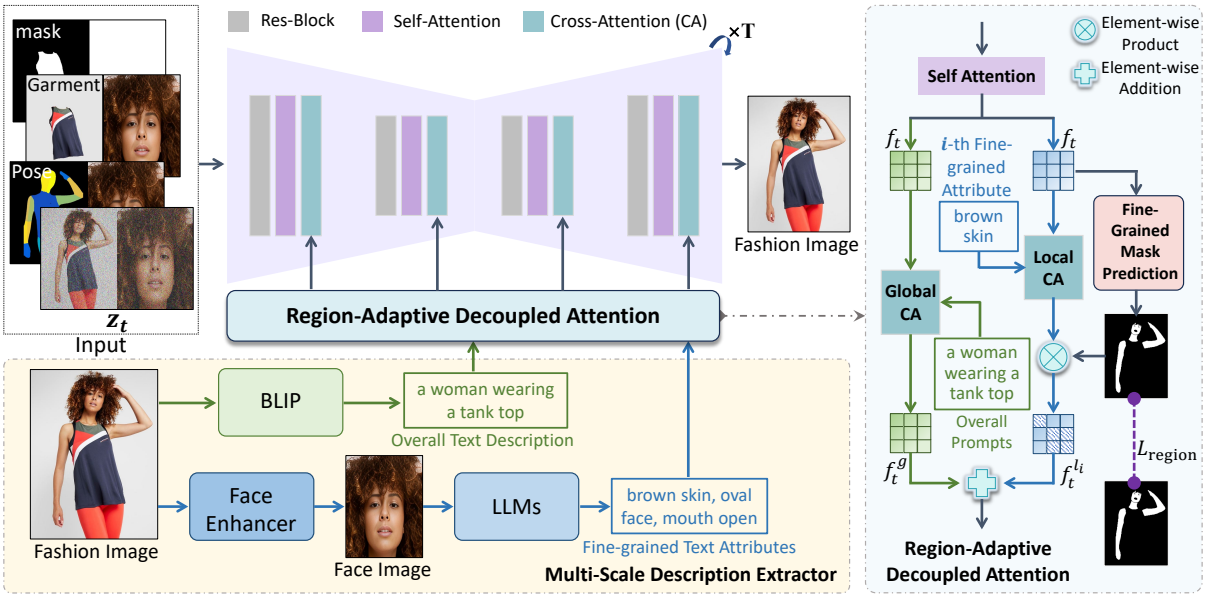

Figure 3: The overview of our framework. It consists of a multi-scale description extractor and a region-adaptive decoupled attention mechanism. After training, our method can take facial images, overall showcase descriptions, or fine-grained text attributes as optional inputs to generate diverse showcase images.

refer to the *Supp.* for details. In the second stage, we build a deformation-free fashion generation model based on latent diffusion models (LDMs) (Rombach et al. 2022).

Deformation-Free Fashion Generation Model

In the second stage, we propose a deformation-free fashion generation model based on LDM to generate fashion showcase images conditioned on the given garment and the predicted pose maps, along with the text prompts and the facial images. The overall architecture is shown in Fig. 3.

In this stage, the denoising U-Net has two kinds of input conditions: 1) conditions that are spatially aligned with the target fashion image, including the garment image, the pose map, and the garment mask indicating the regions to be outpainted, and 2) conditions that are not spatially aligned with the target fashion image, *i.e*. the user-specified facial images of the synthesized fashion models. We leverage a simple yet effective feature fusion operation to deal with multiple conditions. For the conditions that are spatially aligned with the fashion image, we fuse them through channel-wise concatenation. For facial images that are not spatially aligned, we follow the self-attention-based texture transfer operation in (Yang et al. 2024) and concatenate them with the fashion image in the spatial dimension, relying on the self-attention modules to decide which features to preserve.

To enable precise fine-grained control over the generated model's appearance, we leverage information from both global descriptions of the complete fashion image and local attributes (skin tone, hairstyle, expression) that describe the appearance of the model, and explicitly introduce region-aware local structure guidance to ensure the controllability. To this end, we carefully design a multi-scale description extractor to parse a fashion image into hierarchical text prompts, region-adaptive decoupled attention mechanisms that perform global-local feature fusion under region-based spatial priors, and a chained mask injection strategy to propagate structural guidance across timesteps and provide more precise information.

Multi-Scale Description Extractor

To provide effective textual guidance for controllable generation, the multi-scale description extractor (MDE) jointly captures global and fine-grained textual cues from a given fashion image. As shown in Fig. 3, the extractor decouples the textual description into two levels — an overall prompt for the whole image and a set of fine-grained attributes focusing on local facial details. To extract the overall global text prompt, we utilize a pretrained BLIP (Li et al. 2022) model to caption the fashion image. For the fine-grained text prompts, we leverage a face enhancer module to obtain a high-quality facial image. The face enhancer first adopts Yolov5 (Ultralytics 2021) to detect the face region of the fashion image. Then a face restoration model CodeFormer (Zhou et al. 2022) and a super-resolution model RealESRGAN (Wang et al. 2021) are used in a sequential manner to obtain a high-quality facial image F. Since existing fashion datasets are short of detailed facial descriptions of the fashion models, the enhanced face image is fed into a large language model (LLM) (Yang et al. 2025) in a captioning style, prompting it to output detailed appearance attributes such as skin tone, facial shape, expression, *etc*. . These form the fine-grained attribute set that is used for localized guidance during generation.

Region-Adaptive Decoupled Attention

In diffusion-based text-to-image models, the cross-attention layers are responsible for linking textual attributes with corresponding visual regions. However, in practice, certain regions of the denoising feature map may fail to effectively respond to their associated textual tokens.

To further improve the cross-attention layers to better align the visual concepts with fine-grained text attributes, we propose Region-Adaptive Decoupled Attention (RADA) — a dual-branch attention design that explicitly decouples global and fine-grained attribute information and selectively adapts attention responses based on region-aware priors. It is utilized to replace each cross-attention layer of the denoising U-Net (including the encoder and the decoder), as illustrated in Fig. 3. Given an overall global prompt τ_g, a fine-grained prompt τ_l and the denoising feature f_t output by a self-attention layer at timestep t, RADA instantiates two separate cross-attention branches: the Global Cross Attention GCA attends to the overall prompt embedding and the Local Cross Attention LCA attends to individual fine-grained attribute embeddings.

In each local branch, we utilize a lightweight mask prediction head to provide spatial structural priors to guide the attention mechanism during generation. The core idea is to predict soft region masks for each local attribute, indicating where in the image the model should focus its attention for better text-image alignment. Each head has three stacked 3×3 convolution layers followed by a sigmoid activation. We take the output features from the self-attention layer and feed them into the prediction head to generate a set of masks with N channels and resolution matching that block's feature size, where N is the total number of fine-grained attribute classes. We apply a region-level supervision loss L_{region} to optimize the heads for accurate masks. These masks act as explicit spatial cues to enforce structural information in the global-local fusion process.

Then RADA can be represented as:

$$RADA(f_t, \tau_g, \tau_l) = GCA(f_t, \tau_g) + \sum_i^N LCA(f_t, \tau_{l_i}) \odot M^i_{t/t-1}. \quad (1)$$

where i is the i-th fine-grained attribute, $M^i_{t/t-1}$ is the corresponding mask generated by the prediction head at timestep t or $t-1$, which will be exlain in the following section. $\odot$ represents the element-wise product operator. In this way, RADA introduces explicit structural priors into the diffusion attention mechanism by decoupling global and local semantics and leveraging attribute-specific spatial masks. This improves the visual-textual alignment and enhances the fine-grained features.

Chained Mask Injection Strategy

To enhance the structural controllability of image generation, we explore the injection of region-level guidance at different stages of the denoising process. Our empirical findings suggest that incorporating region structure priors across all timesteps consistently improves generation quality. Interestingly, we observe that the benefits are particularly pronounced during early timesteps, where the model is more sensitive to structural cues. This observation aligns with insights from prior work such as Prospect (Zhang et al. 2023),

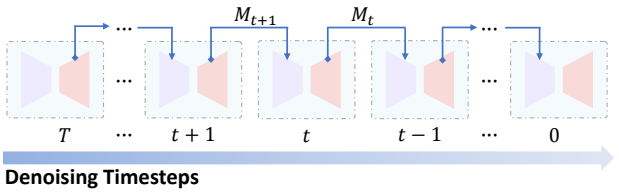

Figure 4: The chained mask injection strategy.

which points out that in the early timesteps of the denoising process, diffusion models primarily focus on recovering global structure and semantic layout.

Furthermore, we investigate the injection locations within the network architecture and find that applying the region priors to both encoder and decoder blocks across all layers yields the best overall performance. However, this introduces a practical challenge: the encoder blocks, being relatively shallow, struggle to generate reliable region masks from highly noisy features in the early denoising steps. In contrast, decoder blocks benefit from deeper feature integration and are capable of producing reasonably accurate masks even under noise conditions.

To fully leverage the potential of the region priors, we propose a chained mask injection (CMI) strategy. During inference, we utilize the predicted mask M_t from the last decoder block at timestep t to guide all encoder blocks at timestep $t-1$ instead of predicting masks from scratch, as illustrated in Fig. 4. In the meantime, for decoder blocks, the region-aware RADA module uses the mask predicted from its corresponding decoder block, since predicted masks from the decoder tend to be refined during the denoising. This chained temporal connection effectively bootstraps the encoder with decoder-refined structure priors, allowing both parts of the network to benefit from more accurate region guidance.

Training Objective

The training objective of the deformation-free fashion generation model comprises two components: a denoising loss $\mathcal{L}_{denoise}$ for the U-Net ϵ_θ and a region loss $\mathcal{L}_{region}$ for supervising the fine-grained mask prediction in RADA.

For the denoising loss, we follow the standard training paradigm of LDM, where the network is trained to predict the added noise given a noisy latent $\mathbf{z_t}$ and conditional inputs $\mathbf{c}$. Specifically, we minimize the mean squared error between the predicted noise and the ground truth noise:

$$\mathcal{L}_{denoise} = \mathbb{E}_{\mathbf{z},\epsilon\sim\mathcal{N}(0,1),t}[||\epsilon - \epsilon_\theta(\mathbf{z_t}, t, \mathbf{c})||_2^2]. \quad (2)$$

For the region supervision, we simply leverage a Euclidean loss to train spatial masks. Given ground truth attribute region masks $\{M^i\}_{i=1}^N$, and predicted masks $\{\hat{M}^i_t\}_{i=1}^N$ of timestep t at resolution corresponding to each decoder block, the loss is employed over all N channels:

$$\mathcal{L}_{region} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{M}^i_t - M^i\right\|_2^2. \quad (3)$$

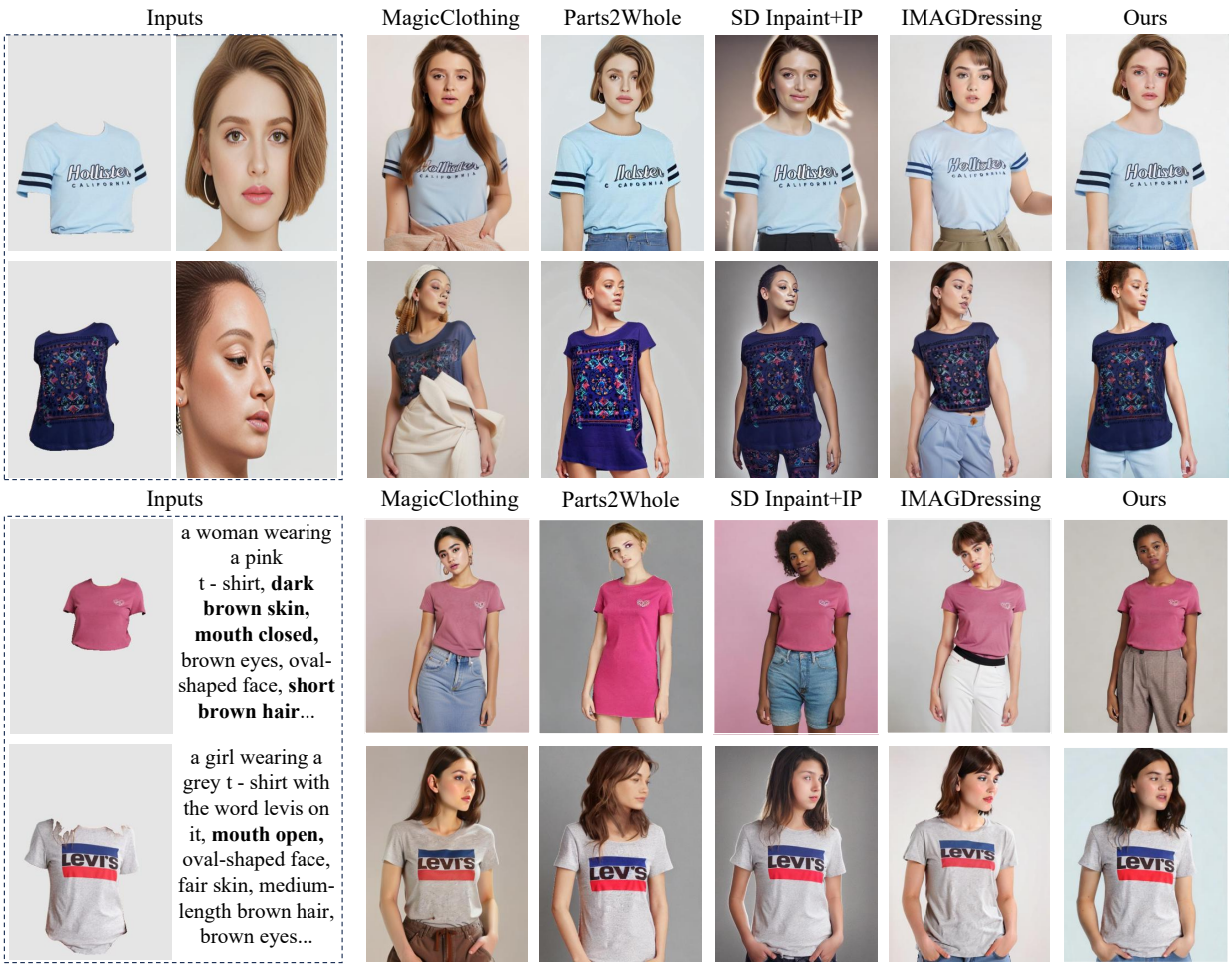

Figure 5: Comparison with the baseline methods. The first two rows show the results with facial image guidance. The last two rows show results with text prompt guidance.

Experiments

Implementation Details

We conduct experiments on the virtual try-on benchmark VITON-HD (Choi et al. 2021), which contains $14,221$ training images and $2,032$ testing images. The model is trained at 576×768. FashionMAC is built on Stable Diffusion 1.4 (Rombach et al. 2022). All experiments are implemented with Pytorch and performed on 4 NVIDIA A100 80G GPUs. Please refer to the *Supp.* for more information and additional results on the IGPair dataset(Shen et al. 2025).

Comparisons

Baselines. We compare FashionMAC with the related inpainting-based and garment-driven image synthesis approaches, including SD Inpaint (Rombach et al. 2022), Parts2Whole (Huang et al. 2024), MagicClothing (Chen et al. 2024), and IMAGDressing (Shen et al. 2025).

Metrics. To evaluate the methods, we calculate the metrics in two settings: with facial images and with text prompts only. We employ LPIPS (Zhang et al. 2018), FID (Heusel et al. 2017), KID (Bińkowski et al. 2018) and DreamSim (Fu et al. 2023) to measure the realism of the generated images from different dimensions. We adopt CLIP-i and CLIP-t (Radford et al. 2021) to estimate the similarity of the CLIP space between the generated images and the ground truths, MP-LPIPS (Chen et al. 2024) to evaluate whether the characteristics of the target garment are well-preserved.

Qualitative Results. Fig. 5 demonstrates the qualitative comparisons between FashionMAC and state-of-the-art baseline methods. The first two rows show the results with facial image guidance, while the last two rows present the results with only text prompts. We observe that our method achieves better results with fewer artifacts in both settings. On one hand, as FashionMAC utilizes a deformation-free outpainting-based framework, the garment details, especially the characters and patterns, are better preserved. In contrast, the garment deformation in baseline methods may introduce undesired distortions or color deviation. FashionMAC can also maintain facial identity and structural integrity more accurately with the face guidance. On the other

Methods	With facial image guidance				With text prompts guidance			
	FID↓	KID↓	CLIP-i↑	MP-LPIPS↓	FID↓	KID↓	CLIP-t↑	MP-LPIPS↓
SD Inpaint	58.58	0.016	0.83	0.119	40.32	0.0186	0.224	0.080
Parts2Whole	41.76	0.031	0.89	0.128	41.17	0.0284	0.214	0.120
MagicClothing	41.74	0.030	0.85	0.117	37.24	0.0303	0.228	0.117
IMAGDressing	29.34	0.021	0.89	0.083	27.80	0.0204	0.227	0.080
Ours	**14.89**	**0.008**	**0.95**	**0.055**	**13.68**	**0.0060**	**0.231**	**0.051**

Table 1: Quantitative comparison with different methods.

Method	FID↓	KID↓	CLIP-t↑
FashionMAC	**13.68**	**0.0060**	**0.231**
w/o CMI	14.91	0.0073	0.221
w/o CMI and RADA	15.21	0.0077	0.222

Table 2: The quantitative results of the ablation study.

hand, our method outperforms others in terms of both visual quality and appearance controllability for the comparisons with only text prompts. The baseline methods tend to ignore some of the attributes, such as 'dark brown skin', 'mouth open', and 'dark brown hair'. In contrast, FashionMAC is capable of maintaining the semantic alignments between the generated fashion images and the specified prompts.

Quantitative Results. Tab. 1 presents the quantitative comparison results between FashionMAC and the baseline methods. Our method clearly outperforms existing state-of-the-art methods with large margins on all the metrics under the guidance of either the facial images or text prompts. Among these metrics, MP-LPIPS numerically validates the effectiveness of FashionMAC for faithful preservation of the intricate garment details. The results on FID and KID verify the superiority of our method in terms of image quality and realism. The CLIP-i score of our method indicates that FashionMAC accurately synthesizes fashion images, while the CLIP-t score demonstrates that our method achieves better visual-textual alignment than baseline methods.

Ablation Study

To assess the effectiveness of the proposed Chained Mask Injection (CMI) strategy and Region-Adaptive Decoupled Attention (RADA) module, we conduct ablation studies on our full model FashionMAC, model without CMI and model without CMI and RADA. As shown in Tab. 2, removing the CMI strategy results in performance degradation across all metrics including FID, KID and CLIP-t, indicating that temporal mask propagation significantly enhances generation quality and text-image alignment. Further removing the RADA module still leads to drops in visual fidelity (FID to 15.21, KID to 0.0077), demonstrating that incorporating spatially aligned region priors is critical for guiding the denoising process. The Fig. 6 demonstrates the qualitative results. Removing the CMI strategy degrades the structural accuracy and visual fidelity. Further removing the RADA module leads to more misalignments between the generated image and the input prompts (e.g. skin color and mouth status). These results validate the complementary roles of CMI and RADA in improving both visual realism and controllability.

Figure 6: The qualitative results of the ablation study.

Conclusion

In this work, we presented FashionMAC, a novel garment-centric fashion image generation framework that eliminates the need for garment deformation and enables fine-grained appearance customization. By directly generating fashion showcase images from dressed garment inputs, our method leverages readily available visual data while avoiding distortion issues introduced by garment deformation. To achieve fine-grained appearance controllability over the synthesized human models (e.g., hairstyle, skin tone, expression, *etc.*), we propose a Region-Adaptive Decoupled Attention (RADA) mechanism coupled with a Chained Mask Injection (CMI) strategy. RADA selectively modulates cross-attention responses based on region-aware priors predicted by fine-grained mask prediction, while CMI progressively propagates structural guidance across timesteps to enhance spatial precision. Extensive experiments demonstrate that FashionMAC outperforms existing baselines in both visual fidelity and appearance controllability, offering a practical solution for e-commerce scenarios.

Acknowledgements

This work was partially supported by Zhejiang Provincial Natural Science Foundation of China (No. LQ23F020009, No. LQN25F020012, No. LD24F020011) and NSFC (No. 62302449, No. 62402439, No. 92570206, No. 62421003).

References

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying MMD GANs. In *International Conference on Learning Representations*.

Chen, W.; Gu, T.; Xu, Y.; and Chen, A. 2024. Magic clothing: Controllable garment-driven image synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 6939–6948.

Choi, S.; Park, S.; Lee, M.; and Choo, J. 2021. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14131–14140.

Choi, Y.; Kwak, S.; Lee, K.; Choi, H.; and Shin, J. 2024. Improving diffusion models for virtual try-on. *arXiv preprint arXiv:2403.05139*.

Dong, J.; Ma, Z.; Mao, X.; Yang, X.; He, Y.; Hong, R.; and Ji, S. 2021. Fine-grained fashion similarity prediction by attribute-specific embedding learning. *IEEE Transactions on Image Processing*, 30: 8410–8425.

Dong, J.; Zhu, J.; Liu, D.; Qu, X.; Bao, C.; Han, Z.; Zhu, J.; and Wang, X. 2025. Open-world fine-grained fashion retrieval with llm-based commonsense knowledge infusion. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 223–232.

Duchon, J. 1977. Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In *Constructive Theory of Functions of Several Variables: Proceedings of a Conference Held at Oberwolfach April 25–May 1, 1976*, 85–100. Springer.

Frühstück, A.; Singh, K. K.; Shechtman, E.; Mitra, N. J.; Wonka, P.; and Lu, J. 2022. Insetgan for full-body image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7723–7732.

Fu, J.; Li, S.; Jiang, Y.; Lin, K.-Y.; Qian, C.; Loy, C. C.; Wu, W.; and Liu, Z. 2022. Stylegan-human: A data-centric odyssey of human generation. In *European Conference on Computer Vision*, 1–19. Springer.

Fu, S.; Tamir, N. Y.; Sundaram, S.; Chai, L.; Zhang, R.; Dekel, T.; and Isola, P. 2023. DreamSim: learning new dimensions of human visual similarity using synthetic data. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 50742–50768.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Gou, J.; Sun, S.; Zhang, J.; Si, J.; Qian, C.; and Zhang, L. 2023. Taming the power of diffusion models for high-quality virtual try-on with appearance flow. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7599–7607.

Han, X.; Zhu, X.; Deng, J.; Song, Y.-Z.; and Xiang, T. 2023. Controllable person image synthesis with pose-constrained latent diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22768–22777.

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. *Advances in neural information processing systems*, 30.

Huang, Z.; Fan, H.; Wang, L.; and Sheng, L. 2024. From Parts to Whole: A Unified Reference Framework for Controllable Human Image Generation. *arXiv preprint arXiv:2404.15267*.

Ju, X.; Zeng, A.; Zhao, C.; Wang, J.; Zhang, L.; and Xu, Q. 2023. Humansd: A native skeleton-guided diffusion model for human image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15988–15998.

Kim, J.; Gu, G.; Park, M.; Park, S.; and Choo, J. 2024. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8176–8185.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, Y.; Huang, C.; and Loy, C. C. 2019. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3693–3702.

Li, Z.; Cao, M.; Wang, X.; Qi, Z.; Cheng, M.-M.; and Shan, Y. 2024. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8640–8650.

Li, Z.; Wei, P.; Yin, X.; Ma, Z.; and Kot, A. C. 2023. Virtual try-on with pose-garment keypoints guided inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22788–22797.

Lin, E.; Zhang, X.; Zhao, F.; Luo, Y.; Dong, X.; Zeng, L.; and Liang, X. 2025. Dreamfit: Garment-centric human generation via a lightweight anything-dressing encoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5218–5226.

Liu, X.; Ren, J.; Siarohin, A.; Skorokhodov, I.; Li, Y.; Lin, D.; Liu, X.; Liu, Z.; and Tulyakov, S. 2023. Hyperhuman: Hyper-realistic human generation with latent structural diffusion. *arXiv preprint arXiv:2310.08579*.

Lu, Y.; Zhang, M.; Ma, A. J.; Xie, X.; and Lai, J. 2024. Coarse-to-fine latent diffusion for pose-guided person image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6420–6429.

Ma, L.; Jia, X.; Sun, Q.; Schiele, B.; Tuytelaars, T.; and Van Gool, L. 2017. Pose guided person image generation. *Advances in neural information processing systems*, 30.

Morelli, D.; Baldrati, A.; Cartella, G.; Cornia, M.; Bertini, M.; and Cucchiara, R. 2023. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8580–8589.

Peng, X.; Zhu, J.; Jiang, B.; Tai, Y.; Luo, D.; Zhang, J.; Lin, W.; Jin, T.; Wang, C.; and Ji, R. 2024. Portraitbooth: A versatile portrait model for fast identity-preserved personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27080–27090.

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 *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Shen, F.; Jiang, X.; He, X.; Ye, H.; Wang, C.; Du, X.; Li, Z.; and Tang, J. 2025. Imagdressing-v1: Customizable virtual dressing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6795–6804.

Shen, F.; Ye, H.; Zhang, J.; Wang, C.; Han, X.; and Yang, W. 2023. Advancing pose-guided image synthesis with progressive conditional diffusion models. *arXiv preprint arXiv:2310.06313*.

Shi, J.; Xiong, W.; Lin, Z.; and Jung, H. J. 2024. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8543–8552.

Ultralytics. 2021. YOLOv5: A state-of-the-art real-time object detection system. https://docs.ultralytics.com. Accessed: insert date here.

Wang, C.; Chen, T.; Chen, Z.; Huang, Z.; Jiang, T.; Wang, Q.; and Shan, H. 2024a. FLDM-VTON: Faithful Latent Diffusion Model for Virtual Try-on. *arXiv preprint arXiv:2404.14162*.

Wang, J.; Sun, Z.; Tan, Z.; Chen, X.; Chen, W.; Li, H.; Zhang, C.; and Song, Y. 2024b. Towards Effective Usage of Human-Centric Priors in Diffusion Models for Text-based Human Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8446–8455.

Wang, X.; Xie, L.; Dong, C.; and Shan, Y. 2021. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1905–1914.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Yang, X.; Ding, C.; Hong, Z.; Huang, J.; Tao, J.; and Xu, X. 2024. Texture-Preserving Diffusion Models for High-Fidelity Virtual Try-On. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7017–7026.

Yang, Z.; Li, S.; Wu, W.; and Dai, B. 2023. 3dhumangan: 3d-aware human image generation with 3d pose mapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23008–23019.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Zeng, J.; Song, D.; Nie, W.; Tian, H.; Wang, T.; and Liu, A.-A. 2024. CAT-DM: Controllable Accelerated Virtual Try-on with Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8372–8382.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, Y.; Dong, W.; Tang, F.; Huang, N.; Huang, H.; Ma, C.; Lee, T.-Y.; Deussen, O.; and Xu, C. 2023. ProSpect: Prompt Spectrum for Attribute-Aware Personalization of Diffusion Models. *ACM Transactions on Graphics (TOG)*, 42(6): 244:1–244:14.

Zhang, Y.; Tzun, T. T.; Hern, L. W.; and Kawaguchi, K. 2024. Enhancing semantic fidelity in text-to-image synthesis: Attention regulation in diffusion models. In *European Conference on Computer Vision*, 70–86. Springer.

Zhou, S.; Chan, K. C.; Li, C.; and Loy, C. C. 2022. Towards Robust Blind Face Restoration with Codebook Lookup TransFormer. In *NeurIPS*.

Zhu, J.; Chen, Y.; Ding, M.; Luo, P.; Wang, L.; and Wang, J. 2024. MoLE: Enhancing Human-centric Text-to-image Diffusion via Mixture of Low-rank Experts. *arXiv preprint arXiv:2410.23332*.

Zhu, L.; Yang, D.; Zhu, T.; Reda, F.; Chan, W.; Saharia, C.; Norouzi, M.; and Kemelmacher-Shlizerman, I. 2023. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4606–4615.