

# Enhanced Multi-Attribute Fashion Image Editing Using Vision Transformer-Guided Diffusion Models

Zeichen Zhao<sup>a</sup>

Madison, School of Computer, Data & Information Sciences, University of Wisconsin, Wisconsin, U.S.A.

**Keywords:** Diffusion Models, Attention-Pooling, Vision Transformer.


**Abstract:** This research explores the application of off-the-shelf diffusion models for fashion imagery generation, aiming to advance attribute-specific image manipulation without the need for manual masking or dataset-specific model training. This study presents a new method that combines a multi-attribute classifier with an attention-pooling mechanism by utilizing the flexibility and generative capabilities of diffusion models. These models were initially trained on large visual datasets such as ImageNet. This method is crucial in directing the diffusion process and enabling specific modifications of various fashion features within a single framework. The classifier's design, based on the Vision Transformer (ViT) architecture, improves the process of manipulating attributes to generate fashion images with greater realism and diversity. The experimental validation confirms that the suggested method outperforms existing generative models in producing fashion images that are of high quality and accurately represent the desired attributes. They provide significant improvements in image quality, attribute integrity, and editing flexibility.

## 1 INTRODUCTION

The fashion industry possesses the capacity to generate a wide array of designs that can stimulate and encourage creativity through the process of creative fashion synthesis, sometimes referred to as fashion image synthesis. Designers can effectively explore a wide range of design concepts and ideas by integrating photos of fashion products with unique characteristics and styles. This reduces the amount of time and financial resources required for prototyping (Sun, 2023). Diffusion models have become a potent alternative to conventional generative models such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). They exhibit greater abilities in creating high-quality, diversified images by means of an iterative refinement process. The models in the fashion industry have the ability to greatly transform design, customisation, and visualization processes. By training on extensive datasets of fashion apparel, diffusion models can potentially offer unparalleled realism and detail in generated fashion images. However, despite their advantages, optimizing these models for the specific challenges of fashion image generation—such as

capturing intricate details, fabric textures, and the dynamic nature of fashion trends—remains an area ripe for exploration.

The evolution of image generation in the fashion industry has primarily been driven by advancements in GANs and VAEs. Studies like Dhariwal and Nichol (Dhariwal, 2021) and Ho et al. (Ho, 2020) have laid the groundwork for diffusion models, highlighting their efficacy in generating complex image distributions. Fashion-focused research, such as the work by Karras et al. (Karras, 2019) and Zhu et al. (Zhu, 2017), has explored GANs for creating realistic clothing images. Diffusion model advancements (Sohl-Dickstein, 2015) (Song, 2020) have opened new avenues for research, particularly in refining image quality and generation efficiency. Among the advancements, Nichol and Dhariwal's optimization of diffusion models for improved image quality represented a significant leap forward (Nichol, 2021). Despite these developments, the specific application to fashion design, particularly in generating realistic garment textures, has seen limited exploration, indicating a gap this paper aims to address.

<sup>a</sup> <https://orcid.org/0009-0008-6188-8038>

This study aims to enhance the realism of clothing generated on models using diffusion models by improving the classifier component. The research focuses on a novel classifier architecture tailored for the intricate fashion domain, capable of efficiently guiding the diffusion process to generate highly realistic and diverse fashion images. The methodology involves fine-tuning a Vision ViT-backed classifier with an attention-pooling mechanism, enabling it to discern and manipulate multiple fashion attributes simultaneously. The classifier's design is optimized for better understanding and manipulating fashion attributes, allowing precise control over the generated imagery. The experimental results illustrate the efficacy of this strategy in generating lifelike clothing images, outperforming current methods in terms of both quality and attribute correctness.

## 2 METHODOLOGIES

### 2.1 Dataset Description and Preprocessing

The foundation of this study is rooted in the utilization of the Shopping100k dataset, a rich and varied compilation of fashion items meticulously categorized across a broad spectrum of attributes. This dataset is instrumental in training and evaluating the classifier-guided diffusion model for nuanced fashion attribute editing. The attributes span across multiple dimensions, including item category, pattern, gender, fabric, and collar, thereby capturing the multifaceted essence of fashion.

Table 1: Details regarding properties and their respective classifications.

Attribute	Number of Classes	Class
Category	15	Shirt, Jean, Tracksuit, Suit, Skirt, Dress, etc.
Fabric	14	Oversize, Regular, Skinny, Slim, etc.
Collar	17	Mandarin, V-neck, Polo, Turndown, Peter Pan, Round, etc.
Gender	2	Male, Female

The preprocessing methods standardized the images to a resolution of 256x256 pixels. It is a vital step in maintaining consistency and aiding efficient processing. The study offers a numerical examination of the characteristics and categories listed in Table 1

(Kenan, 2018). This table presents a picture of the dataset's composition and its extensive coverage of fashion attributes.

### 2.2 Proposed Approach

This research seeks to transform the way fashion images are manipulated by utilizing a pre-trained diffusion model that is readily available for modifying various fashion attributes at the same time, without requiring user intervention through masks. A distinctive feature of this approach is the guidance provided by a domain-specific classifier, enabling the model to efficiently edit a wide array of attributes, from item categories to fabric patterns. The overarching goal is to sustain high-quality image generation while offering a scalable and universally applicable editing framework across a diverse set of attributes using a singular model setup. Figure 1 showcases the model pipeline, detailing the step-by-step progression from an original image through the diffusion process, culminating in an image that mirrors the desired attribute modifications. This figure serves as a visual abstract of the methodology, illustrating the seamless integration of classifier guidance within the diffusion model's workflow.

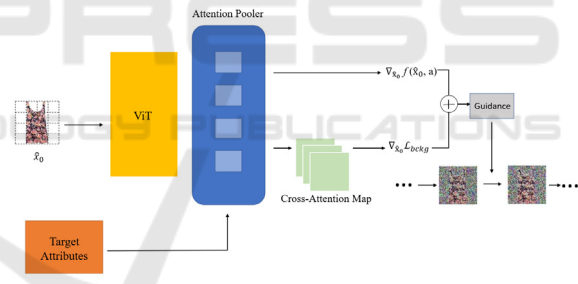


Figure 1: The model's overall pipeline (Picture credit: Original).

#### 2.2.1 ViT for Multi-Attribute Classification

The attribute classifier, central to this framework, is based on a ViT architecture, enhanced for multi-attribute classification through an attention-pooling mechanism. This classifier's ability to discern and focus on specific fashion attributes within an image is pivotal for accurate classification and effective image manipulation. The classifier, detailed in Figure 1, leverages a pre-trained ViT backbone augmented with an attention-pooling layer, optimizing it for fine-grained multi-attribute classification. The figure elaborates on the classifier's workflow, highlighting the attention mechanism's role in identifying and

concentrating on relevant image regions for different attributes, thus underpinning the classifier's precision.

A central aspect of the classifier involves the attention mechanism, crucial for focusing on relevant features within an image. The attention formula defined as follows:

$$Attention(Q, K, V) = \text{soft max}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

The formula utilizes the matrices Q, K, and V, which are derived from the input embeddings, to represent the query, key, and value correspondingly. The term  $d_k$  represents the dimensionality of the key vectors and is used to scale the dot products.

### 2.2.2 Off-the-Shelf Diffusion Model

The image modification framework relies on a pre-trained diffusion model that is readily available and has been developed on broad datasets such as ImageNet. This model's generative capabilities, combined with guidance from the finely tuned classifier, allow for precise attribute manipulation within images. The model leverages a comprehensive understanding of visual semantics, gleaned from its training, to facilitate nuanced edits that align with the specified fashion attributes. For the diffusion model, the denoising process can be described using the reverse diffusion steps, which can be mathematically represented as follows:

$$q(x_{t-1} | x_t) = N(x_{t-1}; \mu(x_t, t), \Sigma(x_t, t)) \quad (2)$$

The above formula represents the conditional distribution of the previous state  $x_{t-1}$  given the current state  $x_t$ . The symbols  $\mu$  and  $\Sigma$  represent the average and spread of the Gaussian distribution, respectively. The characteristics of this distribution are obtained through the training procedure to effectively reverse the diffusion process.

### 2.2.3 Loss Function

The core mechanism of this diffusion model relies on a denoising process that iteratively refines a random noise distribution towards a data distribution of interest, guided by a trained neural network. The technique is formalized by employing a loss function that quantifies the difference between the generated image at a particular diffusion phase and the original, unaltered image. The primary goal is to reduce this loss, thereby improving the quality and precision of

the produced images according to the specified characteristics.

In denoising diffusion probabilistic models (DDPMs), the main loss function commonly used is the mean squared error (MSE) between the original clean image  $x_0$  and the reconstructed image obtained from the diffusion process at a specific step  $t$ , taking into account the added noise  $\epsilon$ . More precisely, this can be stated as:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3)$$

Where  $y_i$  represents the actual value,  $\hat{y}_i$  represents the estimated value, and  $n$  denotes the total number of observations.

$$L = \|\epsilon_0(x_t, t) - \epsilon\|^2 \quad (4)$$

Here (Ho, 2020),  $\epsilon_0(x_t, t)$  represents the predicted noise by the neural network parameterized by  $\theta$  for the noisy image  $x_t$  at timestep  $t$ , and  $\epsilon$  is the true noise that was added to the original image  $x_0$  to obtain  $x_t$ . The expectation is taken over the clean images  $x_0$ , the noise  $\epsilon$ , and the timesteps  $t$ .

This loss function is adapted to ensure not only the fidelity of the generated image to the original one but also its alignment with the desired fashion attributes.

## 2.3 Implementation Details

The implementation of this framework is characterized by the strategic utilization of a pre-trained unconditional diffusion model, explicitly designed for processing images of 256x256 resolution. Adaptation to the specific requirements of the fashion domain is predominantly achieved through the deployment of a multi-attribute classifier, constructed utilizing a Vision Transformer architecture enhanced with an attention-pooling layer. The classifier's fine-tuning on the Shopping100k dataset, with a focused emphasis on a broad spectrum of fashion attributes, has been pivotal. This methodology further incorporates data augmentation techniques to bolster model resilience and hyperparameter optimization strategies to refine performance, ultimately aiming to achieve an optimal equilibrium between precision in attribute manipulation and meticulous background preservation.

Table 2: Finetuning Strategies for the Multi-attribute Classifier.

Initialization	Training Strategy	Category	Fabric	Gender	Collar
Random Init	End to End	30.1	56.6	66.1	31.0
Imagenet-pretrained	Attention-Pool Only	84.5	58.1	94.9	91.3
Imagenet-pretrained	Last2	65.7	52.3	73.8	81.2
Imagenet-pretrained	Last4	42.8	52.3	81.3	75.9
Imagenet-pretrained	Last6	41.1	51.7	77.1	75.6
CLIP-pretrained	Attention-Pool Only	85.8	60.1	95.3	90.9
CLIP-pretrained	Last6	86.6	67.3	97.2	75.1
CLIP-pretrained	Last12	84.9	67.2	96.8	78.1
CLIP-pretrained	Last18	81.8	66.3	95.6	72.8

Table 3: Impact of Data Augmentations.

Model Type	Augmentation	Category	Fabric	Gender	Collar
Imagenet-pretrained	No Aug.	85.3	58.5	95.0	91.4
Imagenet-pretrained	Random Aug.	81.4	57.1	92.4	91.1
CLIP-pretrained	No Aug.	86.2	59.7	96.5	89.3
CLIP-pretrained	Random Aug.	86.2	59.9	95.6	90.9

### 3 RESULTS AND DISCUSSION

Table 2 presents an in-depth quantitative analysis of different fine-tuning strategies for the ViT model, aimed at adapting it for multi-attribute classification within the fashion domain. The table compares the performance across various initialization methods, including random initialization, ImageNet-pretrained, and CLIP-pretrained setups, across a range of fashion attributes. Notably, the CLIP-pretrained initialization with attention-pool only fine-tuning emerges as the most effective strategy, achieving an average classification accuracy significantly higher than other approaches. This underscores the importance of leveraging pre-existing visual semantics knowledge and the effectiveness of attention-pooling in capturing the nuanced distinctions across diverse fashion attributes.

Table 3 delves into the role of data augmentations, presenting a nuanced picture of their effects. The differential impact on models pre-trained with ImageNet versus CLIP points to the intrinsic characteristics imbued by the pretraining data's diversity. For the CLIP-pretrained model, augmentations serve to further generalize the classifier's understanding of fashion attributes, improving manipulation outcomes. This table challenges the one-size-fits-all approach to data augmentation, advocating for a tailored strategy that considers the model's pretraining background.

Table 4 explores the relationship between the ViT model size and its performance in multi-attribute classification.

Table 4: Model Size and Performance.

Model	Category	Fabric	Gender	Collar
B/32	83.7	58.3	94.8	88.3
B/16	84.6	58.7	95.1	88.5
L/14	86.2	59.6	96.2	89.2

The consistent enhancement in classification accuracy as the model size increases (from B/32 to L/14) emphasizes the crucial significance of model capacity in effectively managing the intricacy of fashion features. This trend reaffirms the principle that larger models, capable of encapsulating more nuanced feature representations, are better suited to the demands of multi-attribute classification in the fashion domain.

The integration of classifier-guided diffusion models for fashion attribute editing represents a paradigm shift in the field, addressing the scalability and versatility challenges previously encountered with GAN-based methods. The results delineated in this study provide empirical evidence of the framework's robustness and adaptability across a spectrum of fashion attributes, from item categories to intricate patterns and textures.

A pivotal factor in the framework's success is the strategic finetuning of the ViT for multi-attribute guidance, as detailed in the experimentation with various initialization and finetuning strategies. The adoption of an attention-pooling mechanism enables

the model to discern and manipulate distinct attributes within an image, thereby enhancing the granularity and precision of edits. Furthermore, the elimination of manual region masking through the utilization of classifier attention maps introduces a level of automation and user-friendliness previously unattainable. This aspect of the framework not only simplifies the editing process but also broadens its applicability to non-expert users, potentially revolutionizing how fashion images are manipulated for various applications. The comprehensive evaluation of the framework, including ablation studies and comparative analyses, substantiates its superiority over existing methods. The findings illustrate the framework's capacity to facilitate complex multi-attribute manipulations while maintaining high fidelity and alignment with target attributes, thereby marking a significant advancement in the field of image manipulation.

This study confirms the efficacy and promise of utilizing readily available diffusion models for fashion attribute modification. The suggested framework represents a substantial advancement in meeting the ever-changing requirements of the fashion industry, providing a scalable, adaptable, and user-friendly solution for high-quality editing of fashion images.

## 4 CONCLUSIONS

This research introduces a novel approach to enhancing the realism of clothing in fashion images through the development and application of diffusion models, focusing on the domain-specific intricacies of fashion design. By innovatively combining a pre-trained diffusion model with a newly proposed classifier architecture, this study endeavors to generate high-fidelity and diverse fashion images. The classifier, leveraging a ViT backbone and an attention-pooling mechanism, is finely tuned to efficiently guide the diffusion process across multiple fashion attributes simultaneously. The experimental results demonstrate the superiority of this approach in generating genuine clothes images with remarkable attribute accuracy. This study explores the application of classifier-guided diffusion models in the field of fashion image editing. The primary objective of this study is to address the challenges associated with scalability and the accurate manipulation of qualities within the fashion sector. The research introduces a versatile approach that may be readily adapted to accommodate various image properties. It accomplishes this by employing a pre-existing

diffusion model, a domain-specific classifier, which guides its editing capabilities. The classifier, enhanced with attention-pooling techniques, enables the model to properly process different fashion attributes, resulting in a significant improvement compared to traditional methods that depend on conditional GANs.

The empirical findings of this study illustrate the effectiveness of this framework in producing images of convincing quality and attribute alignment, marking a noteworthy contribution to the field of image manipulation and the broader context of digital fashion design. Future research for both studies will further explore the potential applications and improvements of diffusion models in the fashion industry. The main focus will be on improving the accuracy of classification and the fidelity of attribute manipulation to push the limits of realism in generating fashion images. Additionally, exploration into integrating more complex and nuanced fashion attributes will be pursued to enrich the versatility and applicability of the proposed frameworks, aiming to meet the evolving demands of fashion design and online retail environments.

## REFERENCES

- Sun, Z., Zhou, Y., He, H., and Mok, P.Y., (2023). SGDiff: A Style Guided Diffusion Model for Fashion Synthesis. In Proceedings of the 31st ACM International Conference on Multimedia (MM '23). Association for Computing Machinery, pp: 8433–8442.
- Dhariwal, P., and Nichol, A., (2021). Improved techniques for training score-based generative models. in Proc. of the 34th International Conference on Neural Information Processing Systems (NeurIPS), pp: 1-12.
- Ho, J., Jain, A., and Abbeel, P., (2020). Denoising diffusion probabilistic models. in Proc. of the 33rd International Conference on Neural Information Processing Systems (NeurIPS), pp: 1-11.
- Karras, T., Laine, S., and Aila, T., (2019). A style-based generator architecture for generative adversarial networks," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp: 4401-4410.
- Zhu, J., Park, T., Isola, P., and Efros, A.A., (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks," in Proc. of the IEEE International Conference on Computer Vision (ICCV), pp: 2223-2232.
- Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S., (2015). Deep unsupervised learning using nonequilibrium thermodynamics. in Proc. of the 32nd International Conference on Machine Learning (ICML), pp: 2256-2265.



- Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., and Poole, B., (2020). Score-based generative modeling through stochastic differential equations," in Proc. of the 9th International Conference on Learning Representations (ICLR), pp: 1-18.
- Nichol, A.Q., Dhariwal., (2021). Improved Denoising Diffusion Probabilistic Models," in Proc. of the 38th International Conference on Machine Learning (ICML), PMLR. vol.139, pp: 8162-8171.
- Kenan, A., Joo, Hwee, L., Jo, Yew, T., and Ashraf, Kassim., (2018). Efficient multi-attribute similarity learning towards attribute-based fashion search. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp: 1671–1679.
- Ho, J., Jain, A., and Abbeel, P., (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, vol. 33, pp: 6840–6851.

