Leveraging Off-the-shelf Diffusion Model for Multi-attribute Fashion Image Manipulation

Chaerin Kong[1*]   DongHyeon Jeon[2]   Ohjoon Kwon[2]   Nojun Kwak[1]

[1]Seoul National University   [2]NAVER

veztylord@snu.ac.kr   {donghyeon.jeon, ohjoon.kwon}@navercorp.com   nojunk@snu.ac.kr

Abstract

Fashion attribute editing is a task that aims to convert the semantic attributes of a given fashion image while preserving the irrelevant regions. Previous works typically employ conditional GANs where the generator explicitly learns the target attributes and directly execute the conversion. These approaches, however, are neither scalable nor generic as they operate only with few limited attributes and a separate generator is required for each dataset or attribute set. Inspired by the recent advancement of diffusion models, we explore the classifier-guided diffusion that leverages the off-the-shelf diffusion model pretrained on general visual semantics such as Imagenet. In order to achieve a generic editing pipeline, we pose this as multi-attribute image manipulation task, where the attribute ranges from item category, fabric, pattern to collar and neckline. We empirically show that conventional methods fail in our challenging setting, and study efficient adaptation scheme that involves recently introduced attention-pooling technique to obtain a multi-attribute classifier guidance. Based on this, we present a mask-free fashion attribute editing framework that leverages the classifier logits and the cross-attention map for manipulation. We empirically demonstrate that our framework achieves convincing sample quality and attribute alignments.

1. Introduction

Denoising diffusion models [11, 33, 7, 26, 30] have recently gained great attention from the research community for their impressive synthesis quality, training stability and scalability. They have demonstrated promising performances across diverse tasks and benchmarks spanning unconditional image synthesis [7], text-driven image generation [26, 30, 21], image manipulation [3, 10] and video synthesis [13]. Nevertheless, studies on diffusion models are far from complete; unlike traditional generative model families such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), the true potential of diffusion models are yet to be fully disclosed.

One of the reasons for the popularity of diffusion models is their natural capacity to incorporate conditioning information into the generative process. Conditional diffusion models typically rely on classifier-guidance [7] or classifier-free-guidance [12], where the former requires a separately trained classifier (independent of the diffusion model) while the latter involves condition-aware training of the diffusion model from the beginning. Classifier-guidance, in particular, provides a means to leverage the off-the-shelf diffusion model trained on general visual semantics such as Imagenet [6], which can be particularly useful for domains with insufficient public data. [1]

Fashion domain, being the heart of modern e-commerce, has huge practical upside but retains relatively little publicly available data due to privacy and proprietary issues. To make things worse, the visual semantics are significantly distant from the common generative benchmarks such as Imagenet or FFHQ [14], strongly demanding an adequate adaptation procedure. For these reasons, the field of fashion has been relatively less explored in the deep learning community despite its industrial values.

Image manipulation is a generative task that aims to control the semantics of an input image while preserving the irrelevant details. Fashion image manipulation has great applications in fashion design, interactive online shopping, and personalized marketing. Thus several works [2, 24, 18] have posed this task as fashion attribute editing task, *i.e.,* attribute-guided image manipulation, and delivered promising results. However, their real world applications are significantly restricted as (1) they train a separate generative model (typically GAN) for each fashion dataset, and (2) their editing operations are limited to few predefined attributes such as color or sleeve length, as the *generative model* has to learn to convert these attributes during training. As these limitations render them short for a generic

---

*Work done during internship at NAVER.

[1]Classifiers are generally easier and more straightforward to train or finetune compared to generative models under limited data.

and scalable editing system, we propose to employ an off-the-shelf diffusion model trained on a general semantic domain such as Imagenet, and guide its editing with a domain-specific classifier. To the best of our knowledge, this is the first attempt to introduce diffusion models to the fashion domain, especially fashion attribute editing.

This approach has clear advantage over the prior art for several reasons. First, training a classifier is generally much simpler and easier than training a generative model under limited data. As there is clear shortage of well-annotated fashion images, we present an efficient finetuning strategy that empowers the classifier to reason a wide range of fashion attributes at once with the help of recently proposed attention-pooling technique [36].

Second, as the capacity to understand different fashion attributes has been transferred to the classifier, our manipulation framework can operate in a much greater scale, covering attributes like item category, neckline, fit, fabric, pattern, sleeve length, collar and gender with a single model. This is in clear contrast to previous works that supports only a few editing operations with separate models. We later show that training attribute-editing GANs with such a wide set of attributes leads to total training collapse (Fig. 3). Last, we can edit multiple attributes at once in an integrated manner, since we use the guidance signal of a multi-attribute classifier. This is particularly important for fashion domain as various attributes must be in harmony with one another to yield an attractive output.

Our fashion attribute classifier adopts a pretrained ViT backbone and a finetuned attention pooling layer in order to best perform multi-attribute classification with relatively small training dataset. We use the gradient signal of this classifier to guide the diffusion process as done in [3, 7], thus we can alter more than one attributes at once. For local editing of images, using a user-provided mask that explicitly marks the area to be edited is the most straightforward and widely used approach [3]. However, we leverage the natural capacity of our attribute classifier to attend to the relevant spatial regions during classification, and use the attention signal to suppress *excess modifications* of the original image. This frees the users from the obligation to designate the specific region of interest and simplifies the image manipulation pipeline.

In sum, our contributions can be summarized as

- We introduce classifier-guided diffusion as a simple yet generic and effective framework for fashion attribute editing.

- We empirically present an efficient finetuning scheme to adapt a pretrained ViT for a domain-specific multi-attribute classification setting.

- We demonstrate the effectiveness of our framework with thorough evaluations.

2. Related Works

Diffusion Models [32, 11] are a family of generative models that convert a Gaussian noise into a natural image through an iterative denoising process, which is typically modeled with a learnable neural network such as U-Net [29]. They have gained the attention from both the research community and the public with their state-of-the-art performances in likelihood estimation [11, 22] and sample quality [7, 30, 26]. Specifically, they have demonstrated impressive results in conditional image synthesis, such as class-conditional [7], image-conditional [20] and text-conditional [30, 26] settings. Conditioning a diffusion model is typically done with either the classifier-guidance [7] or the classifier-free-guidance [12], and conditional diffusion models have been shown to be capable of learning an extremely rich latent space [21, 26]. Recently, a line of works [33, 28] focus on improving the sampling speed of diffusion model, by either altering the Markovian noising process or embedding the diffusion steps into a learned latent space. Another group [15, 3, 10] studies the applications of diffusion models such as text-guided image manipulation.

Image Manipulation or image editing has been a long standing challenge in the computer vision community with a wide range of practical applications [39, 23, 16, 17]. Image manipulation with deep models typically accompany editing operations in the latent space to produce semantic and natural modifications. Hence, image manipulation with Generative Adversarial Networks (GANs) poses a new problem of GAN inversion [35, 27], which is the initial step to find a latent vector corresponding to the image that needs to be altered. Recently, as diffusion models rise as prominent alternatives, image manipulation using diffusion models are widely being studied. Blended-diffusion [3] uses a user-provided mask and a textual prompt during the diffusion process to *blend* the target and the existing background iteratively. A concurrent work of ours, prompt-to-prompt [10] captures the text cross-attention structure to enable purely prompt-based scene editing without any explicit masks. In our work, we take blended-diffusion as the starting point, and incorporate a domain-specific classifier and its attention structure for mask-free multi-attribute fashion image manipulation.

Fashion Attribute Editing is a highly practical task for which several works have been studied. AMGAN [2] uses Class Activation Map and two discriminators that identifies both real/fake and the attributes. Fashion-AttGAN [24] improves upon AttGAN [9] to better fit the fashion domain, by including additional optimization objectives. VPTNet [18] aims to handle larger shape changes by posing attribute editing as a two-stage procedure of shape-then-appearance editing. These works all employ the GAN framework, and thus the range of attributes supported for editing is relatively lim-

ited. We explore fashion attribute editing with an off-the-shelf diffusion model that has learned a rich semantic latent space from a set of general images to simplify the generative pipeline and achieve more generic editing capabilities.

3. Approach

3.1. Preliminaries

Denoising Diffusion Probabilistic Models (DDPMs) [11] learn to invert the Markovian noising process which is typically parameterized by isotropic Gaussian. They were shown to achieve the best performance under the reweighted objective [11], and have demonstrated state-of-the-art performances on a wide range of generative benchmarks [7, 30]. We present a brief overview of diffusion models below.

Given a clean data distribution $x_0 \sim q(x_0)$, a Markovian noising process can be defined as:

$$q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}) \qquad (1)$$

$$q(x_1, ..., x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}), \qquad (2)$$

with β_t being the noise scale controlling the diffusion process and $\mathbf{I}$ being the identity matrix. It is shown that with a large enough total time steps T, x_T can be regarded as an isotropic Gaussian variable. As we are gradually adding Gaussian noises, an arbitrary noising step can be directly computed as

$$q(x_t|x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}), \qquad (3)$$

with a scalar $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$.

To sample under the diffusion framework, we *reverse* the forward denoising process. In other words, we begin with a Gaussian noise $x_T \sim \mathcal{N}(0, \mathbf{I})$, and go through a series of posterior sampling steps, $x_{t-1} \sim q(x_{t-1}|x_t)$, which is typically modeled with a neural network with learnable parameters θ:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \qquad (4)$$

[11] shows that with mild assumptions, this posterior can be represented as a Gaussian distribution with a diagonal covariance and the mean $\mu_\theta(x_t, t)$ can be computed from $\epsilon_\theta(x_t, t)$. Thus, the neural network is trained to predict the noise $\epsilon_\theta(x_t, t)$ instead of the mean under the following reconstruction objective:

$$\mathcal{L} = ||\epsilon_\theta(x_t, t) - \epsilon||^2. \qquad (5)$$

Several works [22, 7] recently introduce empirical strategies to produce samples with better quality, such as model architecture and objective weighting.
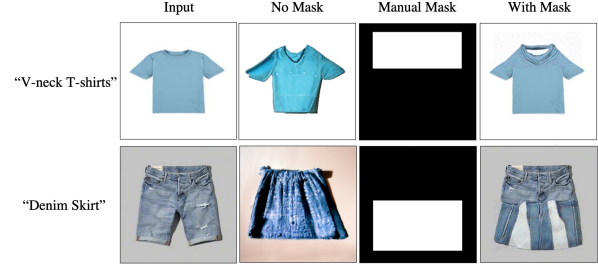

Figure 1. Naive application of blended-diffusion in fashion domain leads to suboptimal results.

Blended-Diffusion [3] adapts the diffusion model for the text-driven image manipulation task. Employing a user-provided regional mask M, a pretrained vision-language encoder, *e.g.,* CLIP [25], provides guidance to form a text-aligned visual semantic in the masked region. At each diffusion step, the masked region guided by CLIP and the unmasked background region are *blended* to form a smooth and natural visual composition. To evade adversarial effects stemming from ascending gradients of CLIP logits, the authors propose *Extending Augmentations* that create multiple views of a given image.

In this work, we propose to tailor this generic editing framework for fashion domain with concise modifications. We first show that naively applying this general framework yields suboptimal results, then present an efficient finetuning technique to prepare a multi-attribute fashion classifier that is capable of reasoning different spatial regions for classification of different attributes. Lastly, we leverage its natural attention map for mask-free fashion attribute editing, completing an extremely simple and generic fashion image manipulation framework. The overall procedure of ours is described in Procedure 1.

3.2. Taming ViT for Multi-attribute Guidance

As shown in Fig. 1, our early experiments show that naively applying the guidance signal from a pretrained model (*e.g.,* CLIP) as in blended-diffusion yields unsatisfactory results in fashion domain due to the severe domain gap. To provide a more fine-grained guidance to our diffusion model, we explore ways to efficiently train a domain-specific classifier.

Vision Transformers (ViTs) [8] have recently shown impressive performances across diverse computer vision tasks, frequently replacing the state-of-the-arts previously held by convolutional neural networks [4, 19]. In conventional ViTs, the global semantic of an input image is aggregated to a single [CLS] token, which is commonly projected with additional learnable layers for downstream tasks. However, in our multi-attribute classification setting, this can lead to *over-simplification* of visual semantics, as different at-
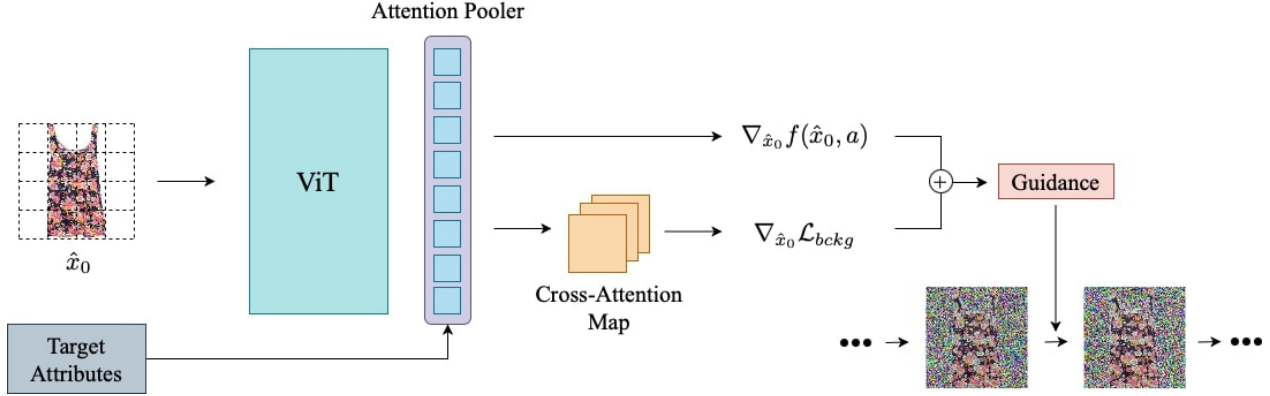
Figure 2. Overall pipeline of our proposed framework. A predicted clean image is first patch-embedded as input to the ViT backbone. The attention-pooling layer on top of the ViT returns attribute logits and the cross-attention maps, which are used to guide the diffusion steps.

Procedure 1 Our proposed fashion attribute editing pipeline, with the off-the-shelf diffusion model θ and the attribute classifier f.

Input: original image X, target attributes A, diffusion step k, number of augmentation views N and attribute loss weight λ

Output: edited output $\hat{X}$

$\quad X_k \sim \mathcal{N}(\sqrt{\bar{\alpha}_k}X_0, (1-\bar{\alpha}_k)\mathbf{I})$

$\quad$**for** t in $k{:}0$ **do**

$\quad\quad \hat{X}_0 \leftarrow \frac{X_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(X_t,t)}{\sqrt{\bar{\alpha}_t}}$

$\quad\quad \hat{X}_{0,aug} \leftarrow ExtendingAugmentations(\hat{X}_0, N)$

$\quad\quad \nabla_{attr} \leftarrow \frac{1}{N}\sum_{i=1}^{N}\nabla_{\hat{X}_{0,aug}}f(\hat{X}_{0,aug}, A)$

$\quad\quad \bar{M} \leftarrow \frac{1}{|A|}\sum_{j=1}^{|A|}AttentionMap(f, \hat{X}_0, A)$

$\quad\quad \nabla_{bckg} \leftarrow -\nabla_{\hat{X}_0}\mathcal{L}_{bckg}(X, \hat{X}_0, \bar{M})$

$\quad\quad \nabla_{total} \leftarrow \lambda\nabla_{attr} + \nabla_{bckg}$

$\quad\quad X_{t-1} \sim \mathcal{N}(\mu_\theta(X_t) + \Sigma_\theta(X_t)\nabla_{total}, \Sigma_\theta(X_t))$

$\quad$**end for**

$\quad$**return** X_{-1}

---

tributes reside in different spatial regions. In other words, we have to look at different parts of a clothing to determine different attributes such as neckline or sleeve length. Hence, we adopt recently proposed attention-pooling mechanism [36] to aggregate the spatial features in a learnable way with an additional cross-attention layer. This enables our model to attend to different spatial regions for classification of different attributes. After the attention pooling, each attribute token is projected for classification, and they are optimized with the classic cross-entropy loss formulated as follows:

$$\mathcal{L}_{cls} = -\sum_{i=1}^{|A|}\log\frac{exp(x_{target}^{(i)})}{\sum_{j=1}^{C^{(i)}}exp(x_j^{(i)})}\cdot\mathbb{1}(y^{(i)}\neq-1). \quad (6)$$

Here, $|A|$ is the number of attributes, $x^{(i)}$ refers to the model prediction for the i-th attribute, $C^{(i)}$ is the number of classes

for the attribute, and $y^{(i)}$ indicates the ground truth label. We mask those with missing attributes, *e.g.,* neckline for a skirt. Note that we drop the mini-batch notation for brevity.

As ViTs are famous for being notoriously data-hungry [8], we choose to finetune a ViT pretrained on a larger dataset. We explore different initialization and fine-tuning methods and present the best practice for our setting in Sec.4.1.

3.3. Local Image Editing with Patch-level Attention

One of the challenges of image editing is to concisely alter the regions to suit the given condition while keeping the irrelevant parts unchanged. Instead of relying on a user-provided mask as done in [3], we employ the attention map of our classifier to determine the relevant regions. This is distinguished from other works that use separate procedures such as Class Activation Map [38] or Grad-CAM [31], as we simply use the attention map of each attribute token (in the attention pooling layer), which is computed during the classifier forward pass.

Formally, for a multi-attribute image editing, we average-pool the spatial attention maps for the target attribute tokens, and impose background preservation loss on the low-attention regions as follows:

$$\mathcal{L}_{bckg} = (1-\bar{M})\odot||\hat{X}-X||_2 + (1-\bar{M})\odot pd(\hat{X}, X), \quad (7)$$

where $\bar{M}$ refers to the average-pooled spatial attention maps, $\odot$ indicates the Hadamard product, $\hat{X}$ is the predicted output from X, and pd is Learned Perceptual Image Patch Similarity (LPIPS) [37], also known as the perceptual distance.

3.4. Overall Framework

We build our framework upon *blended-diffusion* [3], and make modifications to suit the off-the-shelf diffusion model for fashion attribute editing. First, we replace CLIP guidance with our finetuned classifier guidance that supports

fine-grained multi-attribute editing. Then, we get rid of the user mask and enforce background preservation with the classifier attention map. We go through the diffusion steps with the initial input image, where the classifier guidance signal pulls the diffusion process towards the target attributes. The overall pipeline is illustrated in Fig. 2.

4. Experiments

4.1. Datasets and Baselines

We use the commonly used Shopping100k [1] dataset for training and evaluations. As our diffusion model does not need additional training, we only train the classifier on this dataset. For baselines, we compare our method with StarGAN [5] and Fashion-AttGAN [24], two representative attribute manipulation methods of which the latter is further suited for the fashion domain. We train these models on Shopping100k with 8 most common attributes, summing to 100 labels total. The details are specified in Tab. 1.

Attribute	#Classes	Class
Category	16	Coat, Jacket, Suit, Shirt, T-shirt, Jumper, Shorts Trouser, Jean, Swimming, Jumpsuit, Pyjamas, Tracksuit, Bottoms, Tracksuit, Skirt, Dress
Collar	17	Buttondown, Cutaway, High, Hood, Kent, Lapel, Lined, Mandarin, Polo, Round, Shawl, Turndown, V-neck, Peter Pan, Volant, Shirt, Chin
Fabric	14	Canvas, Crocheted, Denim, Fleece, Hardshell, Jersey, Jersey Lace, Lace, Mesh, Mesh Jersey, Rib, Softshell, Sweat, Leather
Fit	15	Bootcut, Flared, High waist, Jeggings, Large, Loose, Low, Oversize, Regular, Skinny, Slim, Small, Straight, Tailered, Tapered
Gender	2	Male, Female
Neckline	11	Boat, Backless, Cache-coeur, Henley, Low v-neck, Low round, Off-the-shoulder, Round, Square, V-neck, Envelope
Pattern	16	Animal, Burnout, Camouflage, Checked, Marl, Color gradient, Colorful, Floral, Herringbone, Paisley, Photo, Pinstriped, Plain, Polka dot, Print, Striped
Sleeve Length	9	3/4, Spaghetti, Sleeveless, Elbow, Extra long, Extra short, Long, Short, Strapless

Table 1. Details on the attributes and their corresponding classes. We use the attribute annotations of Shopping100k dataset.

4.2. Implementation Details

We leverage the unconditional diffusion model pretrained on 256x256 scale Imagenet [7]. Surprisingly enough, despite the fact that the diffusion model has never seen any fashion-specific images, the classifier alone can guide fashion attribute editing in the rich latent space of the diffusion model. For the classifier, we test three initialization methods: random, Imagenet ViT and CLIP ViT. ViT-Large model was used for all experiments, and we added an attention pooling layer on top of the final transformer block. This attention pooler is essentially a cross-attention layer with 8 queries, each corresponding to an attribute. These tokens are linearly projected for final classification.
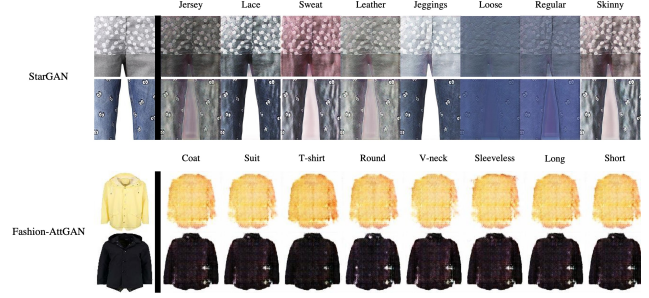

Figure 3. Attribute-editing GANs fail to handle generic manipulation scenario with more attributes and display training collapse.

4.3. Collapse of Attribute-editing GANs under Wider Attribute Space

Previous methods using attribute-editing GANs typically focus on one or two attributes (*e.g.,* color or sleeve length), as these methods enforce the generator itself to learn to reason about different fashion attributes. Hence, we first observe how these conventional GANs perform in the more generic manipulation setting, where we tackle most of the widely used fashion attributes, ranging from item category to fit and neckline, with a single framework.

Fig. 3 clearly illustrates that naively applying attribute-editing GANs [5, 24] leads to severe training collapse. As these methods aim to embed attribute manipulation capacity into the *generator*, they have inherent disadvantage for scalability. For multi-attribute editing, *e.g.,* editing the neckline *and* the sleeve length, multiple generative models have to be trained to perform a sequence of manipulation operations, which greatly limits their practical applications.

4.4. ViT Finetuning

As demonstrated in the previous section, training an attribute-aware generative model is not a scalable option. Therefore, we investigate ways to leverage a well trained generative model with a domain-specific attribute-aware classifier, deferring the attribute reasoning capacity to the classifier for the sake of a generic and scalable attribute editing system.

We first present empirical findings from the classifier training. As annotated data is relatively scarce in the fashion domain, we regard the initialization scheme to be a key factor in the final classifier performance. Hence, we explore three settings: (1) random initialization, (2) Imagenet-pretrained initialization and (3) CLIP-pretrained initialization. For the last two, we further compare different fine-tuning approaches, *i.e.,* how many layers or blocks to set as learnable (unfreeze) during the adaptation period.

Tab. 2 shows the result. We observe that with relatively limited data, freezing most of the parameters while training a learnable attention pooling layer yields the best

		Category	Fabric	Sleeve Length	Pattern	Gender	Fit	Collar	Neckline	Average
Random Init.										
	End-to-End	30.1	56.6	51.8	50.3	66.1	57.1	31.0	65.1	51.0
Imagenet-pretrained										
	Attention-Pool Only	85.4	58.7	84.8	76.1	95.0	66.4	**91.4**	**84.4**	80.3
	Last2	67.0	52.8	78.9	48.5	74.6	59.7	81.6	78.1	67.6
	Last4	44.8	52.8	60.8	45.0	82.0	58.6	76.1	76.2	62.0
	Last6	43.1	52.1	73.0	44.0	77.4	54.9	76.0	71.1	61.5
CLIP-pretrained										
	Attention-Pool Only	86.3	60.2	**84.9**	80.4	95.8	69.9	91.0	83.3	**81.5**
	Last6	**87.2**	**67.9**	84.4	**87.2**	**97.4**	**70.9**	75.3	78.0	81.0
	Last12	85.6	67.8	83.2	86.4	97.1	69.9	78.5	76.5	80.6
	Last18	82.3	66.8	79.7	83.2	95.9	68.4	72.9	73.7	77.8
	Last24	51.7	61.9	70.4	61.5	84.7	61.9	41.6	65.1	62.3

Table 2. Quantitative evaluations on different finetuning strategies. We present finetuning results on Shopping100k [1], and all models adopt ViT-L for comparison.

		Category	Fabric	Sleeve Length	Pattern	Gender	Fit	Collar	Neckline	Average
Imagenet-pretrained										
	No Aug.	**85.4**	**58.7**	**84.8**	**76.1**	**95.0**	**66.4**	**91.4**	**84.4**	**80.3**
	Random Aug.	81.5	57.3	83.3	73.7	92.7	65.2	91.2	82.0	78.4
CLIP-pretrained										
	No Aug.	86.3	59.8	**85.5**	79.6	**96.6**	67.7	89.5	82.5	80.9
	Random Aug.	86.3	**60.2**	84.9	**80.4**	95.8	**69.9**	**91.0**	**83.3**	**81.5**

Table 3. Ablation on random data augmentations. Only the attention pooling layer is finetuned with Shopping100k dataset. Note that in Tab. 2, we report the numbers from *No Aug.* model for Imagenet-pretrained initialization.

Model	Category	Fabric	Sleeve Length	Pattern	Gender	Fit	Collar	Neckline	Average
B/32	83.9	58.4	84.3	77.9	95.0	66.0	88.4	79.5	79.2
B/16	84.8	58.8	**85.5**	78.6	95.3	**67.8**	88.6	80.2	80.0
L/14	**86.3**	**59.8**	**85.5**	**79.6**	**96.6**	67.7	**89.5**	**82.5**	**80.9**

Table 4. Model size ablation. We observe steady improvements in the overall performance as the model gets bigger, which agrees with recent findings.

overall performance. This indicates that the pretrained ViT backbone retains a reasonable level of general visual reasoning, and the additional attention pooling layer successfully extracts diverse visual information to perform the fine-grained multi-attribute classification. Moreover, initializing with pretrained weights helps, and CLIP pretraining covers a wider range of visual semantic compared to Imagenet, possibly leading to better outcomes. Lastly, as different attributes demand different aspects of visual reasoning, the classification accuracy trend is not always consistent. We basically choose the model with the highest average score for the best overall guidance.

Data augmentations have been shown to affect the performance of visual classifier models [34]. Thus, we explore the influence of data augmentations in our framework. From Tab. 3, we gain conflicting insights; random augmentation boosts the performance for CLIP-pretrained ViT but damages it for the Imagenet-pretrained variant. We note that this result is obtained from finetuning the attention pooling layer on Shopping100k dataset, so it is possible to observe different trends in different settings. In our situa-

tion, we hypothesize that as we freeze most of the model parameters, imposing strong augmentations could result in *over-pressuring* the limited learnable parameters. As CLIP-pretraining prepares the model for a much greater diversity (compared to Imagenet, in terms of data scale), this potential adversarial effect only realizes for the less robust model variant.

Lastly, we explore the performance trend depending on the model size. Tab. 4 shows that bigger models (or longer input sequence) consistently yield better outcomes, even under a strict finetuning scheme. Hence, we adopt the best performing model variant, CLIP-pretrained ViT-L/14, as our diffusion guidance in the following sections.

4.5. Generic Fashion Attribute Editing

Thanks to the domain-specific multi-attribute classifier, we can leverage a pretrained off-the-shelf diffusion model (not necessarily adapted for fashion domain) for generic fashion attribute editing. Fig. 4 shows qualitative manipulation results on diverse attributes. We note that these editing operations are all done with a single finetuned classifier and

Figure 4. Qualitative evaluations on generic fashion attribute editing. We display the original source images on first and fifth column, and the rest are edited samples. Our framework handles both texture manipulation and shape deformation proficiently, producing plausible samples across various attributes.

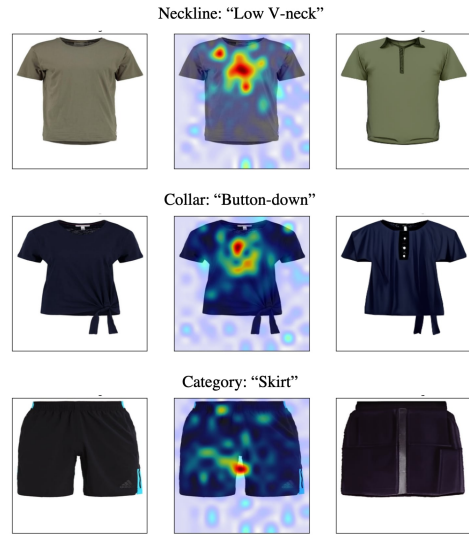Figure 5. Qualitative results from multi-attribute editing.

Figure 6. Classifier attention-map visualizations.

an off-the-shelf diffusion model. Unlike attribute-editing GANs that collapse as the number of attributes increases (see Fig. 3), our framework shows satisfactory editing capabilities across various attributes as the classifier can handle a much wider set of attributes with little instability when

compared to the generator. Especially, to our knowledge, attributes like item category or fabric have not been directly explored in the previous works on attribute editing mainly due to their difficulties. Since we employ a powerful diffusion model trained with a large dataset, providing the *right*
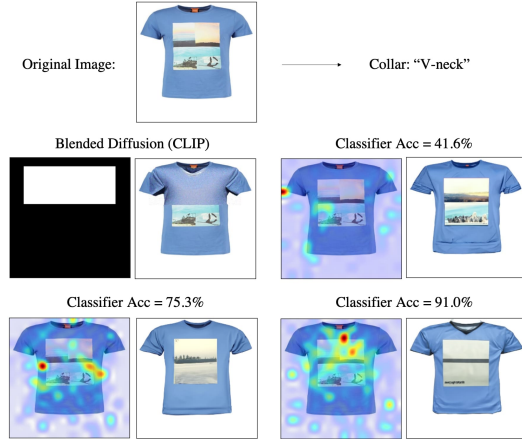
Figure 7. Classification accuracy and the guidance quality. The better classifier better guides the diffusion model.

Figure 8. Ablation on loss components. We set "v-neck collar" as the target attribute. λ refers to the weight multiplied to our classifier attribute guidance during the diffusion process.

guidance yields impressive outcomes.

Moreover, it is noteworthy how the diffusion model controls both the texture and the shape in a stable manner. The left side of Fig. 4 displays shape deformation, and the right side shows texture manipulations, where samples with convincing qualities are consistently delivered. The whole editing pipeline does not require any texture- or shape-specific modules, which is distinguished from the prior art [18], making our framework extremely simple and generic.

In Fig. 5, we present multi-attribute editing results, *i.e.,* manipulating multiple attributes at once. This task is apparently more challenging as it requires the editing of different attributes to be in harmony with one another. We observe that our framework produces reasonable outcomes across different attribute combinations, sometimes deforming the original image to a significant extent to meet the complex target condition.

As our method uses the attribute-wise classifier attention-map for background preservation, the ability to generate adequate spatial attention-maps is critical. Fig. 6 displays the classifier attention maps for different attribute editing operations, where we see that the classifier is capable of attending to the relevant regions with the learned attention-pooling layer. This component plays a vital role in removing the manual region masking that is necessary in *blended diffusion* [3], rendering our manipulation pipeline simple and compact.

4.6. Ablations

Classifier-guidance is the key component in our framework that injects conditioning information (target attribute) to the diffusion process. In Fig. 7, we illustrate how the capacity of the classifier affects the editing process. We set blended diffusion as the baseline, as CLIP guidance can be reg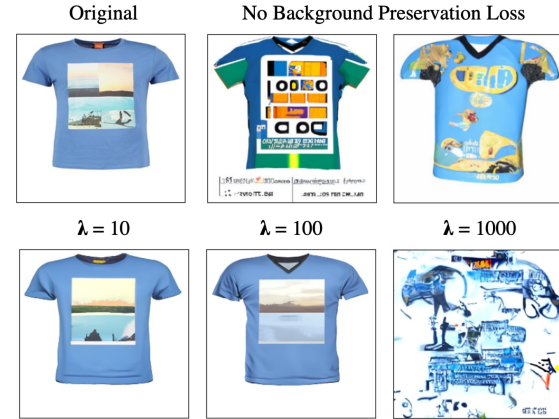arded as the coarsest level of guidance. Then we present three classifier variants with different classification accuracies. We observe that the classification performance clearly affects the spatial attention quality and the final editing performance, highlighting the importance of effective classifier finetuning.

Fig. 8 shows qualitative ablations on the core loss components and their hyperparameters. From the first row, we observe that background preservation loss is crucial for our framework, as there is no separate spatial mask that explicitly marks the region to be edited. Hence, without this term, the diffusion model generates random images that satisfy the given condition, *e.g.,* random pictures of v-neck t-shirts.

The second row illustrates how the guidance weight affects the sample quality. With small guidance scale (low λ value), we obtain realistic samples that are *not* aligned with the attribute condition. When the guidance scale is too big, the diffusion process yields unsatisfactory outputs, possibly due to the train-test mismatch pointed out by [30]. We find the sweet spot at around $\lambda = 100$, where a good balance between the fidelity and the alignment is achieved. Hence, all of our experiments are conducted in this setting.

5. Conclusion

In this paper, we have explored the prominent task of fashion attribute editing. As conventional approaches fall short in terms of scalability, we propose classifier-guided diffusion as a simple yet effective alternative. To that end, we train a multi-attribute classifier equipped with attention-pooling layer, and use its signal to guide the diffusion process. Empirical validations demonstrate the effectiveness of our framework as a generic attribute editing pipeline.

References

[1] Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim. Efficient multi-attribute similarity learning towards attribute-based fashion search. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1671–1679. IEEE, 2018. 5, 6

[2] Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim. Attribute manipulation generative adversarial networks for fashion images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10541–10550, 2019. 1, 2

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 1, 2, 3, 4, 8

[4] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 3

[5] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 5

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1, 2, 3, 5

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 4

[9] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE transactions on image processing*, 28(11):5464–5478, 2019. 2

[10] 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. 1, 2

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2, 3

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1, 2

[13] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 1

[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[15] Gwanghyun Kim and Jong Chul Ye. Diffusionclip: Text-guided image manipulation using diffusion models. 2021. 2

[16] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*, 2019. 2

[17] Chaerin Kong, Jeesoo Kim, Donghoon Han, and Nojun Kwak. Smoothing the generative latent space with mixup-based distance learning. *arXiv preprint arXiv:2111.11672*, 2021. 2

[18] Youngjoong Kwon, Stefano Petrangeli, Dahun Kim, Haoliang Wang, Viswanathan Swaminathan, and Henry Fuchs. Tailor me: An editing network for fashion attribute shape manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3831–3840, 2022. 1, 2, 8

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 3

[20] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 2

[21] Alex 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. *arXiv preprint arXiv:2112.10741*, 2021. 1, 2

[22] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2, 3

[23] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 2

[24] Qing Ping, Bing Wu, Wanying Ding, and Jiangbo Yuan. Fashion-attgan: Attribute-aware fashion editing with multi-objective gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 1, 2, 5

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3

[26] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image gen-

eration with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2

[27] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021. 2

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[30] 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. 1, 2, 3, 8

[31] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 4

[32] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2

[33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 2

[34] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021. 6

[35] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 2

[36] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2, 4

[37] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, Los Alamitos, CA, USA, jun 2018. IEEE Computer Society. 4

[38] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 4

[39] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2