AnyDressing: Customizable Multi-Garment Virtual Dressing via Latent Diffusion Models

Xinghui Li[1] Qichao Sun[1] Pengze Zhang[1] Fulong Ye[1] Zhichao Liao[2]
Wanquan Feng[1†] Songtao Zhao[1†] Qian He[1]
[1]Bytedance Intelligent Creation [2]Tsinghua University
https://crayon-shinchan.github.io/AnyDressing/

Figure 1. **Customizable virtual dressing results** of our AnyDressing. **Reliability:** AnyDressing is well-suited for a variety of scenes and complex garments. **Compatibility:** AnyDressing is compatible with LoRA [16] and plugins such as ControlNet [55] and FaceID [54].

Abstract

Recent advances in garment-centric image generation from text and image prompts based on diffusion models are impressive. However, existing methods lack support for various combinations of attire, and struggle to preserve the garment details while maintaining faithfulness to the text prompts, limiting their performance across diverse scenarios. In this paper, we focus on a new task, i.e., ***Multi-Garment Virtual Dressing****, and we propose a novel* ***AnyDressing*** *method for customizing characters conditioned on any combination of garments and any personalized text prompts. AnyDressing comprises two primary networks named GarmentsNet and DressingNet, which are respectively dedicated to extracting detailed clothing features and generating customized images. Specifically, we propose an efficient and scalable module called Garment-Specific Feature Extractor in GarmentsNet to individually encode garment textures in parallel. This design prevents garment confusion while ensuring network efficiency. Meanwhile,*

†Corresponding author.

we design an adaptive Dressing-Attention mechanism and a novel Instance-Level Garment Localization Learning strategy in DressingNet to accurately inject multi-garment features into their corresponding regions. This approach efficiently integrates multi-garment texture cues into generated images and further enhances text-image consistency. Additionally, we introduce a Garment-Enhanced Texture Learning strategy to improve the fine-grained texture details of garments. Thanks to our well-craft design, AnyDressing can serve as a plug-in module to easily integrate with any community control extensions for diffusion models, improving the diversity and controllability of synthesized images. Extensive experiments show that AnyDressing achieves state-of-the-art results.

1. Introduction

In recent years, the field of image generation has experienced transformative advancements [3, 9, 21], particularly with methods based on Latent Diffusion Models (LDMs) achieving remarkable success in text-to-image generation tasks [10, 15, 37, 38, 40, 41, 45]. Considering only textual information is inadequate for image customization, numerous approaches incorporate reference images with textual descriptions for image generation [24, 39, 49]. Specially, the Virtual Dressing (VD) task of generating garment-centric human images based on the reference garments has sparked considerable research interest [4, 42, 48], due to its significant potential in e-commerce and creative design.

VD is used to be regarded as a subtask of traditional subject-driven image customization, prior approaches [11, 19, 24, 30, 39, 43, 46, 57] simply integrate the features of reference image into the text embeddings without fully exploiting the information from the reference image. Several subsequent works [32, 54] more comprehensively utilize the features of the reference image by training additional cross-attention layers to integrate reference image features into the diffusion model. However, these methods struggle to preserve the intricate textures of the garment. Recently, some methods [4, 42, 48] focus on garment-centric image generation. Most of them leverage a full copy of diffusion U-Net as the garment encoder named ReferenceNet to maintain fine-grained garment information. DreamFit [29] proposes a lightweight encoder, which utilizes trainable LoRA layers to extract garment features instead of finetuning a full copy of the UNet. Nevertheless, these methods are tailored exclusively to single items of clothing and lack support for multiple conditions, thus hindering the ability to freely dress in any combination of various garments.

In this work, our focus is on a new task ***Multi-Garment Virtual Dressing***, personalizing a character wearing any combination of target garments according to the customized text prompt or other controls. The task poses several challenges, including: 1) Garment fidelity: preventing confusion among multiple garments while preserving the intricate textures of each; 2) Text-Image consistency: minimizing the influence of multiple garments on irrelevant regions to ensure the faithfulness of the generated images to the text prompts; 3) Plugin compatibility: enabling seamless integration with community control plugins for LDMs.

To address the aforementioned issues, we propose ***AnyDressing***, a novel approach that customizes characters conditioned on any combination of garments and any personalized text prompts. AnyDressing primarily comprises two primary networks named GarmentsNet and DressingNet. The GarmentsNet leverages a core Garment-Specific Feature Extractor (GFE) module to extract multi-garment detailed features, which utilizes parallelized self-attention layers within a shared U-Net architecture to individually encode garment textures. And we employ LoRA mechanism within the self-attention layers to further reduce the parameter increase associated with the added garments. The GFE module not only avoids clothing blending but also ensures network efficiency, allowing for easy scalability to any number of garments. The DressingNet employs a Dressing-Attention (DA) mechanism to seamlessly integrate multi-garment features into the denoising process. To ensure that each garment instance focuses specifically on its corresponding region, we further introduce a novel Instance-Level Garment Localization (IGL) learning strategy in DA. This avoids influencing other irrelevant regions in the synthetic image, thus improving fidelity to customized text prompts. Additionally, to enhance texture details, we design a Garment-Enhanced Texture Learning (GTL) strategy that strengthens the supervision of attire by imposing constraints from perceptual features and high-frequency information.

Extensive experiments show that AnyDressing has certain advantages in the quantitative and qualitative results compared to state-of-the-art methods. Especially, AnyDressing can serve as a plugin compatible with various fine-tuned LDMs, customized LoRAs [16], and other extensions such as ControlNet [55] and IP-Adapter [54], enhancing the diversity and controllability of synthetic images. In summary, our contributions are as follows:

- We propose a novel GarmentsNet to efficiently capture multi-garment textures in parallel by employing a core Garment-Specific Feature Extractor.
- We design a novel DressingNet incorporating a Dressing-Attention mechanism and an Instance-Level Garment Localization Learning strategy to accurately inject multi-garment features into their corresponding regions.
- We introduce a Garment-Enhanced Texture Learning strategy to effectively enhance the fine-grained texture details in synthetic images.
- Our framework can seamlessly integrate with any community control plugins for diffusion models. Both quan-

titative and qualitative experimental results demonstrate the superiority of our AnyDressing.

2. Related Work

Latent Diffusion Models. Latent Diffusion Models (LDMs) [33, 38] have become widely used in text-to-image generation tasks. Recent advancements have focused on making generated content more stable and controllable. For instance, ControlNet [55] and T2I Adapter [36] introduced additional conditioning modules injecting control into the denoising U-net via extra branches, such as edges and pose. Additionally, large model fine-tuning methods like LoRA [16] have significantly enhanced LDMs' generative capabilities in specific scenarios. In this work, we can integrate with various fine-tuned LDMs and customized LoRAs to enhance the diversity of generated images.

Subject-Driven Image Generation. Subject-driven generation aims to produce content that aligns with the visual features of a reference image. Methods for this task can be categorized into Tuning-based methods [11, 13, 24, 39] and Tuning-free methods [18, 22, 26, 28, 32, 50, 51, 54, 56]. Tuning-based methods, such as DreamBooth [39] and Custom-Diffusion [24] require optimizing specific text tokens to represent target concepts using a limited set of subject images. On the other hand, Tuning-free methods generally encode the reference image into feature embeddings. FastComposer [51] integrates image features into text embeddings, while IP-Adapter[54] and SSR-Encoder[56] integrate image features into the denoising U-net through a decoupled cross-attention mechanism. However, these methods struggle to preserve the fine-grained texture.

Virtual Try-On. Virtual Try-On (VTON) aims to synthesize an image of a specific person wearing a desired garment. Early methods [5, 14, 20, 25, 27, 34, 47, 52] utilize generative adversarial networks (GANs) with two-stage strategy, which rely on an explicit warping module and struggle to handle complex backgrounds. Recent studies [6, 12, 23, 35, 53] have used pre-trained LDMs as priors for VTON tasks. LADI-VTON [35] and DCI-VTON [12] explicitly deform the clothes and then use diffusion models to fuse and refine them. Rencent works [6, 23, 53] employ parallel U-Nets for clothing feature extraction and inject features into a denoising U-Net. However, VTON is essentially a localized image editing task and requires an existing model image, lacking flexibility in application scenarios.

Virtual Dressing. Virtual Dressing (VD) [4, 42, 48] aims to generate freely editable human images with reference garments and optional conditions. StableGarment [48] and IMAGDressing [42] leverage a garment U-Net for extracting fine-grained clothing features and a denoising U-Net with a hybrid attention module to incorporate garment features into denoising process. Magic Clothing [4] additionally proposes a joint classifier-free guidance to balance the control of garment features and text prompts. DreamFit [29] proposes a lightweight garment encoder based on trainable LoRA layers to streamline model complexity and memory usage. However, existing approaches are limited to processing single items of clothing, and difficult to maintain fidelity to text prompts. In contrast, our method allows for freely dressing multiple garments and produces coherent and attractive images following customized text prompts.

3. Preliminaries

Stable Diffusion. The Diffusion Model belongs to a class of generative models that generate data through iterative denoising from random noise. In this paper, we specifically employ Stable Diffusion [38]. Stable Diffusion is a latent diffusion model that operates in the latent space of an autoencoder $\mathcal{D}(\mathcal{E}(\cdot))$, where $\mathcal{E}$ and $\mathcal{D}$ represent the encoder and decoder, respectively. For a given image $\mathbf{x}_0$ with its corresponding latent feature $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$, the diffusion forward process is defined as:

$$\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{z}_0 + \sqrt{1-\alpha_t}\epsilon, \tag{1}$$

where $\alpha_t = \prod_{s=1}^{t}(1-\beta_s)$, $\epsilon \sim \mathcal{N}(0,1)$, and β_s is the pre-defined variance schedule at timestep s.

In the diffusion backward process, a U-Net ϵ_θ is trained to predict the noise. Given the textual condition $\mathbf{C}$, the training objective $\mathcal{L}_{LDM}$ is defined as follows:

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathbf{z}_0,\epsilon,\mathbf{C},t}\|\epsilon - \epsilon_\theta(\mathbf{z}_t, \mathbf{C}, t)\|_2. \tag{2}$$

4. Methodology

Given N target garments, the proposed AnyDressing aims to generate a new image x_{dr}, showcasing a customized character dressed in multiple target garments across various scenes, styles and actions based on the text prompt. As illustrated in Fig. 2, AnyDressing comprises two primary networks: GarmentsNet and DressingNet. The GarmentsNet leverages the GFE module to extract detailed features from multiple garments (Sec. 4.1). Meanwhile, the DressingNet integrates these features for virtual dressing using a DA module and an IGL learning mechanism (Sec. 4.2). Additionally, a GTL strategy is designed further to enhance crucial texture details in the synthesis images (Sec. 4.3). Next, we will introduce the aforementioned modules, along with training and inference processes (Sec. 4.4), in detail.

4.1. GarmentsNet

Previous methods [4, 42, 48] leverage a full copy of diffusion U-Net [2, 17] as garment encoding network, ensuring precise preservation of clothing details. However, these methods are limited to handling a single garment and face

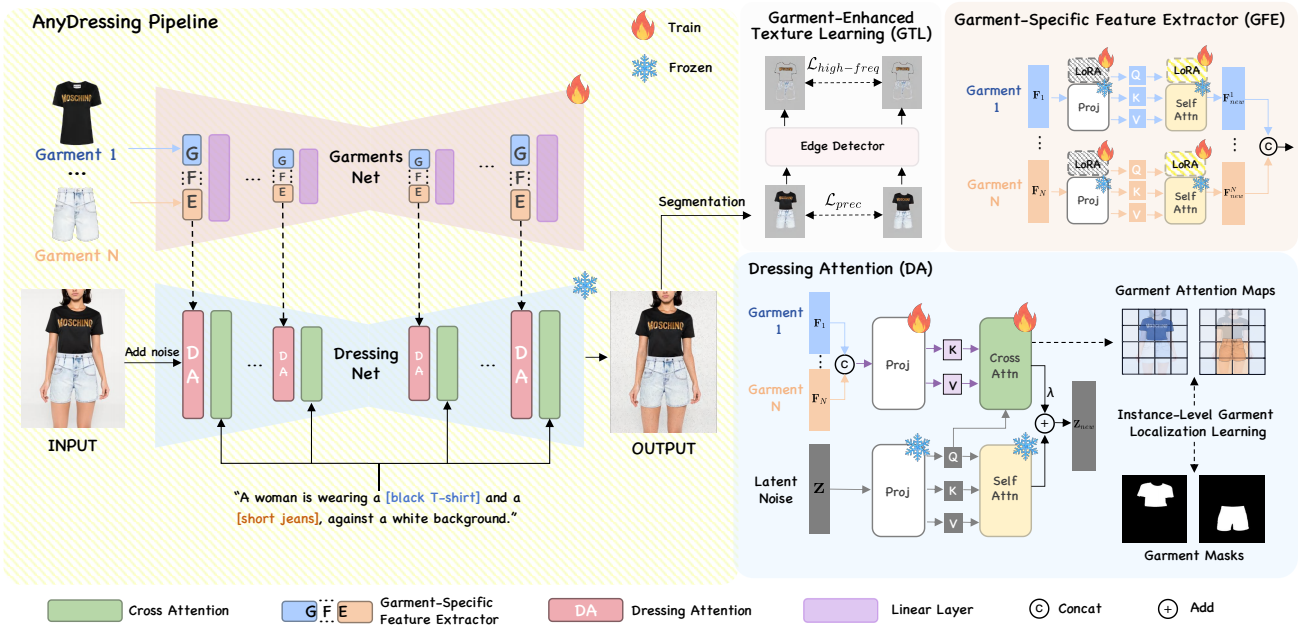

Figure 2. **Overview of AnyDressing.** Given N target garments, AnyDressing customizes a character dressed in multiple target garments. The GarmentsNet leverages the Garment-Specific Feature Extractor (GFE) module to extract detailed features from multiple garments. The DressingNet integrates these features for virtual dressing using a Dressing-Attention (DA) module and an Instance-Level Garment Localization Learning mechanism. Moreover, the Garment-Enhanced Texture Learning (GTL) strategy further enhances texture details.

significant garment confusion issues when applied to multi-garment virtual dressing, as shown in Fig. 3. A straightforward approach to dress multiple garments is to simply duplicate several garment encoding networks to manage different conditions. However, this method would result in a substantial increase in the number of parameters, making it computationally impractical.

Drawing inspiration from the successful practice of the aforementioned reference mechanisms, we observe that self-attention layers are crucial for the implicit warping of features, enabling the effective matching of input garments to the appropriate body parts. Meanwhile, other layers are typically responsible for general feature extraction and can be shared across different garments without compromising the model's performance. Building on this insight, we innovatively design a simple yet effective architecture named GarmentsNet, which employs a core GFE module to encode features for each garment utilizing individual self-attention layers within a shared U-Net framework. Inspired by [29], we integrate LoRA [16] mechanism into self-attention layers, minimizing the increase in parameters associated with the added garments. As a result, this design significantly avoids garment blending while maintaining network efficiency. As illustrated in Fig. 2, the GFE module employs a parallelized self-attention mechanism to extract detailed features of multiple garments. Specifically, for each garment condition $\mathbf{F}_i$, we define the proprietary self-attention LoRA matrix $\triangle\hat{\mathbf{W}}_i = \{\triangle\hat{\mathbf{W}}_q^i, \triangle\hat{\mathbf{W}}_k^i, \triangle\hat{\mathbf{W}}_v^i\}$, where $\triangle\hat{\mathbf{W}}_q^i$, $\triangle\hat{\mathbf{W}}_k^i$ and $\triangle\hat{\mathbf{W}}_v^i$ represent LoRA layers for the query, key and value projections of self-attention layers. We then concatenate self-attention results of each garment condition to obtain the aggregated garment features $\mathbf{F}_{new}$:

$$\mathbf{F}_{new}^i = Softmax(\frac{\mathbf{Q}_i(\mathbf{K}_i)^\top}{\sqrt{d}})\mathbf{V}_i, \tag{3}$$

$$\mathbf{F}_{new} = Concat(\mathbf{F}_{new}^1, \mathbf{F}_{new}^2, \cdots, \mathbf{F}_{new}^N), \tag{4}$$

where $\mathbf{Q}_i = \mathbf{F}_i(\hat{\mathbf{W}}_q + \triangle\hat{\mathbf{W}}_q^i)$, $\mathbf{K}_i = \mathbf{F}_i(\hat{\mathbf{W}}_k + \triangle\hat{\mathbf{W}}_k^i)$, $\mathbf{V}_i = \mathbf{F}_i(\hat{\mathbf{W}}_v + \triangle\hat{\mathbf{W}}_v^i)$, only $\triangle\hat{\mathbf{W}}$ is trainable and N represents the number of reference garments.

Thanks to the multi-garment parallel processing design of our GFE module, GarmentsNet can seamlessly scale to any number of garments. Notably, this expansion requires only some newly added LoRA matrix $\triangle\hat{\mathbf{W}}$, and significantly minimizes both training and inference time compared with duplicating the entire garment encoding network. Considering the capability of GFE module to individually encode each garment, we excise the cross-attention modules in GarmentsNet to further reduce redundancy.

4.2. DressingNet

To incorporate multi-garment features during the diffusion process, we meticulously design the DressingNet, which serves as the main denoising net and primarily includes an adaptive Dressing-Attention mechanism and an Instance-Level Garment Localization Learning strategy.

4.2.1 Adaptive Dressing-Attention

In the VD task, the main denoising network is typically kept frozen during training [4, 42] to preserve its original editing and generation capabilities as much as possible. To incorporate reference garment features into latent features, we design an adaptive Dressing-Attention (DA) mechanism to efficiently integrate multi-garment texture cues into synthetic images, inspired by [54]. As shown in Fig. 2, the Dressing-Attention module includes a frozen self-attention module and a learnable cross-attention module. Let $\{\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_N\}$ denote N garment features output by the GarmentsNet at corresponding positions, we first concatenate these features along the spatial dimension to obtain the final garment features: $\mathbf{F}_{all} = Concat(\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_N)$. We then introduce two trainable linear projection layers $\mathbf{W}'_k$ and $\mathbf{W}'_v$ to align garment features with latent feature $\mathbf{Z}$. Formally, the output of Dressing-Attention $\mathbf{Z}_{new}$ is:

$$\mathbf{Z}_{new} = Softmax(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}})\mathbf{V} + \lambda * Softmax(\frac{\mathbf{Q}(\mathbf{K}')^\top}{\sqrt{d}})\mathbf{V}' \tag{5}$$

where λ is a hyperparameter ensuring the flexibility of incorporating garment features, and $\mathbf{Q} = \mathbf{Z}\mathbf{W}_q$, $\mathbf{K} = \mathbf{Z}\mathbf{W}_k$, $\mathbf{V} = \mathbf{Z}\mathbf{W}_v$, $\mathbf{K}' = \mathbf{F}_{all}\mathbf{W}'_k$, $\mathbf{V}' = \mathbf{F}_{all}\mathbf{W}'_v$. Here, $\mathbf{W}_q$, $\mathbf{W}_k$ and $\mathbf{W}_v$ are frozen self-attention layers. To accelerate the coverage, we initialize the $\mathbf{W}'_k$, $\mathbf{W}'_v$ with $\mathbf{W}_k$, $\mathbf{W}_v$.

4.2.2 Instance-Level Garment Localization Learning

Although the above Dressing-Attention (DA) mechanism facilitates the integration of multi-garment features, it may result in text-image inconsistency. We argue that this results from the garment's attention map covering the entire image in the DA module, thereby injecting garment cues into the other irrelevant regions incorrectly. To tackle this issue, we introduce an Instance-Level Garment Localization (IGL) learning strategy, ensuring that each garment instance focuses solely on its corresponding region. Specifically, for each garment feature, we obtain its attention map A with the latent noise in each layer of the DA module:

$$P = Softmax(\mathbf{Q}(\mathbf{K}')^\top / \sqrt{d}), \tag{6}$$

$$A = \sum_{j=1}^{L} P_j, \tag{7}$$

where L denotes the length of corresponding garment features. Then, a regularization term L_{loc} is applied to explicitly learn attention localization for each garment instance:

$$\mathcal{L}_{loc} = \frac{1}{N}\sum_{k=1}^{N} \|A_k - M_k\|_2, \tag{8}$$

where N is the number of garments in the reference image, and M_k represents the reference garment's segmentation mask. It is worth noting that the proposed IGL learning strategy is applied exclusively during the training phase and does not introduce any additional cost during inference.

4.3. Garment-Enhanced Texture Learning

Generally, diffusion models are merely optimized relying on the mean-squared loss defined in Eqn. 2, which treats all regions of the synthetic image equally, resulting in a struggle to maintain garment consistency, especially in cases of small text and intricate patterns. To synthesize fine-grained textures, we design a Garment-Enhanced Texture Learning (GTL) strategy to strengthen the supervision of attire details in image space, incorporating a perceptual loss $\mathcal{L}_{perc}$ and a high-frequency loss $\mathcal{L}_{high-freq}$.

Before introducing the proposed two losses, we define the generated image as: $\hat{I} = \mathcal{D}(\hat{\mathbf{z}}_0)$, where $\mathcal{D}$ denotes the VAE decoder, and $\hat{\mathbf{z}}_0$ is estimated through a single step of inference from the latent $\mathbf{z}_t$:

$$\hat{\mathbf{z}}_0 = \frac{\mathbf{z}_t - \sqrt{1-\alpha_t}\epsilon_\theta}{\sqrt{\alpha_t}}. \tag{9}$$

Considering the one-step inference may produce noisy and flawed images, the proposed losses are only applied at less noisy timestep ($t \leq \eta$). To sum up, GTL can be defined as:

$$\mathcal{L}_{texture} = \begin{cases} \mathcal{L}_{perc} + \mathcal{L}_{high-freq}, & t \leq \eta \\ 0, & t > \eta \end{cases}. \tag{10}$$

Perception Loss To simultaneously enhance structural consistency and pattern similarity with reference garments, we employ a perceptual loss based on the Deep Image Structure and Texture Similarity (DISTS) metric [7]. Specifically, we use the reference garment's segmentation mask to isolate the attire in both the generated and ground truth images, averaging their structural and textural inconsistencies within a perceptual feature space, defined as:

$$\mathcal{L}_{prec} = \frac{1}{N}\sum_{k=1}^{N} \mathcal{DISTS}(\hat{I} \odot M_k, I \odot M_k), \tag{11}$$

where $\odot$ signifies element-wise multiplication.

High-Frequency Loss As intricate details in the dressing garments typically appear as high-frequency components with rich edge information, we use edge detection to extract this high-frequency information, aiming to strengthen the constraints on detailed patterns. Specifically, we utilize Canny operator [8] to capture these rich-texture regions, the high-frequency loss $\mathcal{L}_{high-freq}$ is defined as:

$$\mathcal{L}_{high-freq} = \frac{1}{N}\sum_{k=1}^{N} \|\hat{I} \odot M'_k - I \odot M'_k\|_2, \tag{12}$$

where $M'_k = M_k \odot P$, P is the extracted edge map of I.

Method	Single Grament								Multiple Graments			
	VITON-HD [5]				Proprietary Dataset				Dressing-Pair			
	CLIP-T ↑	CLIP-I ↑	CLIP-AS ↑	DINO ↑	CLIP-T ↑	CLIP-I ↑	CLIP-AS ↑	DINO ↑	CLIP-T ↑	CLIP-I* ↑	CLIP-AS ↑	DINO* ↑
IP-Adapter [54]	0.268	0.644	5.674	0.500	0.272	0.632	5.678	0.460	0.277	0.523	5.795	0.350
StableGarment [48]	0.285	0.583	5.781	0.522	0.281	0.587	5.648	0.510	0.284	0.556	5.735	0.412
MagicClothing [4]	0.288	0.640	5.703	0.363	0.298	0.619	5.784	0.340	0.266	0.583	5.540	0.290
IMAGDressing [42]	0.202	0.734	5.077	0.553	0.230	0.684	5.133	0.453	0.242	0.614	5.291	0.378
Ours	0.289	0.741	5.881	0.571	0.296	0.710	5.931	0.559	0.296	0.734	5.874	0.674

Table 1. **Quantitative comparisons** with baseline methods for both single-garment and multi-garment evaluation.

4.4. Training and Inference

In training, we average $\mathcal{L}_{loc}$ across all m layers and define overall loss $\mathcal{L}$ as follows:

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathbf{z}_0,\epsilon,\mathbf{C}_t,\mathbf{C}_g,t}\|\epsilon - \epsilon_\theta(\mathbf{z}_t, \mathbf{C}_t, \mathbf{C}_g, t)\|_2, \quad (13)$$

$$\mathcal{L} = \mathcal{L}_{LDM} + \frac{\lambda_1}{m}\mathcal{L}_{loc} + \lambda_2\mathcal{L}_{texture}, \quad (14)$$

where $\mathbf{C}_t$ and $\mathbf{C}_g$ represent text condition and clothing condition respectively. In the inference stage, we apply classifier-free guidance during the denoising process:

$$\hat{\epsilon}_\theta(\mathbf{z}_t, \mathbf{C}_t, \mathbf{C}_g, t) = \omega\epsilon_\theta(\mathbf{z}_t, \mathbf{C}_t, \mathbf{C}_g, t) + (1-\omega)\epsilon_\theta(\mathbf{z}_t, t). \quad (15)$$

5. Experiments

5.1. Setup

Dataset. Notably, a dataset comprising image triplets that include model images paired with multiple laid-out garments is currently lacking. Therefore, we utilize a HumanParsing model to extract clothing items from DressCode [34] and an additional proprietary dataset collected from the internet, forming triplet data pairs (upper garment, lower garment, person image). In these triplets, one garment is an original laid-out image, while the other is a segmented image from the person's image. Finally, we construct 26,114 public triplets from Dresscode and 37,065 triplets from proprietary dataset to train AnyDressing. For model evaluation, we introduce two benchmarks to evaluate the model on single-garment and multi-garment dressing respectively. Specifically, for single-garment evaluation, we select 300 reference garments from VITON-HD [5] encompassing various styles and colors, and additionally collect 300 diverse garments with intricate textures from the internet. For multi-garment evaluation, we meticulously gather 25 lowers from the internet and pair each with 10 different uppers, resulting in 250 pairs named Dressing-Pair.

Implementation Details. In our experiments, we initialize the weights of GarmentsNet and DressingNet with the weights of the U-Net in Stable Diffusion v1.5 [38]. Our model is trained on paired images at the resolution of 768×576. The trainable parameters are GarmentsNet and the cross-attention layers in Dressing-Attention module. During training, We adopt AdamW [31] optimizer with a fixed learning rate of 5e-5. The model is trained for 100k steps on 8 NVIDIA A100 GPUs with a batch size of 4. During inference, we use DDIM [44] sampler with 30 steps and set guidance scale ω to 6.0. Please refer to the supplementary materials for more details.

Method	Texture Consistency ↑	Align with Prompt ↑	Image Quality ↑	Comprehensive Evaluation ↑
IP-Adapter [54]	0.45%	6.65%	11.95%	2.20%
StableGarment [48]	1.60%	4.85%	2.65%	2.05%
MagicClothing [4]	2.05%	9.00%	9.70%	3.75%
IMAGDressing [42]	2.10%	2.50%	3.90%	1.70%
Ours	93.80%	77.00%	71.80%	90.30%

Table 2. **User study** with baseline methods.

Baselines. We compare our method against the following state-of-the-art image synthesis method: IP-Adapter [54], MagicClothing [4], StableGarment [48] and IMAGDressing [42]. We use the official model parameters from their official implementations. For a fair comparison, all experiments are conducted with the resolution of 768×576.

Evaluation Metrics. We follow previous methods to adopt three metrics for evaluation: CLIP-T for text-image similarity, CLIP-I and DINO for garment consistency, and CLIP Aesthetic Score (CLIP-AS) for overall generation quality. Especially, to better evaluate multi-garment dressing, we introduce new metrics CLIP-I* and DINO* to assess texture consistency by leveraging OpenPose [1] to obtain the matching partitions of the reference garments in the synthesized image and averaging their corresponding metrics.

5.2. Qualitative Analysis

Since the compared methods lack multi-garment support, we obtain baseline results by concatenating multiple garments along the spatial dimension as input. Fig. 3 presents visual comparisons between our method and baseline approaches. AnyDressing maintains superior consistency in clothing style and texture, and exhibits better text fidelity, while other methods struggle to balance garment preservation and prompt faithfulness. In particular, baselines encounter significant background contamination and garment confusion in multi-garment dressing results, whereas our method demonstrates exceptional reliability, which is attributed to our designed GarmentsNet and DressingNet architectures. And Fig. 4 presents the results of AnyDressing as a plug-in module combined with other extensions and

Figure 3. **Qualitative comparisons** with state-of-the-art methods. Please zoom in for more details.

customized LoRAs, demonstrating its powerful compatibility. Please refer to the supplementary for more results.

5.3. Quantitative Comparisons

Metric Evaluation. Tab. 1 shows the quantitative results of our methods against baselines. For single-garment evaluation, extensive experiments conducted on VITON-HD [5] and proprietary dataset prove the superiority of AnyDressing compared with all baselines. And our method significantly surpasses all baselines across all metrics in multi-garment virtual dressing results, fully demonstrating AnyDressing's reliability in handling both single-garment and multi-garment virtual dressing tasks.

User Study. We conduct a user study to evaluate the generation quality of our model. We use all test garments and prompts in our dataset and randomly show the users 25 single-garment results and 25 multi-garment results from the baselines and our method. Each participant is asked to select the ***most*** preferred result under four criteria: texture consistency, alignment with the text prompt, image quality and comprehensive evaluation. In the end, we receive valid responses from 40 users. The collected preferences are reported in Tab. 2. In terms of four criteria, our method is preferred by most participants, with percentages reaching

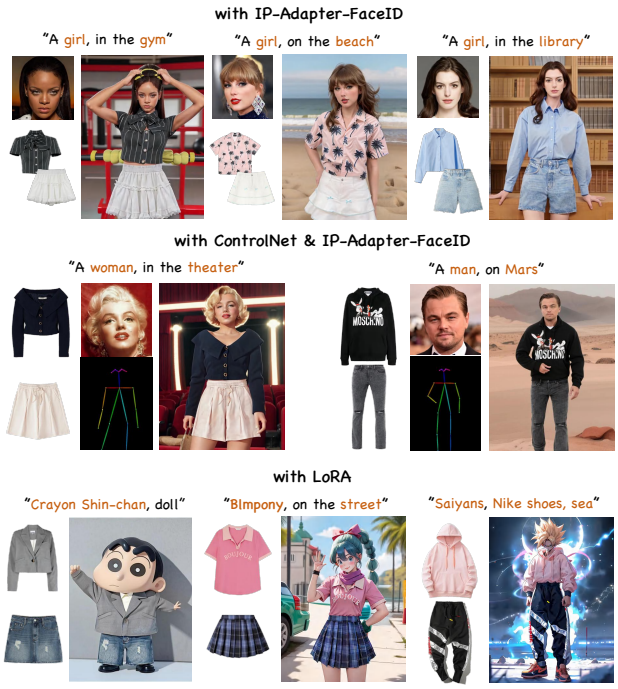

Figure 4. **Examples of plug-in** results of AnyDressing.

GFE	IGL	GTL	CLIP-T ↑	CLIP-I* ↑	CLIP-AS ↑
✗	✗	✗	0.260	0.625	5.572
✓	✗	✗	0.265	0.718	5.627
✓	✓	✗	0.289	0.722	5.790
✓	✓	✓	**0.296**	**0.734**	**5.874**

Table 3. **Ablation study** of AnyDressing.

93.80%, 77.00%, 71.80% and 90.30% respectively.

5.4. Ablation Studies

GFE & IGL. To validate the effectiveness of our proposed architecture, we employ traditional ReferenceNet [17] to encode multiple garments concurrently and then incorporate them into the denoising U-Net similar to [42] as our base model. As illustrated in Fig. 5, Base+GFE significantly reduces garment confusion and improves garment consistency compared to Base, which is attributed to the multi-garment parallel processing design of the GFE module. Base+GFE+IGL shows better fidelity to the text prompts and further mitigates background contamination, which demonstrates IGL mechanism effectively constrains garment features to attend to the correct regions. The quantitative comparison in Tab. 3 further proves the effectiveness of each module, with GFE primarily improving the CLIP-I* and IGL enhancing both CLIP-T and CLIP-AS.

GTL. Fig. 6 intuitively demonstrates the effectiveness of our proposed GTL strategy, encouraging the model to enhance detail preservation, particularly in small text and intricate patterns. And quantitative result in Tab. 3 also verifies that our designed GTL improves texture consistency.

Figure 5. **Ablation results** on GFE and IGL modules.

Figure 6. **Ablation results** on GTL module.

6. Conclusion

This paper presents AnyDressing comprising two core networks named GarmentsNet and DressingNet to focus on a new task, i.e., Multi-Garment Virtual Dressing. The GarmentsNet employs a Garment-Specific Feature Extractor module to efficiently encode multi-garment features in parallel. The DressingNet integrates these features for virtual dressing using a Dressing-Attention module and an Instance-Level Garment Localization Learning mechanism. Additionally, we design a Garment-Enhanced Texture Learning strategy to further enhance texture details. Our approach can seamlessly integrate with any community control plugins. Extensive experiments show that AnyDressing achieves state-of-the-art results.

References

[1] Z Cao, G Hidalgo, T Simon, SE Wei, and Y Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2020. 6

[2] Di Chang, Yichun Shi, Quankai Gao, Jessica Fu, Hongyi Xu, Guoxian Song, Qing Yan, Xiao Yang, and Mohammad Soleymani. Magicdance: Realistic human dance video generation with motions & facial expressions transfer. *arXiv preprint arXiv:2311.12052*, 2023. 3

[3] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2

[4] Weifeng Chen, Tao Gu, Yuhao Xu, and Chengcai Chen. Magic clothing: Controllable garment-driven image synthesis. *arXiv preprint arXiv:2404.09512*, 2024. 2, 3, 5, 6

[5] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14131–14140, 2021. 3, 6, 7

[6] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for virtual try-on. *arXiv preprint arXiv:2403.05139*, 2024. 3

[7] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 5

[8] Lijun Ding and Ardeshir Goshtasby. On the canny edge detector. *Pattern recognition*, 34(3):721–725, 2001. 5

[9] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835, 2021. 2

[10] Kunyu Feng, Yue Ma, Bingyuan Wang, Chenyang Qi, Haozhe Chen, Qifeng Chen, and Zeyu Wang. Dit4edit: Diffusion transformer for image editing. *arXiv preprint arXiv:2411.03286*, 2024. 2

[11] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2, 3

[12] Junhong Gou, Siyu Sun, Jianfu Zhang, Jianlou Si, Chen Qian, and Liqing Zhang. 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*, pages 7599–7607, 2023. 3

[13] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[14] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3470–3479, 2022. 3

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 1, 2, 3, 4

[17] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 3, 8

[18] Qihan Huang, Siming Fu, Jinlong Liu, Hao Jiang, Yipeng Yu, and Jie Song. Resolving multi-condition confusion for finetuning-free personalized image generation. *arXiv preprint arXiv:2409.17920*, 2024. 3

[19] Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin CK Chan, and Ziwei Liu. Reversion: Diffusion-based relation inversion from images. *arXiv preprint arXiv:2303.13495*, 2023. 2

[20] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzenes. Do not mask what you do not need to mask: a parser-free virtual try-on. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 619–635. Springer, 2020. 3

[21] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134, 2023. 2

[22] Chanran Kim, Jeongin Lee, Shichang Joung, Bongmo Kim, and Yeul-Min Baek. Instantfamily: Masked attention for zero-shot multi-id image generation. *arXiv preprint arXiv:2404.19427*, 2024. 3

[23] Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. 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*, pages 8176–8185, 2024. 3

[24] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 2, 3

[25] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European Conference on Computer Vision*, pages 204–219. Springer, 2022. 3

[26] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[27] Kedan Li, Min Jin Chong, Jeffrey Zhang, and Jingen Liu. Toward accurate and realistic outfits visualization with attention to details. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15546–15555, 2021. 3

[28] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8640–8650, 2024. 3

[29] Ente Lin, Xujie Zhang, Fuwei Zhao, Yuxuan Luo, Xin Dong, Long Zeng, and Xiaodan Liang. Dreamfit: Garment-centric human generation via a lightweight anything-dressing encoder. *arXiv preprint arXiv:2412.17644*, 2024. 2, 3, 4

[30] Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: Customizable image synthesis with multiple subjects. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 57500–57519, 2023. 2

[31] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[32] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 2, 3

[33] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. *arXiv preprint arXiv:2406.01900*, 2024. 3

[34] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2231–2235, 2022. 3, 6

[35] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8580–8589, 2023. 3

[36] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 3

[37] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[38] 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, 3, 6

[39] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2, 3

[40] Fei Shen and Jinhui Tang. Imagpose: A unified conditional framework for pose-guided person generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2

[41] Fei Shen, Hu Ye, Jun Zhang, Cong Wang, Xiao Han, and Wei Yang. Advancing pose-guided image synthesis with progressive conditional diffusion models. *arXiv preprint arXiv:2310.06313*, 2023. 2

[42] Fei Shen, Xin Jiang, Xin He, Hu Ye, Cong Wang, Xiaoyu Du, Zechao Li, and Jinghui Tang. Imagdressing-v1: Customizable virtual dressing. *arXiv preprint arXiv:2407.12705*, 2024. 2, 3, 5, 6, 8

[43] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8543–8552, 2024. 2

[44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 6

[45] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[46] Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. Concept decomposition for visual exploration and inspiration. *ACM Transactions on Graphics (TOG)*, 42(6): 1–13, 2023. 2

[47] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 589–604, 2018. 3

[48] Rui Wang, Hailong Guo, Jiaming Liu, Huaxia Li, Haibo Zhao, Xu Tang, Yao Hu, Hao Tang, and Peipei Li. Stablegarment: Garment-centric generation via stable diffusion. *arXiv preprint arXiv:2403.10783*, 2024. 2, 3, 6

[49] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953, 2023. 2

[50] Zhichao Wei, Qingkun Su, Long Qin, and Weizhi Wang. Mm-diff: High-fidelity image personalization via multi-modal condition integration. *arXiv preprint arXiv:2403.15059*, 2024. 3

[51] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, pages 1–20, 2024. 3

[52] Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. Gp-

vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23550–23559, 2023. 3

[53] Y Xu, T Gu, W Chen, and C Chen. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. arxiv 2024. *arXiv preprint arXiv:2403.01779*, 2024. 3

[54] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1, 2, 3, 5, 6

[55] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1, 2, 3

[56] Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, et al. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8069–8078, 2024. 3

[57] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6027–6037, 2023. 2