Controllable Human Image Generation with Personalized Multi-Garments

Yisol Choi[1] Sangkyung Kwak[1,3] Sihyun Yu[1] Hyungwon Choi[2] Jinwoo Shin[1]
[1]KAIST [2]OMNIOUS.AI [3]Scaled Foundations
{yisol.choi, skkwak9806, sihyun.yu, jinwoos}@kaist.ac.kr, hyungwon.choi@omnious.com

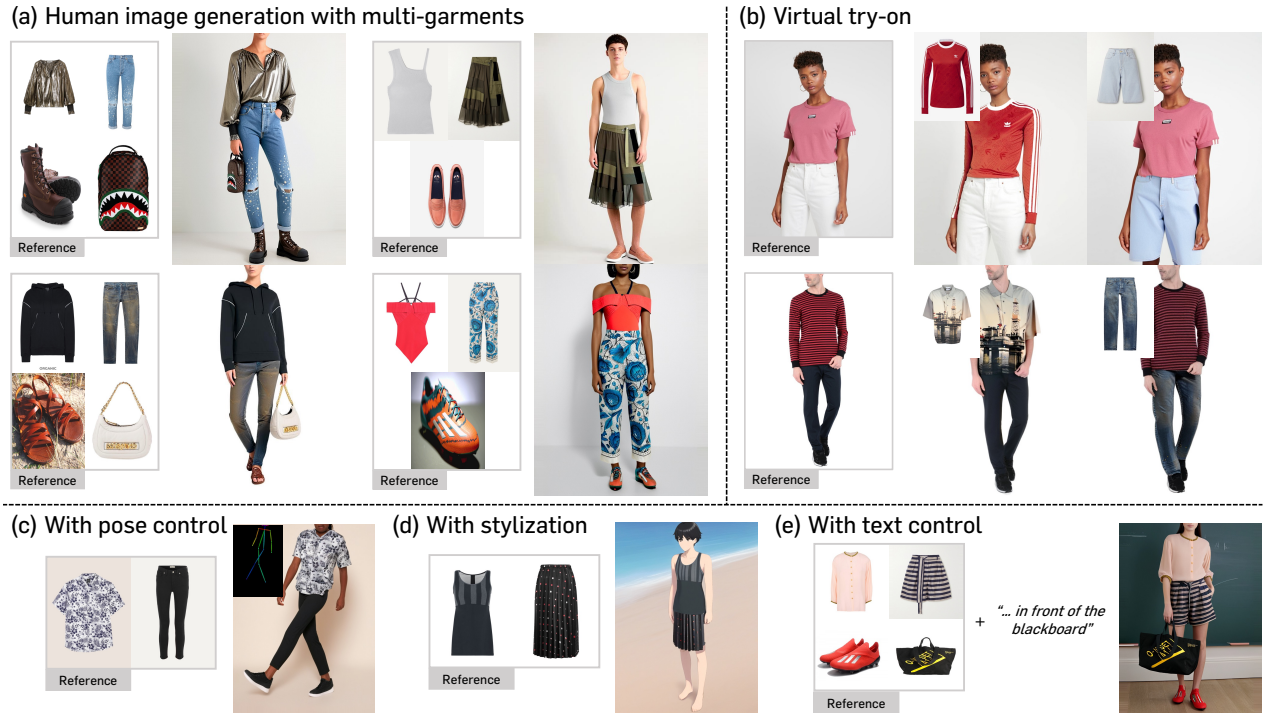

Figure 1. **Generated images by BootComp.** (a) BootComp generates high-quality human images wearing multiple reference garments, with support for extended categories such as bag, shoes, even in unusual garment combinations (*e.g.*, swimming suit with soccer cleats). We show BootComp's generalization capability through various conditional image generations, such as (b) virtual try-on, (c) pose guided generation, (d) stylization, and (e) text guided generation, even though BootComp is not directly trained or fine-tuned for each task.

Abstract

We present BootComp, a novel framework based on text-to-image diffusion models for controllable human image generation with multiple reference garments. Here, the main bottleneck is data acquisition for training: collecting a large-scale dataset of high-quality reference garment images per human subject is quite challenging, i.e., ideally, one needs to manually gather every single garment photograph worn by each human. To address this, we propose a data generation pipeline to construct a large synthetic dataset, consisting of human and multiple-garment pairs, by introducing a model to extract any reference garment images from each human image. To ensure data quality, we also propose a filtering strategy to remove undesirable generated data based on measuring perceptual similarities between the garment presented in human image and extracted garment. Finally, by utilizing the constructed synthetic dataset, we train a diffusion model having two parallel denoising paths that use multiple garment images as conditions to generate human images while preserving their fine-grained details. We further show the wide-applicability of our framework by adapting it to different types of reference-based generation in the fashion domain, including virtual try-on, and controllable human image generation with other conditions, e.g., pose, face, etc.

Project page: https://omnious.github.io/BootComp

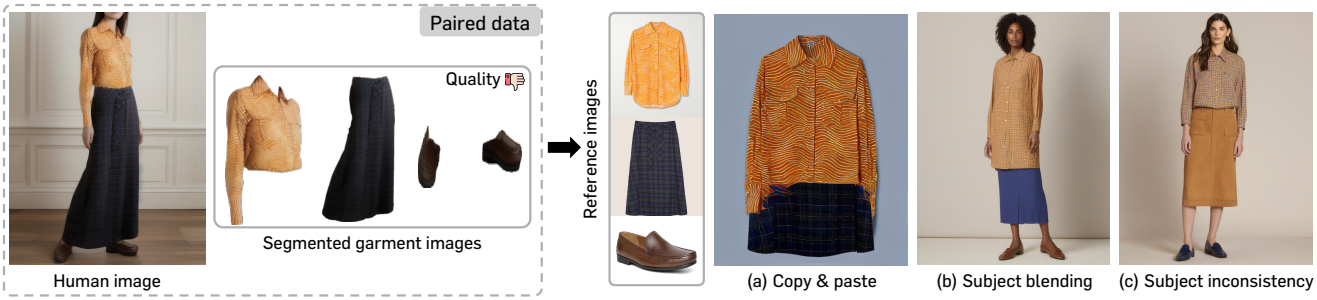

Figure 2. **Limitations of previous data curation approaches used in controllable generation.** Previous approaches on controllable generation often use a paired dataset consisting of low-quality segmented garments and human images for training. It leads to several undesirable artifacts as shown in right (generated with baselines). For example, garments are directly replicated from the reference images in (a), shirts and skirts are blended together in (b), and generated skirts fail to resemble the reference in (c).

1. Introduction

Recent advances in text-to-image (T2I) diffusion models [7, 37, 40] have shown great progress in numerous challenging real-world scenarios, such as personalized generation [23, 41], style transfer [11, 47], image editing [1, 10, 29], and compositional image generation [32, 35, 46]. These remarkable successes have provided great potential to aid users in a variety of creative pursuits [22].

Among them, *controllable human image generation* using T2I diffusion models [16] can provide lots of intriguing use cases in real-world scenarios. Specifically, by training a model capable of creating human images conditioned on a variety of garments, one can enjoy diverse applications such as outfit recommendations for users, generating fashion models for clothing brands, or virtual try-on [6, 21, 31], through a *single unified framework*.

One can consider fine-tuning T2I models and image encoders [34, 38] using curated paired image datasets that consist of condition garments and the target human images [16]. However, hand-collecting multiple garment photographs worn by human is labor-intensive. Prior works [15, 35, 51] have attempted to obtain the pair images by extracting all reference objects from real images, segmenting out each object from the original images. However, this data curation protocol makes curated garments have exactly the same shape with their appearance in the target human image. Thus, generated images are likely to suffer from copy-and-paste problem: they easily generate exactly the same image in generated samples without altering pose or appearance (see (a) in Fig. 2). To mitigate this issue, several works propose to curate data from videos by doing segmentation from different video frames that contain the same objects [4, 46]. However, collecting such paired datasets in large amounts is challenging and often results in low-quality reference images; thereby, the trained model fails to generalize and suffers from subject blending or inconsistency within the images [46] (see (b), (c) in Fig. 2). Such drawbacks become more critical in practical scenarios related to human image generation, as the model must generate human images with diverse poses while accurately preserving the details of each garment.

Contribution. We address the aforementioned shortcomings by presenting *Bootstrapping paired data for Compositional and controllable human image generation* (BootComp), a novel framework for controllable human image generation using T2I diffusion models. Specifically, it is a two-stage framework (see Fig. 3 for illustration):

- *Synthetic data generation*: We first propose a high-quality synthetic data generation pipeline for training controllable human image generation model. We achieve this by introducing a decomposition module, which is a mapping from a single garment worn by a human to a product view of the garment image. We train this model with a paired dataset of *single* reference garment and human image (*e.g.*, shirts and human wearing those shirts), which is easy to collect [5, 24, 30]. Using this model, we bootstrap synthetic paired data at scale from a large number of human images; thus, each pair consists of a human image and all garment images that the human is wearing. To ensure high-quality data, we also present a filtering strategy that further improves the data quality based on measuring the perceptual similarities between the original segmentation results and the data generated from the decomposition module.
- *Composition module*: We also present a fine-tuning scheme of T2I diffusion models for our goal using the synthetic dataset. We use two T2I diffusion models: one serves as an image encoder to extract garment features and the other one functions as a generator to create human images. We only train the encoder model, employing an extended self-attention mechanism to generator for conditioning garment images. Since we keep the generator frozen during the training, BootComp can be attached to various adapter modules or replaced with pre-trained models specialized to generate images with different styles. This enables BootComp to provide various applications (*e.g.*, pose-guided or cartoon-style generation) for free without requiring any additional fine-tuning.

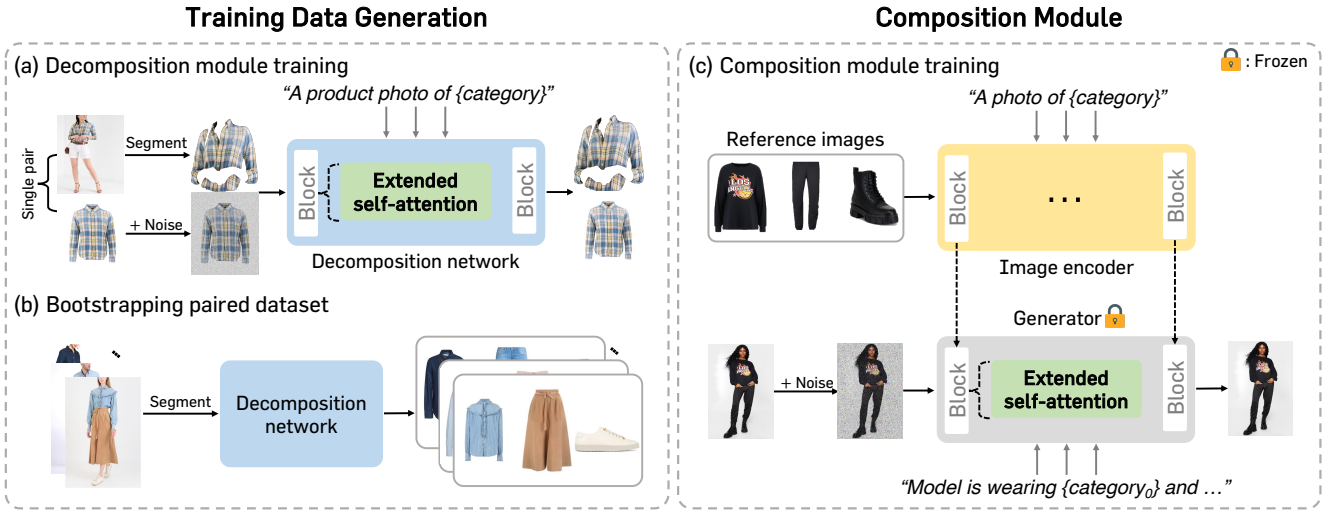

Figure 3. **Overview of BootComp.** We propose a two-stage framework: synthetic data generation and composition module training for controllable human image generation. (a) We train a decomposition network that maps from a segmented garment image to a product garment image. (b) We bootstrap synthetic paired data of human and multiple garment images. (c) We finally train our composition module with the synthetic paired dataset enabling it to generate human images with multiple reference garment images.

We demonstrate the effectiveness of BootComp in terms of garment fidelity and compositionality through extensive experiments. For example, BootComp shows 30% improvement on MP-LPIPS [3] than the previous state-of-the-art methods. Moreover, our BootComp is extensively applied to various conditional human image generations in the fashion domain, such as virtual try-on and controllable human image generation with other conditions, such as faces and poses. We also highlight the generalization capabilities of BootComp across different image domains, generating human images in various styles like cartoons.

2. Background

2.1. Diffusion Models

Diffusion models [14, 19, 20, 44] are a type of generative model consisting of a forward process and a reverse process. Specifically, diffusion models learn the reverse process of the forward process, where the forward process is defined as a Markov chain that gradually adds Gaussian noise to data. Starting from Gaussian noise, The sampling is done with a learned reverse process of this forward process.

Formally, let $\mathbf{x}_0$ represent a data instance (*e.g.*, an image or a latent vector from an autoencoder's output [40]). Diffusion models consider a pre-defined forward process $q(\mathbf{x}_t|\mathbf{x}_0)$ given a closed form as a normal distribution $\mathcal{N}(\alpha_t\mathbf{x}_0, \sigma_t^2\mathbf{I})$, so the sampling can be done from Gaussian distribution $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ using reparametrization to have $\mathbf{x}_t = \alpha_t\mathbf{x}_0 + \sigma_t\boldsymbol{\epsilon}$. Here, $\{\alpha_t\}_{t=1}^T$ and $\{\sigma_t\}_{t=1}^T$ are pre-defined decreasing and increasing noise scheduling sequences (respectively) for $t = 1, \dots, T$ that let $p(\mathbf{x}_T)$ converge a distribution close to Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

Learning the reverse process $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ of a diffusion model is equivalent to learning a score function of perturbed data distribution (through score matching [17]), typically achieved via an ϵ-noise prediction loss [14] by training a denoising autoencoder. Specifically, one can formulate the training objective of the diffusion model as:

$$\mathcal{L}_{\text{DM}} = \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}),\, t\sim\mathcal{U}[0,T]}\left[\omega(t)\|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t; t) - \boldsymbol{\epsilon}\|_2^2\right],$$

where $\omega(t) > 0$ is a weight function at each timestep t and $\mathcal{U}[0, T]$ denotes a uniform distribution.

After training, data sampling can be done using the learned reverse process. Specifically, starting from $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \sigma_T^2\mathbf{I})$, the model gradually denoises $\mathbf{x}_t$ to $\mathbf{x}_{t-1}$ for each t, until $\mathbf{x}_0$ is drawn from the data distribution.

2.2. Text-to-Image (T2I) Diffusion Models

Text-to-image (T2I) diffusion models [7, 40, 42] are text-conditional diffusion models $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t; \mathbf{c}, t)$ that generate an image $\mathbf{x}_0$ conditioned on a given text prompt $\mathbf{c}$. This prompt is usually provided as a text representation encoded by pre-trained text encoders, such as T5 [39] or CLIP [38]. Commonly, T2I diffusion models employ convolutional U-Net architectures combined with attention layers [14, 45] to condition the model on texts. Among T2I diffusion models, Stable Diffusion [SD; 40] is one of the de-facto T2I diffusion models that generates high-quality images. We mainly use Stable Diffusion XL (SDXL) [37], one of the SD variants. However, our framework is model-agnostic and can be adapted to any other T2I diffusion models.

3. Method

Let $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ be a set of $N \gg 1$ reference garment images (*e.g.*, shirt, pants, *etc.*) and $\mathbf{y}$ be a human image that is wearing $\mathbf{x}_1, \dots, \mathbf{x}_N$. Our goal is to learn a conditional distribution $p(\mathbf{y}|\mathbf{X})$—we train a conditional generative model $g_\theta(\mathbf{X}) = \mathbf{y}$ that generates human image $\mathbf{y}$ wearing arbitrary garment images $\mathbf{X}$ given as a condition.

One straightforward direction is to train the model g_θ using a paired dataset $\mathcal{D} = \{(\mathbf{X}^i, \mathbf{y}^i)\}_{i=1}^d$ with a dataset size $d > 0$, where each $\mathbf{X}^i = \{\mathbf{x}_1^i, \dots, \mathbf{x}_{N_i}^i\}$ consists of N_i different number of reference images. However, this approach suffer from data acquisition problem: collecting all of the reference garment images of a given human image is wearing is challenging. In practice, there usually exists a single reference image, *i.e.*, N_i mostly becomes 1 (*e.g.*, a human and pants that he/she is wearing). Thus, the model trained with this data easily lacks compositional generalization capability at inference time, *i.e.*, the trained model g_θ fails to generate the human image with large number of garments.

To tackle this data curation problem, we introduce an additional decomposition network f_ϕ that can extract reference images from a given human image. By doing so, we generate a synthetic dataset $\tilde{\mathcal{D}}$, where each $(\tilde{\mathbf{X}}^i, \mathbf{y}^i) \in \tilde{\mathcal{D}}$ satisfies $|\tilde{\mathbf{X}}^i| \gg 1$ and $\mathbf{y}^i$ is in the original dataset. We then train the conditional generative model g_θ using this synthetic dataset. Here, we also introduce a filtering strategy to improve the quality of the synthetic dataset $\tilde{\mathcal{D}}$ generated from f_ϕ, by removing low-quality extraction results.

In the rest of this section, we explain our BootComp in detail. In Section 3.1, we describe the training data generation process, introducing our decomposition network f_ϕ, which is used for synthetic data generation, and explaining our data filtering strategy. Finally, in Section 3.2, we explain the details of our network for our original goal of controllable generation trained with the synthetic dataset.

3.1. Training data generation

Decomposition module. Our decomposition module generates a *single* garment image in a product view, denoted as $\mathbf{x}$, from a garment of category $\mathbf{m}$ that human $\mathbf{y}$ is wearing. We consider this mapping as an image-to-image translation problem: generating the reference garment image $\mathbf{x}$ from the portion of person image $\mathbf{y}$ that falls into category $\mathbf{m}$.

To achieve this, we initialize a diffusion model f_ϕ as a pre-trained text-to-image diffusion model and fine-tune it with the following objective:

$$\mathcal{L}(\phi) := \mathbb{E}\Big[\omega(t)\big|\big|f_\phi(\mathbf{x}_t; \mathbf{c}, t, \mathbf{x}^s) - \boldsymbol{\epsilon}\big|\big|_2^2\Big], \tag{1}$$

where $\mathbf{x}^s = S(\mathbf{y}, \mathbf{m})$ is a segmented garment part using an off-the-shelf human parsing model S [50], and we let a text prompt $\mathbf{c}$ be "A product photo of $\{$*category*$\}$" to extensively leverage the prior knowledge of the T2I diffusion model.

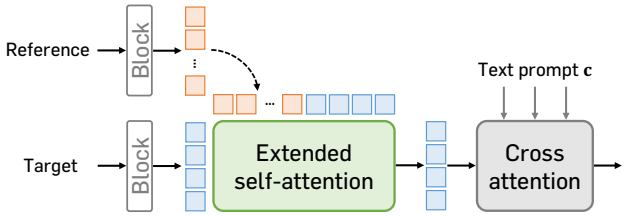

Figure 4. **Extended self-attention architecture.** In a extended self-attention layer, reference hidden states are concatenated with the target hidden states in the key and value matrices. This architecture enables injecting reference image features within the target image. Note that decomposition module also uses same structure but works within a single network.

To condition the model on an image $\mathbf{x}^s$, we utilize the pretrained diffusion model as an image encoder, which can extract rich features and can preserve the fine-details (*e.g.*, small logos). Specifically, for each self-attention layer in the model, we concatenate the corresponding key and value vectors computed with $\mathbf{x}^s$, so the self-attention in the forwarding path of $\mathbf{x}_t$ can be conditioned on $\mathbf{x}^s$ (see Fig. 4).

Finally, note that training f_ϕ can be done with the dataset $\mathcal{D}$ which consists of a pair of *single* reference garment and a human image, because we train the model to extract a *single* reference garment from the human image.

Synthetic data generation with filtering. After training the decomposition module, one can use it for extracting all of the reference images $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ from each human image $\mathbf{y}$. It results in a synthetic dataset $\tilde{\mathcal{D}}$, which can be used for the conditional generative model g_ϕ for our goal of controllable generation. However, we find that the decomposition network f_ϕ sometimes generates low-quality reference images, especially when the prediction results from the parsing model S are incorrect, which might harm the performance of g_ϕ (see Fig. 5).

Thus, we introduce a simple filtering strategy to improve the quality of our synthetic dataset $\tilde{\mathcal{D}}$. Specifically, we measure the image similarity score between the generated garment image $\tilde{\mathbf{x}} = f_\phi(\mathbf{y}, \mathbf{m})$ and the segmentation results $\mathbf{x}^s$. We discard pair sets if any garment in the set has a similarity score below the threshold value $\tau > 0$, namely:

$$d(\mathbf{x}^s, \tilde{\mathbf{x}}) < \tau \tag{2}$$

For the scoring function for image similarity, we empirically find that dreamsim [8] aligns the most with human perception (See Appendix **??** for details).

3.2. Composition module

Our composition module consists of two diffusion models: one for a generation and the other one for an image encoder, denoted by g_{θ^-} and g_θ, respectively. Both networks

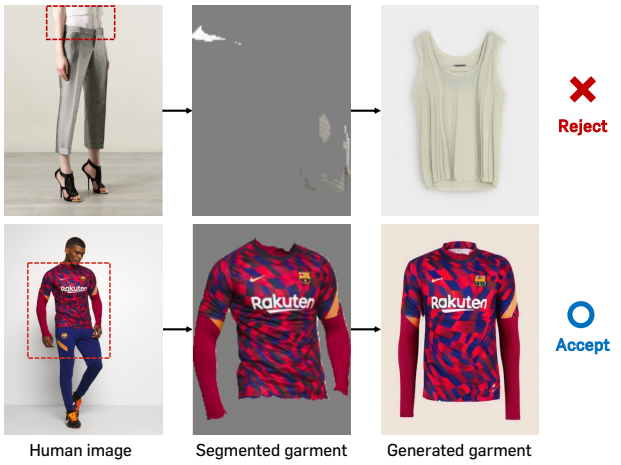

Figure 5. **Examples of high&low-quality generated garments.** When human parsing results are not precise, the decomposition network struggles to generate product garment images accurately, resulting in low-quality garment images. We filter out these cases.

are initialized with the same pre-trained T2I diffusion models, where we freeze g_{θ^-} used as a generator and only train the encoder network g_θ using the synthetic dataset $\tilde{\mathcal{D}}$. In particular, the encoder g_θ is used to provide conditioning of garments $\tilde{\mathbf{X}}$ to the generator g_{θ^-}.

To condition $\tilde{\mathbf{X}}$ to the generation model g_{θ^-}, we concatenate the key and value vectors in each self-attention layer computed with each $\tilde{\mathbf{x}} \in \tilde{\mathbf{X}}$ and corresponding category $\mathbf{m}_{\tilde{\mathbf{x}}}$ using the encoder model g_θ. By doing so, generator g_θ can be conditioned on $\tilde{\mathbf{X}}$ through its attentions. In particular, query, key, and value vectors of each of the attention layer in g_θ are computed with the following vectors

$$\text{query} := \mathbf{h}_{\mathbf{y}}, \quad \text{key, value} := [\mathbf{h}_{\mathbf{y}}, \mathbf{h}_{\tilde{\mathbf{x}}_1}, \dots, \mathbf{h}_{\tilde{\mathbf{x}}_N}], \tag{3}$$

where $\mathbf{h}_{\mathbf{y}}$ and $[\mathbf{h}_{\tilde{\mathbf{x}}_1}, \dots, \mathbf{h}_{\tilde{\mathbf{x}}_N}]$ are hidden states before the self-attention layer computed with the generation model g_{θ^-} and the encoder model g_θ, respectively. To compute each $\mathbf{h}_{\tilde{\mathbf{x}}}$ we provide the text caption "A photo of $\{category\}$" to the encoder model g_θ, where $\{category\}$ is a type of garment $\tilde{\mathbf{x}}$.

Thus, we fine-tune the encoder g_θ through the diffusion model objective of the generator g_{θ^-}:

$$\mathcal{L}(\theta) := \mathbb{E}\Big[\omega(t)\big|\big|g_{\theta^-}(\mathbf{y}_t; \mathbf{c}, t, \tilde{\mathbf{X}}) - \boldsymbol{\epsilon}\big|\big|_2^2\Big], \tag{4}$$

where we employ synthetic text description for human image generated by vision-language model [27] for $\mathbf{c}$.

4. Experiments

We validate the effectiveness of BootComp and the effect of the proposed components through extensive experiments. In particular, we investigate the following questions:

- Can BootComp generate authentic human images wearing multiple garments while preserving details? (Tab. 1, Fig. 6)
- Is our data generation pipeline effective and scalable, ensuring the model's performance? (Tabs. 2 and 3, Fig. 9)
- Can BootComp be used for a wide range of downstream tasks? (Fig. 7)

4.1. Experiment Setup

We explain some important experimental setups in this section. We include more details in Appendix **??**.

Implementation details. We use Stable Diffusion XL (SDXL) [37] for model initializations. We collect human-*single* reference garment paired datasets from VITON-HD [5], DressCode [30] and LAION-Fashion [24] for training the decomposition module. The dataset consists of 25,210 upper garments, 7,151 lower garments, 27,677 dresses, 5,675 bags, 1,599 shoes, 825 scarf, and 159 hats, resulting 68,296 single reference pairs on different categories. We train the decomposition module for 140K iterations with a total batch size of 32 on 4 H100 GPUs. For the data generation phase, we process 240K human images obtained from VITON-HD, DressCode, LAION-Fashion, and DeepFashion [28] datasets, thereby collecting 240K paired data of human image and *multiple* garment images at resolution 512×384. We obtain and use 54K high-quality paired data after applying our filtering strategy with the threshold value $\tau = 0.4$. For the composition module, we train for 115K iterations with a total batch size of 48 on 8 H100 GPUs. For inference, we use the DDPM sampler [14] with a sampling step of 50, where we apply classifier-free guidance (CFG; [13]) with a guidance scale of $w = 2.0$.

Baselines. First, we consider MIP-Adapter [15] as baselines, which is a recent generic controllable generation method with multiple conditions. We also compare BootComp with FromParts2Whole [16], the most relevant baseline for our task that aims for controllable human image generation with multiple reference garments. We use the official model parameters from their official implementations. We employ "A model wearing upper garment and lower garment and shoes" as the text prompt to both models.

Evaluation metric. We report Frenchét Inception Distance (FID) [12], MP-LPIPS [3], and two different image similarities metrics [15, 46] using DINOv2 [34] (DINO and M-DINO). First, we use the FID score to measure the fidelity of generated human images, *i.e.*, whether multiple garments are harmonized in the generated images. Next, we employ MP-LPIPS to evaluate the consistency of the target image to the source ground-truth garment. Finally, DINO and M-DINO measure the semantic similarity between each reference garment image and the respective garment present in the generated human image.

Figure 6. **Qualitative comparison of human image generation with multiple garments.** BootComp generates realistic human images with multiple reference garments even with non-straightforward combinations of garments without losing details of each reference. For example, Parts2Whole replaces reference soccer cleats with stilettos, while ours accurately generates each reference (left, middle row).

Table 1. **Quantitative comparisons.** We compare BootComp with baselines on garment similarity and image fidelity. We see that BootComp outperforms other methods, preserving fine-details of garments and naturally generating human images.

Method	MP-LPIPS ↓	DINO ↑	M-DINO ↑	FID ↓
MIP-Adapter [15]	0.276	0.308	0.025	59.99
Parts2Whole [16]	0.267	0.362	0.036	28.39
BootComp (ours)	**0.187**	**0.379**	**0.046**	**27.63**

Evaluation datasets. We manually collect a dataset for evaluation as there are no common datasets for evaluating controllable human image generation. To evaluate MP-LPIPS, DINO, and M-DINO, we curate 5,000 garment image sets of three representative garment categories for human images (upper and lower garments and shoes). We randomly take upper and lower garment images from the test dataset of DressCode [30] dataset and shoe images from a public dataset.[1] Next, for evaluation using FID, we gather 30,000 human images wearing various garments in different poses from the test dataset of DressCode, VITON-HD, and Deepfashion to use them as reference image sets.

[1] https://www.kaggle.com/datasets/noobyogi0100/shoe-dataset

4.2. Results

Qualitative results. We provide qualitative comparisons of our method (BootComp) with other baseline methods in Fig. 6. As shown in this figure, BootComp generates more realistic human images in various poses, faithfully preserving details of reference garment images, while other methods often generate human images wearing garments inconsistent with the references. Moreover, this result shows that BootComp generates creative combinations of garments. For instance, in the first example of the second row, BootComp generates a human image with uncommon combination of garments (*e.g.*, trousers with soccer cleats) but Parts2Whole or MIP-Adapter fails to achieve this: they either undesirably replace the cleats to trousers or struggle with generating high-fidelity garments (respectively). We provide more visualizations in Appendix **??**.

Quantitative results. We report quantitative evaluation results of BootComp and baselines in Tab. 1. BootComp outperforms both MIP-Adapter and Parts2Whole across all of four evaluate metrics. In particular, BootComp achieves a 30% improvement in MP-LPIPS score over the baselines, demonstrating its effectiveness in preserving garment details. Moreover, BootComp shows its capabilities in authentic image generation for human images, as indicated by a better FID values than baselines.

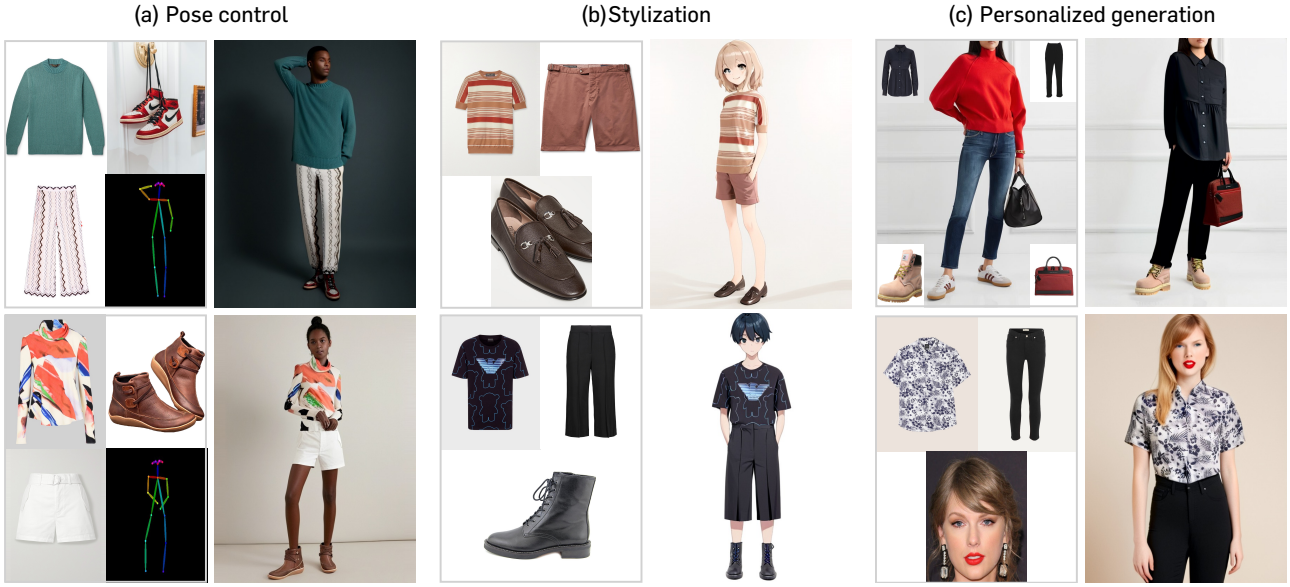

Figure 7. **More applications of BootComp.** We showcase the extensive applications of our method, BootComp. BootComp creates human images by controlling the (a) poses and (b) styles of the generated human images. BootComp also enables (c) personalized human image generation by taking user's images as conditions (*e.g.*, face, full body).

More applications. In Fig. 7, we apply BootComp to several downstream tasks and visualize their results. First, we show that BootComp can generate human images conditioned on the pose. In Fig. 7 (a), BootComp generates human images in diverse poses following the extra conditions even with reference garments of intricate patterns, demonstrating its generalization capability. We also show that BootComp can generate human images with different stylizations such as cartoons in Fig. 7 (b). Finally, we show that BootComp can be used for personalized human image generation such as virtual try-on, *i.e.*, changing garments on a given human image to reference garments. In Fig. 7 (c), BootComp replaces garments on a given human image with the reference garment images and enables personalized generation conditioning face image.

Note that this can be done without any additional task-specific fine-tuning as we freeze the generator in the composition module during training. This enables BootCompto be easily integrated with other modules, *e.g.*, IP-Adapter [51] or ControlNet [53], that provides controllability with additional condition inputs. We provide additional generation results for each application in Appendix **??**.

4.3. Analysis and ablation studies

Finally, we conduct several analyses on synthetic data to validate our data generation pipeline, including its scalability and the impact compared with a naïve use of a segmented paired dataset. To reduce the computation cost, we use Stable Diffusion v1.5 for all analyses while we strictly follow the other setups used in the main experiments.

Figure 8. **Visualization of segmented paired data and our synthetic paired data.** We provide a visual comparison between segmented and synthetic pairs. Given a single garment and a human image pair, we segment out other garments from the human image in the segmented paired data.

Table 2. **Comparison on dataset construction methods.** The model trained on the segmented paired dataset shows worse performance compared to one trained on our synthetic paired dataset both in garment similarity and image fidelity.

Dataset	MP-LPIPS ↓	DINO ↑	M-DINO ↑	FID ↓
Segmented	0.374	0.284	0.025	59.27
Synthetic	**0.197**	**0.365**	**0.043**	**29.41**

Effect of data generation. We first show the effect of our data generation scheme. We demonstrate this by constructing a dataset by segmenting out all garment images from the human except the given one in the dataset (see Fig. 8), and train the composition module on this dataset. As shown in Tab. 2, the model trained on the segmented paired dataset achieves worse performance across all evaluation metrics. Also, Fig. 9 visualizes undesirable generated images by the model trained on the segmented dataset. This indicates the model struggles to generate desirable human images, highlighting the effectiveness of our data generation scheme.

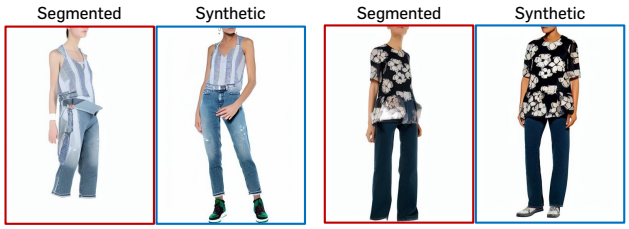

Figure 9. **Visual comparison on data construction methods.** Visual comparison between generated human images where each model is trained on segmented and synthetic pairs. The model trained on segmented pair data struggles to generate naturally harmonized human images (red).

Table 3. **Comparison on dataset scale.** Training with a larger datatset (after filtered) improves the model's overall performance in both garment similarity and image fidelity.

Dataset size	DINO ↑	M-DINO ↑	FID ↓
5K	0.337	0.248	34.15
15K	0.338	0.251	32.32
30K	0.344	0.261	26.99
50K	**0.360**	**0.285**	**25.88**

Scalability of the data generation scheme. Next, we investigate the scalability of our data generation scheme by exploring the effect of dataset size to the performance. We observe that using a larger dataset for training always improves the model's performance in both garment fidelity and image fidelity, as shown in Tab. 3.

Table 4. **Ablation study for threshold value τ on filtering.** The data quality improves with a stricter threshold value, leading to better performance. We adopt $\tau = 0.4$ when applying the filtering.

τ	0.4	0.5	0.6	0.7	1.0
DINO↑	**0.360**	0.347	0.343	0.342	0.338

Ablation study: threshold value τ. Finally, we conduct an ablation study on the threshold value τ used in our dataset filtering strategy. In Table. 4, we report similarity score (DINO) of the models trained with different datasets by varying values of τ from 0.4 to 1.0, where 1.0 indicates no filtering is applied. We observe that more strict data filtering can provide more performance gain to the model.

5. Related Work

Controllable image generation. In addition to using text prompts as conditions, recent works have attempted to improve the controllability of text-to-image (T2I) diffusion models by incorporating additional inputs (*e.g.*, images). In particular, many works focus on generating images that preserve the identity of subjects in the source image by proposing additional modules to the model [25, 35, 51]. Despite their effort, they have struggled to generalize with multiple subjects and suffer from several issues, such as subject blending. To mitigate this issue, several approaches such as MS-Diffusion [46] and FastComposer [49] introduce an additional regional information for each subject. Our framework also tries to improve image generation with multiple subjects, but we focus on human image generation and propose a novel data generation pipeline to improve the quality.

Virtual try-on. Inspired by the great progress of T2I diffusion models, recent works have explored their application to various tasks on fashion domain such as virtual try-on [2, 6, 21, 36, 54–56] and virtual dressing [3, 43]. However, most of them are limited to single-garment based generation as they rely on existing public datasets [5, 30] consisting of single-paired data. While several works [36, 54, 56] address multi-garment virtual try-on, they depend on proprietary datasets, which limits scalability and its capability to support a few garment categories. Our data generation pipeline tackles this data acquisition bottleneck and supports multi-garment based generation with a wide range of categories.

Improving diffusion models with self-data generation. Recent works have tried to improve the performance of the pre-trained model itself on the specific tasks [1, 9, 18, 48, 52] using generated image data from the same model. For example, JeDi [52] generates same-subject images using LLMs and pretrained T2I diffusion models. They are used to fine-tune T2I diffusion models for personalized generation [41] without additional tuning at inference.

However, these approaches are not suitable for the case of images with multiple subjects, such as controllable human generation, as most T2I diffusion models still lack the capability to accurately generate images with multiple subjects [26, 33]. As a result, synthetic image data with multiple subjects generated with T2I models often exhibit low-quality results, and thus fine-tuning with this dataset does not lead to the improvement. Thus, rather than generating multi-subject images from T2I models, existing approaches have curated data through a segmentation from the multi-subject images [16]. However, these models suffer from the copy-and-paste and subject inconsistency problems. Our method bridges the former and latter approaches to improve data quality used for controllable human generation.

6. Conclusion

In this paper, we present BootComp, a novel framework for controllable human image generation with multiple garments given as image conditions. Our pipelines for synthetic paired data generation and controllable generation enabled creating human images wearing multiple reference garments. We show the broad applicability of BootComp by adapting it to various types of tasks in the fashion domain.

Acknowledgement

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.RS-2019-II190075 Artificial Intelligence Graduate School Program(KAIST); No.RS-2021-II212068, Artificial Intelligence Innovation Hub).

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2, 8

[2] Mengting Chen, Xi Chen, Zhonghua Zhai, Chen Ju, Xuewen Hong, Jinsong Lan, and Shuai Xiao. Wear-any-way: Manipulable virtual try-on via sparse correspondence alignment. In *European Conference on Computer Vision*, 2024. 8

[3] Weifeng Chen, Tao Gu, Yuhao Xu, and Chengcai Chen. Magic clothing: Controllable garment-driven image synthesis. *arXiv preprint arXiv:2404.09512*, 2024. 3, 5, 8

[4] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6593–6602, 2023. 2

[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*, 2021. 2, 5, 8

[6] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. In *European Conference on Computer Vision*, 2024. 2, 8

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 2, 3

[8] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *Advances in Neural Information Processing Systems*, 36, 2024. 4

[9] Rinon Gal, Or Lichter, Elad Richardson, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Lcm-lookahead for encoder-based text-to-image personalization. *arXiv preprint arXiv:2404.03620*, 2024. 8

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *International Conference on Learning Representations*, 2023. 2

[11] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785, 2024. 2

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[13] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 5

[15] 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. 2, 5, 6

[16] Zehuan Huang, Hongxing Fan, Lipeng Wang, and Lu Sheng. From parts to whole: A unified reference framework for controllable human image generation. *arXiv preprint arXiv:2404.15267*, 2024. 2, 5, 6, 8

[17] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005. 3

[18] Sangwon Jang, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. Identity decoupling for multi-subject personalization of text-to-image models. *arXiv preprint arXiv:2404.04243*, 2024. 8

[19] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in neural information processing systems*, pages 26565–26577, 2022. 3

[20] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 3

[21] 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. 2, 8

[22] Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. Large-scale text-to-image generation models for visual artists' creative works. In *Proceedings of the 28th international conference on intelligent user interfaces*, 2023. 2

[23] Kyungmin Lee, Sangkyung Kwak, Kihyuk Sohn, and Jinwoo Shin. Direct consistency optimization for robust customization of text-to-image diffusion models. *Advances in neural information processing systems*, 2024. 2

[24] Simon Lepage, Jérémie Mary, and David Picard. Lrvs-fashion: Extending visual search with referring instructions. *arXiv:2306.02928*, 2023. 2, 5

[25] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *Advances in Neural Information Processing Systems*, 2023. 8

[26] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 8

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2023. 5

[28] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 5

[29] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2

[30] 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. 2, 5, 6, 8

[31] 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 ACM International Conference on Multimedia*, 2023. 2

[32] Weili Nie, Sifei Liu, Morteza Mardani, Chao Liu, Benjamin Eckart, and Arash Vahdat. Compositional text-to-image generation with dense blob representations. In *International Conference on Machine Learning*, 2024. 2

[33] Weili Nie, Sifei Liu, Morteza Mardani, Chao Liu, Benjamin Eckart, and Arash Vahdat. Compositional text-to-image generation with dense blob representations. *arXiv preprint arXiv:2405.08246*, 2024. 8

[34] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 2, 5

[35] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-G: Generating images in context with multimodal large language models. In *International Conference on Learning Representations*, 2024. 2, 8

[36] Soonchan Park and Jinah Park. Full-body virtual try-on using top and bottom garments with wearing style control. *Computer Vision and Image Understanding*, 251:104259, 2025. 8

[37] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *International Conference on Learning Representations*, 2024. 2, 3, 5

[38] 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. 2, 3

[39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 3

[40] 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

[41] 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*, 2023. 2, 8

[42] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 3

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

[44] 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. 3

[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. 3

[46] X Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209*, 2024. 2, 5, 8

[47] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7677–7689, 2023. 2

[48] Daniel Winter, Matan Cohen, Shlomi Fruchter, Yael Pritch, Alex Rav-Acha, and Yedid Hoshen. Objectdrop: Bootstrapping counterfactuals for photorealistic object removal and insertion. *arXiv preprint arXiv:2403.18818*, 2024. 8

[49] 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*, 2024. 8

[50] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in neural information processing systems*, 2021. 4

[51] 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. 2, 7, 8

[52] Yu Zeng, Vishal M Patel, Haochen Wang, Xun Huang, Ting-Chun Wang, Ming-Yu Liu, and Yogesh Balaji. Jedi: Joint-image diffusion models for finetuning-free personalized text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6786–6795, 2024. 8

[53] 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. 7

[54] Xujie Zhang, Ente Lin, Xiu Li, Yuxuan Luo, Michael Kampffmeyer, Xin Dong, and Xiaodan Liang. Mmtryon: Multi-modal multi-reference control for high-quality fashion generation. *arXiv preprint arXiv:2405.00448*, 2024. 8

[55] Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4606–4615, 2023.

[56] Luyang Zhu, Yingwei Li, Nan Liu, Hao Peng, Dawei Yang, and Ira Kemelmacher-Shlizerman. Mm vto: Multi-garment virtual try-on and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 8