

# PromptDresser: Improving the Quality and Controllability of Virtual Try-On via Generative Textual Prompt and Prompt-aware Mask

Jeongho Kim    Hoiyeong Jin    Sunghyun Park    Jaegul Choo  
KAIST, Daejeon, South Korea

{rlawjdghek, hy.jin, psh01087, jchoo}@kaist.ac.kr

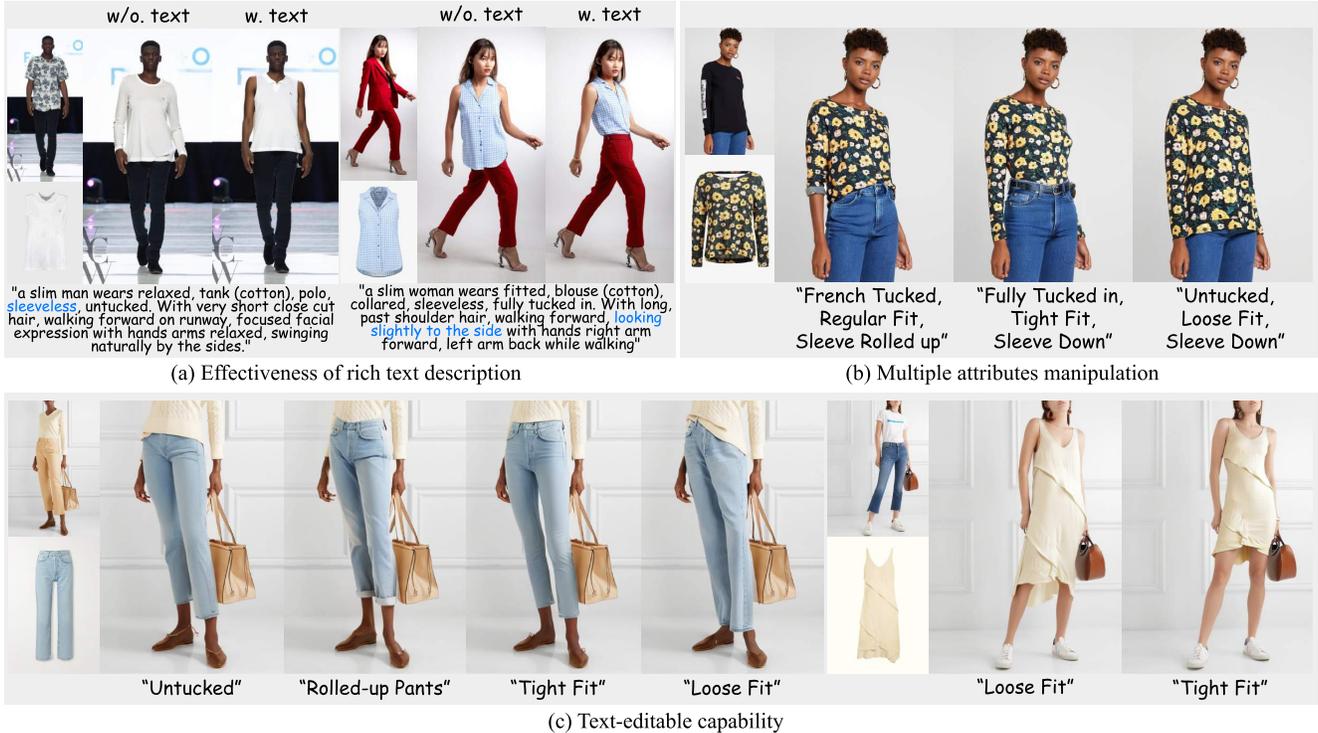


Figure 1. Generated results of PromptDresser: PromptDresser (a) enables high-quality virtual try-on through rich text description, (b) highlights multiple attributes manipulation simultaneously, and (c) generates versatile outputs across multiple clothing categories, including bottoms and dresses.

## Abstract

Recent virtual try-on approaches have advanced by fine-tuning the pre-trained text-to-image diffusion models to leverage their powerful generative ability. However, the use of text prompts in virtual try-on is still underexplored. This paper tackles a text-editable virtual try-on task that changes the clothing item based on the provided clothing image while editing the wearing style (e.g., tucking style, fit) according to the text descriptions. In the text-editable virtual try-on, three key aspects exist: (i) designing rich text descriptions for paired person-clothing data to train the model, (ii) addressing the conflicts where textual in-

formation of the existing person’s clothing interferes the generation of the new clothing, and (iii) adaptively adjust the inpainting mask aligned with the text descriptions, ensuring proper editing areas while preserving the original person’s appearance irrelevant to the new clothing. To address these aspects, we propose PromptDresser, a text-editable virtual try-on model that leverages large multi-modal model (LMM) assistance to enable high-quality and versatile manipulation based on generative text prompts. Our approach utilizes LMMs via in-context learning to generate detailed text descriptions for person and clothing images independently, including pose details and editing attributes using minimal human cost. Moreover, to ensure the editing areas, we adjust the inpainting mask depending on the text prompts adaptively. We found that our ap-

*proach, utilizing detailed text prompts, not only enhances text editability but also effectively conveys clothing details that are difficult to capture through images alone, thereby enhancing image quality. Extensive experiments demonstrate that our method outperforms the baselines significantly while showing versatile manipulation capabilities based on text prompts. Our code is available at <https://github.com/rlawjdghek/PromptDresser>.*

## 1. Introduction

Virtual try-on [21] provides an advanced technique for personalized fashion previews, removing the need for physical trials. Moreover, the capability to control wearing styles, such as tucking style or fit, can enhance overall user experience by allowing users to visualize different outfit options.

Thanks to the advanced generative capabilities of diffusion models [14, 25, 41], virtual try-on have shown significant improvements compared to earlier generative adversarial network-based approaches [3, 11, 17–19, 22, 29, 47, 52]. These recent models leverage the powerful priors embedded in large-scale diffusion models [9, 39, 42, 44], resulting in improvements in clothing detail representation and generalization performance. Furthermore, recent studies have enhanced generative quality by incorporating text inputs with advanced text-to-image (T2I) models [12, 34] and have added an editing capability via click or drag [10]. However, although these approaches are built on text-based generative models [39, 42], they typically use simple textual prompts of clothing attributes, such as ‘short sleeve t-shirt.’ The potential for using rich textual information to achieve higher performance and extensive text-driven editability is still largely unexplored.

In this paper, we tackle a text-editable virtual try-on task that changes the clothing item based on the provided clothing image and edits the wearing style according to the text descriptions. Text-editable virtual try-on has unique and challenging aspects. (i) It is crucial to construct rich and well-aligned text descriptions for wearing the new clothing to the person image, (ii) avoiding textual conflicts between the original and new clothing during sampling. Since the wearing styles (*e.g.*, untucked, fully tucked in) vary according to the the text descriptions, (iii) an adaptive mask aligned with the text prompt is necessary to preserve regions irrelevant to the clothing and to minimize the influence of the original clothing shape.

To address these challenges, we introduce PromptDresser, a novel virtual try-on model that leverages rich textual information to achieve high-quality and versatile manipulation. To obtain the text prompts for paired person-clothing data, we instruct the large multimodal models (LMMs) to describe the person and clothing images, respectively. However, LMMs often produce excessively di-

verse text descriptions on the provided images, leading to potential misalignment with the images or insufficient details. Therefore, we leverage in-context learning that conditions expressive layouts and clothing, effectively specifying attributes to focus on the inpainting regions with minimal human cost. This approach significantly outperforms the model that relies on holistic descriptions of the entire person and clothing image, providing a scalable and efficient solution for versatile virtual try-on. Furthermore, we found that detailed text descriptions produced by LMMs not only enhance text editability but also effectively maintain clothing details that are difficult to capture through clothing images alone, thereby generating high-fidelity images.

Following the previous approaches [10–12, 28, 29, 38], we also utilize an inpainting mask to determine the preserved and generated regions. The traditional inpainting mask often constrains the model to the shape of the existing clothing, leading to unnatural results when adapting to different clothing types. To mitigate such constraints, we applied random dilation mask augmentation, enabling the model to learn from a wide range of mask sizes, from broad to narrow. Expanding the masked area effectively removes details related to the original clothing’s length and shape, but it also poses the risk of erasing areas that should be preserved (*e.g.*, pants when generating a top). Therefore, we propose to use an expanded inpainting mask to obtain an approximate clothing region aligned with the text prompt. Next, we generate a refined mask by combining this expanded mask with a fine mask that removes areas where clothing will be applied (*e.g.*, arms or torso) to preserve as much of the original person and background as possible. By employing a refined mask that is agnostic to the existing clothing and aligned with the text prompts, our model preserves details irrelevant to the worn clothing, enabling a wide range of text-based manipulation. Through extensive experiments, we demonstrate that PromptDresser achieves superior image quality compared to the existing virtual try-on methods while effectively controlling the wearing styles using diverse text prompts.

In summary, our contributions are as follows:

- We propose a text-editable virtual try-on model that, with the assistance of a large multimodal model (LMM) with in-context learning, achieves rich, well-aligned text descriptions for both person and clothing, ensuring no textual conflicts with any clothing provided.
- To mitigate the issue of following the original clothing’s attributes, such as shape and length, we propose random dilation mask augmentation. Our prompt-aware mask generation enhances diversity in virtual try-on results while effectively preserving the person’s original appearance.
- Our approach achieves state-of-the-art performance across multiple datasets with and enables versatile ma-

manipulation capabilities, highlighting the effectiveness of generative textual prompt for virtual try-on.

## 2. Related Work

### 2.1. Image-based Virtual Try-On

Early approaches often relied on two-stage frameworks, combining explicit warping networks with GAN-based generators [3, 11, 17, 29, 50, 52]. However, these methods continue to struggle with error accumulation across multiple stages, leading to a shift towards diffusion models which have shown impressive generative capabilities across various domains [6, 43, 54, 55, 58]. Due to the challenges of dataset acquisition and the advantages of leveraging prior knowledge, most approaches build upon large-scale text-to-image diffusion models [9, 39, 42, 44], inherently benefiting from their robust inpainting capabilities [19] or utilizing textual inversion techniques [38]. Notably, recent research [12, 28] propose attention-based, end-to-end virtual try-on models that preserve fine details of clothing while achieving the generalization performance.

Another line of progress in image-based virtual try-on is controllability. LC-VTON [53], for instance, introduced new segmentation labels to incorporate clothing length, enabling the generation of high-fidelity images. Additionally, some studies [10, 31] have utilized landmarks to introduce point-based controllability. While these spatial condition-based approaches offer a fine degree of controllability, they are limited in their ability to perform comprehensive editing such as adjusting fit and overall appearance. On the other hand, recent research leveraging text-based controllability has tried to address such challenges. However, most existing studies still incorporate text but fail to consider the existing areas (*e.g.*, background) that need to be preserved [59] or are limited to captioning solely for clothing [12, 34].

In this paper, we propose a virtual try-on model that uses rich text descriptions to harness the capabilities of large-scale text-to-image diffusion models. By independently extracting captions for both the person and the clothing with LMMs, we enhance generalization performance and enable manipulation. Moreover, we propose a novel adaptive mask for further flexible manipulation while preserving the original person’s appearance.

### 2.2. LMMs for Multimodal Data Augmentation

Recent advancements in large multimodal models (LMMs) [2, 30, 35, 46, 57] have demonstrated their powerful visual understanding [36, 48], achieving impressive performance across various vision tasks. Harnessing the capabilities of LMMs, they are utilized to tune image editing models [7] and to enhance captioning performance through image-caption fusion [4]. However, their application to virtual try-on remains still underexplored. In

this paper, we leverage off-the-shelf LMMs to augment image captions from virtual try-on datasets, enabling the generation of higher-fidelity images. Furthermore, we allow users to wear diverse styles through text-based manipulation according to their preferences.

## 3. Method

### 3.1. Preliminary: Latent Diffusion Model

Our approach builds on pre-trained text-to-image latent diffusion models (LDMs) [39, 42], which consist of three main components: a variational auto-encoder (VAE) with an encoder  $\mathcal{E}(\cdot)$  and decoder  $\mathcal{G}(\cdot)$ , a text encoder  $\tau(\cdot)$  and main U-Net  $\epsilon_\theta(\cdot)$ . The pre-trained VAE encodes an image  $\mathbf{x}$  into a low-dimensional latent space as  $\mathbf{z}_0 = \mathcal{E}(\mathbf{x})$  and reconstructs it back into RGB space as  $\hat{\mathbf{x}} = \mathcal{G}(\mathbf{z}_0)$ . The main U-Net is trained to predict the  $\mathbf{z}_0$  from the perturbed latent variable  $\mathbf{z}_t$ , defined as  $\mathbf{z}_t = \mathcal{N}(\mathbf{z}_t; \sqrt{\bar{\alpha}_t}\mathbf{z}_0, (1 - \bar{\alpha}_t)\mathbf{I})$ . Here,  $\bar{\alpha}_t = \prod_{s=1}^t(1 - \beta_s)$ , where  $(\beta_t)_{t=0}^T$  is a decreasing sequence [25]. The loss function of LDMs is given by:

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathbf{z}_0 \sim \mathcal{E}(\mathbf{x}), \epsilon \sim \mathcal{N}(0, \mathbf{I}), \mathbf{y}, t} [\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \tau(\mathbf{y}))\|_2^2]. \quad (1)$$

This objective function directs the model to reduce the difference between the added noise,  $\epsilon$ , and the noise predicted by the U-Net. The predicted noise is then used in the reverse diffusion process to approximate the original latent representation.

### 3.2. Overall Framework

Our method, PromptDresser, aims to enable text-editable virtual try-on of reference clothing  $\mathbf{x}^c$  onto a target person  $\mathbf{x}^p$  with additional generative textual prompt provided by LMMs. An overview of the proposed method is illustrated in Fig. 2. Following the existing virtual try-on works [19, 28, 38], we adopt an inpainting framework, which reconstructs the target person image from a masked version, conditioned on the reference clothing image. Specifically, the main U-Net generates the target person image based on the input including a noise image ( $\mathbf{z}_t$ ), a resized dilated clothing-agnostic mask ( $\mathcal{R}(\mathbf{m}_d)$ ), and a latent agnostic map ( $\mathcal{E}(\mathbf{x}_a^p)$ ). Here, the dilated clothing-agnostic mask  $\mathbf{m}_d$  is produced by a random dilation mask augmentation.

To preserve fine clothing details, we leverage a frozen U-Net as a feature extractor [39, 45], referred to as the reference U-Net. We then integrate the clothing features of reference U-Net by concatenating the key and value from self-attention layers of the reference U-Net with those of corresponding layers in the main U-Net [51].

To obtain rich text descriptions, we introduce an LMM-driven captioning mechanism. Merely providing a detailed description on person images can result in redundancy, as

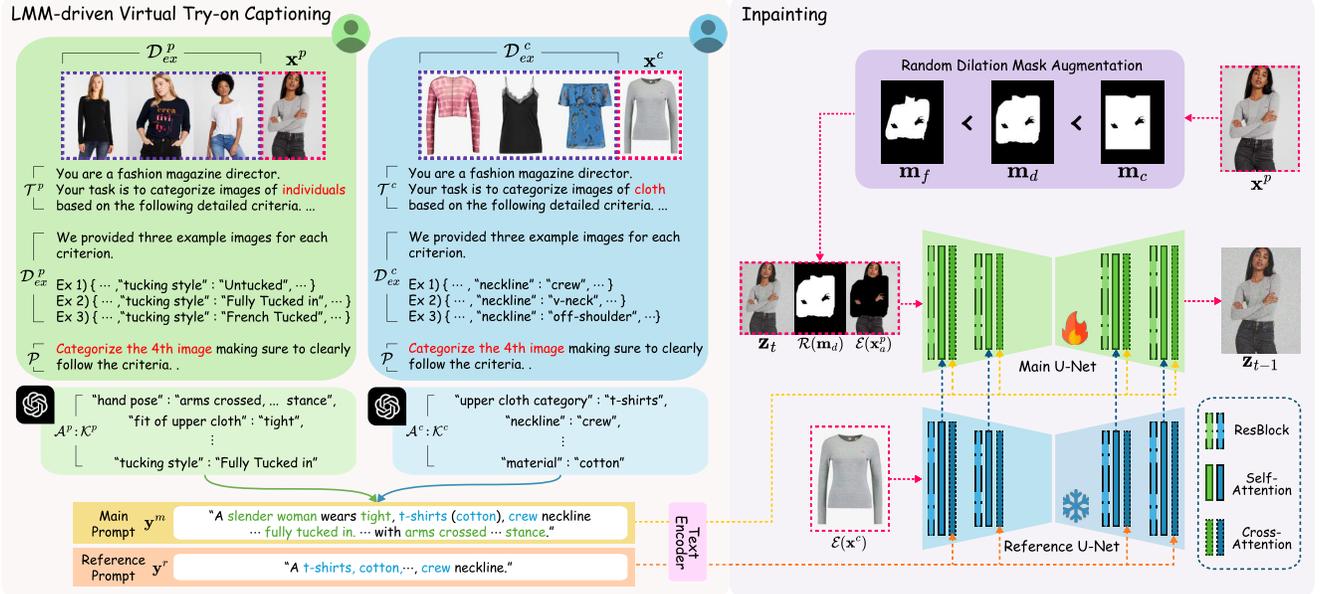


Figure 2. Overview of PromptDresser. By using LMM with in-context learning, we generate two types of captions specific to the person and clothing images. The reference prompt consisting of clothing captions is provided to the reference U-Net, while a main prompt, including both clothing and person, is input into the main U-Net. To preserve the details of clothing, we use a frozen U-Net as a feature extractor. To enable learning across a diverse range of masked images, we randomly dilate the inpainting mask.

it may encompass areas that remain unmasked. Furthermore, this approach fails to address the unique nature of unpaired person-clothing data in virtual try-on, where the information of the person’s original clothing and the new clothing can become entangled. Therefore, we designed our approach to separate information by pre-defining distinct attributes for the person and clothing.

The existing clothing-agnostic person representation [11, 29] fails to precisely remove structural features like clothing length, causing the generated images to conform closely to the shape of the original clothing, which limits the flexibility for manipulation. On the other hand, overly expanded mask regions make it difficult to accurately reconstruct the original person’s information. Therefore, we introduce random dilation mask augmentation, enabling the model to learn a range of mask images from coarse to fine. This approach allows for a prompt-aware mask generation (PMG) during the inference, providing preservation of the original person’s appearance irrelevant to the reference clothing.

### 3.3. LMM-driven Virtual Try-on Captioning

In this work, we propose to improve the quality and controllability of virtual try-on via generative textual prompt. The primary challenges are: 1) *ensuring that the text description focuses on the masked region* and 2) *excluding textual information about the existing clothing in the inference process*. A naive approach would be to instruct the LMM to describe the person image, which results in detailed infor-

mation about visible features in the unmasked regions, such as expression and hairstyle, as in “*The image showcases a bold fashion statement . . . large hoop earrings and curly, voluminous hair enhance the overall stylish and confident look.*”. Moreover, describing the entire image in this way can introduce textual conflicts when virtually fitting new clothing, as information about the person’s existing outfit may also be included.

To address this, we propose designing pre-defined attributes and tasking the LMM with generating captions based on these attributes. Using the LMM, we listed the attributes of both person and clothing images and carefully selected  $n^{\{p,c\}}$  representative attributes  $\mathcal{A}^{\{p,c\}} = \{a_1^{\{p,c\}}, \dots, a_{n^{\{p,c\}}}^{\{p,c\}}\}$ . Specifically, for person-related attributes, we prioritized those within the masked region, including hand pose, body shape, and tucking style, along with attributes that support editability. Due to the limited token length of the backbone’s text encoder (*i.e.*, CLIP), we focused on selecting global clothing features, such as category and material, rather than local details like logos. The main advantages of using the LMM to generate captions based on pre-defined attributes for both person and clothing are: 1) it enables the generation of **informative text descriptions specifically for the inpainting areas** and 2) the separation of information for person and clothing allows us to **create adequate prompts aligned to unpaired person-clothing scenarios**, enabling descriptions of individuals wearing new clothing. Furthermore, we instruct the LMM to provide a detailed description of the pose. By us-

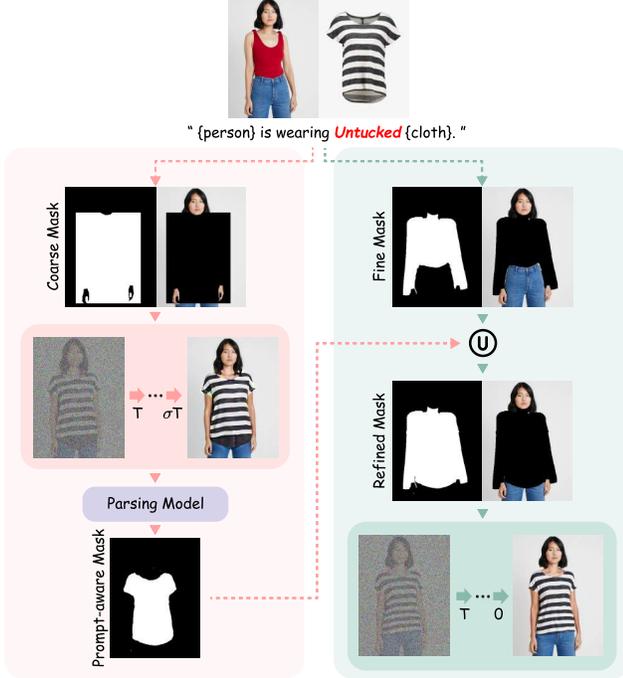


Figure 3. Prompt-aware mask generation for text-based manipulation. During the inference stage, PromptDresser takes a coarse mask as input and generates a prompt-aware mask. A refined mask for inpainting is then obtained by performing a union operation with a fine mask, ensuring minimal alteration to regions irrelevant to the clothing in the person image.

ing pose descriptions instead of DensePose [20], we retain the text-to-image backbone architecture and effectively reduce the errors associated with pose networks used in previous works, particularly in complex samples.

As depicted on the left of Fig. 2, we then utilize the in-context learning capability [8] of LMM models to generate rich, free-form captions for pre-defined attributes. Based on the observation that multi-modal models are proficient at predicting head categories such as gender [33], we carefully select  $N$  few-shot exemplar images, each exhibiting different subtle details such as tucking or rolling style. Human annotators then label each caption to capture these specific details accurately.

Therefore, for an arbitrary person or clothing image  $\mathbf{x} \in \{\mathbf{x}^p, \mathbf{x}^c\}$  and a LMM  $\mathcal{M}$ , the predicted captions  $\mathcal{K}^{\{p,c\}}$  is obtained through in-context learning as:

$$\mathcal{K}^{\{p,c\}} = \{k_1^{\{p,c\}}, \dots, k_n^{\{p,c\}}\} = \mathcal{M}(\mathbf{x} | \mathcal{P}, \mathcal{D}_{ex}^{\{p,c\}}, \mathcal{T}^{\{p,c\}}), \quad (2)$$

where  $\mathcal{P}$  is the input prompt,  $\mathcal{D}_{ex}^{\{p,c\}}$  is the in-context learning dataset consisting of few-shot examples for captioning, and  $\mathcal{T}^{\{p,c\}}$  is the task description.

Using the predicted captions, we construct textual prompts for both the reference U-Net and the main U-Net. For the reference U-Net, the reference prompt  $\mathbf{y}^r$  includes

only the clothing-specific captions  $\mathcal{K}^c$ . In contrast, the main prompt  $\mathbf{y}^m$  combines both the person-specific captions  $\mathcal{K}^p$  and the clothing-specific captions  $\mathcal{K}^c$ . In practice, we use the following format as the main prompt: “a {body shape ( $a_1^p$ )} {gender ( $a_2^p$ )} wears {cloth category ( $a_1^c$ )}, {material ( $a_2^c$ )}, ..., with {hand pose ( $a_n^p$ )}.”, where green and blue color denote person and clothing attributes, respectively. Therefore, our approach can generate the main prompt for the resulting image, even when arbitrary clothing is provided.

Each prompt is processed through a text encoder before being input to its respective U-Net. Additional details on the exemplar dataset, task descriptions, and templates are provided in the supplementary material.

### 3.4. Enhancing Adaptability via Mask Refinement

**Random Dilation Mask Augmentation.** We train a virtual try-on model via generative textual prompt that allows for text-based editing. However, we note the limitations in the commonly used masking approach, known as *clothing-agnostic person representation* [11], frequently used in virtual try-on methods. While such a masking approach effectively preserves the original person’s appearance, it also retains certain features from the original clothing such as length and fit. This constrained mask region causes the reference clothing to fit too closely to the mask boundaries during training, making the new clothing mimic the shape of the original clothing and creating potential conflicts during manipulation.

To address these issues, we propose random dilation mask augmentation. As illustrated in Fig 2, we introduce a coarse mask  $\mathbf{m}_c$  and a fine mask  $\mathbf{m}_f$  to enable learning across a diverse range of masked images. We randomly dilate the fine mask, ensuring it does not extend beyond the boundaries of the coarse mask. The dilated mask  $\mathbf{m}_d$  used for training is represented as follows:

$$\mathbf{m}_d = (\mathbf{m}_f \oplus^n \mathbf{b}) \cap \mathbf{m}_c, \quad (3)$$

where  $\oplus^n$  denotes  $n$ -iterated dilation with a structuring element  $\mathbf{b}$  [23], up to a sufficiently large but finite  $n$ .

**Prompt-aware Mask Generation.** For the inference stage, we introduce a novel coarse-to-fine generation approach to effectively preserve the original person’s appearance while allowing flexible text-based image manipulation. As illustrated on the left side of Fig. 3, we begin inference with a coarse mask, aiming to create an initial approximation of the clothing region aligned with the text prompt. For efficiency, we apply early stopping in the denoising process, running it only from timestep  $T$  to  $\sigma T$ , where  $\sigma \in [0, 1)$ . We then approximate  $\hat{\mathbf{z}}_0$ , decode it to  $\hat{\mathbf{x}}_0$ , and segment the region of interest using an off-the-shelf human parsing model. By taking the union of this mask with an existing

clothing-agnostic person representation (fine mask), we acquire a refined mask, which subsequently serves as the inpainting mask for generating the final output. This method achieves a balance between precision and efficiency, improving the alignment of the output with the text prompt.

## 4. Experiment

**Benchmarks.** We train PromptDresser separately on VITON-HD [11] and DressCode [37] datasets, and SHHQ-1.0 is used to evaluate the generalizability of the model trained on VITON-HD [28].

We compare our model to two GAN-based models (HR-VITON [29], GP-VTON [50]) and four diffusion-based models (LADI-VTON[38], DCI-VTON [19], Stable-VITON [28], and IDM-VTON [12]). We use pre-trained weights if available; otherwise, we re-implement them using official code. LADI-VTON, DCI-VTON, and Stable-VITON, all based on Stable Diffusion 1.5, generate images at  $512 \times 384$  resolution. To ensure a fair comparison, we upscale the outputs to  $2 \times$  using Real-ESRGAN [49].

**Implementation Details.** We utilize a frozen SDXL [39] and SDXL inpainting model [26] as the reference and main U-Net, respectively. During inference, we set the denoising step as 30 with  $\sigma$  set to 0.5 for prompt-aware mask generation. To maintain overall pose consistency, we retain hand and foot details within the inpainting mask by Sapiens [27]. Additionally, we use GPT-4o [1] to automatically generate high-quality captions for pre-defined attributes across all experimental datasets.

### 4.1. Comparison with Baselines

**Qualitative Results** Fig. 4 shows a comparative analysis of our model and baseline models trained on the VITON-HD [11] dataset (first row), evaluated both on the VITON-HD test dataset and the SHHQ-1.0 [16] dataset (second row). The baselines either closely follow the original clothing shape or lack detailed clothing features (especially on SHHQ-1.0). In contrast, our model accurately captures the details and shape of the reference clothing while preserving the features of the original person image, such as background and pose. Additionally, we compared images of the upper body, lower body, and dresses on the DressCode [37] dataset in Fig. 5. Across these images, the baselines tend to follow to the original clothing shape. Specifically in the lower body, baselines generate unnatural jeans by following the existing pants shape even when provided with a short skirt; similarly, for dresses, models including OOTDiffusion and IDM-VTON produce excessively long dresses. This demonstrates that well-aligned and rich textual information enhances model generalization performance, and our adaptive mask disregard the original clothing shape but retain the person’s appearance.

Dataset Method	VITON-HD				SHHQ-1.0	
	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	KID $\downarrow$	FID $\downarrow$	KID $\downarrow$
HR-VITON	<b>0.8767</b>	0.1153	12.24	3.74	53.12	31.3
GP-VTON	<u>0.8763</u>	0.1279	9.95	1.80	-	-
LADI-VTON	0.8630	0.1393	9.95	2.20	26.23	6.91
DCI-VTON	0.8712	0.1245	9.46	1.61	26.39	7.37
StableVITON	0.8757	0.1253	9.84	2.07	23.97	7.30
OOTDiffusion	0.8424	0.1200	9.36	<u>1.04</u>	24.10	<u>6.20</u>
IDM-VTON	0.8626	<u>0.1023</u>	9.20	1.27	24.76	8.06
IDM-VTON + our text	0.8650	<u>0.1025</u>	<u>9.15</u>	1.26	23.78	7.46
Ours <sub>pose</sub>	0.8623	<b>0.1013</b>	<u>9.15</u>	1.21	<u>23.65</u>	6.79
Ours	0.8530	0.1165	<b>8.64</b>	<b>0.75</b>	<b>23.46</b>	<b>6.18</b>

Table 1. Quantitative comparisons trained on VITON-HD. **Bold** and underline denote the best and second best result, respectively.

**Quantitative Results** For quantitative evaluation, we use four metrics: SSIM [13], LPIPS [56] for the paired setting, and FID [24], and KID [5] for the unpaired setting. Tables 1 and 2 show the comparison results between PromptDresser and the baseline models on the VITON-HD and DressCode datasets, respectively. Our model consistently and significantly outperforms existing models in the unpaired setting (*i.e.*, FID and KID). Notably, applying the generative textual prompt to IDM-VTON, we observed performance improvements across all metrics except LPIPS, highlighting our method’s effectiveness and the benefit of leveraging textual information about person and clothing attributes to enhance the realism of the generated images.

While we include pose descriptions and hand as hints, there is a slight performance drop in the paired setting due to minor misalignment in the pose. We also experimented with expanding the first convolution layer of our model to 13 channels to include DensePose as an additional input [28], named as Ours<sub>pose</sub>. Ours<sub>pose</sub> achieved the highest LPIPS score, as shown in Table 1. However, a slight decrease in performance is observed in the unpaired setting due to modifications in the backbone architecture, which, along with pose-related errors, indicates a drop in generalization performance. As shown in qualitative evaluations, visual differences were minimal, so we selected the model without spatial pose conditioning as our final choice, given its superior scores in the unpaired setting.

### 4.2. Further Analysis on PromptDresser

**Evaluation on Text Alignment** To validate our method’s editing capability, we generated edited versions of 2,032 test images from VITON-HD by fixing a specific attribute (*e.g.*, “tucking style”) to a caption (*e.g.*, “untucked”) and then evaluated whether the captions generated by the LMM for these edited images matched the intended caption. We conducted experiments with two settings: (i) setting the tucking style to “untucked” (*i.e.*, the top is worn outside the pants) and (ii) setting the clothing fit to “tight fit”. The “Base Ratio” is the proportion of the 2,032 test images in which the LMM (*i.e.*, GPT-4o) identified a specific caption. For example, if the LMM identifies 1,016 out of the 2,032 test images to have an “untucked” tucking style, the base



Figure 4. Qualitative comparison with baselines trained on VITON-HD dataset (first row: VITON-HD, second row: SHHQ-1.0).

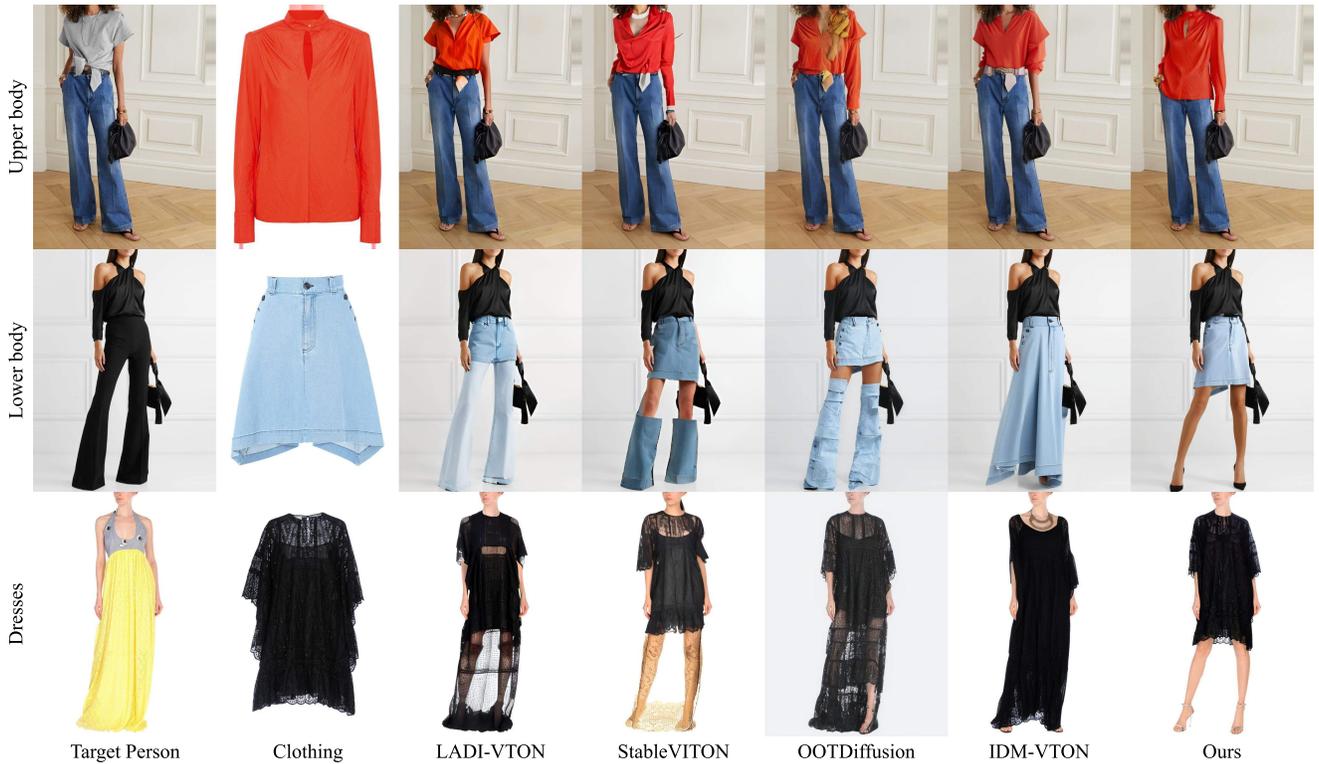


Figure 5. Qualitative comparison with baselines trained on DressCode dataset.

Dataset Metric	Upper body				Lower body				Dresses			
	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	KID $\downarrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	KID $\downarrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	KID $\downarrow$
LADI-VTON	0.9116	0.0995	15.06	3.78	<u>0.8999</u>	0.1072	14.78	2.97	0.8686	0.1348	14.70	3.35
StableVITON	0.9119	0.0925	14.13	3.34	0.8916	0.1104	14.91	3.10	<u>0.8745</u>	0.1194	12.87	2.55
OOTDiffusion	<u>0.9065</u>	0.0755	15.04	<u>2.96</u>	0.8990	0.0741	14.05	<u>2.66</u>	0.8554	0.1134	16.40	4.20
IDM-VTON	<b>0.9267</b>	<b>0.0522</b>	<u>11.91</u>	1.75	<b>0.9106</b>	<b>0.0608</b>	<u>13.89</u>	2.69	<b>0.8787</b>	<b>0.0916</b>	<u>11.36</u>	<u>1.38</u>
Ours	0.9160	<u>0.0606</u>	<b>11.30</b>	<b>0.79</b>	0.8990	<u>0.0703</u>	<b>12.71</b>	<b>1.44</b>	0.8660	<u>0.0945</u>	<b>11.04</b>	<b>1.12</b>

Table 2. Quantitative comparisons trained on DressCode dataset. **Bold** and underline denote the best and the second best result, respectively.

Methods	“Untucked”	“Tight fit”
Base Ratio	44.64%	23.13%
LADI-VTON	50.78%	37.5%
IDM-VTON	46.31%	43.85%
Ours <sub>pose</sub>	62.84%	44.09%
Ours	<b>89.42%</b>	<b>66.98%</b>

Table 3. Evaluation on Text Alignment.

Methods	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	KID $\downarrow$
Ours w/ holi.	0.8440	0.1345	9.41	1.33
Ours w/o PMG	0.8262	0.1538	9.11	0.98
Ours	<b>0.8530</b>	<b>0.116</b>	<b>8.64</b>	<b>0.75</b>

Table 4. Comparison of qualitative results with ablated methods.



Figure 6. Visual comparisons for ablation studies.

ratio for ‘untucked’ would be 50%. In addition to two baselines, LADI-VTON and IDM-VTON, we compared our method with Ours<sub>pose</sub> to examine the effect of DensePose information on text-editability with the same text prompt.

Table 3 shows that our model achieved significantly higher accuracy for both attributes. Notably, the baseline models scored 50.78% and 46.31% for the “untucked”, performing similarly to the base ratio. This result underscores the limitations of conventional agnostic masks that restrict manipulation to pre-defined areas, especially for adjusting clothing length. Furthermore, for the “tight fit” attribute, IDM-VTON and ours<sub>pose</sub> both achieved around 44%, while our method reached 66.98%, indicating that DensePose can hinder accurate text-based editing. Therefore, the results demonstrate that our approach addresses the limitations of conventional agnostic mask, enabling for accurate manipulation based on text prompts.

**Ablation Study** We investigate the key design choices of our method through ablation studies on the VITON-HD dataset, as shown in Fig. 6 and Table 4. Our comparisons include: (i) using a holistic text description (*i.e.*, a overall description of the image in detail) for both person and clothing images with the LMM and (ii) excluding the prompt-aware mask generation (PMG) during inference.

Fig. 6 shows that using a holistic description can fail to capture complex poses accurately. Moreover, our method with holistic description shows the lowest FID and KID scores in Table 4. These results indicate that text prompts describing the images using LMM without in-context learning are insufficient to represent fine-grained details including poses. Since Ours without PMG employs only the

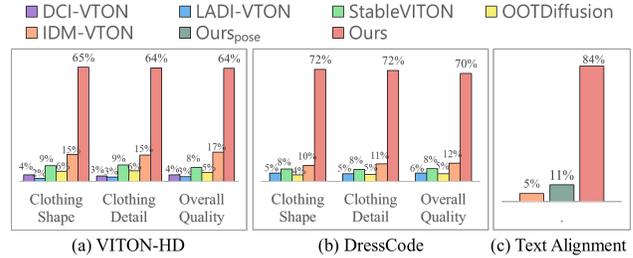


Figure 7. User Study Results. We requested users to identify the most appropriate samples for each criterion from (a) the VITON-HD, SHHQ, and (b) DressCode dataset. We performed (c) a qualitative assessment of text alignment using the VITON-HD.

coarse masks, the regions that need to be preserved are often generated differently, as shown in Fig. 6 (*e.g.*, a belt absent on the target person). Table 4 demonstrate that PMG improves the performance significantly in all metrics.

**User Study** We further assess our method and other baselines through a user study involving 40 participants, using VITON-HD and DressCode datasets. As shown in Fig. 7 (a) and (b), we first compare five and six models on VITON-HD and DressCode datasets, respectively. The participants selected the best results based on three aspects: (i) clothing shape, (ii) clothing detail, and (iii) overall quality. As shown in Figure 7, our method shows a significantly high preference across three aspects. To evaluate the text editability of the models, we compare our method with IDM-VTON which uses SDXL-based architecture, and Ours<sub>pose</sub> (with added pose information). We asked participants to select the sample that best matched the conditions of “untucked”, “tight fit”, and “sleeve rolled up” for each model. As shown in Fig. 7 (c), our method achieved an 84% preference, visually demonstrating superior text editability.

**Limitations and Future Work** This paper demonstrated the utility of generative text prompts using LMMs with in-context learning. However, due to the limitation of SDXL that accepts only up to 77 text tokens, we manually constructed in-context dataset. Inspired by recent studies [32] on configuring in-context sequences to enhance in-context learning performance, exploring in-context data construction for virtual try-on can be a promising research direction. Stable Diffusion 3 [15] has employed a T5 text encoder [40] capable of covering long context lengths. Utilizing well-configured in-context learning on such models is expected to further improve the performance of virtual try-on.

## 5. Conclusion

This paper introduces a novel virtual try-on model that leverages generative text prompt and advanced masking methods. Along with rich text descriptions from a large multimodal model, our approach not only improves performance but also enables versatile manipulation of virtual try-on results. Our dilated mask addresses the issue of gen-

erated clothing following too closely to the person’s existing clothing, and allows for a more natural overlay. We propose a prompt-aware mask generation technique, which enhances diversity while preserving the person’s original appearance. Our method achieves state-of-the-art results across multiple datasets, highlighting the effectiveness of generative textual prompt for virtual try-on.

## References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 6
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 3
- [3] Shuai Bai, Huiling Zhou, Zhikang Li, Chang Zhou, and Hongxia Yang. Single stage virtual try-on via deformable attention flows. In *European Conference on Computer Vision*, pages 409–425. Springer, 2022. 2, 3
- [4] Simone Bianco, Luigi Celona, Marco Donzella, and Paolo Napoletano. Improving image captioning descriptiveness by ranking and llm-based fusion. *arXiv preprint arXiv:2306.11593*, 2023. 3
- [5] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 6
- [6] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 3
- [7] 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. 3
- [8] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 5
- [9] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart- $\alpha$ : Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 2, 3
- [10] 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. *arXiv preprint arXiv:2403.12965*, 2024. 2, 3
- [11] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *CVPR*, pages 14131–14140, 2021. 2, 3, 4, 5, 6, 1
- [12] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for virtual try-on. *arXiv preprint arXiv:2403.05139*, 2024. 2, 3, 6
- [13] Nicki Skafte Detlefsen, Jiri Borovec, Justus Schock, Ananya Harsh Jha, Teddy Koker, Luca Di Liello, Daniel Stancil, Changsheng Quan, Maxim Grechkin, and William Falcon. Torchmetrics-measuring reproducibility in pytorch. *Journal of Open Source Software*, 7(70):4101, 2022. 6
- [14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2
- [15] 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. 2024. 8
- [16] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen-Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. *arXiv preprint*, arXiv:2204.11823, 2022. 6, 2
- [17] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *CVPR*, pages 8485–8493, 2021. 2, 3
- [18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [19] 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. *arXiv preprint arXiv:2308.06101*, 2023. 2, 3, 6
- [20] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, pages 7297–7306, 2018. 5
- [21] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *CVPR*, pages 7543–7552, 2018. 2
- [22] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *ICCV*, pages 10471–10480, 2019. 2
- [23] Robert M Haralick, Stanley R Sternberg, and Xinhua Zhuang. Image analysis using mathematical morphology. *IEEE transactions on pattern analysis and machine intelligence*, (4):532–550, 1987. 5
- [24] 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. 2017. 6
- [25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3
- [26] Sdxl inpainting 0.1. <https://huggingface.co/diffusers/stable-diffusion-xl-1.0-inpainting-0.1>. 6
- [27] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and

- Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pages 206–228. Springer, 2025. 6
- [28] Jeongho Kim, Guojung Gu, Minh 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, 3, 6
- [29] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *ECCV*, pages 204–219. Springer, 2022. 2, 3, 4, 6
- [30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3
- [31] Kedan Li, Jeffrey Zhang, Shao-Yu Chang, and David Forsyth. Controlling virtual try-on pipeline through rendering policies. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5866–5875, 2024. 3
- [32] Li Li, Jiawei Peng, Huiyi Chen, Chongyang Gao, and Xu Yang. How to configure good in-context sequence for visual question answering. In *CVPR*, pages 26710–26720, 2024. 8
- [33] Shikai Li, Jianglin Fu, Kaiyuan Liu, Wentao Wang, Kwan-Yee Lin, and Wayne Wu. Cosmicman: A text-to-image foundation model for humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6955–6965, 2024. 5
- [34] Yuhan Li, Hao Zhou, Wenxiang Shang, Ran Lin, Xuanhong Chen, and Bingbing Ni. Anyfit: Controllable virtual try-on for any combination of attire across any scenario. *arXiv preprint arXiv:2405.18172*, 2024. 2, 3
- [35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3, 1
- [36] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Da Li, Pengcheng Lu, Tao Wang, Linmei Hu, Minghui Qiu, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023. 3
- [37] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *ECCV*, pages 2231–2235, 2022. 6, 2
- [38] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. *arXiv preprint arXiv:2305.13501*, 2023. 2, 3, 6
- [39] 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. *arXiv preprint arXiv:2307.01952*, 2023. 2, 3, 6
- [40] 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. 8
- [41] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2
- [42] 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
- [43] 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. 3
- [44] 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. 2, 3
- [45] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 3
- [46] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3
- [47] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *ECCV*, pages 589–604, 2018. 2
- [48] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [49] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 6
- [50] 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 *CVPR*, pages 23550–23559, 2023. 3, 6
- [51] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of*

- the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024. 3
- [52] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *CVPR*, pages 7850–7859, 2020. 2, 3
- [53] Jinliang Yao and Haonan Zheng. Lc-vton: Length controllable virtual try-on network. *IEEE Access*, 2023. 3
- [54] Hu Ye, Jun Zhang, Sibao 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. 3
- [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. 3
- [56] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6
- [57] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3
- [58] 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*, pages 4606–4615, 2023. 3
- [59] Luyang Zhu, Yingwei Li, Nan Liu, Hao Peng, Dawei Yang, and Ira Kemelmacher-Shlizerman. M&m vto: Multi-garment virtual try-on and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1346–1356, 2024. 3

# PromptDresser: Improving the Quality and Controllability of Virtual Try-On via Generative Textual Prompt and Prompt-aware Mask

## Supplementary Material

### A. Details of LMM-driven Virtual Try-on Captioning

We provide detailed explanations of the exemplar datasets, task descriptions, and templates for the categories of upper body, lower body, and dresses in Fig. 8, 9, and 10, respectively. We first gave the LMM model the instruction to identify and list detailed attributes of a given person image, including components, such as the facial expression, skin color, clothing logos. We then selected attributes related to the masked regions of the person image or associated with the style of wearing the clothing, such as pose, hair length, and tucking style. For clothing images, we excluded attributes describing fine details, such as logo shapes or patterns, but instead focused on high-level attributes, such as the clothing category or sleeve.

### B. Additional Details on User Study

In our user study, we recruited 40 participants to evaluate the images generated by the baselines across 30 questions. For each question, participants selected the model that best addressed the specified criteria.

For questions 1-25, participants compared the images from multiple datasets. Questions 1–10 featured images from six models (*i.e.*, DCI-VTON, LADI-VTON, Stable-VITON, OOTDiffusion, IDM-VTON, and Ours), based on the VITON-HD and SHHQ-1.0 datasets. Questions 11-25 include three categories of DressCode dataset: upper-body clothing (Questions 11-15), lower-body clothing (Questions 16-20), and dresses (Questions 21-25). For these questions, images from five models were compared, excluding DCI-VTON.

Participants answered the following three questions for each image set:

- Clothing shape: Select the image that best reflects the length and shape of the given garment.
- Clothing details: Select the image that best reflects the text, texture, and pattern of the given garment.
- Overall quality: Select the image of the best overall quality.

For questions 26-30, participants evaluated images generated using the VITON-HD dataset and selected the one that best matched the style described as “untucked, tight fit, and sleeve rolled up.”

Methods	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	KID $\downarrow$
Ours w/ LLaVA	<b>0.8533</b>	0.1222	8.96	1.04
Ours w/ GPT-4o	0.8530	<b>0.116</b>	<b>8.64</b>	<b>0.75</b>

Table 5. Ablation Results on VITON-HD [11] dataset.

### C. Additional Experimental Results

**Comparison to other LMMs.** In this paper, we utilize GPT-4o to generate captions for all experimental datasets. To investigate whether our model exhibits a high dependency on GPT-4o in test time, we evaluated it using text prompts generated by an open-source LMM called LLaVA [35]. As shown in Table 5, prompts from LLaVA exhibit slightly degraded performances in the unpaired setting (*i.e.*, FID and KID) but achieve comparable scores in a paired setting (*e.g.*, SSIM and LPIPS), compared to GPT-4o. This demonstrates that the proposed textual prompt can effectively be generated by open-source LMMs such as LLaVA, other than GPT-4o.

$\sigma$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	KID $\downarrow$
0.8	0.855	0.1170	8.71	0.78
0.7	0.855	0.1166	8.65	0.78
0.6	<b>0.856</b>	0.1166	8.66	0.78
0.5	0.853	0.1165	8.64	<b>0.75</b>
0.4	<b>0.856</b>	0.1162	8.65	0.78
0.3	<b>0.856</b>	<b>0.1161</b>	<b>8.63</b>	<b>0.75</b>

Table 6. Ablation results for  $\sigma$  values.

**Ablation Study on  $\sigma$ .** In this paper, we introduce a novel prompt-aware mask to preserve the original person’s appearance and enable flexible text-based image manipulation. In generating the mask, we apply early stopping for computational efficiency and adjust the number of inference steps through a hyper-parameter  $\sigma$ . As the value of  $\sigma$  increases, the generation time for the prompt-aware mask decreases. We set the number of denoising steps to 30 across all configurations. Table 6 shows the performance behavior based on different  $\sigma$  values. The lowest  $\sigma$  value (0.3) results in more accurate refined masks, achieving the best performance across all metrics. However, slightly reduced performance can be traded off for efficient inference times. In this paper, we adopt  $\sigma = 0.5$ , which offers inference efficiency while maintaining FID and KID values comparable to those achieved with  $\sigma = 0.3$ .

**Additional Qualitative Comparisons.** We present addi-

tional qualitative results in Fig. 11 and 12. The first three rows in Fig. 11 depict generated images on the VITON-HD [11] dataset using a model trained on the same dataset, while the fourth and fifth rows show generated images on the SHHQ-1.0 [16] dataset. Our model consistently generates the most realistic images, even for complex poses (rows 1 and 2), and addresses the issue of following the shape of the original clothing (rows 3 and 4). Notably, in the third row, only our model accurately captures the shape of the given cropped top. Moreover, Fig. 12 shows additional results for the upper body, lower body, and dresses categories on the DressCode [37] dataset. Similar to the results on the VITON-HD dataset, our PromptDresser accurately generate the length of the clothing and mitigate the constraint the model follows the original clothing’s shape, highlighting the effectiveness of our rich text prompts and a novel mask refinement process.

**Additional text-based editing Results.** Fig. 13 and 14 demonstrate the text-editing capability of our PromptDresser on VITON-HD and lower body category of the DressCode datasets, respectively. Fig. 13 shows variations in tucking styles on the VITON-HD dataset, where the given clothing is generated based on the text prompts “fully tucked in”, “untucked”, and “french tuck”. Fig. 14 presents variations on the DressCode dataset, including “loose fit,” “tight fit,” and “pants rolled up.” The generated results demonstrate accurate and text-based editing capability of PromptDresser.

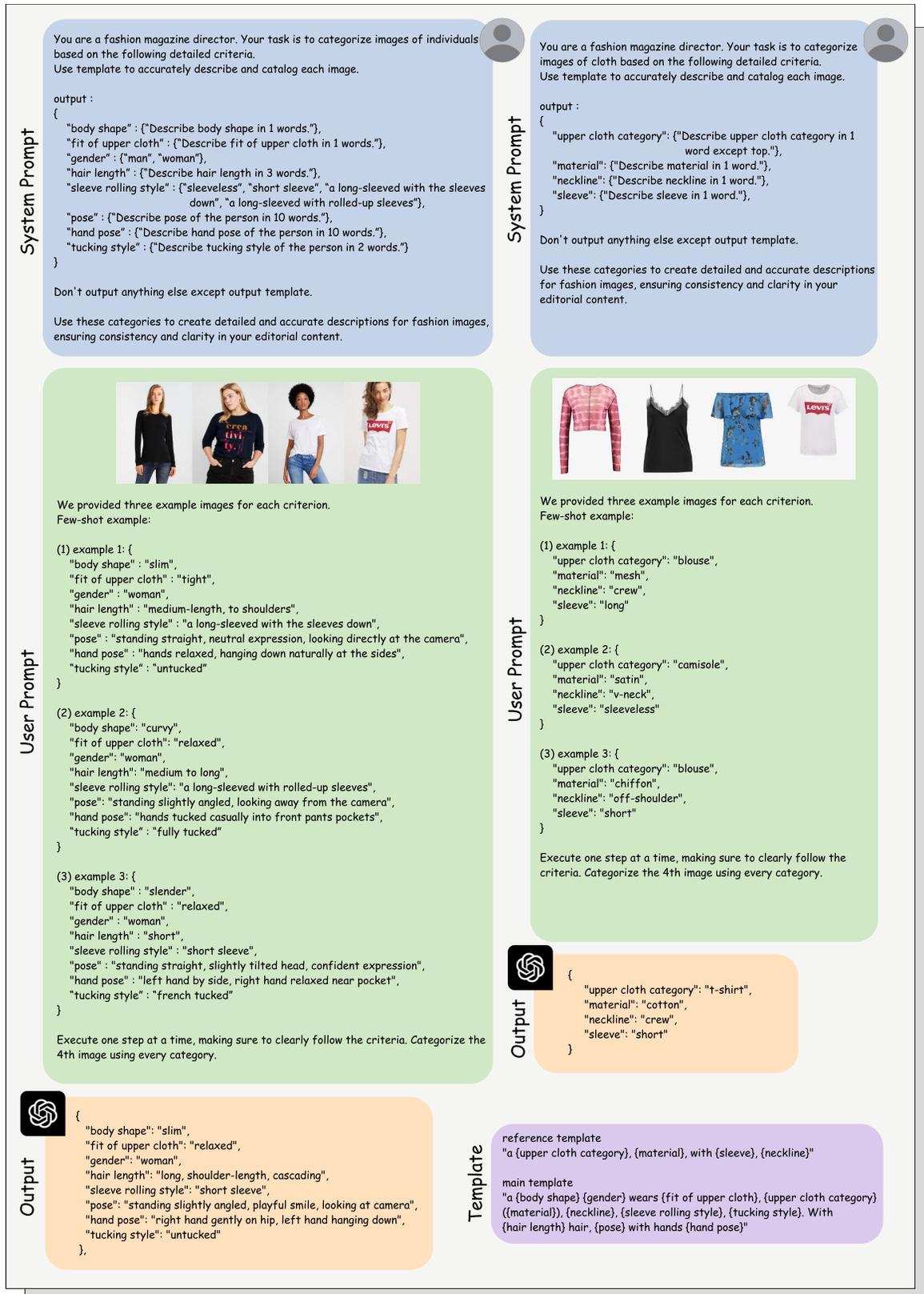


Figure 8. Detailed explanation of the exemplar dataset, task description, and templates for the upper body category.

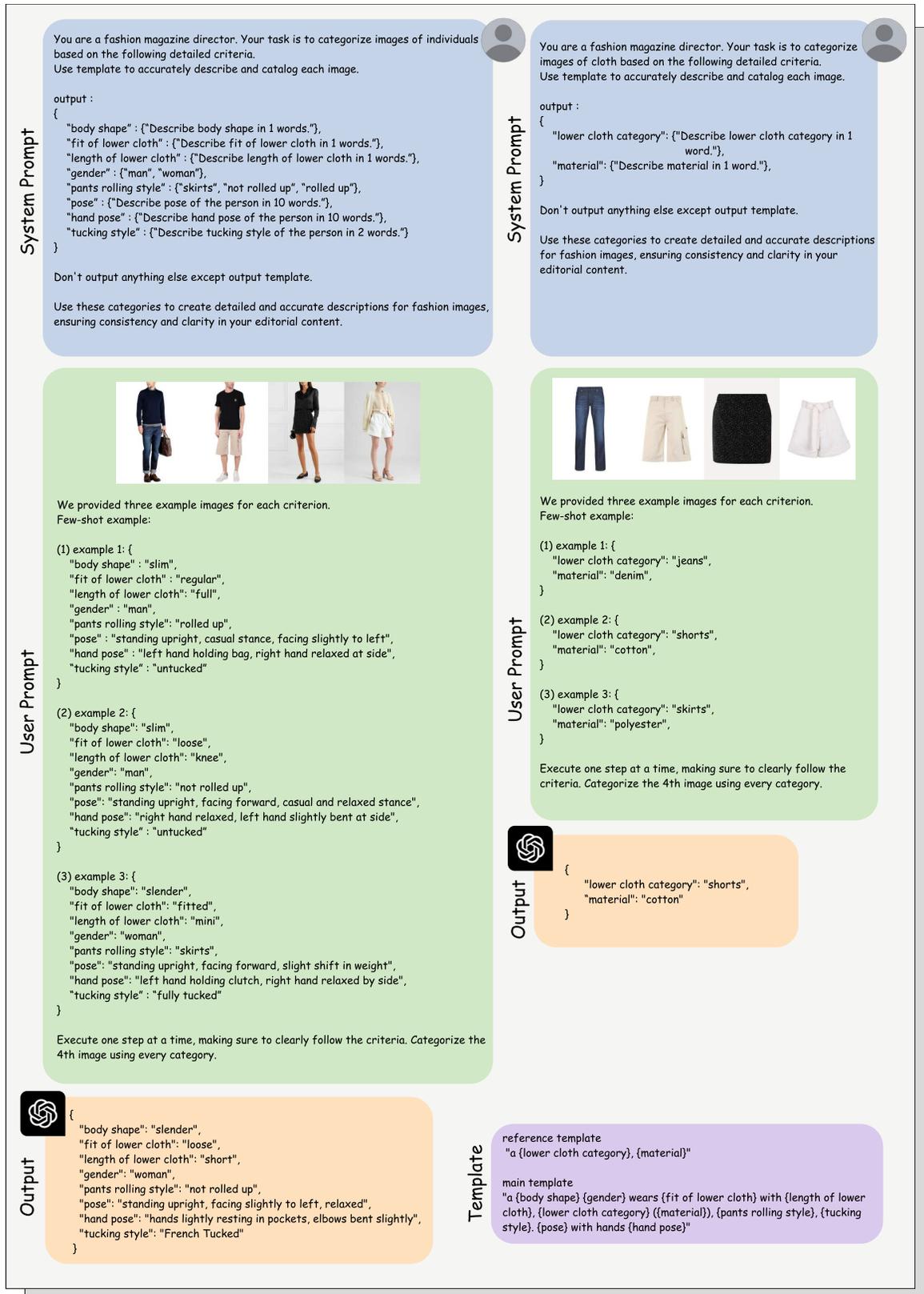


Figure 9. Detailed explanation of the exemplar dataset, task description, and templates for the lower body category.

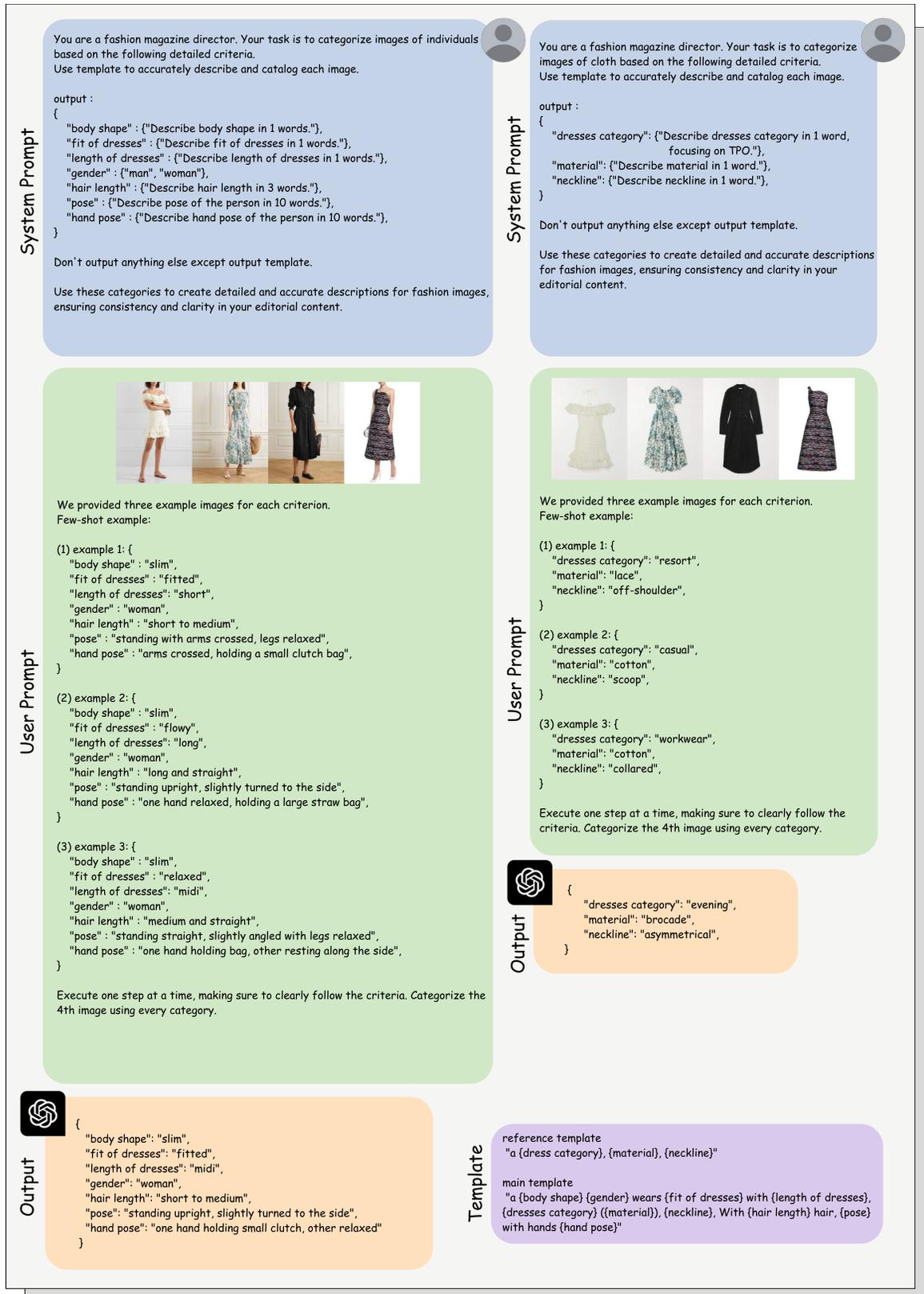


Figure 10. Detailed explanation of the exemplar dataset, task description, and templates for the dresses category.



Target Person    Clothing    HR-VTON    LADI-VTON    DCI-VTON    StableVITON    OOTDiffusion    IDM-VTON    Ours

Figure 11. Qualitative comparison with baselines trained on VITON-HD dataset (first / second / third row: VITON-HD, fourth / fifth row: SHHQ-1.0)



Figure 12. Qualitative comparison with baselines trained on DressCode dataset.



Figure 13. Additional text-based editing results for the upper body category of the VITON-HD dataset.



Figure 14. Additional text-based editing results for the lower body category of the DressCode dataset.