(Article begins on next page)

Multimodal Garment Designer:
Human-Centric Latent Diffusion Models for Fashion Image Editing

Alberto Baldrati[1,3,*], Davide Morelli[2,3,*], Giuseppe Cartella[2], Marcella Cornia[2],
Marco Bertini[1], Rita Cucchiara[2,4]

[1]University of Florence, Italy   [2]University of Modena and Reggio Emilia, Italy
[3]University of Pisa, Italy   [4]IIT-CNR, Italy

[1]{name.surname}@unifi.it   [2]{name.surname}@unimore.it

Figure 1: In this work, we propose a novel multimodal garment designer framework based on latent diffusion models that can generate a novel fashion image conditioned on text, human keypoints, and a garment sketch.

Abstract

Fashion illustration is used by designers to communicate their vision and to bring the design idea from conceptualization to realization, showing how clothes interact with the human body. In this context, computer vision can thus be used to improve the fashion design process. Differently from previous works that mainly focused on the virtual try-on of garments, we propose the task of multimodal-conditioned fashion image editing, guiding the generation of human-centric fashion images by following multimodal prompts, such as text, human body poses, and garment sketches. We tackle this problem by proposing a new architecture based on latent diffusion models, an approach that has not been used before in the fashion domain. Given the lack of existing datasets suitable for the task, we also extend two existing fashion datasets, namely Dress Code and VITON-HD, with multimodal annotations collected in a semi-automatic manner. Experimental results on these new datasets demonstrate the effectiveness of our proposal, both in terms of realism and coherence with the given multimodal inputs. Source code and collected multimodal annotations are publicly available at: https://github.com/aimagelab/multimodal-garment-designer.

1. Introduction

Computer Vision research has always paid much attention both to the human person and to fashion-related problems, especially working on the recognition and retrieval of clothing items [11, 24], the recommendation of similar garments [8, 18, 41], and the virtual try-on of clothes and accessories [7, 13, 29, 30, 50, 55]. In the last years, some research efforts have been dedicated to the text-conditioned image editing task where, given a model image and a textual description of a garment, the goal is to generate the

---

*Equal contribution.

input model wearing a new clothing item corresponding to the given textual description. In this context, only a few works [19, 35, 59] have been proposed, exclusively employing GAN-based approaches for the generative step.

Recently, diffusion models [10, 17, 32, 44] have attracted more and more attention due to their outstanding generation capabilities, allowing the improvement of a variety of downstream tasks in several domains, while their applicability to the fashion domain is still unexplored. Many different solutions have been introduced and can roughly be identified based on the denoising conditions used to guide the diffusion process, which can enable greater control of the synthesized output. A particular type of diffusion model has been proposed in [39] that, instead of applying the diffusion process in the pixel space, defines the forward and the reverse processes in the latent space of a pre-trained autoencoder, becoming one of the leading choices thanks to its reduced computational cost. Although this solution can generate highly realistic images, it does not perform well in human-centric generation tasks and can not deal with multiple conditioning signals to guide the generation phase.

In this work, we address an extended and more general framework and define the new task of *multimodal-conditioned fashion image editing*, which allows guiding the generative process via multimodal prompts while preserving the identity and body shape of a given person (Fig. 1). To tackle this task, we introduce a new architecture, called Multimodal Garment Designer (MGD), that emulates the process of a designer conceiving a new garment on a model shape, based on preliminary indications provided through a textual sentence or a garment sketch. In particular, starting from Stable Diffusion [39], we propose a denoising network that can be conditioned by multiple modalities and also takes into account the pose consistency between input and generated images, thus improving the effectiveness of human-centric diffusion models.

To address the newly proposed task, we present a semi-automatic framework to extend existing datasets with multimodal data. Specifically, we start from two famous virtual try-on datasets (*i.e.* Dress Code [30] and VITON-HD [7]) and extend them with textual descriptions and garment sketches. Experimental results on the two proposed multimodal fashion benchmarks show both quantitatively and qualitatively that our proposed architecture generates high-quality images based on the given multimodal inputs and outperforms all considered competitors and baselines, also according to human evaluations.

To sum up, our contributions are as follows: (1) We propose a novel task of multimodal-conditioned fashion image editing, which entails the use of multimodal data to guide the generation. (2) We introduce a new human-centric generative architecture based on latent diffusion models, capable of following multimodal prompts while preserving the

model's characteristics. (3) To tackle the new task, we extend two existing fashion datasets with textual sentences and garment sketches devising a semi-automatic annotation framework. (4) Extensive experiments demonstrate that the proposed approach outperforms other competitors in terms of realism and coherence with multimodal inputs.

2. Related Work

Text-Guided Image Generation. Creating an image that faithfully reflects the provided textual prompt is the goal of text-to-image synthesis. In this context, early approaches were based on GANs [48, 54, 56, 58], while most recent solutions exploit the effectiveness of diffusion models [33, 37, 39]. In the fashion domain, only a few attempts of text-to-image synthesis have been proposed [19, 35, 59]. Specifically, Zhu *et al.* [59] presented a GAN-based solution that generates the final image conditioned on both textual descriptions and semantic layouts. A different approach is the one introduced in [35], where a latent code regularization technique is employed to augment the GAN inversion process by exploiting CLIP textual embeddings [36] to guide the image editing process. Instead, Jiang *et al.* [19] proposed an architecture that synthesizes full-body images by mapping the textual descriptions of clothing items into one-hot vectors, limiting however the expressiveness capability of the conditioning signal.

Multimodal Image Generation with Diffusion Models. A related line of works aims to condition existing diffusion models on different modalities thus enabling greater control over the generation process [5, 6, 27, 31, 51]. For example, Choi *et al.* [6] proposed to refine the generative process of an unconditional denoising diffusion probabilistic model [32] by matching each latent variable with the given reference image. On a different line, the approach introduced in [27] adds noise to a stroke-based input and applies the reverse stochastic differential equation to synthesize images, without additional training. Wang *et al.* [51], instead, proposed to learn a highly semantic latent space and perform conditional finetuning for each downstream task to map the guidance signals to the pre-trained space. Other recent works proposed to add sketches as additional conditioning signals, either concatenating them with the model input [5] or training an MLP-based edge predictor to map latent features to spatial maps [49].

Among contemporary works that aim to condition pre-trained latent diffusion models, ControlNet [57] proposes to extend the Stable Diffusion model [39] with an additional conditioning input. This process involves creating two versions of the original model's weights: one that remains fixed and unchanged (locked copy) and another that can be updated during training (trainable copy). The purpose of this is to allow the trainable version to learn the newly introduced condition while the locked version retains the origi-
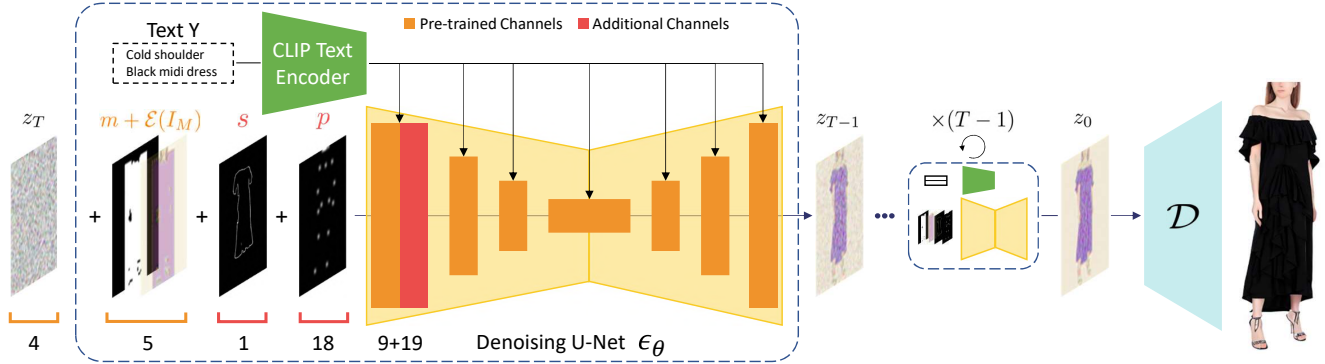
Figure 2: Overview of the proposed Multimodal Garment Designer (MGD), a human-centric latent diffusion model conditioned on multiple modalities (*i.e.* text, human pose, and garment sketch).

nal model knowledge. On the other hand, T2I-Adapter [31] learns modality-specific adapter modules that enable Stable Diffusion conditioning on new modalities.

In contrast, we focus on the fashion domain and propose a human-centric architecture based on latent diffusion models that directly exploits the conditioning of textual sentences and other modalities such as human body poses and garment sketches.

3. Proposed Method

In this section, we propose a novel task to automatically edit a human-centric fashion image conditioned on multiple modalities. Specifically, given the model image $I \in \mathbb{R}^{H \times W \times 3}$, its pose map $P \in \mathbb{R}^{H \times W \times 18}$ where the channels represent the human keypoints, a textual description Y of a garment, and a sketch of the same $S \in \mathbb{R}^{H \times W \times 1}$, we want to generate a new image $\tilde{I} \in \mathbb{R}^{H \times W \times 3}$ that retains the information of the input model while substituting the target garment according to the multimodal inputs. To tackle the task, we propose a novel latent diffusion approach, called Multimodal Garment Designer (MGD), that can effectively combine multimodal information when generating the new image $\tilde{I}$. Our proposed architecture is a general framework that can be easily extended to other modalities such as texture and 3d information. We strongly believe this task can foster research in the field and enhance the design process of new fashion items with greater customization. An overview of our model is shown in Fig. 2.

3.1. Preliminaries

While diffusion models [44] are latent variable architectures that work in the same dimensionality of the data (*i.e.* in the pixel space), latent diffusion models (LDMs) [39] operate in the latent space of a pre-trained autoencoder achieving higher computational efficiency while preserving the generation quality. In our work, we leverage the Stable Diffusion model [39], a text-to-image implementation of LDMs as a starting point to perform multimodal condition-

ing for human-centric fashion image editing. Stable Diffusion is composed of an autoencoder with an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$, a text-time-conditional U-Net denoising model ϵ_θ, and a CLIP-based text encoder T_E taking as input a text Y. The encoder $\mathcal{E}$ compresses an image I into a lower-dimensional latent space defined in $\mathbb{R}^{h \times w \times 4}$, where $h = H/8$ and $w = W/8$. The decoder $\mathcal{D}$ performs the opposite operation, decoding a latent variable into the pixel space. For the sake of clarity, we define the ϵ_θ convolutional input (*i.e.* z_t in this case) as spatial input γ because of the property of convolutions to preserve the spatial structure and the attention conditioning input as ψ. The denoising network ϵ_θ is trained according to the following loss:

$$L = \mathbb{E}_{\mathcal{E}(I),Y,\epsilon \sim \mathcal{N}(0,1),t} \left[\| \epsilon - \epsilon_\theta(\gamma, \psi) \|_2^2 \right], \qquad (1)$$

where t is the diffusing time step, $\gamma = z_t$, $\psi = [t; T_E(Y)]$, and $\epsilon \sim \mathcal{N}(0,1)$ is the Gaussian noise added to $\mathcal{E}(I)$.

3.2. Human-Centric Image Editing

Our task aims to generate a new image $\tilde{I}$, by replacing in the input image I the target garment using multimodal inputs, while preserving the model's identity and physical characteristics. As a natural consequence, this task can be identified as a particular type of inpainting tailored for human body data. Instead of using a standard text-to-image model, we perform inpainting concatenating along the channel dimension of the denoising network input z_t an encoded masked image $\mathcal{E}(I_M)$ and the relative resized binary inpainting mask $m \in \{0,1\}^{h \times w \times 1}$, which is derived from the original inpainting mask $M \in \{0,1\}^{H \times W \times 1}$. Since here, the spatial input of the denoising network is $\gamma = [z_t; m; \mathcal{E}(I_M)], \gamma \in \mathbb{R}^{h \times w \times 9}$. Thanks to the fully convolutional nature of the encoder $\mathcal{E}$ and the decoder $\mathcal{D}$, this LDMs-based architecture can preserve the spatial information in the latent space. Exploiting this feature, our method can thus optionally add conditioning constraints to the generation. In particular, we propose to add two generation constraints in addition to the textual information: the

3

model pose map P to preserve the original human pose of the input model and the garment sketch S to allow the final users to better condition the garment generation process.

Pose Map Conditioning. In most cases [23, 26, 47], inpainting is performed with the objective of either removing or entirely replacing the content of the masked region. However, in our task, we aim to remove all information regarding the garment worn by the model while preserving the model's body information and identity. Thus, we propose to improve the garment inpainting process by using the bounding box of the segmentation mask along with pose map information representing body keypoints. This approach enables the preservation of the model's physical characteristics in the masked region while allowing the inpainting of garments with different shapes. Differently from conventional inpainting techniques, we focus on selectively retaining and discarding specific information within the masked region to achieve the desired outcome. To enhance the performance of the denoising network with human body keypoints, we modify the first convolution layer of the network by adding 18 additional channels, one for each keypoint. Adding new inputs usually would require retraining the model from scratch, thus consuming time, data, and resources, especially in the case of data-hungry models like the diffusion ones. Therefore, we propose to extend the kernels of the pre-trained input layer of the denoising network with randomly initialized weights sampled from a uniform distribution [14] and retrain the whole network. This consistently reduces the number of training steps and enables training with less data. Our experiments show that such improvement enhances the consistency of the body information between the generated image and the original one.

Incorporating Sketches. Fully describing a garment using only textual descriptions is a challenging task due to the complexity and ambiguity of natural language. While text can convey specific attributes like style, color, and patterns of a garment, it may not provide sufficient information about its spatial characteristics, such as shape and size. This limitation can hinder the customization of the generated clothing item other than the ability to accurately match the user's intended style. Therefore, we propose to leverage garment sketches to enrich the textual input with additional spatial fine-grained details. We achieve this following the same approach described for pose map conditioning. The final spatial input of our denoising network is $\gamma = [z_t; m; \mathcal{E}(I_M); p; s]$, $[p; s] \in \mathbb{R}^{h \times w \times (18+1)}$, p and s are obtained by resizing P and S to match the latent space dimensions. In the case of sketches, we only condition the early steps of the denoising process as the final steps have little influence on the shapes [2].

Mask Composition. To preserve the model identity when performing human-centric inpainting, we perform mask composition as the final step of the proposed approach.

Defining $\hat{I} = \mathcal{D}(z_0) \in \mathbb{R}^{H \times W \times 3}$ as the output of the decoder $\mathcal{D}$ and $M_{\text{head}} \in \{0, 1\}^{H \times W \times 1}$ as the model face binary mask of the image I, the final output image $\tilde{I}$ is obtained as follows: $\tilde{I} = M_{\text{head}} \odot I + (1 - M_{\text{head}}) \odot \hat{I}$, where $\odot$ denotes the element-wise multiplication operator.

3.3. Training and Inference

As in standard latent diffusion models, given an encoded input $z = \mathcal{E}(I)$, the proposed denoising network is trained to predict the noise stochastically added to z. The corresponding objective function can be specified as

$$L = \mathbb{E}_{\mathcal{E}(I), Y, \epsilon \sim \mathcal{N}(0,1), t, \mathcal{E}(I_M), m, p, s} \left[\|\epsilon - \epsilon_\theta(\gamma, \psi)\|_2^2 \right], \quad (2)$$

where $\gamma = [z_t; m; \mathcal{E}(I_M); p; s]$ and $\psi = [t; T_E(Y)]$.

Classifier-Free Guidance. Classifier-free guidance is an inference technique that requires the denoising network to work both conditioned and unconditioned. This method modifies the unconditional model predicted noise moving it toward the conditioned one. Specifically, the predicted diffusion process at time t, given the generic condition c, is computed as follows:

$$\hat{\epsilon}_\theta(z_t|c) = \epsilon_\theta(z_t|\emptyset) + \alpha \cdot (\epsilon_\theta(z_t|c) - \epsilon_\theta(z_t|\emptyset)), \quad (3)$$

where $\epsilon_\theta(z_t|c)$ is the predicted noise at time t given the condition c, $\epsilon_\theta(z_t|\emptyset)$ is the predicted noise at time t given the null condition, and the guidance scale α controls the degree of extrapolation towards the condition.

Since our model deals with three conditions (*i.e.* text, pose map, and sketch), we use the fast variant multi-condition classifier-free guidance proposed in [1]. Instead of performing the classifier-free guidance according to each condition probability, it computes the direction of the joint probability of all the conditions $\Delta^t_{\text{joint}} = \epsilon_\theta(z_t|\{c_i\}_{i=1}^{i=N}) - \epsilon_\theta(z_t|\emptyset)$:

$$\hat{\epsilon}_\theta(z_t|\{c_i\}_{i=1}^{i=N}) = \epsilon_\theta(z_t|\emptyset) + \alpha \cdot \Delta^t_{\text{joint}}. \quad (4)$$

This reduces the number of feed-forward executions from $N + 1$ to 2.

Unconditional Training. Ensuring the ability of the denoising model to work both with and without conditions is achieved by replacing at training time the condition with a null one according to a fixed probability. This approach allows the model to learn from both conditional and unconditional samples, resulting in improved mode coverage and sample fidelity. Moreover, this technique also allows the model to optionally use the control signals at prediction time. Since our approach considers multiple conditions, we propose to extend the input masking to each condition independently. Experiments show that tuning this parameter can effectively affect the quality of the final result.

4. Collecting Multimodal Fashion Datasets

Currently available datasets for fashion image generation often contain low-resolution images and lack all the required multimodal information needed to perform the task previously described. For this reason, the collection of new multimodal datasets for the fashion domain plays a crucial role to advance research in the field. To this aim, we start from two recent high-resolution fashion datasets introduced for the virtual try-on task, namely Dress Code [30] and VITON-HD [7], and extend them with textual sentences and garment sketches. Both datasets include image pairs with a resolution of 1024×768, each composed of a garment image and a reference model wearing the given fashion item. In this section, we introduce a framework to semi-automatically annotate fashion images with multimodal information and provide a complete description of how to enrich Dress Code and VITON-HD with garment-related text and sketches. We call our extended versions of these datasets Dress Code Multimodal and VITON-HD Multimodal, respectively. Sample images and multimodal data of the collected datasets can be found in Fig. 3.

4.1. Dataset Collection and Annotation

Data Preparation. We start the annotation from the Dress Code dataset, which contains more than 53k model-garment pairs of multiple categories. As a first step, we need to associate each garment with a textual description containing fashion-specific and non-generic terms which are sufficiently detailed but not extremely lengthy to be exploited for constraining the generation. Motivated by recent findings in the field showing that humans tend to describe fashion items using only a few words [3], we propose to use noun chunks (*i.e.* short textual sentences composed of a noun along with its modifiers) that can effectively capture important information while reducing unnecessary words or details. Given that manually annotating all the images would be time-consuming and resource-intensive[1], we propose a novel framework to semi-automatically annotate the dataset using noun chunks. Firstly, domain-specific captions are collected from two available fashion datasets, namely FashionIQ [53] and Fashion200k [12], standardizing them with word lemmatization and eventually reducing each word to its root form with the NLTK library[2]. Then, we extract noun chunks from the captions, filtering the results by removing all textual items that start with or contain special characters. After this pre-processing stage, we obtain more than 60k unique noun chunks, divided into three different categories (*i.e.* upper-body clothes, lower-body clothes, and dresses).

---

[1]Since the Dress Code dataset consists of over 53k fashion items and assuming that each annotation requires approximately 5 minutes, a single annotator working 8 hours per day, 5 days a week, and 260 working days per year would take more than 2 years to complete the annotation task.
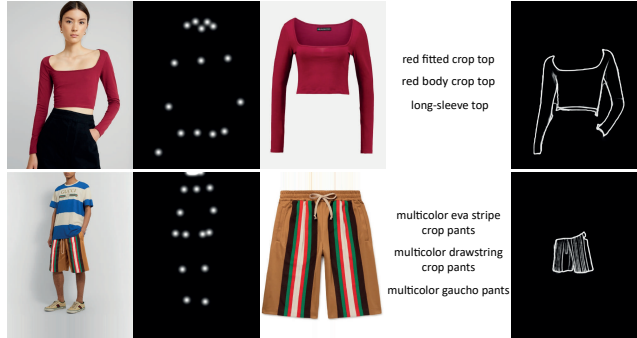
[2]https://www.nltk.org/

Figure 3: Sample images and multimodal data from our newly collected datasets.

Dataset	Text	Pose	Sketch	# Images	# Products	# Unique Texts	# Unique Words
VITON-HD [7]	✗	✓	✗	27,358	13,679	-	-
Dress Code [30]	✗	✓	✗	107,584	53,792	-	-
Be Your Own Prada [59]	✓	✓	✗	78,979	N/A	3,972	445
DF-Multimodal [19]	✓	✓	✗	44,096	N/A	10,253	77
VITON-HD Multimodal	✓	✓	✓	27,358	13,679	5,143	1,613
Dress Code Multimodal	✓	✓	✓	107,584	53,792	25,596	2,995

Table 1: Comparison of Dress Code and VITON-HD Multimodal with other fashion datasets with multimodal annotations.

To determine the most relevant noun chunks for each garment, we employ the CLIP model [36] and its open-source adaptation (*i.e.* OpenCLIP [52]). We select the VIT-L14@336 and RN50×64 models for CLIP, and the VIT-L14, ViT-H14, and ViT-g14 models for OpenCLIP. Prompt ensembling is performed to improve the results and, for each image, we select 25 noun chunks based on the top-5 noun chunks per model rated by cosine similarity between image and text embeddings, avoiding repetitions.

Fine-Grained Textual Annotation. To ensure the accuracy and representativeness of our annotations, we manually annotate a significant portion of Dress Code images. In particular, we select the three most representative noun chunks, among the 25 automatically associated, with each garment image. To minimize the annotation time, we develop a custom annotation tool that constrains the annotation time to an average time of 60 seconds per item and allows the annotator to manually insert noun chunks in the case that none of the automatically extracted ones are suitable for the image. Overall, we manually annotate 26,400 different garments (8,800 for each category) out of the 53,792 products included in the dataset, ensuring to include all fashion items of the original test set [30].

Coarse-Grained Textual Annotation. To complete the annotation, we first finetune the OpenCLIP ViT-B32 model, pre-trained on the English portion of the LAION5B dataset [42], using the newly annotated image-text pairs. We then use this model and the collected set of noun chunks

to automatically tag all the remaining elements of the Dress Code dataset with the three most similar noun chunks, always determined via cosine similarity between multimodal embeddings. We employ the same strategy also to automatically annotate all garment images of the VITON-HD dataset. In this case, since this dataset only contains upper-body clothes, we limit the table noun chunks to the ones describing upper-body garments.

Extracting Sketches. The introduction of garment sketches can provide valuable design details that are not easily discernible from text alone. In this way, the dataset can provide a more accurate and comprehensive representation of the garments, leading to improved quality and better control of the generated design details. To extract sketches for both Dress Code and VITON-HD datasets, we employ PiDiNet [46], a pre-trained edge detection network.

Given that the selected datasets have originally been introduced for virtual try-on, they consist of both paired and unpaired test sets. While for the paired set we can directly use the human parsing mask to extract the garment of interest worn by the model and then feed it to the edge detection network, for the unpaired set we need to first create a warped version of the in-shop garment matching the body pose and shape of the target model. Following virtual try-on methods [50,55], we train a geometric transformation module that performs a thin-plate spline transformation [38] of the input garment and then refines the warped result using a U-Net model [40]. From each warped garment, we extract the sketch image enabling the use of the proposed solution even in unpaired settings.

4.2. Comparison with Other Datasets

The only two text-to-image generation datasets available in the fashion domain [19, 59] are both based on images from the DeepFashion dataset [24]. While the dataset introduced in [59] contains short textual descriptions, DeepFashion-Multimodal [19] is annotated with attributes (*e.g.* category, color, fabric, etc.) that can be composed in longer captions. In Table 1, we summarize the main statistics of the publicly available datasets textual annotations compared with those of our newly extended datasets. As can be seen, our datasets contain more variety in terms of textual items and words, confirming the appropriateness of our annotation procedure and enabling a more personalized control of the generation process. Also, it is worth noting that the other datasets have no in-shop garment images making them difficult to employ in our case.

5. Experimental Evaluation

5.1. Implementation Details and Competitors

Training and Inference. All models are trained on the original splits of the Dress Code Multimodal and VITON-HD Multimodal datasets on a single NVIDIA A100 GPU for 150k steps, using a batch size of 16, a learning rate of 10^{-5} with a linear warmup for the first 500 iterations, and AdamW [25] as optimizer with weight decay 10^{-2}. To speed up training and save memory, we use mixed precision [28]. We set both the fraction of steps conditioned by the sketch and the portion of masked conditions during training to 0.2. During inference, we employ the DDIM [45] with 50 steps as noise scheduler and set the classifier-free guidance parameter α to 7.5.

Baselines and Competitors. As first competitor, we use the out-of-the-box implementation of the inpainting Stable Diffusion pipeline[3] provided by Huggingface. Moreover, we adapt two existing models, namely FICE [35] and SDEdit [27], to work on our setting. In particular, we retrain all main components of the FICE model on the newly collected datasets. We employ the same resolution used by the authors (*i.e.* 256×256), downsampling each image to 256×192 and applying padding to match the desired size (which is then removed during evaluation). To compare our model with a different conditioning strategy, we employ the approach proposed in [27] using our model trained only with text and human poses as input modalities and perform the sketch guidance using as starting latent variable the sketch image with added random noise. Following the original paper instructions, we use 0.8 as the strength parameter.

5.2. Evaluation Metrics

To assess the realism of generated images, we employ the Fréchet Inception Distance (FID) [16] and the Kernel Inception Distance (KID) [4]. For both metrics, we adopt the implementation proposed in [34]. Instead, to evaluate the adherence of the image to the textual conditioning input, we employ the CLIP Score (CLIP-S) [15] provided in the TorchMetrics library [9], using the OpenCLIP ViT-H/14 model as cross-modal architecture. We compute the score on the inpainted region of the generated output pasted on a 224×224 white background.

Pose Distance (PD). We propose a novel pose distance metric that measures the coherence of human body poses between the generated image and the original one estimating the distance between the human keypoints extracted from the original and generated images. Specifically, we employ the OpenPifPaf [22] human pose estimation network and compute the ℓ_2 distance between each pair of real-generated corresponding estimated keypoints. We only consider the keypoints involved in the generation (*i.e.* that falls in the mask M) and weigh each keypoint distance with the detector confidence to take into account any estimation errors.

Sketch Distance (SD). To quantify the adherence of the generated image to the sketch constraint, we propose a

---

		Modalities			Dress Code Multimodal					VITON-HD Multimodal				
Model	Resolution	Text	Pose	Sketch	FID↓	KID↓	CLIP-S↑	PD↓	SD↓	FID↓	KID↓	CLIP-S↑	PD↓	SD↓
Paired setting														
Stable Diffusion [39]	256×192	✓			17.05	9.28	28.71	4.62	-	15.18	6.38	30.40	5.04	-
FICE [35]	256×192	✓	✓		30.63	23.54	28.72	6.87	-	49.44	44.74	29.26	6.37	-
MGD (ours)	256×192	✓	✓		**5.57**	**1.67**	**31.33**	**2.37**	-	**10.11**	**3.14**	**31.85**	**2.90**	-
Paired setting														
Stable Diffusion [39]	512×384	✓			17.43	9.48	29.18	9.24	0.467	16.28	6.56	30.70	10.78	0.410
SDEdit [27]	512×384	✓	✓	✓	10.19	5.03	29.21	5.41	0.398	13.07	4.66	30.58	6.76	0.306
MGD (ours)	512×384	✓	✓	✓	**5.74**	**2.11**	**31.68**	**4.72**	**0.374**	**10.60**	**3.26**	**32.39**	**5.94**	**0.253**
Unpaired setting														
Stable Diffusion [39]	256×192	✓			19.11	10.69	27.53	5.07	-	17.37	7.55	28.40	5.50	-
FICE [35]	256×192	✓	✓		34.14	26.86	26.03	7.15	-	52.74	48.58	25.94	6.58	-
MGD (ours)	256×192	✓	✓		**7.01**	**2.19**	**29.58**	**2.96**	-	**11.54**	**3.18**	**29.95**	**3.30**	-
Unpaired setting														
Stable Diffusion [39]	512×384	✓			19.55	10.80	28.02	9.89	0.582	18.45	7.87	28.74	11.60	0.561
SDEdit [27]	512×384	✓	✓	✓	11.38	5.69	27.10	**6.16**	0.509	15.12	5.67	28.61	7.35	0.406
MGD (ours)	512×384	✓	✓	✓	**7.73**	**2.82**	**30.04**	6.79	**0.458**	**12.81**	**3.86**	**30.75**	**7.22**	**0.317**

Table 2: Quantitative results on the Dress Code Multimodal and VITON-HD Multimodal datasets for both paired and unpaired settings .

	Modalities			Dress Code Multimodal				
Model	Text	Pose	Sketch	FID↓	KID↓	CLIP-S↑	PD↓	SD↓
	✓			6.19	2.15	**31.79**	6.16	0.411
	✓	✓		6.31	2.33	31.67	5.31	0.405
MGD (ours)	✓	✓	✓	**5.74**	**2.11**	31.68	**4.72**	**0.374**

	Modalities			VITON-HD Multimodal				
Model	Text	Pose	Sketch	FID↓	KID↓	CLIP-S↑	PD↓	SD↓
	✓			11.39	3.52	32.16	7.83	0.339
	✓	✓		11.07	3.36	32.27	6.77	0.318
MGD (ours)	✓	✓	✓	**10.60**	**3.26**	**32.39**	**5.94**	**0.253**

Table 3: Performance analysis on the paired setting of both datasets as input modalities vary.

	Modalities			Realism			Multimodal Coherence		
	Text	Pose	Sketch	Stable Diff.	FICE	SDEdit	Stable Diff.	FICE	SDEdit
Dress Code M.	✓			70.82	-	-	65.32	-	-
	✓	✓		70.73	96.26	-	65.15	84.48	-
	✓	✓	✓	70.29	-	52.54	65.38	-	66.23
VITON HD M.	✓			67.03	-	-	57.76	-	-
	✓	✓		66.17	93.84	-	73.73	83.46	-
	✓	✓	✓	60.71	-	53.44	69.47	-	59.34

Table 4: User study results on the unpaired setting of both datasets. We report the percentage of times an image from MGD is preferred against a competitor. Comparisons with FICE [35] are performed at 256×192 resolution.

novel sketch distance metric. To compute the score, we extract the segmentation map of the original and generated garments using an off-the-shelf clothing segmentation network[4]. We then use the segmented garment area to extract garment sketches using the PIDInet [46] edge detector network. The final score is the mean squared error between these sketches, weighting the per-pixel results on the inverse pixel frequency of the activated pixels. More details about these proposed metrics can be found in the supplementary.

5.3. Experimental Results

Comparison with Other Methods. We test our proposal for the paired and unpaired settings of the considered datasets. In the former, the conditions (*e.g.* text, sketch) refers to the garment the model is wearing, while in the latter, the in-shop garment differs from the worn one. In Table 2, we report the quantitative results on Dress Code Multimodal and VITON-HD Multimodal in comparison with the aforementioned competitors. As can be seen, the proposed MGD model consistently outperforms competitors, in terms of realism (*i.e.* FID and KID) and coherency with input modalities (*i.e.* CLIP-S, PD, and SD). In particular, when considering low-resolution results, we notice that FICE [35] can produce images fairly consistent with the text conditioning, albeit less realistic than other methods. While Stable Diffusion [39] enhances image realism, it fails to preserve the input model's pose due to the lack of pose information in the inputs. It is noteworthy that in this case, we compare the results of our model only using text and pose map as conditioning since both considered competitors are not conditioned on sketches. For this reason, we do not report the results in terms of sketch distance for low-resolution images.

In the high-resolution setting, we evaluate instead our MGD method using all multimodal conditions (*i.e.* text, pose map, and sketch) as input. In this case, we compare MGD with Stable Diffusion [39] plus SDEdit [27], where we use our text-pose conditioned denoising network as SDEdit backbone. Our findings indicate that Stable Dif-

---

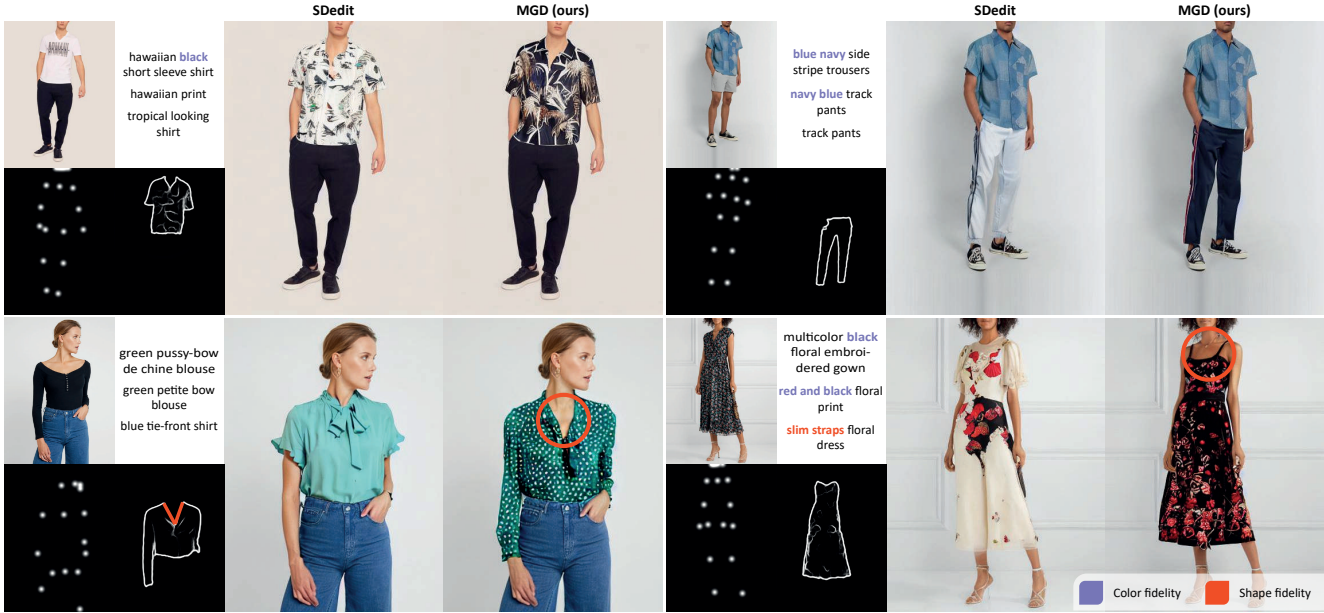[4]https://github.com/levindabhi/cloth-segmentation

Figure 4: Sample generated images on Dress Code Multimodal and VITON-HD Multimodal (bottom left) using all multimodal inputs.

fusion performs worse in terms of the pose distance than both SDEdit and MGD, owing to the lack of pose information in the inputs. It is noteworthy that SDEdit performs worse than our model in all metrics. We attribute this behavior to the way sketch conditioning happens. In SDEdit, it occurs only at the beginning by initializing z_t using the sketch image with added noise according to the conditioning strength, while our model conditions the denoising process in multiple steps, depending on the sketch conditioning parameter. Qualitative results reported in Fig. 4 highlight how our model better follows the given conditions and generate high-realistic images.

To validate our results based on human judgment, we conduct a user study that evaluates both the realism of the generation and the adherence to multimodal inputs. Overall we collect about 7k evaluations involving more than 150 users. Additional details are reported in the supplementary. Table 4 shows the user study results. Also in this case our model outperforms the competitors, thus confirming the effectiveness of our proposal.

Varying Input Modalities. In Table 3, we study the behavior of our MGD model when the input modalities are masked (*i.e.* where we feed the model with a zero tensor instead of the considered modality). In particular, we focus on the CLIP-S for text adherence and on the newly proposed pose and sketch distances for the pose and sketch coherency, respectively. Notice that the text input anchors the CLIP-S metrics of all experiments and makes them comparable in all cases. Starting from the fully conditioned model (*i.e.* text, pose, sketch), we mask the sketch. As the decrease

		Dress Code Multimodal				
Uncond. Portion	Sketch Cond.	FID ↓	KID ↓	CLIP-S ↑	PD ↓	SD ↓
0.1	1.0	9.64	3.76	30.24	7.66	0.459
0.2	1.0	8.62	3.24	29.06	7.51	0.430
0.3	1.0	10.93	4.78	28.47	7.69	0.432
0.2	0.8	8.56	3.28	29.31	7.32	0.433
0.2	0.6	8.43	3.21	29.51	7.32	0.436
0.2	0.4	8.11	3.00	29.79	7.13	0.440
0.2	0.2	7.73	2.82	30.04	6.79	0.458
0.2	0.0	7.82	2.85	29.93	6.26	0.519

Table 5: Ablation analysis of our complete model varying the unconditioning portion during training and the sketch conditioning steps. Results refer to the unpaired setting.

of the sketch distance in Table 3 confirms, this input actually influences the generation process of our model in both the considered datasets. Also, this modality slightly affects the pose distance as the sketch implicitly contains information about the model's body pose. We further mask the pose map input and compare the output with previous results. In this case, we can also notice a consistent difference with the text-only conditioned model, according to all metrics except CLIP-S as expected. These results confirm that our MGD model can effectively deal with the conditions in a disentangled way, making them optional.

Unconditional Training and Sketch Conditioning. In Table 5, we inquire about the fully conditioned network performance according to the variance of the portion of unconditional training. Additionally, we evaluate the results by varying the fraction of sketch conditioning steps. As can be seen, the best results are achieved by using 0.2 for both pa-

rameters. In particular, for unconditional training, we train three different models (*i.e.* with 0.1, 0.2, and 0.3). When evaluating the sketch conditioning parameter, we test our model with values between 0 and 1 with a stride of 0.2. It is worth noting that the sketch distance consistently decreases as the number of sketch conditioning steps increases, showing the robustness of the approach.

6. Conclusion

The Multimodal Garment Designer proposed in this paper is the first latent diffusion model defined for human-centric fashion image editing, conditioned by multimodal inputs such as text, body pose, and sketches. The novel architecture, trained on two new semi-automatically annotated datasets and evaluated with standard and newly proposed metrics, as well as by user studies, is very promising. The result is one of the first successful attempts to mimic the designers' job in the creative process of fashion design and could be a starting point for a capillary adoption of diffusion models in creative industries, oversight by human input.

Acknowledgments

References

[1] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. SpaText: Spatio-Textual Representation for Controllable Image Generation. In *CVPR*, 2023. 4

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 4

[3] Federico Bianchi, Jacopo Tagliabue, and Bingqing Yu. Query2Prod2Vec: Grounded word embeddings for eCommerce. In *NAACL*, 2021. 5

[4] Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *ICLR*, 2018. 6, 15

[5] Shin-I Cheng, Yu-Jie Chen, Wei-Chen Chiu, Hung-Yu Tseng, and Hsin-Ying Lee. Adaptively-Realistic Image Generation from Stroke and Sketch with Diffusion Model. In *WACV*, 2023. 2

[6] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *ICCV*, 2021. 2

[7] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization. In *CVPR*, 2021. 1, 2, 5

[8] Guillem Cucurull, Perouz Taslakian, and David Vazquez. Context-aware visual compatibility prediction. In *CVPR*, 2019. 1

[9] Nicki Skafte Detlefsen, Jiri Borovec, Justus Schock, Ananya Harsh Jha, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, and William Falcon. TorchMetrics-Measuring Reproducibility in Py-Torch. *Journal of Open Source Software*, 7(70):4101, 2022. 6, 13

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. In *NeurIPS*, 2021. 2

[11] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. Where to buy it: Matching street clothing photos in online shops. In *ICCV*, 2015. 1

[12] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *ICCV*, 2017. 5, 11

[13] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *CVPR*, 2018. 1

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015. 4

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*, 2021. 6, 13

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a Nash equilibrium. In *NeurIPS*, 2017. 6

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *NeurIPS*, 2020. 2

[18] Wei-Lin Hsiao and Kristen Grauman. Creating capsule wardrobes from fashion images. In *CVPR*, 2018. 1

[19] Yuming Jiang, Shuai Yang, Haonan Qju, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics*, 41(4):1–11, 2022. 2, 5, 6

[20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 12

[21] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015. 12

[22] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Open-PifPaf: Composite Fields for Semantic Keypoint Detection and Spatio-Temporal Association. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):13498–13511, 2021. 6, 14

[23] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. MAT: Mask-Aware Transformer for Large Hole Image Inpainting. In *CVPR*, 2022. 4

[24] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *CVPR*, 2016. 1, 6

[25] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019. 6, 12

[26] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting Using Denoising Diffusion Probabilistic Models. In *CVPR*, 2022. 4

[27] Chenlin Meng, Yutong He adnd Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. 2, 6, 7, 15, 20, 21

[28] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed Precision Training. In *ICLR*, 2018. 6, 12

[29] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On. In *ACM Multimedia*, 2023. 1

[30] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress Code: High-Resolution Multi-Category Virtual Try-On. In *ECCV*, 2022. 1, 2, 5

[31] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models. *arXiv preprint arXiv:2302.08453*, 2023. 2, 3

[32] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 2

[33] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *ICML*, 2022. 2

[34] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On Aliased Resizing and Surprising Subtleties in GAN Evaluation. In *CVPR*, 2022. 6

[35] Martin Pernuš, Clinton Fookes, Vitomir Štruc, and Simon Dobrišek. FICE: Text-Conditioned Fashion Image Editing With Guided GAN Inversion. *arXiv preprint arXiv:2301.02110*, 2023. 2, 6, 7, 15, 22, 23

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021. 2, 5, 11

[37] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[38] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *CVPR*, 2017. 6, 12

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *CVPR*, 2022. 2, 3, 7, 15, 16, 20, 21, 22, 23

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*, 2015. 6, 12

[41] Rohan Sarkar, Navaneeth Bodla, Mariya I Vasileva, Yen-Liang Lin, Anurag Beniwal, Alan Lu, and Gerard Medioni. OutfitTransformer: Learning Outfit Representations for Fashion Recommendation. In *WACV*, 2023. 1

[42] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 5

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 12

[44] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2, 3

[45] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *ICLR*, 2021. 6

[46] Zhuo Su, Wenzhe Liu, Zitong Yu, Dewen Hu, Qing Liao, Qi Tian, Matti Pietikäinen, and Li Liu. Pixel difference networks for efficient edge detection. In *ICCV*, 2021. 6, 7, 14

[47] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *WACV*, 2022. 4

[48] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis. In *CVPR*, 2022. 2

[49] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-Guided Text-to-Image Diffusion Models. *arXiv preprint arXiv:2211.13752*, 2022. 2

[50] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *ECCV*, 2018. 1, 6, 12

[51] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is All You Need for Image-to-Image Translation. *arXiv preprint arXiv:2205.12952*, 2022. 2

[52] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *CVPR*, 2022. 5, 11

[53] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback. In *CVPR*, 2021. 5, 11

[54] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. In *CVPR*, 2018. 2

[55] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wang-meng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *CVPR*, 2020. 1, 6

[56] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *CVPR*, 2021. 2

[57] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2, 15, 16

[58] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis. In *CVPR*, 2019. 2

[59] Shizhan Zhu, Raquel Urtasun, Sanja Fidler, Dahua Lin, and Chen Change Loy. Be Your Own Prada: Fashion Synthesis with Structural Coherence. In *ICCV*, 2017. 2, 5, 6

A. Dress Code Multimodal and VITON-HD Multimodal Datasets

In this section, we give additional details about the dataset collection and annotation process and provide statistics and further examples of the collected datasets.

A.1. Data Preparation

Before extracting noun chunks from the textual sentences of FashionIQ [53] and Fashion200k [12], we perform word lemmatization to reduce each word to its root form. Such pre-processing stage is crucial for the FashionIQ dataset, as the captions do not describe a single garment but instead express the properties to modify in a given image to match its target. Fig. 5 shows two examples of FashionIQ annotations.

We use the spaCy NLP toolkit[5] to extract noun chunks from textual sentences. To facilitate prompt engineering at a later stage, we remove the articles at the beginning of each noun chunk. Subsequently, we filter out all noun chunks starting with or containing special characters and keep unique elements. Table 6 reports detailed statistics about the number of unique captions and extracted noun chunks from which we start the annotation.

Textual Prompts. As described in the main paper, we rely on the cosine similarity between CLIP-based image and text embeddings to associate each garment with the 25 most representative noun chunks. We exploit prompt ensembling to perform such zero-shot association as it is shown in [36] that this technique improves performance.

The employed textual prompts are:
- a photo of a [noun chunk],
- a photo of a nice [noun chunk],
- a photo of a cool [noun chunk],
- a photo of an expensive [noun chunk],
- a good photo of a [noun chunk],

Figure 5: Examples of FashionIQ data type.

Dataset	Unique Captions			Unique Noun Chunks		
	Upper	Lower	Dresses	Upper	Lower	Dresses
FashionIQ [53]	27,339	0	15,101	7,801	0	3,592
Fashion200k [12]	25,959	11,022	16,694	22,898	13,420	15,890

Table 6: Number of unique captions and noun chunks for each category of the FashionIQ and Fashion200k datasets.

- a bright photo of a [noun chunk],
- a fashion studio shot of a [noun chunk],
- a fashion magazine photo of a [noun chunk],
- a fashion brochure photo of a [noun chunk],
- a fashion catalog photo of a [noun chunk],
- a fashion press photo of a [noun chunk],
- a yoox photo of a [noun chunk],
- a yoox web image of a [noun chunk],
- a high-resolution photo of a [noun chunk],
- a cropped photo of a [noun chunk],
- a close-up photo of a [noun chunk],
- a photo of one [noun chunk].

A.2. Annotation Tool for Fine-Grained Annotation

We develop a custom annotation tool using the Django and Angular web frameworks to ease and speed up the fine-grained annotation process. Fig. 6 depicts the user interface. In the annotation phase, users are provided with both model's image and the corresponding in-shop garment and should select the three most representative noun chunks per item (Fig. 6a). If the automatic selection process fails to suggest three correct noun chunks, the user can manually insert them (Fig. 6b).

A.3. Coarse-Grained Annotation

After completing the manual annotation process on Dress Code, we obtain 26,400 different model-garment pairs (with 8,800 items per category), each associated with three different noun chunks. To annotate the remaining 27,392 items of Dress Code Multimodal and the 13,679 items of VITON-HD Multimodal, we leverage the manually annotated image-text pairs and finetune the OpenCLIP ViT-B/32 [52] model pre-trained on the English portion of the LAION-5B dataset.

CLIP Finetuning. We finetune both encoders of the Open-CLIP model using a single NVIDIA A100 GPU for 400 steps, with a batch size of 2048 and a learning rate of 10^{-6}.

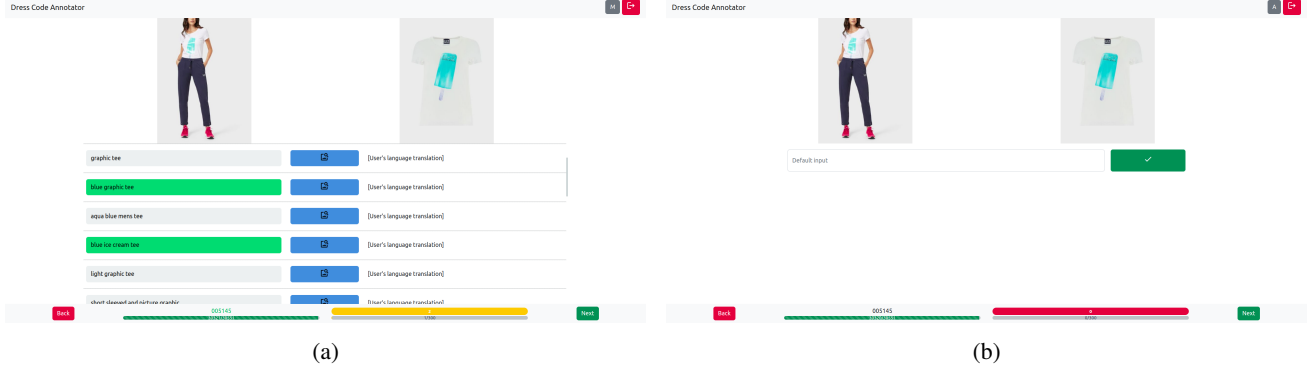[5]https://spacy.io/

*Equal contribution.

(a)   (b)

Figure 6: User interface of the custom annotation tool. In (a) the user can select the noun chunks among the proposed ones, while in (b) the user can manually annotate the garment.

As optimizer, we use AdamW [25] with a weight decay of 0.2. We use mixed precision [28] to speed up training and save memory. During the training process, we monitor the model performance using the top-3 accuracy metric on the test split of the Dress Code Multimodal dataset. We choose this metric intending to associate each image with three distinct noun chunks. The out-of-the-box model achieves a top-3 accuracy of 12.95%, which improves to 16.60% after finetuning. The OpenCLIP ViT-g/14 model instead achieves a top-3 accuracy of 16.21%, while being computationally heavier than the ViT-B/32 version. Since the ViT-g/14 model predicts the set of noun chunks from which we extract the ground-truth, the actual difference in performance between the finetuned ViT-B/32 model and the out-of-the-box ViT-g/14 model could be even higher.

A.4. Extracting Sketches

As mentioned in the main paper, we train a warping module to generate input sketches for the unpaired setting (*i.e.* when we give as input the multimodal information corresponding to a garment different from the one originally worn by the model). In particular, our method involves the transformation of a given in-shop garment $C \in \mathbb{R}^{H \times W \times 3}$ into a warped image of the same garment that fits the model of a target image I. We employ the warping module proposed in [50], refining the results with a U-Net based component [40].

The warping module computes a correlation map between the encoded representations of the in-shop garment C and a cloth-agnostic person representation composed of the pose map $P \in \mathbb{R}^{H \times W \times 18}$ and the masked model image $I_M \in \mathbb{R}^{H \times W \times 3}$. We use two separate convolutional networks to obtain these encoded representations. Based on the computed correlation map, we predict the spatial transformation parameters θ of a thin-plate spline geometric transformation [38] (*i.e.* TPS_θ). We then use the θ parameters to compute the coarse warped garment $\hat{C}$ starting from the

			Images			Unique Noun Chunks		
Dataset	Ann.	Split	Upper	Lower	Dresses	Upper	Lower	Dresses
Dress Code M.	F	Train	7,000	7,000	7,000	4,751	5,914	4,410
		Test	1,800	1,800	1,800	2,337	2,861	2,144
		∪	8,800	8,800	8,800	5,284	6,509	4,915
		∩	-	-	-	1,804	2,266	1,639
Dress Code M.	C	Train	6,563	151	20,666	7,198	320	8,650
		Test	0	0	0	0	0	0
		∪	6,563	151	20,666	7,198	320	8,650
		∩	-	-	-	0	0	0
Dress Code M.	F+C	Train	13,563	7,151	27,666	9,163	6,037	9,465
		Test	1,800	1,800	1,800	2,337	2,861	2,144
		∪	15,363	8,951	29,466	9431	6,597	9,568
		∩	-	-	-	2,069	2,301	2,041
VITON-HD M.	C	Train	11,647	-	-	4,823	-	-
		Test	2,032	-	-	2,149	-	-
		∪	13,679	-	-	5,143	-	-
		∩	-	-	-	1,829	-	-

Table 7: Number of images and unique noun chunks per category for both Dress Code Multimodal and VITON-HD Multimodal. (F) indicates the fine-grained annotation while (C) indicates the coarse-grained annotation.

in-shop garment C as follows:

$$\hat{C} = \text{TPS}_\theta(C). \tag{5}$$

To refine the result, we employ a U-Net model that takes as input the concatenation of the coarse warped garment $\hat{C}$, the pose map P, and the masked model image I_M, and predicts the refined warped garment $\tilde{C}$.

We train this model on the training set of both Dress Code Multimodal and VITON-HD Multimodal using a combination of an L1 loss between generated and target in-shop garments and a perceptual loss (also known as VGG loss [20]) to compute the difference between the feature maps of generated and target garments extracted with a VGG-19 [43]. We train with a resolution of 256×192, Adam [21] as optimizer with $\beta_1 = 0.5, \beta_2 = 0.99$, and a learning rate equal to 10^{-4}. We train the network on the VITON-HD dataset for 30 epochs, while the training on the Dress Code dataset converges after 80 epochs.
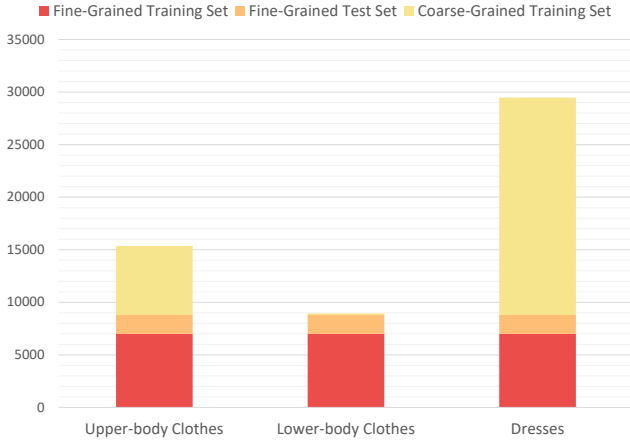
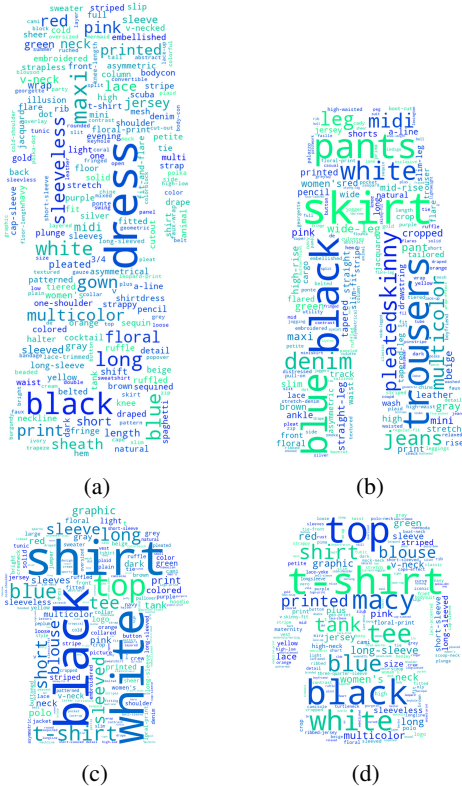Figure 7: Annotated images per category on Dress Code Multimodal.

Figure 8: Vocabulary of the frequent words scaled by frequency for dresses (a), lower-body clothes (b), upper-body clothes (c) of Dress Code Multimodal and clothing items of VITON-HD Multimodal (d).

A.5. Additional Statistics and Annotated Samples

Table 7 summarizes the number of images and unique noun chunks for each category of Dress Code Multimodal and VITON-HD Multimodal. The table shows that the datasets share noun chunks between the train and test set (∩). This behavior is likely due to the limited capacity of the textual modality to represent the whole semantic information of the image. Fig. 7 instead shows the number of samples for each category highlighting the different annotation strategies on Dress Code Multimodal.

In Fig. 8, we report the word clouds extracted from the textual annotations, representing the most frequently used words in the collected noun chunks for each category of Dress Code Multimodal and VITON-HD Multimodal. From this visualization, we can notice that the frequency of the terms varies according to the garment category, and the semantic richness of our annotations is consistent across different garment types.

In Fig. 11 and Fig. 12, we report samples from the fine-grained and coarse-grained subsets of Dress Code Multimodal, respectively. Instead, in Fig. 13, we show additional examples extracted from VITON-HD Multimodal.

B. Evaluation Metrics

This section provides additional details about the evaluation metrics used in our experiments. We first give some clarifications about the CLIP-S metric and then present an accurate formulation of the proposed sketch distance and pose distance metrics.

CLIP-S. The CLIP score [15] is a well-known metric to evaluate the similarity between images and textual sentences. In our setting, we employ this metric to assess the coherence of the generated images with respect to the corresponding textual inputs used to condition the generation

process. As mentioned in the main paper, our implementation relies on the CLIP-S of the TorchMetrics library [9] and adopts the ViT-H/14 trained on LAION-2B as the CLIP model. Specifically, we crop the generated image using the bounding box used to mask it and paste the resulting crop on a white background, obtaining a final resolution equal to 224×224. The adopted metric is defined as follows:

$$\text{CLIP-S}(I, Y) = \max(100 * cos(E_{\tilde{I}}, E_Y), 0), \quad (6)$$

where $E_{\tilde{I}}$ represents the CLIP embedding of the generated portion of the image $\tilde{I}$ pasted on white background, E_Y represents the CLIP embedding for the caption Y, and cos is the cosine similarity. We calculate the cosine similarity between the image and caption embeddings and scale the result by a factor of 100. If the cosine similarity is negative, then CLIP-S is zero.

Pose Distance (PD). To measure the coherence of human-body poses between the generated image and the original one, we propose a novel pose distance metric that estimates the distance between human keypoints extracted from the original and the generated images. Given a ground-truth image I and a generated image $\tilde{I}$, we extract human keypoints from each of them using the keypoint extraction network $\mathcal{K}$

(*i.e.* in our case, we use OpenPifPaf [22]) and identify the set of keypoints falling in the mask M as $\mathcal{K}(\cdot)_M$. We compute the final score with an ℓ_2 distance between each pair of real-generated corresponding keypoints (*i.e.* $k \in \mathcal{K}(I)_M$ and $\tilde{k} \in \mathcal{K}(\tilde{I})_M$, respectively), weighting each keypoint distance with the detector confidence to consider possible estimation errors. Formally, our pose distance metric is defined as follows:

$$\text{PD}(I, \tilde{I}) = \frac{\displaystyle\sum_{\substack{k \in \mathcal{K}(I)_M \\ \tilde{k} \in \mathcal{K}(\tilde{I})_M}} \sqrt{(k_x - \tilde{k}_x)^2 + (k_y - \tilde{k}_y)^2} \cdot \text{CF}_{k\tilde{k}}}{\sum_{k\tilde{k}} \text{CF}_{k\tilde{k}}},$$

(7)

where, for each pair of real-generated keypoints, $\text{CF}_{k\tilde{k}}$ is 1 if the confidence of the detector $\mathcal{K}$ on both keypoints is greater or equal to 0.5, and 0 otherwise.

Sketch Distance (SD). To evaluate the adherence of the generated images to the constraints imposed by the input sketch, we propose a new sketch distance metric. To compute the metric, we first extract the ground-truth and the generated garments label maps using an off-the-shelf semantic segmentation model[6]. We segment the garment according to its category and paste it on a white background of shape 512×384. We refer to these new images with I_S and $\tilde{I}_S$, respectively. Then, we extract the garment sketches of both the ground-truth and the generated images using an edge detector network $Edge$ (*i.e.* PIDInet [46]). Finally, we compute the mean squared error between the extracted sketches, weighting the per-pixel results on the inverse frequency of the activated pixels. Formally, the introduced sketch distance metric is defined as follows:

$$\text{SD}(I_S, \tilde{I}_S) = \text{MSE}\left(Edge(I_S), Edge(\tilde{I}_S)\right) * p, \quad (8)$$

where p is the inverse pixel frequency. It is noteworthy that sketch thresholding could be applied before distance computation. Nevertheless, we argue that avoiding thresholding enables an effective comparison of hand-drawn ground-truth grayscale sketches. This approach can facilitate the evaluation of methods that generate images conditioned using the sketch. Therefore, we think the proposed metric can be a valuable tool for comparing sketch-guided generative architectures.

C. User Study

As mentioned in the main paper, we conduct a user study to evaluate the realism of generated images and their adherence to the given multimodal inputs, comparing our results with those from the considered competitors. To this aim, we develop a custom web interface presenting two different surveys. The former (Fig. 9a) assesses the realism of the

---
[6]https://github.com/levindabhi/cloth-segmentation

(a)

(b)
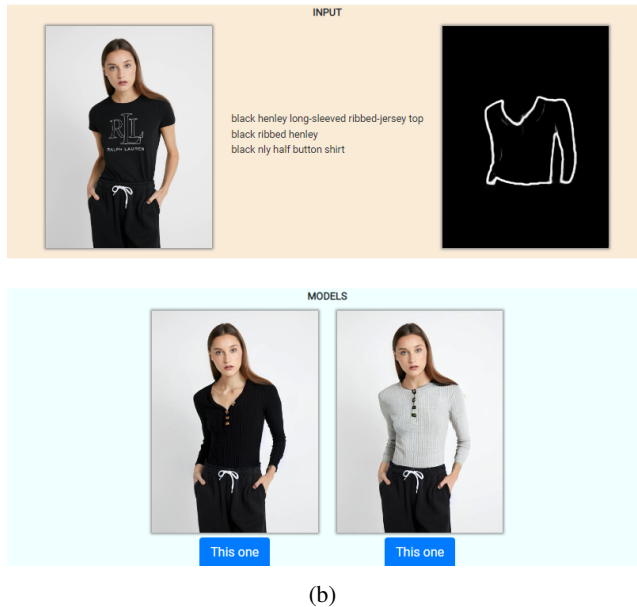
Figure 9: User study interface, where (a) corresponds to the realism evaluation and (b) refers to the coherence analysis between generated images and the given multimodal inputs.

generated output asking the user to select for each comparison the image that seems more realistic. In the latter (Figure 9b), given the model's image, the set of noun chunks describing the garment, and the sketch, the user is asked to select which of the two proposed outputs looks more coherent with the multimodal inputs also taking into account the model's body pose. Overall, we collect around 7k evaluations, 3.5k for each test, and involving more than 150 users.

D. Additional Results

In this section, we provide additional experimental results to understand the strengths and limitations of our approach. Table 8 extends Table 2 of the main paper showing quantitative results on each garment category of Dress Code Multimodal. Since each category contains only 1,800 images, the FID score presents a high variance in the re-

Model	Resolution	Modalities			Upper-body					Lower-body					Dresses				
		Text	Keypoints	Sketch	FID↓	KID↓	CLIP-S↑	PD↓	SD↓	FID↓	KID↓	CLIP-S↑	PD↓	SD↓	FID↓	KID↓	CLIP-S↑	PD↓	SD↓
Paired setting																			
Stable Diff. [39]	256×192	✓			22.86	9.73	28.31	4.29	-	28.78	13.93	26.41	4.97	-	36.31	20.74	27.84	5.67	-
FICE [35]	256×192	✓	✓		46.41	32.26	28.58	7.46	-	41.68	27.22	28.14	7.54	-	34.06	20.58	29.47	6.06	-
MGD (ours)	256×192	✓	✓		**11.88**	**2.82**	**31.48**	**1.91**	-	**10.24**	**1.55**	**30.50**	**2.58**	-	**11.87**	**2.03**	**32.05**	**2.57**	-
Paired setting																			
Stable Diff. [39]	512×384	✓			21.00	8.59	30.17	7.95	0.310	28.40	14.48	28.02	9.96	0.345	33.12	17.39	29.36	9.86	0.450
SDEdit [27]	512×384	✓	✓	✓	15.78	5.52	29.73	4.21	0.222	16.64	6.07	29.00	6.51	0.256	21.53	9.02	28.89	5.67	0.270
MGD (ours)	512×384	✓	✓	✓	**12.42**	**3.71**	**31.90**	**3.72**	**0.190**	**10.70**	**2.01**	**31.10**	**5.70**	**0.210**	**11.38**	**1.89**	**32.02**	**4.93**	**0.194**
Unpaired setting																			
Stable Diff. [39]	256×192	✓			22.86	9.73	28.31	4.29	-	28.78	13.93	26.41	4.97	-	36.31	20.74	27.84	5.67	-
FICE [35]	256×192	✓	✓		49.77	35.37	26.48	7.64	-	44.94	30.39	25.42	7.84	-	39.04	25.27	26.14	6.39	-
MGD (ours)	256×192	✓	✓		**14.50**	**3.48**	**29.24**	**2.39**	-	**13.70**	**2.48**	**29.09**	**3.32**	-	**13.72**	**2.50**	**30.37**	**3.17**	-
Unpaired setting																			
Stable Diff. [39]	512×384	✓			24.23	10.39	28.64	8.59	0.413	30.90	15.38	27.03	10.43	0.453	35.96	19.94	28.37	10.60	0.609
SDEdit [27]	512×384	✓	✓	✓	17.86	6.50	27.36	**4.78**	0.357	19.16	6.85	27.08	**7.53**	0.399	22.97	9.98	26.85	**6.42**	0.411
MGD (ours)	512×384	✓	✓	✓	**15.99**	**4.50**	**29.76**	5.41	**0.291**	**14.82**	**2.81**	**29.96**	7.96	**0.289**	**14.71**	**3.63**	**30.41**	7.15	**0.252**

Table 8: Category-wise quantitative results on the Dress Code Multimodal dataset.

Sketch Cond.	Dress Code Multimodal				
	FID↓	KID↓	CLIP-S↑	PD↓	SD↓
1.0	5.44	1.82	31.03	4.43	0.363
0.8	5.65	1.96	31.17	4.42	0.364
0.6	5.73	2.11	31.31	4.50	0.365
0.4	5.80	2.17	31.44	4.51	0.368
0.2	5.74	2.11	31.68	4.72	0.374
0.0	6.31	2.33	31.67	5.31	0.405

Table 9: Ablation study by varying the sketch conditioning steps on the paired setting of Dress Code Multimodal.

Sketch Cond.	VITON-HD Multimodal				
	FID↓	KID↓	CLIP-S↑	PD↓	SD↓
1.0	13.01	4.00	30.32	7.05	0.225
0.8	12.75	3.73	30.46	7.11	0.250
0.6	12.76	3.75	30.53	7.13	0.263
0.4	12.71	3.67	30.56	7.12	0.280
0.2	12.81	3.86	30.75	7.22	0.317
0.0	12.40	3.36	30.34	7.53	0.435

Table 10: Ablation study by varying the sketch conditioning steps on the unpaired setting of VITON-HD Multimodal.

Model	Modalities			Dress Code Multimodal				
	Text	Pose	Sketch	FID↓	KID↓	CLIP-S↑	PD↓	SD↓
	✓			**7.61**	**2.54**	**30.17**	7.22	0.527
	✓	✓		7.82	2.85	29.93	**6.26**	0.519
MGD (ours)	✓	✓	✓	7.73	2.82	30.04	6.79	**0.458**

Model	Modalities			VITON-HD Multimodal				
	Text	Pose	Sketch	FID↓	KID↓	CLIP-S↑	PD↓	SD↓
	✓			12.73	3.59	30.24	8.64	0.643
	✓	✓		**12.40**	**3.36**	30.34	7.53	0.435
MGD (ours)	✓	✓	✓	12.81	3.86	30.75	7.22	**0.317**

Table 11: Performance analysis on the unpaired setting of both datasets as input modalities vary.

Code contains full-body target models.

In Table 11, we show the performance of our MGD model when masking different input modalities. In this case, we report the results on the unpaired setting of both datasets. As it can be seen, evaluation metrics measuring the realism of the generation (*i.e.* FID and KID) are comparable among different cases, while the pose distance and sketch distance metrics correlate in general with the given input (*i.e.* with the pose map and the garment sketch, respectively). Moreover, in this case, the warped in-shop garment not fitting the model's body shape affects the pose distance metric for the Dress Code Multimodal dataset.

Finally, in Table 12 we report a comparison with the concurrent work ControlNet [57] adapted to work with the Stable Diffusion inpaint denoising network. Following the original paper, we only condition ControlNet on text plus an additional modality (*i.e.* pose or sketch). It is worth noting that across all configurations, MGD outperforms ControlNet by a significant margin.

Qualitative results. We also show additional qualitative results for both datasets. Specifically, in Fig. 14 and Fig. 15, we compare images generated by our approach and competitors using a resolution of 512×384, for Dress Code Multimodal and VITON-HD Multimodal, respectively. In-

sults [4], while the KID metric presents more accurate results. Nevertheless, our method outperforms all competitors in all metrics except for the pose metrics in the unpaired setting. This behavior is due to the imperfect match of the predicted warped unpaired sketches and the model's body shape and pose. In fact, from the analysis of the sketch conditioning steps in the unpaired setting (Table 5 of the main paper), we can see that the pose distance directly correlates with the sketch conditioning parameter, while in the paired one (Table 9) the pose distance metric decreases as the number of sketch conditioning steps increases. Instead, when evaluating the results on VITON-HD Multimodal, the pose distance metric in the unpaired setting decreases (Table 10). We believe this behavior relates to the size of the worn garment in this last dataset, which facilitates garment warping. In fact, VITON-HD features half-body images, while Dress

		Modalities			Dress Code Multimodal					VITON-HD Multimodal				
Model	Resolution	Text	Pose	Sketch	FID↓	KID↓	CLIP-S↑	PD↓	SD↓	FID↓	KID↓	CLIP-S↑	PD↓	SD↓
Paired setting														
ControlNet [57]	512×384	✓	✓		18.36	9.82	29.00	7.46	0.462	19.08	9.35	30.03	7.72	0.392
MGD (ours)	512×384	✓	✓		**6.31**	**2.33**	**31.67**	**5.31**	**0.405**	**11.07**	**3.36**	**32.27**	**6.77**	**0.318**
ControlNet [57]	512×384	✓		✓	27.23	19.01	27.07	7.54	0.436	25.44	17.05	28.31	8.16	0.298
MGD (ours)	512×384	✓		✓	**5.72**	**2.15**	**31.69**	**4.94**	**0.373**	**10.64**	**3.26**	**32.31**	**6.18**	**0.255**
Unpaired setting														
ControlNet [57]	512×384	✓	✓		20.66	11.58	27.57	8.15	0.577	21.03	10.34	28.11	8.38	0.534
MGD (ours)	512×384	✓	✓		**7.82**	**2.85**	**29.93**	**6.26**	**0.519**	**12.40**	**3.36**	**30.34**	**7.53**	**0.435**
ControlNet [57]	512×384	✓		✓	29.61	20.83	25.75	9.74	0.544	27.41	18.66	26.63	9.53	0.416
MGD (ours)	512×384	✓		✓	**7.65**	**2.70**	**30.21**	**7.50**	**0.456**	**12.65**	**3.59**	**30.69**	**7.49**	**0.320**

Table 12: Performance comparison with ControlNet on the Dress Code Multimodal and VITON-HD Multimodal datasets for both paired and unpaired settings.

Figure 10: Time window conditioned examples on Dress Code Multimodal. We report qualitative results fixing the sketch conditioning steps to around a third of diffusion steps and varying the starting conditioning timestep (*i.e.* $t_{start} = 0, 16, 34$).

family [39] and derives from the high spatial compression nature of the latent space ($8\times$ for each spatial dimension). Instead, the third example of the first row and the first two samples of the second row highlight the dependence of our model performance from the given sketch. When the geometric warping module fails to generate a sketch able to fit the model's shape, the generation task fails as well, creating unwanted artifacts (*e.g.* a sketch may be smaller than the model's body shape as in the third example of the first row, resulting in an artifact near the model's left hand).

stead, in Fig. 16 and Fig. 17, we report low-resolution qualitative comparisons. Fig. 19 shows some qualitative results varying the sketch conditioning parameter. Increasing the number of sketch conditioning steps leads to images that better follow the given sketch while slightly reducing the realism of the generated garments. Finally, we investigate the conditioning contribution in various time windows in Fig. 10. We perform this experiment by fixing the sketch conditioning steps to around a third of diffusion steps and varying the starting conditioning timestep (*i.e.* $t_{start} = 0, 16, 34$). Qualitative results show that starting the sketch conditioning in the central (*i.e.* $t_{start} = 16$, $t_{end} = 34$) or final denoising steps (*i.e.* $t_{start} = 34$, $t_{end} = 50$) leads the model to generate images that do not follow the input sketch and present artifacts.

Limitations and failure cases. Fig. 20 shows some failure cases of the proposed approach. In the first row, the first two examples show that our model sometimes fails to generate hands accurately when they occupy a limited area within the source image. This behavior is intrinsic in LDMs

Figure 11: Sample images and multimodal data from our newly collected Dress Code Multimodal dataset (fine-grained textual annotations).

Figure 12: Sample images and multimodal data from our newly collected Dress Code Multimodal dataset (coarse-grained textual annotations).

pink printed blouse

pink wisteria flo-
ral-print blouse'

women's oversized
palm jacquard top

red printed
three-quarter sleeve
red floral print shirt
red floral-print top

multicolor flo-
ral-print t-shirt only
macy
black t-shirt floral
print
black flower-print
tee

orange tee
boxy fit t-shirt'
bright orange t-shirt

black saint tropez
blouse
black lace-up peas-
ant blouse
black plus size
cold-shoulder peas-
ant

Figure 13: Sample images and multimodal data from our newly collected VITON-HD Multimodal dataset (coarse-grained textual annotations).
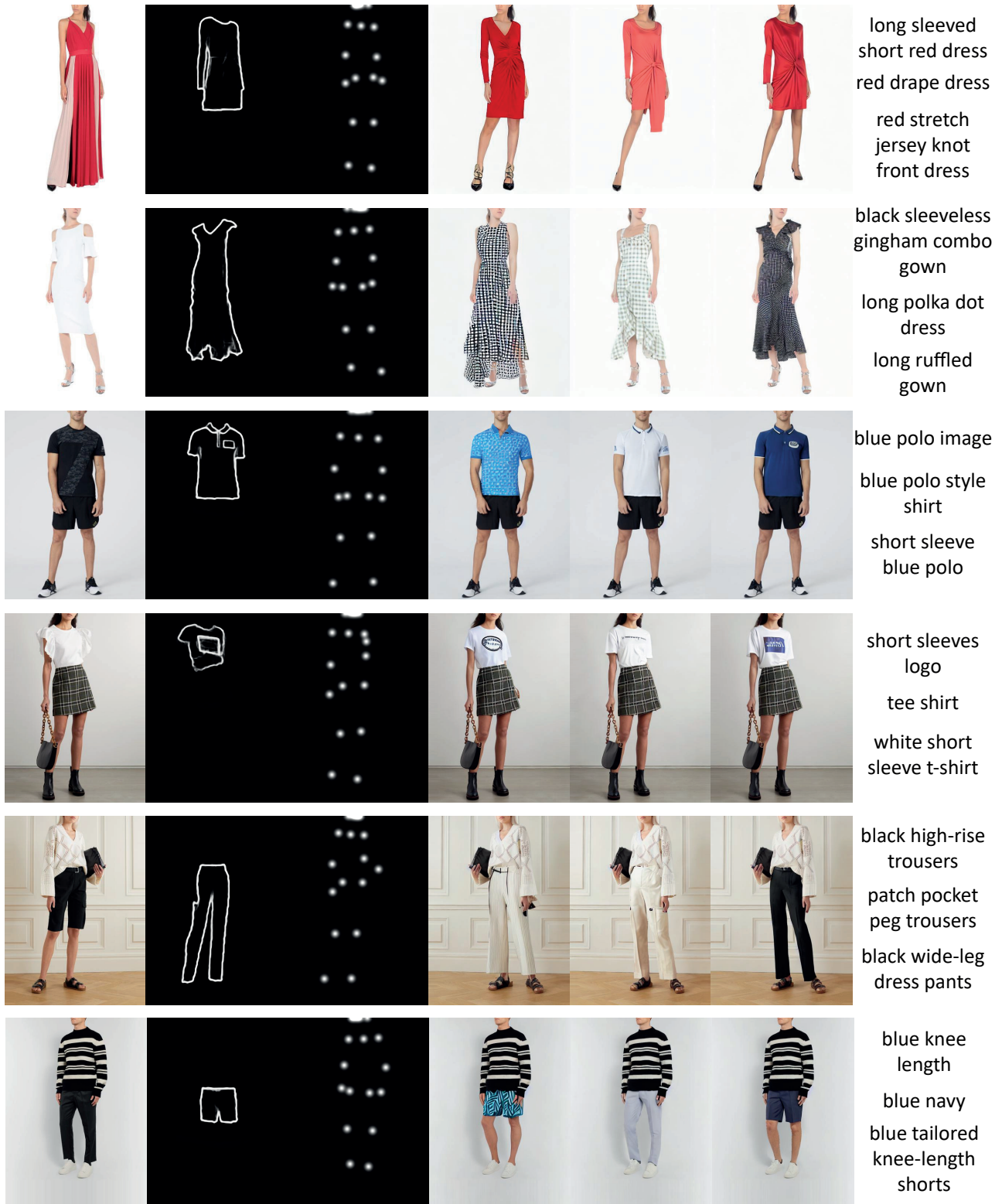
Figure 14: Qualitative comparison on Dress Code Multimodal. From left to right: model's image, input sketch, pose map, image generated by Stable Diffusion [39], image generated by SDedit [27], image generated by MGD (ours), and noun chunks.
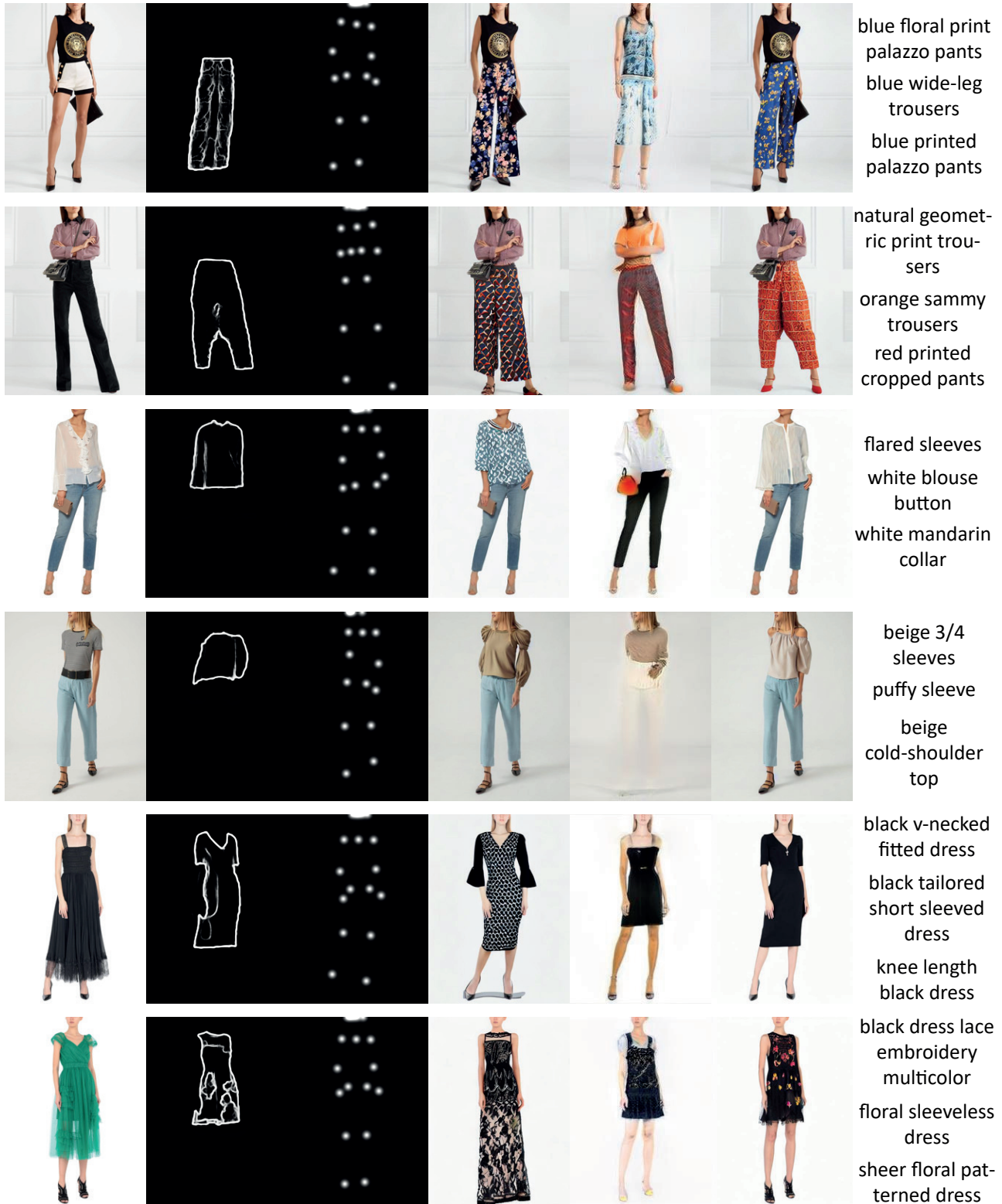
Figure 15: Qualitative comparison on VITON-HD Multimodal. From left to right: model's image, input sketch, pose map, image generated by Stable Diffusion [39], image generated by SDedit [27], image generated by MGD (ours), and noun chunks.

Figure 16: Qualitative comparison with low-resolution images on Dress Code Multimodal. From left to right: model's image, input sketch, pose map, image generated by Stable Diffusion [39], image generated by FICE [35], image generated by MGD (ours), and noun chunks.

22

Figure 17: Qualitative comparison with low-resolution images on VITON-HD Multimodal. From left to right: model's image, input sketch, pose map, image generated by Stable Diffusion [39], image generated by FICE [35], image generated by MGD (ours), and noun chunks.
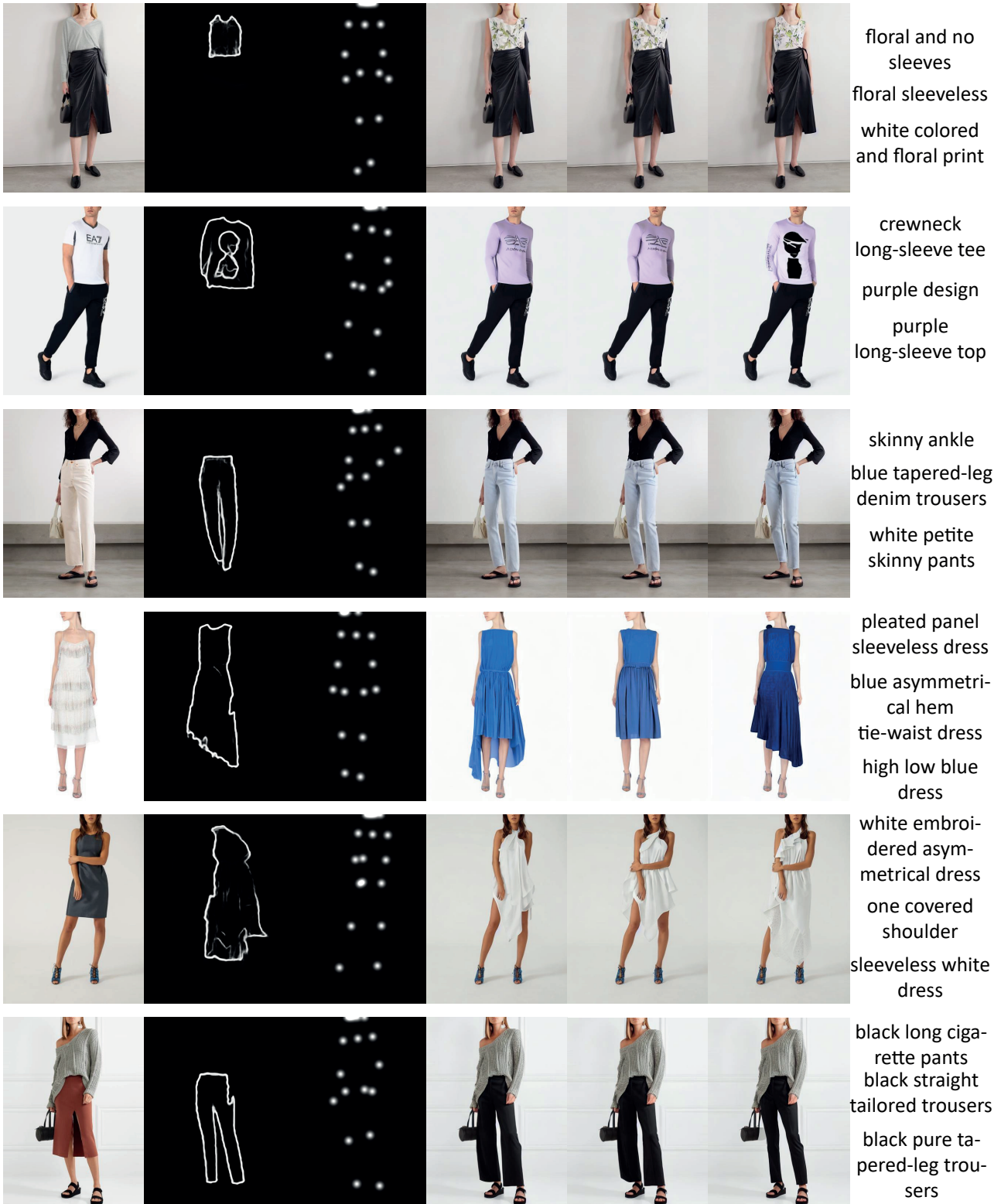
Figure 18: Qualitative comparison of images generated by our model on Dress Code Multimodal using different conditioning modalities. From left to right: model's image, input sketch, pose map, image generated using only text, image generated using text and pose map, image generated with all input modalities (*i.e.* text, pose map, and sketch).

Figure 19: Qualitative results generated by MGD increasing the sketch conditioning steps.

Figure 20: Failure cases on Dress Code Multimodal (first row) and VITON-HD Multimodal (second row).