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

Supplementary Material

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

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 [20] and Fashion200k [3], 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[1] 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 [12] 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],`
- `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].`

Figure 5: Examples of FashionIQ data type.

	Unique Captions			Unique Noun Chunks		
Dataset	**Upper**	**Lower**	**Dresses**	**Upper**	**Lower**	**Dresses**
FashionIQ [20]	27,339	0	15,101	7,801	0	3,592
Fashion200k [3]	25,959	11,022	16,694	22,898	13,420	15,890

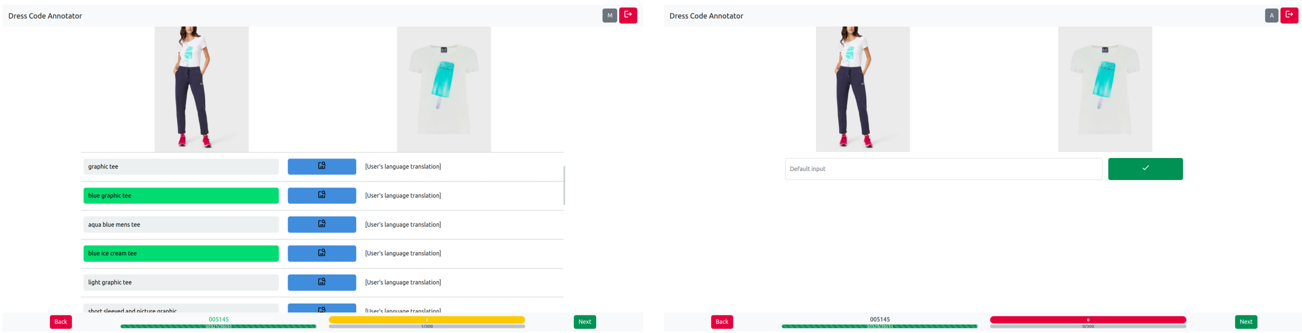
Table 6: Number of unique captions and noun chunks for each category of the FashionIQ and Fashion200k datasets.

A.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 inter-

[1]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.

face. 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 [19] model pre-trained on the English portion of the LAION-5B dataset.

CLIP Finetuning. We finetune both encoders of the OpenCLIP model using a single NVIDIA A100 GPU for 400 steps, with a batch size of 2048 and a learning rate of 10^{-6}. As optimizer, we use AdamW [8] with a weight decay of 0.2. We use mixed precision [10] 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 [18], refining the results with a U-Net based component [15].

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 [13] (*i.e.* TPS_θ). We then use the θ parameters to compute the coarse warped garment $\hat{C}$ starting from the 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 [5]) to compute the difference between the feature maps of generated and target garments extracted with a VGG-19 [16]. We train with a resolution of 256×192, Adam [6] 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.

Dataset	Ann.	Split	Images			Unique Noun Chunks		
			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.

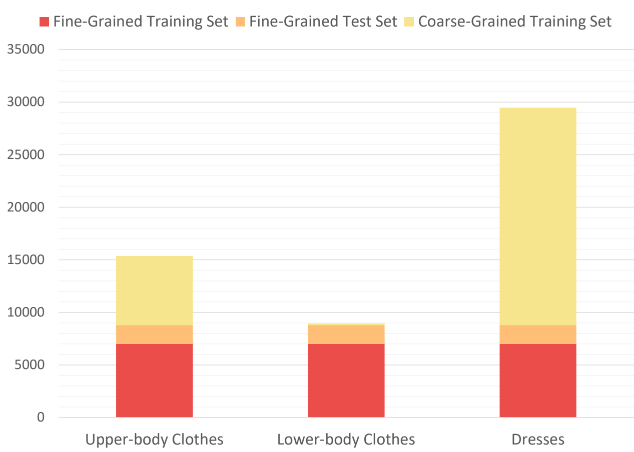


Figure 7: Annotated images per category on Dress Code Multimodal.

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.

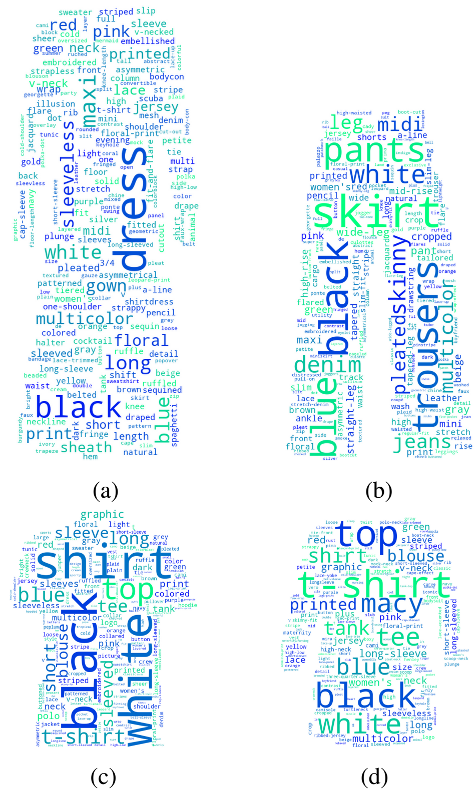


Figure 8: Vocabulary of the frequent words scaled by frequency for dresses (a), lower-body clothes (b), upper-body clothes (c) of Dress Code Multimodal and clothing items of VITON-HD Multimodal (d).

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 [4] 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 [2] 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 [7]) 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:

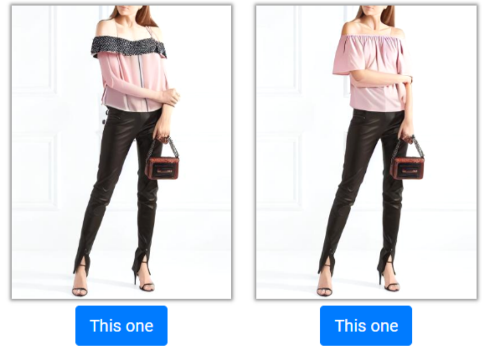
$$\text{PD}(I, \tilde{I}) = \frac{\sum\limits_{\substack{k \in \mathcal{K}(I)_M \\ \tilde{k} \in \mathcal{K}(\tilde{I})_M}} \sqrt{(k_x - \tilde{k}_x)^2 + (k_y - \tilde{k}_y)^2} \cdot \text{CF}_{k\tilde{k}}}{\sum_{k\tilde{k}} \text{CF}_{k\tilde{k}}}, \quad (7)$$

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.

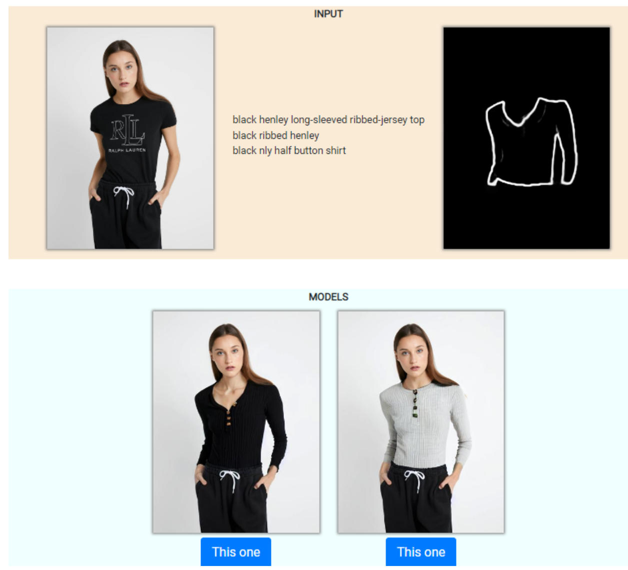
Sketch Distance (SD). To evaluate the adherence of the generated images to the constraints imposed by the input sketch, we propose a new sketch distance metric. To compute the metric, we first extract the ground-truth and the generated garments label maps using an off-the-shelf semantic segmentation model[2]. 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 [17]). 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:

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

[2]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.

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

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. [14]	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 [11]	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. [14]	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 [9]	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. [14]	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 [11]	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. [14]	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 [9]	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.

surveys. The former (Fig. 9a) assesses the realism of the 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 results [1], 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 Code contains full-body target models.

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**
	Modalities			**VITON-HD Multimodal**				
Model	**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.

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 com-

		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 [21]	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 [21]	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 [21]	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 [21]	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.

parable 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 [21] 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. Instead, 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

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.* $t_{start} = 0, 16, 34$).

within the source image. This behavior is intrinsic in LDMs family [14] and derives from the high spatial compression nature of the latent space (8× 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).

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).

Figure 13: Sample images and multimodal data from our newly collected VITON-HD Multimodal dataset (coarse-grained textual annotations).

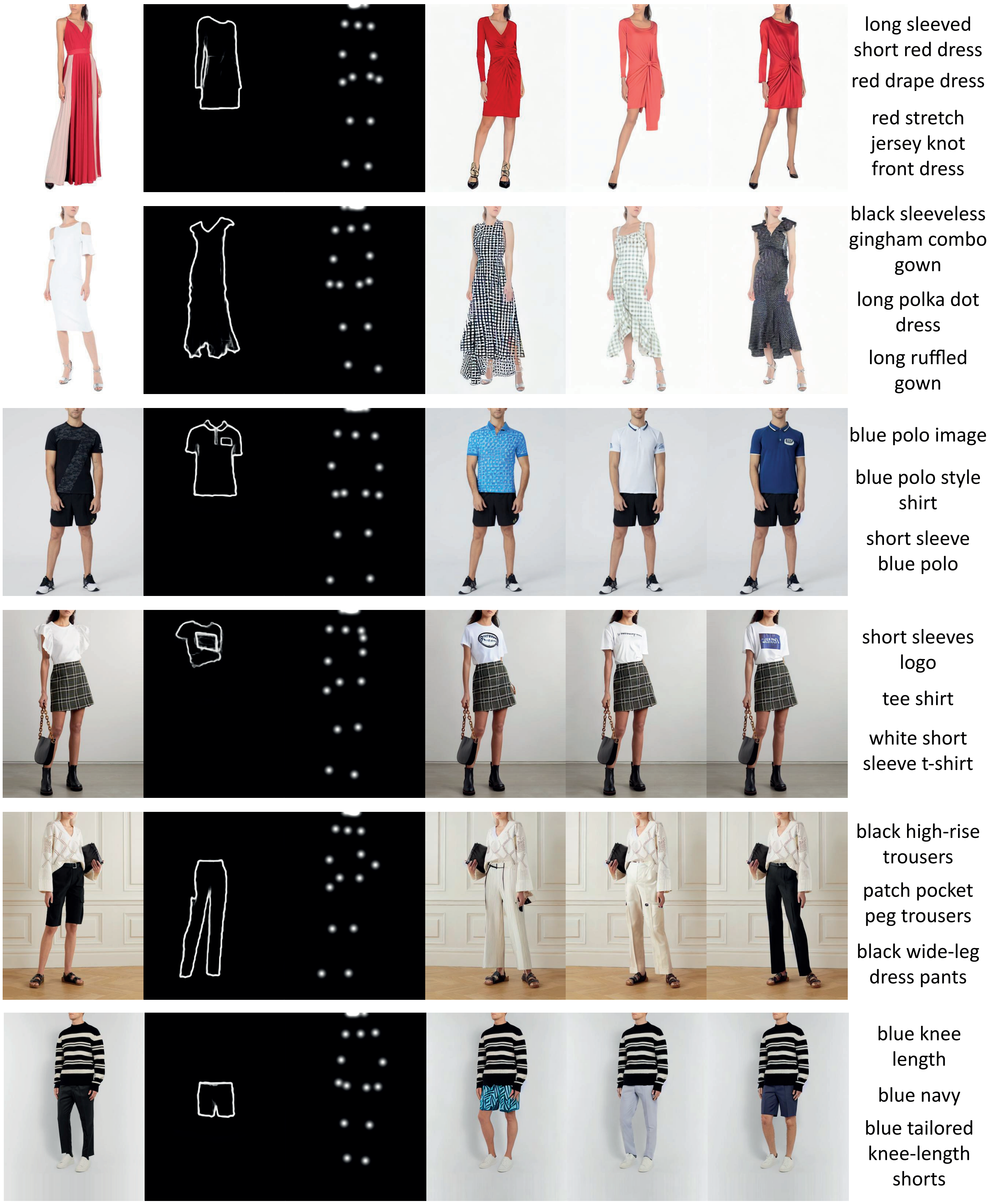


Figure 14: Qualitative comparison on Dress Code Multimodal. From left to right: model's image, input sketch, pose map, image generated by Stable Diffusion [14], image generated by SDedit [9], image generated by MGD (ours), and noun chunks.

Figure 15: Qualitative comparison on VITON-HD Multimodal. From left to right: model's image, input sketch, pose map, image generated by Stable Diffusion [14], image generated by SDedit [9], 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 [14], image generated by FICE [11], image generated by MGD (ours), and noun chunks.

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 [14], image generated by FICE [11], image generated by MGD (ours), and noun chunks.

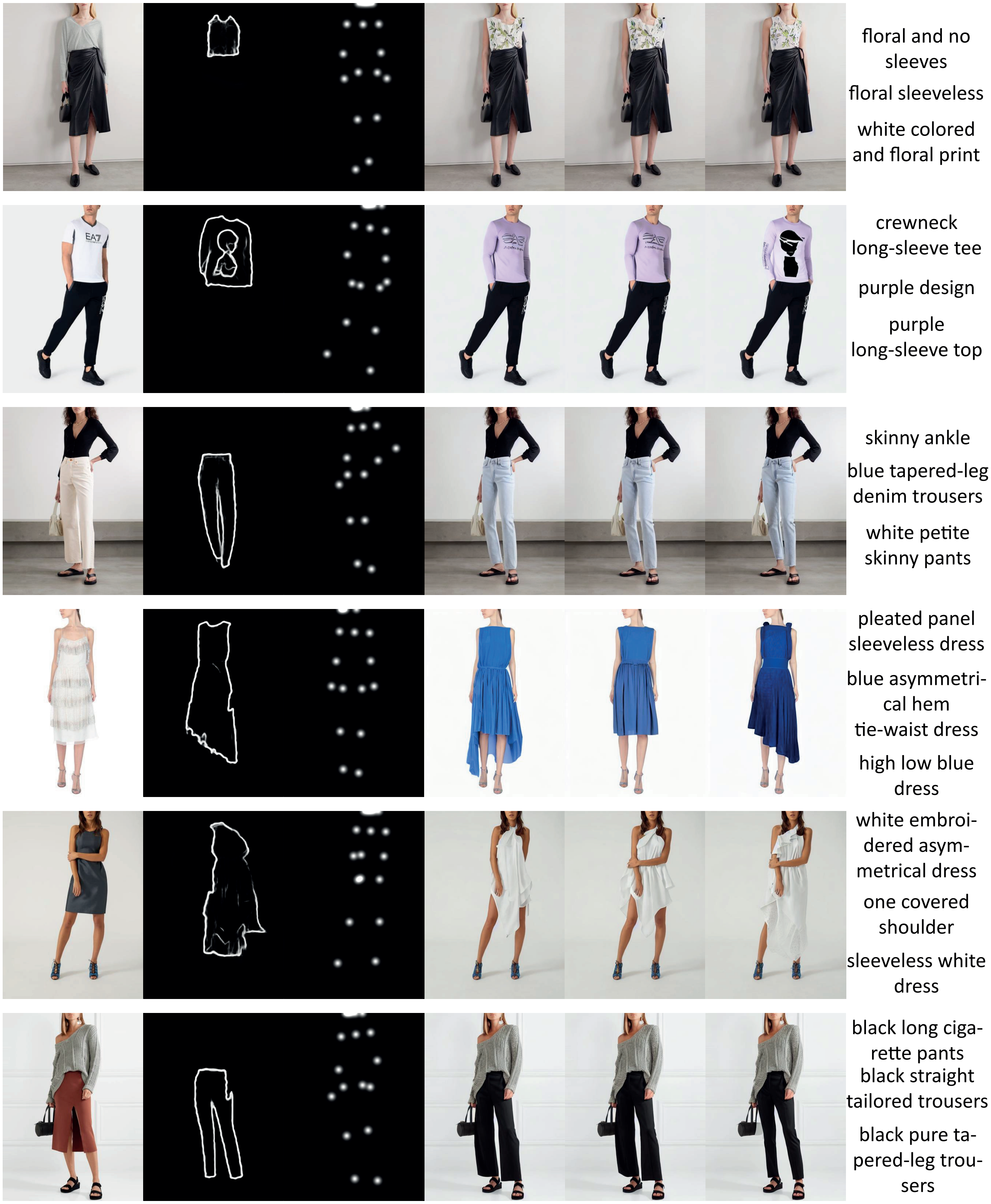


Figure 18: Qualitative comparison of images generated by our model on Dress Code Multimodal using different conditioning modalities. From left to right: model's image, input sketch, pose map, image generated using only text, image generated using text and pose map, image generated with all input modalities (*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).

References

[1] Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *ICLR*, 2018. 5

[2] 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 PyTorch. *Journal of Open Source Software*, 7(70):4101, 2022. 3

[3] 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. 1

[4] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*, 2021. 3

[5] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 2

[6] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015. 2

[7] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. OpenPifPaf: Composite Fields for Semantic Keypoint Detection and Spatio-Temporal Association. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):13498–13511, 2021. 4

[8] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019. 2

[9] 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. 5, 10, 11

[10] 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. 2

[11] 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. 5, 12, 13

[12] 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. 1

[13] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *CVPR*, 2017. 2

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *CVPR*, 2022. 5, 6, 10, 11, 12, 13

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*, 2015. 2

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 2

[17] 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. 4

[18] 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. 2

[19] 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. 2

[20] 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. 1

[21] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 6