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].

\(^1\)https://spacy.io/

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Unique Captions</th>
<th>Unique Noun Chunks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Upper Lower</td>
<td>Upper Lower Dresses</td>
</tr>
<tr>
<td>FashionIQ [20]</td>
<td>27,339 0 15,101</td>
<td>7,801 0 3,592</td>
</tr>
</tbody>
</table>

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-
The 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.

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 $\theta$ of a thin-plate spline geometric transformation [13] (i.e. $\text{TPS}_\theta$). We then use the $\theta$ parameters to compute the coarse warped garment $\hat{C}$ starting from the in-shop garment $C$ as follows:

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

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 $256 \times 192$, 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.
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 ($\cap$). 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 [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 back-
ground, obtaining a final resolution equal to $224 \times 224$. The adopted metric is defined as follows:

$$\text{CLIP-S}(I, Y) = \max(100 \times \cos(E_I, E_Y), 0),$$

(6)

where $E_I$ 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 $K$ (i.e. in our case, we use OpenPifPaf [7]) and identify the set of keypoints falling in the mask $M$ as $K(\cdot)_{M}$. We compute the final score with an $\ell_2$ distance between each pair of real-generated corresponding keypoints (i.e. $k \in K(I)_{M}$ and $\tilde{k} \in K(\tilde{I})_{M}$, respectively), weighting each keypoint distance with the detector confidence to consider possible estimation errors. Formally, our pose distance metric is defined as follows:

$$\text{PD}(I, \tilde{I}) = \frac{\sum_{k \in K(I)_{M}} \sum_{\tilde{k} \in K(\tilde{I})_{M}} \sqrt{(x_k - x_{\tilde{k}})^2 + (y_k - y_{\tilde{k}})^2} \cdot \text{CF}_{k\tilde{k}}}{\sum_{k\tilde{k}} \text{CF}_{k\tilde{k}}},$$

(7)

where, for each pair of real-generated keypoints, $\text{CF}_{k\tilde{k}}$ is 1 if the confidence of the detector $K$ 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 model². We segment the garment according to its category and paste it on a white background of shape $512 \times 384$. 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 $\text{Edge}$ (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:

$$\text{SD}(I_S, \tilde{I}_S) = \text{MSE}(\text{Edge}(I_S), \text{Edge}(\tilde{I}_S)) \cdot p,$$

(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

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.
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.

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Resolution</th>
<th>Text</th>
<th>Keypoints</th>
<th>Sketch</th>
<th>FID ↓</th>
<th>KID ↓</th>
<th>CLIP-S ↑</th>
<th>PD ↓</th>
<th>SD ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stable Diff. [14]</td>
<td>256×192</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>22.86</td>
<td>9.73</td>
<td>31.81</td>
<td>7.52</td>
<td>0.425</td>
</tr>
<tr>
<td>FICE [11]</td>
<td>256×192</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>46.41</td>
<td>32.26</td>
<td>28.58</td>
<td>7.46</td>
<td>-</td>
</tr>
<tr>
<td>MGD (ours)</td>
<td>256×192</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>11.88</td>
<td>2.82</td>
<td>31.48</td>
<td>1.91</td>
<td>-</td>
</tr>
</tbody>
</table>

| Stable Diff. [14] | 512×384 | ✓ | ✓ | ✓ | 21.00 | 8.59 | 30.17 | 7.95 | 0.310 |
| SDEdit [9] | 512×384 | ✓ | ✓ | ✓ | 15.78 | 5.82 | 29.73 | 4.21 | 0.222 |
| MGD (ours) | 512×384 | ✓ | ✓ | ✓ | 12.82 | 3.71 | 31.90 | 0.372 | 0.190 |

| MGD (ours) | 256×192 | ✓ | ✓ | ✓ | 22.86 | 9.73 | 31.81 | 7.52 | 0.425 |
| FICE [11] | 256×192 | ✓ | ✓ | ✓ | 49.77 | 35.37 | 26.48 | 7.64 | - |
| MGD (ours) | 256×192 | ✓ | ✓ | ✓ | 14.50 | 3.48 | 29.24 | 2.39 | - |

In Table 9, 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 input modalities vary.

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

<table>
<thead>
<tr>
<th>Sketch Cond.</th>
<th>FID ↓</th>
<th>KID ↓</th>
<th>CLIP-S ↑</th>
<th>PD ↓</th>
<th>SD ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>5.44</td>
<td>1.82</td>
<td>31.03</td>
<td>4.43</td>
<td>0.363</td>
</tr>
<tr>
<td>0.8</td>
<td>5.65</td>
<td>1.96</td>
<td>31.17</td>
<td>4.42</td>
<td>0.364</td>
</tr>
<tr>
<td>0.6</td>
<td>5.75</td>
<td>2.11</td>
<td>31.31</td>
<td>4.50</td>
<td>0.365</td>
</tr>
<tr>
<td>0.4</td>
<td>5.80</td>
<td>2.17</td>
<td>31.44</td>
<td>4.51</td>
<td>0.368</td>
</tr>
<tr>
<td>0.2</td>
<td>5.75</td>
<td>2.11</td>
<td>31.68</td>
<td>4.72</td>
<td>0.374</td>
</tr>
<tr>
<td>0.0</td>
<td>6.31</td>
<td>2.33</td>
<td>31.67</td>
<td>5.31</td>
<td>0.405</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Sketch Cond.</th>
<th>FID ↓</th>
<th>KID ↓</th>
<th>CLIP-S ↑</th>
<th>PD ↓</th>
<th>SD ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>13.01</td>
<td>4.00</td>
<td>30.32</td>
<td>7.05</td>
<td>0.225</td>
</tr>
<tr>
<td>0.8</td>
<td>12.75</td>
<td>3.73</td>
<td>30.46</td>
<td>7.11</td>
<td>0.250</td>
</tr>
<tr>
<td>0.6</td>
<td>12.75</td>
<td>3.75</td>
<td>30.53</td>
<td>7.13</td>
<td>0.263</td>
</tr>
<tr>
<td>0.4</td>
<td>12.71</td>
<td>3.67</td>
<td>30.56</td>
<td>7.12</td>
<td>0.280</td>
</tr>
<tr>
<td>0.2</td>
<td>12.81</td>
<td>3.86</td>
<td>30.75</td>
<td>7.22</td>
<td>0.317</td>
</tr>
<tr>
<td>0.0</td>
<td>12.40</td>
<td>3.36</td>
<td>30.34</td>
<td>7.53</td>
<td>0.435</td>
</tr>
</tbody>
</table>

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.

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Text</th>
<th>Pose</th>
<th>Sketch</th>
<th>FID ↓</th>
<th>KID ↓</th>
<th>CLIP-S ↑</th>
<th>PD ↓</th>
<th>SD ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

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. As it can be seen, evaluation metrics measuring the realism of the generation (i.e. FID and KID) are comparable.

In fact, VITON-HD features half-body images, while Dress Code contains full-body target models. 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.

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.

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Resolution</th>
<th>Modalities</th>
<th>Dress Code Multimodal</th>
<th>VITON-HD Multimodal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>FID ↓ KID ↓ CLIP-S ↑</td>
<td>PD ↓ SD ↓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paired setting</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ControlNet [21]</td>
<td>512×384</td>
<td>✓ ✓</td>
<td>18.36 9.82 29.00</td>
<td>7.46 0.462 19.08 9.35 30.03 7.72 0.392</td>
</tr>
<tr>
<td>MGD (ours)</td>
<td>512×384</td>
<td>✓ ✓</td>
<td>6.31 2.33 31.67</td>
<td>5.31 0.405 11.07 3.36 32.27 6.77 0.318</td>
</tr>
<tr>
<td>ControlNet [21]</td>
<td>512×384</td>
<td>✓ ✓</td>
<td>27.23 19.01 27.07</td>
<td>7.54 0.436 25.44 17.05 28.31 8.16 0.298</td>
</tr>
<tr>
<td>MGD (ours)</td>
<td>512×384</td>
<td>✓ ✓</td>
<td>5.72 2.15 31.69</td>
<td>4.94 0.373 10.64 3.26 32.31 6.18 0.255</td>
</tr>
</tbody>
</table>

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 |

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. \( t_{\text{start}} = 0, 16, 34 \)). Qualitative results show that starting the sketch conditioning in the central (i.e. \( t_{\text{start}} = 16, t_{\text{end}} = 34 \)) or final denoising steps (i.e. \( t_{\text{start}} = 34, t_{\text{end}} = 50 \)) 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 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.
blue floral print palazzo pants
blue wide-leg trousers
blue printed trousers
natural geometric print trousers
orange sammy trousers
red printed cropped pants
flared sleeves
white blouse button
white mandarin collar
beige 3/4 sleeves
puffy sleeve
beige cold-shoulder top
black v-necked fitted dress
black tailored short sleeved dress
knee length black dress
black dress lace embroidery
multicolor floral sleeveless dress
sheer floral patterned dress

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


[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


