

Magic Clothing: Controllable Garment-Driven Image Synthesis

Supplementary Material

1 MP-LPIPS DETAILS

Table 1 provides the detailed parameter settings for Matched-Points-LPIPS (MP-LPIPS), which are omitted in the paper for brevity. We first uniformly sample points \mathbb{P}_G inside the garment area of the garment image I_G given the garment mask from the VITON-HD test dataset [1] to ensure the sample distance $d_s = 40$ pixels between each point and its nearest neighbour. For sampled points \mathbb{P}_G on I_G , we retrieve their corresponding matching points \mathbb{P}_C on the target character I_C using the diffusion features [4] at time step $t = 41$ and 11-th layer inside UNet upsampling blocks. We eventually calculate MP-LPIPS as the average LPIPS distance of the patches $\mathbb{P}_G, \mathbb{P}_C$ centred on matched points \mathbb{P}_G and \mathbb{P}_C , where the patch size $s = 33 \times 33$ pixels and we black out the area except for the garment part. If any of \mathbb{P}_C fall outside the garment area of I_C by more than $\tau_s = 17$ pixels, we view those points as mismatched and set their LPIPS to 0.6 as the mismatch penalty. Here the garment area of I_C is obtained by applying HumanParsing [2] on I_C .

Name	Value
time step t	41
layer l	11
patch size s	33×33
sample distance d_s	40
sample threshold τ_s	17
mismatch penalty p	0.6

Table 1: Detailed Settings for the MP-LPIPS.

2 FAILURE CASES

In Figure 1 we present several failure cases as discussed in the paper. While our model achieves remarkable results in garment-driven image synthesis and is compatible with various finetuned LDMs, the styles of generated images are highly dependent on the base diffusion models. In practice, we apply our model to Stable Diffusion V1.5 [3] finetuned on the photorealistic data (i.e., Realistic Vision V4.0¹) and Stable Diffusion V1.5 finetuned on the anime-style data (i.e., Counterfeit V3.0²), respectively. As shown in the first row of Figure 1, our model tends to generate realistic characters with Realistic Vision V4.0 and anime-style characters with Counterfeit V3.0, regardless of the target styles specified by the text prompts. Another limitation is that due to limited garment types of training samples in the VITON-HD dataset, our model may not perform perfectly on complicated garments. As can be seen in the second row of figure 1, our model fails to preserve the middle layer of the given down jacket and slightly reduces the size of the overcoat.



Figure 1: Failure cases of our Magic Clothing. Based on Realistic Vision V4.0, it fails to generate an anime-style character (1st row left), and vice versa for Counterfeit V3.0 (1st row right). Example results of complicated garments like down jackets and coats are shown in the second row.

REFERENCES

- [1] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. 2021. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14131–14140.
- [2] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. 2020. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 6 (2020), 3260–3271.
- [3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
- [4] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. 2023. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems* 36 (2023), 1363–1389.

¹<https://huggingface.co/SG161222>

²<https://huggingface.co/gsd/Counterfeit-V3.0>