Learning Flow Fields in Attention for Controllable Person Image Generation

Supplementary Material

In the supplementary material, we provide additional experimental details along with qualitative and quantitative results. Additionally, we discuss our diffusion-based baseline and *Leffa* loss, along with their limitations.

A. Experimental Details

A.1. Datasets

In this section, we provide a detailed introduction to the three datasets used in our study.

VITON-HD [6] dataset is the most commonly used dataset for virtual try-on task. The training set contains 11,647 person and garment image pairs, and the test set contains 2,032 pairs. All images are front-view, upper-body garments with a resolution of 1024×768.

DressCode [34] dataset is composed of various types of garments, comprising 48,392 person and garment image pairs in the training set. This includes 13,563 upper-body, 7,151 lower-body, and 27,678 dress garment pairs. The test set contains 5,400 pairs, evenly distributed across 1,800 pairs each for upper-body, lower-body, and dress garments. All images have a resolution of 1024×768.

DeepFashion [29] dataset includes high resolution 52,712 person images in the fashion domain. Following Zhu et al. [77], we split the dataset into training and test subsets with 101,966 and 8,570 pairs, respectively. Each pair includes the same person in the same garment but with different poses.

B. More Results for DressCode Dataset

To further validate our performance, we conduct separate evaluations on different garment categories for the DressCode dataset. In Tab. 5, results show that our method significantly outperforms previous methods across all garment categories. Specifically, for the upper body category, it achieves an FID reductions of -2.33 (paired) and -0.91 (unpaired); for the lower body category, -2.94 (paired) and -2.55 (unpaired); and for dresses, -1.25 (paired) and -1.71 (unpaired).

C. More Ablation Studies

Unless stated otherwise, the ablation studies are conducted on the VITON-HD dataset for the virtual try-on task, in alignment with the main paper.

Effect of $\mathcal{L}_{\text{leffa}}$. To validate the effectiveness of *Leffa* loss, we conduct ablation studies on the DressCode dataset for virtual try-on and the DeepFashion dataset for pose transfer. i) As shown in Tab. 6, our *Leffa* reduces FID by -1.58 in the paired setting and -1.5 in the unpaired setting. ii) Tab. 7 further shows an FID reduction of -1.49 on pose transfer task. These results confirm that *Leffa* loss significantly improves controllable person image generation, enhancing both appearance and pose control.

Qualitative impact of our *Leffa* loss $\mathcal{L}_{\text{leffa}}$. In Fig. 8, we visualize attention maps with varying λ_{leffa} to assess its impact on model training. The first row shows generated results with different λ_{leffa} values, with $\lambda_{\text{leffa}} = 0$ indicating no $\mathcal{L}_{\text{leffa}}$. Rows 2 to 5 highlight reference key regions attended by the target query, marked by arrows of various colors. Without $\mathcal{L}$leffa, attention is dispersed. Increasing λ_{leffa} improves focus, guiding the query to correct regions.

Method	paired				unpaired	
	FID ↓	KID ↓	SSIM ↑	LPIPS ↓	FID ↓	KID ↓
upper body						
FS-VTON [19]	11.29	3.65	0.941	0.035	16.34	5.93
HR-VITON [26]	15.36	5.27	0.916	0.071	16.82	5.70
GP-VTON [62]	7.38	0.74	0.945	0.039	12.21	1.19
LADI-VTON [35]	9.53	0.20	0.928	0.049	13.26	2.67
DCI-VTON [15]	7.47	1.07	0.942	0.041	11.64	0.86
StableVITON [23]	9.94	0.12	0.937	0.039	-	-
OOTDiffusion [63]	11.03	0.29	-	-	-	-
***Leffa* (Ours)**	**5.05**	**0.02**	**0.949**	**0.021**	**10.73**	**0.77**
lower body						
FS-VTON [19]	11.65	3.82	0.934	0.053	22.43	9.81
HR-VITON [26]	11.41	3.20	0.937	0.045	16.39	4.31
GP-VTON [62]	7.73	0.71	0.938	0.042	16.70	2.89
LADI-VTON [35]	8.52	1.04	0.922	0.051	14.80	3.13
DCI-VTON [15]	7.97	0.96	0.939	0.045	15.45	1.60
***Leffa* (Ours)**	**4.79**	**0.05**	**0.941**	**0.024**	**12.25**	**1.66**
dresses						
FS-VTON [19]	13.04	4.44	0.888	0.070	20.95	8.96
HR-VITON [26]	16.82	4.89	0.865	0.113	18.81	5.41
GP-VTON [62]	7.44	**0.32**	0.881	0.073	12.64	1.83
LADI-VTON [35]	9.07	1.12	0.868	0.089	13.40	2.50
DCI-VTON [15]	8.48	1.08	0.887	0.070	12.35	1.36
***Leffa* (Ours)**	**6.19**	**0.32**	**0.891**	**0.044**	**10.64**	**0.59**

Table 5. Quantitative results comparison with other methods on the DressCode dataset for virtual try-on. Our *Leffa* achieves state-of-the-art results across all categories of garment.

Method	$\mathcal{L}_{\text{leffa}}$	paired				unpaired	
		FID ↓	KID ↓	SSIM ↑	LPIPS ↓	FID ↓	KID ↓
Ours	✗	3.64	0.33	0.911	0.040	5.98	1.42
	✓	**2.06**	**0.07**	**0.924**	**0.031**	**4.48**	**0.62**

Table 6. Ablation study on DressCode dataset for virtual try-on. Our *Leffa* loss significantly improves model performance.

Method	$\mathcal{L}_{\text{leffa}}$	FID ↓	SSIM ↑	LPIPS ↓
Ours (512×352)	✗	5.72	0.744	0.153
	✓	**4.23**	**0.755**	**0.119**

Table 7. Ablation study on the DeepFashion dataset for pose transfer. Our *Leffa* loss significantly improves model performance.

Method	paired				unpaired	
	FID ↓	KID ↓	SSIM ↑	LPIPS ↓	FID ↓	KID ↓
Leffa (Ours)	**4.54**	**0.05**	**0.899**	**0.048**	**8.52**	**0.32**
Leffa w/o average A	6.02	0.74	0.863	0.072	9.78	0.98
Leffa w/o upsample $\mathcal{F}$	4.94	0.32	0.888	0.064	9.33	0.78

Table 8. Ablation study for the proposed *Leffa* loss.

Method	paired				unpaired	
	FID ↓	KID ↓	SSIM ↑	LPIPS ↓	FID ↓	KID ↓
Our baseline	**5.31**	**0.30**	0.885	0.058	**9.38**	**0.91**
freeze Reference UNet	6.42	0.77	0.863	0.066	10.63	1.32
+ CLIP visual feature	5.33	0.31	**0.886**	**0.056**	9.40	0.95
+ CLIP textual feature	5.37	0.40	0.876	0.060	9.45	0.98

Table 9. Ablation study for our diffusion-based baseline. Making both the generative and reference UNets trainable is key to performance improvement for our diffusion-based baseline.

However, when $\lambda_{\text{leffa}} > 10^{-3}$, the attention becomes overly narrow and less accurate, hindering image generation.

Effect of averaging attention map across multi-head. Inspired by [11], we average the attention maps across all heads before computing the *Leffa* loss. As shown in Tab. 8 (*Leffa* w/o average A), computing the loss for each head individually degrades performance, emphasizing the role of redundancy in attention maps for better generalization.

Effect of upsampling flow fields. To evaluate the impact of upsampling the flow fields to image resolution, we experiment with retaining their latent resolution for the *Leffa*, requiring the ground truth image to be downsampled. This resizing reduces detail, hindering accurate supervision. As shown in Tab. 8 (*Leffa* w/o upsample $\mathcal{F}$), not upsampling the flow fields results in a performance drop, further supporting our viewpoint.

D. Discussion

Why does our diffusion-based baseline perform comparably to state-of-the-art methods? Our baseline is similar to existing virtual try-on and pose transfer methods [2, 7, 18, 63], but we find that complex designs are unnecessary for strong performance. The key lies in making both UNets in the dual architecture fully trainable, as freezing one significantly degrades results. As shown in Tab. 9, freezing the reference UNet (as done in Choi et al. [7]) leads to a significant performance drop (FID reduction of -1.11/-1.25 for the paired/unpaired settings). In contrast, adding CLIP visual and textual features [7, 63] results in only a slight performance decline. This highlights that the key to improving the baseline lies in making both UNets trainable, while adding more complex designs (*e.g.*, add CLIP, textual information) is unnecessary.

Why not add $\mathcal{L}_{\text{leffa}}$ from the first training stage? At the start of training, the attention maps are not well learned, and applying our *Leffa* loss too early forces the target query to prematurely attend to the reference key, hindering convergence rather than accelerating it. Instead, introducing our *Leffa* loss in a subsequent training stage significantly enhances performance, demonstrating its ability to correct inaccurate attention and guide the model toward more effective learning.

E. Limitation

While *Leffa* significantly improves controllable person image generation in appearance and pose control, it has several limitations. First, it requires multi-stage training with the *Leffa* loss applied only in the final stage. In future work, we aim to design a single-stage model to simplify the training process. Second, appearance control relies on garment segmentation, which impacts performance when segmentation is inaccurate. We plan to develop a mask-free approach to ensure high quality generation and preserve fine-grained details without distortion. Third, our method struggles to preserve extremely fine-grained details, such as small text, due to the $8\times$ resolution compression brought by the latent encoder in the latent diffusion model. It is worth noting that the issues mentioned above are not unique to our method but are also present in other methods.

F. More Qualitative Comparison

To further demonstrate the effectiveness of our *Leffa*, we conduct qualitative comparisons with other methods on the VITON-HD, DressCode, and DeepFashion datasets, as shown in Figs. 9–14. The results indicate that our *Leffa* not only generates images with higher overall quality but also significantly alleviates fine-grained detail distortion.

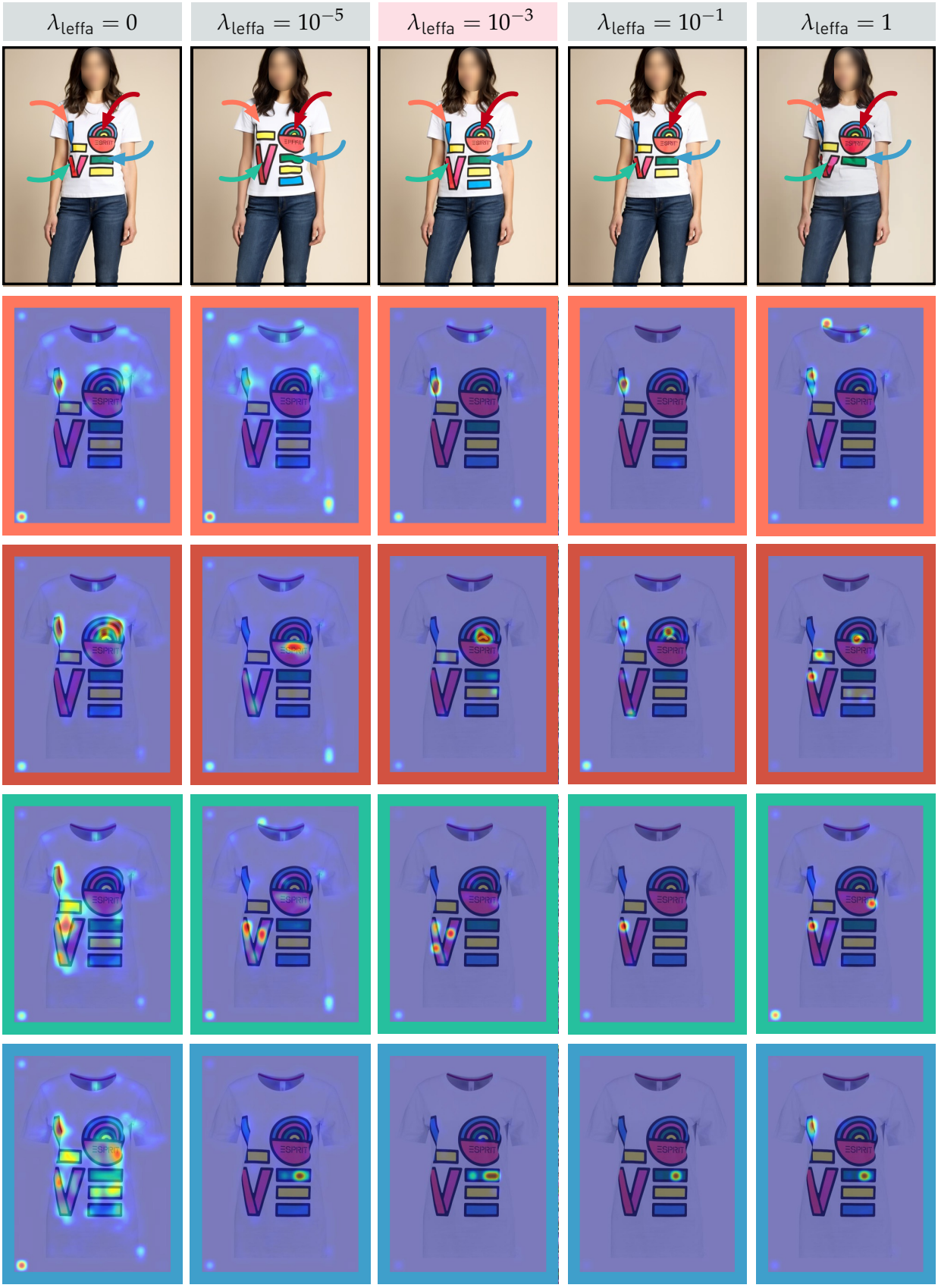

Figure 8. More visualizations of feature maps to assess the impact of *Leffa* loss. The third column is the optimal setting used in our paper.