DEFT-VTON: Efficient Virtual Try-On with Consistent Generalised H-Transform

Supplementary Material

1. Ablation study

We provide ablation studies on the coefficients of $\mathcal{L}_{DSM}$ as well as the number of steps.

$\mathcal{L}_{DSM}$ coefficient While the table shows better SSIM and LPIPS scores for fewer steps (15) with changing $\mathcal{L}_{DSM}$ coefficients, the FID and KID scores, which assess perceptual quality, are worse. Empirically, we observe that the consistency finetuned models better preserve garment colors and complex text/graphics, as shown in Figure 1, on some challenging tasks involving complex text/graphics preservation, unclear reference images, and tasks that are rare in the training dataset, 15 steps sampling of the consistency finetuned model qualitatively outperforms the one only finetuned with DEFT loss, reaffirming the quantitative observation that consistency finetuning improves sampling with fewer steps.

We also observe that, when the coefficient on $\mathcal{L}_{DSM}$ is overly small, the DEFT-VTON model loses its VTO abilities.

Number of sampling steps Table 1 shows that the SSIM and LPIPS scores exhibit an initial rise and subsequent decline as sampling steps increase in the consistency finetuned DEFT-VTON model, implying an optimal performance at approximately 15 steps. We also observe that the FID and KID score consistently improve as we increase the number of sampling steps. Although this shows improvements in human perception, it does not always translate to better VTO results, with rising cases of hallucinations.

Models	VITON-HD			
	SSIM ↑	LPIPS ↓	FID ↓	KID ↓
StableVTON	0.8543	0.0905	11.054	3.914
LaDI-VTON	0.8603	0.0733	14.648	8.754
IDM-VTON	0.8499	0.0603	9.842	1.123
OOTDiffusion	0.8187	0.0876	12.408	4.680
CatVTON	0.8704	0.0565	9.015	1.091
DEFT, $\mathcal{L}_{DSM}$, 25 steps	0.9118	<u>0.0533</u>	**8.3351**	**0.5212**
DEFT, $\mathcal{L}_{DSM}$, 15 steps	0.9098	**0.0521**	8.6339	0.7916
DEFT, $0.2\mathcal{L}_{DSM} + \mathcal{L}_{CM}$, 12 steps	0.9063	0.0782	25.7948	15.52
DEFT, $0.2\mathcal{L}_{DSM} + \mathcal{L}_{CM}$, 15 steps	0.9064	0.0798	27.8843	18.92
DEFT, $0.2\mathcal{L}_{DSM} + \mathcal{L}_{CM}$, 25 steps	0.9034	0.0853	33.2223	24.53
DEFT, $0.4\mathcal{L}_{DSM} + \mathcal{L}_{CM}$, 12 steps	<u>0.9135</u>	0.0553	9.1671	1.2612
DEFT, $0.4\mathcal{L}_{DSM} + \mathcal{L}_{CM}$, 15 steps	**0.9136**	0.0552	8.7922	0.9970
DEFT, $0.4\mathcal{L}_{DSM} + \mathcal{L}_{CM}$, 25 steps	0.9098	0.0581	<u>8.4704</u>	0.6649
DEFT, $0.6\mathcal{L}_{DSM} + \mathcal{L}_{CM}$, 12 steps	0.9122	0.0560	9.0136	1.1998
DEFT, $0.6\mathcal{L}_{DSM} + \mathcal{L}_{CM}$, 15 steps	0.9132	0.0553	8.6611	0.9104
DEFT, $0.6\mathcal{L}_{DSM} + \mathcal{L}_{CM}$, 25 steps	0.9100	0.0579	8.5988	0.6713
DEFT, $0.8\mathcal{L}_{DSM} + \mathcal{L}_{CM}$, 12 steps	0.9112	0.0571	9.0765	1.2336
DEFT, $0.8\mathcal{L}_{DSM} + \mathcal{L}_{CM}$, 15 steps	0.9126	0.0561	8.7394	0.9378
DEFT, $0.8\mathcal{L}_{DSM} + \mathcal{L}_{CM}$, 25 steps	0.9101	0.0592	8.4160	<u>0.5474</u>
DEFT, $1.0\mathcal{L}_{DSM} + \mathcal{L}_{CM}$, 10 steps	0.8935	0.0862	12.7324	3.5322
DEFT, $1.0\mathcal{L}_{DSM} + \mathcal{L}_{CM}$, 11 steps	0.9111	0.0573	9.2023	1.3684
DEFT, $1.0\mathcal{L}_{DSM} + \mathcal{L}_{CM}$, 12 steps	0.9114	0.0571	8.9859	1.1134
DEFT, $1.0\mathcal{L}_{DSM} + \mathcal{L}_{CM}$, 15 steps	0.9125	0.0561	8.7113	0.8937
DEFT, $1.0\mathcal{L}_{DSM} + \mathcal{L}_{CM}$, 20 steps	0.9081	0.0617	8.8301	0.8030
DEFT, $1.0\mathcal{L}_{DSM} + \mathcal{L}_{CM}$, 25 steps	0.9091	0.0599	8.5058	0.5952

Table 1. Results of baseline comparisons with DEFT-VTON on VITON-HD dataset. Bold texts indicate best models, underlined texts indicate second best models.

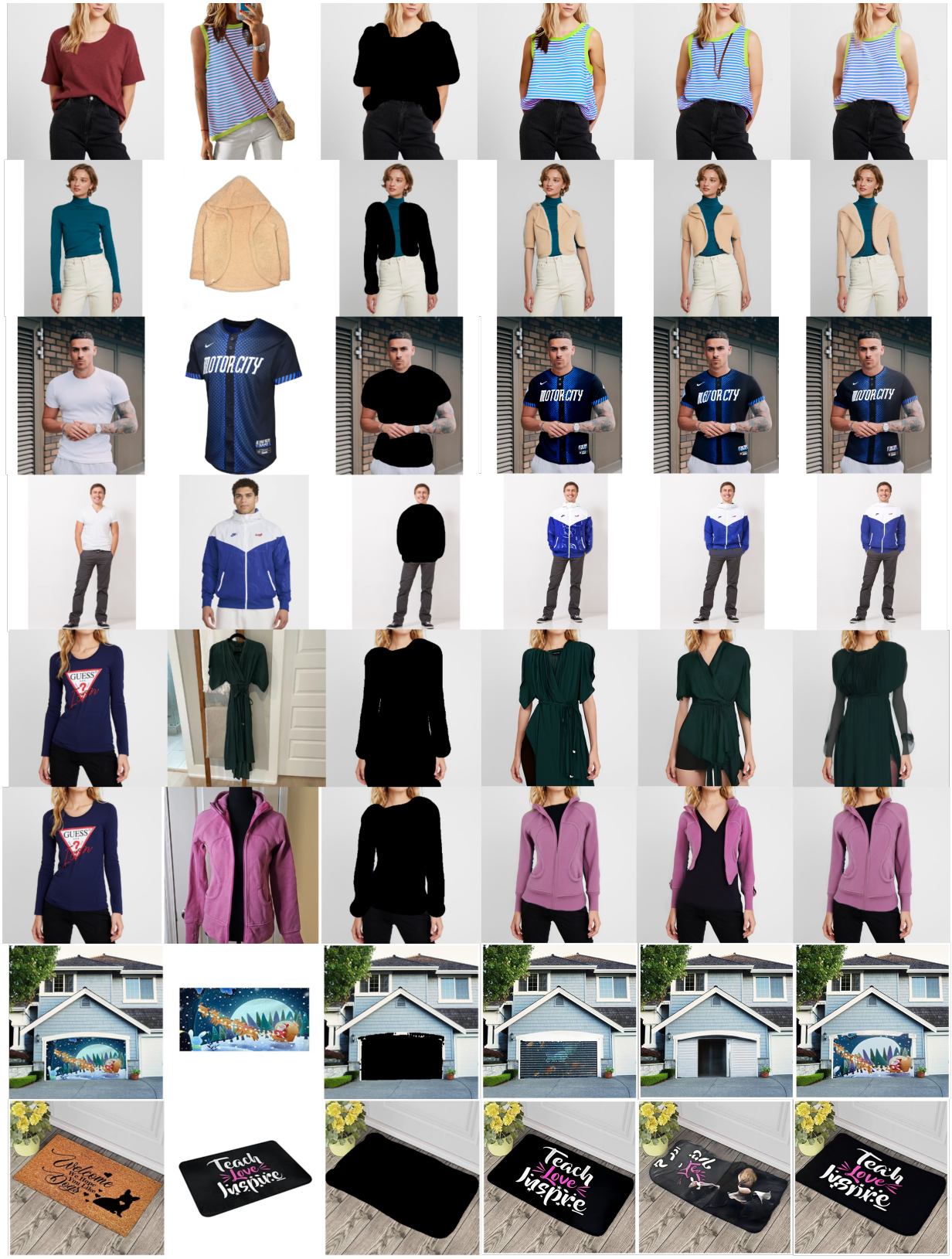

Figure 1. Results from test cases involving complex text and graphics preservation, unclear reference images, and unusual tasks. Each row shows a single task, starting with the original image and progressing through garment image, masked original image, 25-step sampling results (before consistency fine tuning), 15-step sampling results (before consistency fine tuning), and finally, 15-step sampling results (after consistency fine tuning).