Auto-encoder	Pre-trained	`sd-vae-ft-mse`
	In channels	3
	Latent channels	4
	Block channels	[128, 256, 512, 512]
	Down-sample ratio	8
	Layers per block	2
	Norm groups	32
U-Net	Pre-trained	✗
	In channels	8
	Out channels	4
	Block channels	[128, 256, 512, 512]
	Attention channels	[128, 256, 512, 512]
	Layers per block	2
	Head nums	8
	Filter nums	64
	Norm groups	32
Source image encoder	Pre-trained	✗
	In channels	4
	Block channels	[128, 256, 512, 512]
	Layers per block	1
	Norm groups	32
	Time embedding	✗
Noise scheduler	β schedule	Scaled linear
	β start	0.00085
	β end	0.012
Data augmentation	RandomCrop	✗
	RandomFlip	✓
Training setting	Iterations	$600k$
	Batch size	32
	Initial LR	5e-5
	Warm-up scheme	Linear
	Warm-up iterations	$1k$
	Warm-up starting	0
	Optimizer	Adam (0.9, 0.999)
	Weight decay	0.01
	Gradient clip	0.1
	Precision	fp16
	CFG probability	10%
Sampling setting	Scheduler	PNDM [3]
	CFG scale	5
	Steps	50
Hardware	GPU	$2 \times$ V100 (32 GB)
	Training duration	7 days

Table 1. Details for training and sampling PoCoLD.

A. Implementation details

We list all hyper-parameters used for training and sampling our PoCoLD in Tab. 1, including model architecture details, training recipe, and sampling setting.

B. Additional experiment results

Impact of CFG values. While tuning CFG is indeed useful, it alone is insufficient to achieve SOTA performance along with vanilla cross attention, as reflected in Tab. 2. We empirically found that PIDM's CFG strategy is not suited for our case and exploited our well-tuned CFG strategy

Attention	CFG Type	CFG Values	FID ↓	SSIM ↑	LPIPS ↓
Vanilla	Disentangled	$\omega_p, \omega_s = 2$	19.0473	0.5327	0.3620
Vanilla	Dual CFG [1]	$\omega_p, \omega_s = 2$	11.9200	0.6601	0.2440
Vanilla	Dual CFG [1]	$\omega_p, \omega_s = 5$	8.2903	0.7095	0.1783
Ours	Dual CFG [1]	$\omega_p, \omega_s = 5$	**8.0667**	**0.7310**	**0.1642**

Table 2. Quantitative results of tuning different CFG values.

Figure 1. Qualitative comparison between our PoCoLD and the variant which replaces DensePose with pose skeleton.

(Dual CFG [2] with $\omega_p, \omega_s = 5$) as the default setting for all experiments. The proposed attention is designed for efficiently leveraging DensePose, resulting in further performance improvements on the basis of already using the best CFG, and achieving SOTA results.

Impact of DensePose. We try to replace DensePose by using the body skeleton in the latent space (channel-wise) while keeping all training recipes intact. This variant gives 14.7362/0.6315/0.2550 in FID/SSIM/LPIPS, *vs.* 8.0667/0.7310/0.1642 by the original variant. Along with the qualitative results shown in Fig. 1, this verifies again that: (1) DensePose offers more comprehensive structural information, which is helpful to mitigate ambiguity; and (2) DensePose facilitates spatial alignment with the target image under proper regularization (*e.g.*, the proposed pose constraints).

High-resolution visualization results. We provide some high-resolution visualization results in Fig. 2 to better understand the performance of our PoCoLD in a qualitative way. We mainly compare our PoCoLD with prior diffusion-based art, *i.e.*, PIDM [1]. In each row, the sequence of images is as follows, from left to right: source image, target pose, ground truth, generation by PIDM, and our result. Our PoCoLD exhibits enhanced preservation of both texture and shape. Moreover, it demonstrates greater stability in generating results in certain infrequent scenarios, *e.g.*, enlarged person/garment in the source image.

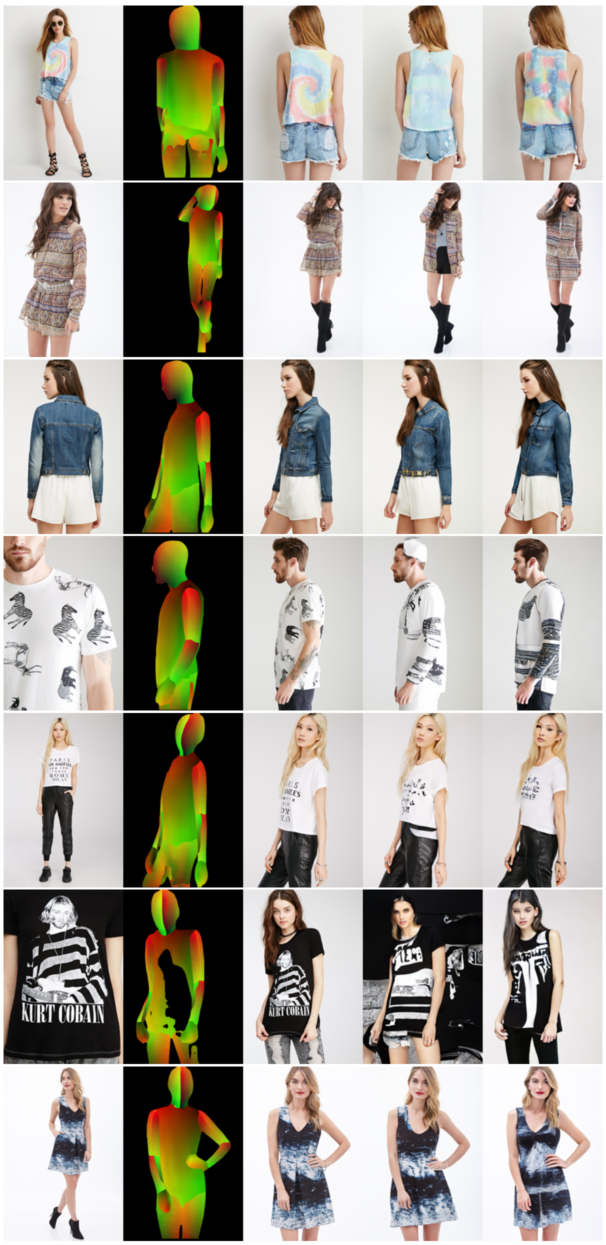

Figure 2. High-resolution qualitative result (from left to right: source image, target pose, ground truth, PIDM, and our PoCoLD).

References

[1] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *CVPR*, 2023. 1, 2

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 1

[3] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *ICLR*, 2022. 1