Method	Venue	FID↓	LPIPS↓	SSIM↑	PSNR↑
GAN-based Methods					
GFLA	CVPR 20'	19.740	0.2815	0.2808	14.337
XingGAN	ECCV 20'	22.520	0.3058	0.3044	14.446
SPGNet	CVPR 21'	23.057	0.2777	0.3139	14.489
DPTN	CVPR 22'	18.995	0.2711	0.2854	14.521
Diffusion-based Methods					
CFLD (Ours)		**11.972**	**0.2636**	**0.3173**	**14.861**
VAE Reconstructed		6.028	0.0164	0.9883	36.625
Ground Truth		4.845	0.0000	1.0000	$+\infty$

Table 4. Quantitative comparisons with the state of the arts on Market-1501 [5] dataset.

Evaluation on Market-1501. Since none of the diffusion-based methods including PIDM [1], PoCoLD [2] and concurrent PCDMs [4] have released generated images or checkpoints on Market-1501 [5], we make fair comparisons with available GAN-based methods in Tab. 4. From these results, CFLD still outperforms across different metrics faithfully, which validates our robustness.

Ablation on classifier-free strategy. In the Tab. 5 we vary the choices of Eq.(7) on DeepFashion [3]. The results show that appropriate reinforcement of both appearance and pose information (i.e., increase guidance weights) can effectively improve the quality of generated images.

Additional qualitative results. To further evaluate the generalization ability of our method, we generate person images at arbitrary poses randomly selected from the test set following in Figs. 8 to 10. The results show that our method consistently generate high-quality person images while preserving the appearance in the source image. Even if the target pose differs significantly from the source image, or if invisible areas of the source image are required, the generated images are still free of distortion. With the guidance of coarse-grained prompts, our method has a high-level understanding and does not suffer from overfitting such as forcing the texture details of the source image to be aligned. On this basis, our embedded hybrid-granularity attention only supplements the necessary fine-grained appearance features, thus enabling more realistic and natural textures.

More discussion of over-fitting and biasing. Our observation is that previous diffusion-based methods would fit the spatially convolutional features of source image into noisy sample directly. But this doesn't make sense in practice, because the texture details of source image probably shouldn't be present in the same position of target sample, especially in the exaggerated pose transition case. Since the model is actually performing copy-and-paste, the generations are distorted and unnatural, which we call this phenomenon **overfitting** and lack of generalization ability.

Strategy	w_{pose}	w_{app}	FID↓	LPIPS↓	SSIM↑	PSNR↑
disabled	1.0	1.0	8.143	0.2000	0.7055	15.753
appearance only	1.0	2.0	8.334	0.1921	0.7131	16.429
pose only	2.0	1.0	7.580	0.1770	0.7256	17.611
both	**2.0**	**2.0**	**6.804**	**0.1519**	**0.7378**	**18.235**
both	3.0	3.0	7.423	0.1746	0.7250	17.706

Table 5. Ablation on classifier-free strategy.

To circumvent it, we made three efforts: 1) We introduce pre-trained text-to-image diffusion as foundation model to improve generalization ability since it has been exposed to billions of image-text pairs. This empowers the model to speculate on some regions of the target pose that are not visible in the source image. 2) Note that textual description for PGPIS task is not available. To promote efficient fine-tuning without loss of generalization, we freeze most parameters in diffusion model (98.8%) and thus forcing the proposed PRD to learn coarse-grained semantics just as what the CLIP text encoder provide. 3) To decouple the fine-grained appearance and pose information as opposed to previous approaches, we endeavour to encode the multi-scale convolutional features as bias terms into cross-attention. The multi-scale **biasing** would be necessary since the coarse-grained prompts learned solely by the PRD may lack the preservation of texture details, given that the conditional prompt is the same for each scale in U-Net blocks. We leave the biased queries ($\boldsymbol{Q}$ in Eq.(4)) untrained and adopt zero convolution designs both in order to reduce the learning velocity of the HGA module thereby promoting a coarse-to-fine appearance control as stated in manuscript.

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*, page 5968–5976, 2023. 1

[2] Xiao Han, Xiatian Zhu, Jiankang Deng, Yi-Zhe Song, and Tao Xiang. Controllable person image synthesis with pose-constrained latent diffusion. In *ICCV*, page 22768–22777, 2023. 1

[3] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, page 1096–1104, 2016. 1

[4] Fei Shen, Hu Ye, Jun Zhang, Cong Wang, Xiao Han, and Wei Yang. Advancing pose-guided image synthesis with progressive conditional diffusion models. In *ICLR*, 2024. 1

[5] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, page 1116–1124, 2015. 1

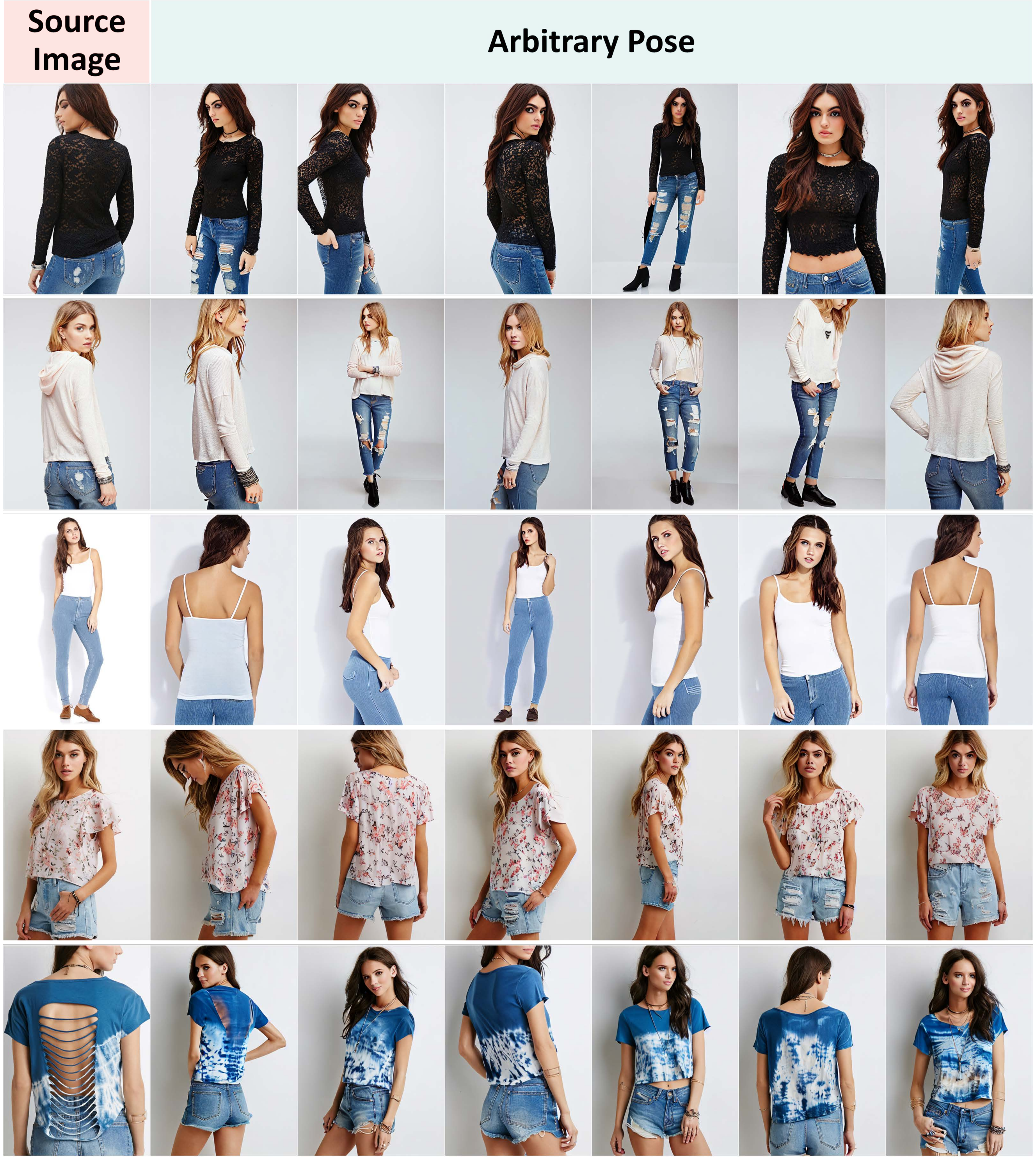

Figure 8. Additional results on arbitrary poses from the test set.

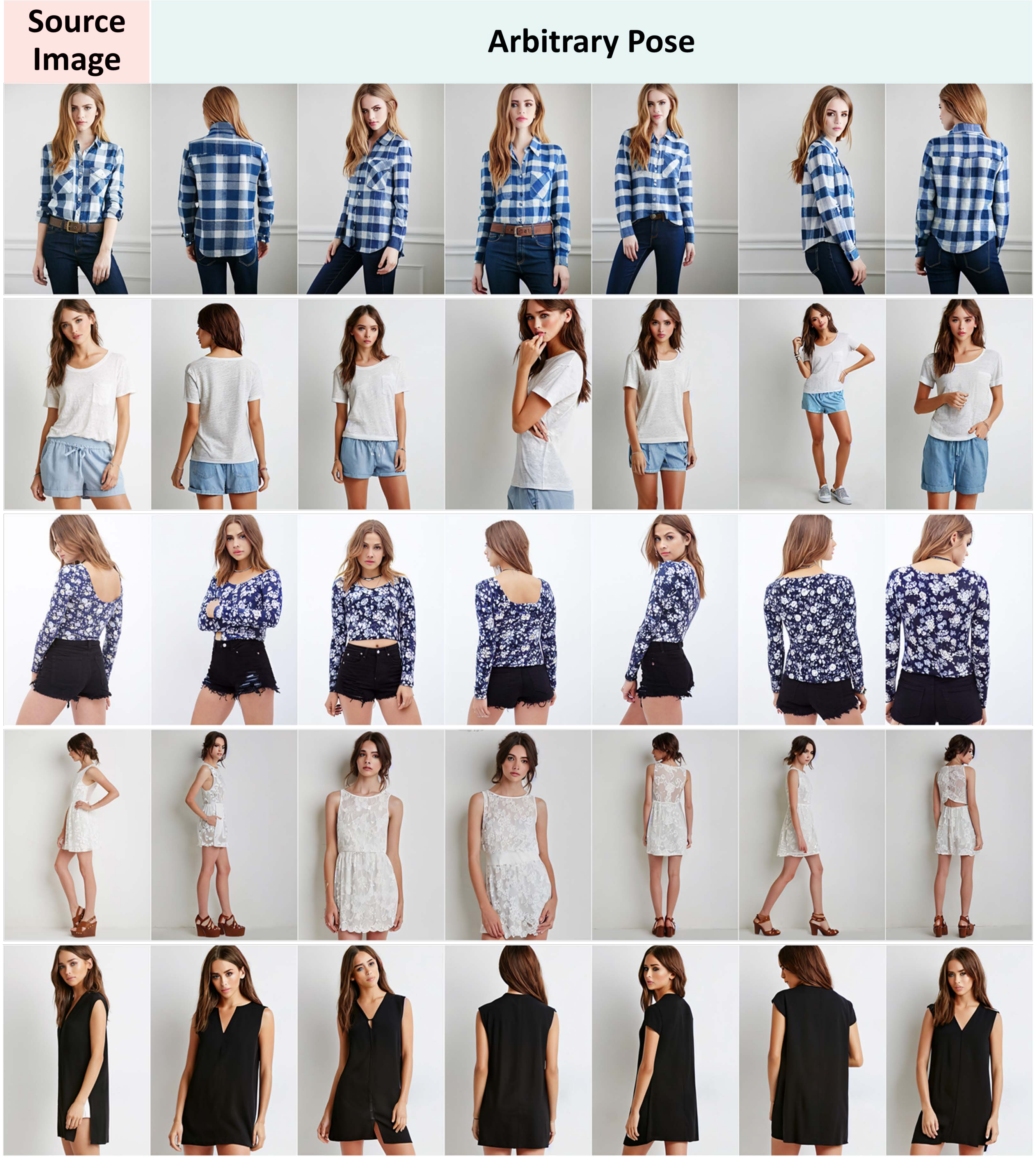

Figure 9. Additional results on arbitrary poses from the test set.

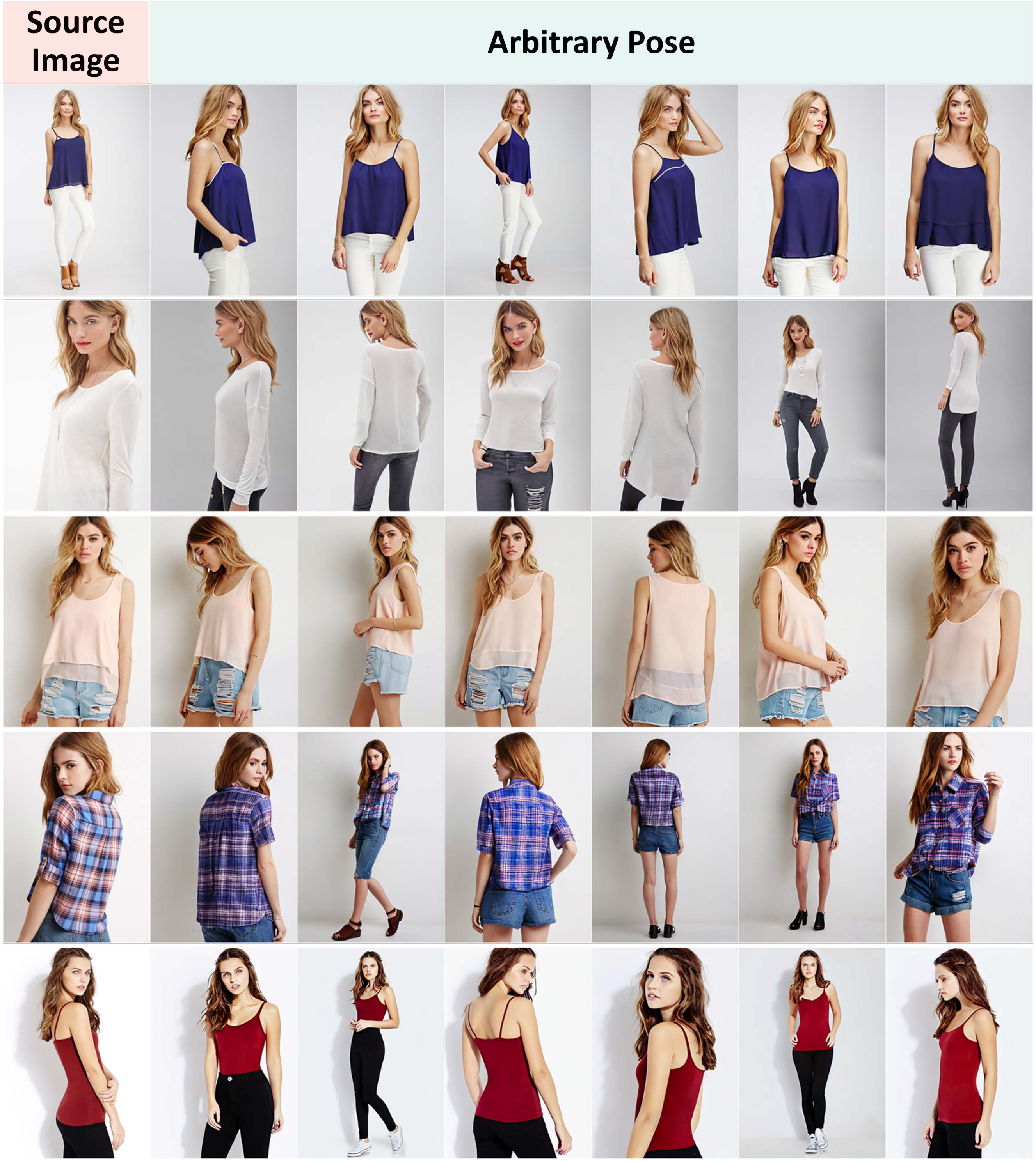

Figure 10. Additional results on arbitrary poses from the test set.