

DreamVTON: Customizing 3D Virtual Try-on with Personalized Diffusion Models – Supplementary Materials –

Anonymous Authors

1 INTRODUCTION

In supplementary materials, we first provide additional experiment details in Section 2, including the details of the user study and visual examples of templates used for training. We then provide more visual results in Section 3, including more qualitative comparisons with the existing 3D human generation methods and additional videos of generated 3D results. Finally, we present the details of DreamVTON’s architecture in Section 4.

2 ADDITIONAL EXPERIMENT DETAILS

2.1 Details of User Study

For user study, we design three questionnaires to separately evaluate the **Accuracy** (as shown in Figure 2 (a)), **Realism** (as shown in Figure 2 (b)), and **Geometry Smoothness** (as shown in Figure 2 (c)) of the 3D try-on results generated by various methods. To be specific, **Accuracy** is used to measure whether the try-on results can preserve a particular person’s identity and clothes characteristics, while **Realism** and **Geometry Smoothness** are used to evaluate the quality of generated texture and geometry, respectively.

Each questionnaire is composed of 9 assignments, and the amounts of volunteers for these three questionnaires (i.e., (a), (b), (c) in Figure 2) are 31, 26, and 21, respectively. For each assignment in the questionnaire, given the person and clothes images, volunteers are asked to select the best 3D try-on result (presented in video format) out of four options, which are generated by our DreamVTON and the other baseline methods (i.e., DreamWaltz [2], TEXTure [4], TeCH [3]). Besides, the order of the generated results in each assignment is randomly shuffled. Figure 2 shows the interface for each questionnaire. For analysis of the user study results, please refer to Section 4.4 in the main text.

2.2 Visual Examples of Training Templates

As mentioned in Section 3.2 in the main text, our DreamVTON employs various generated template images for geometry and texture optimization. For geometry optimization, DreamVTON uses eight mask templates $\{\hat{\mathbf{m}}_i\}_{i=1}^8$ (i.e., uniformly distributed views around the human body) to constrain the geometry shape learning, and uses two normal templates $\{\hat{\mathbf{n}}_i\}_{i=1}^2$ (i.e., front view and back view) to enhance the geometry detail learning. Figure 1 (a) and (b) display examples of mask and normal templates for one try-on use case. For texture optimization, DreamVTON uses three RGB templates $\{\hat{\mathbf{x}}_i\}_{i=1}^3$ (i.e., front and back full body view, and front face view) to enhance the texture detail learning. Figure 1 (c) displays examples of RGB templates for one try-on use case.

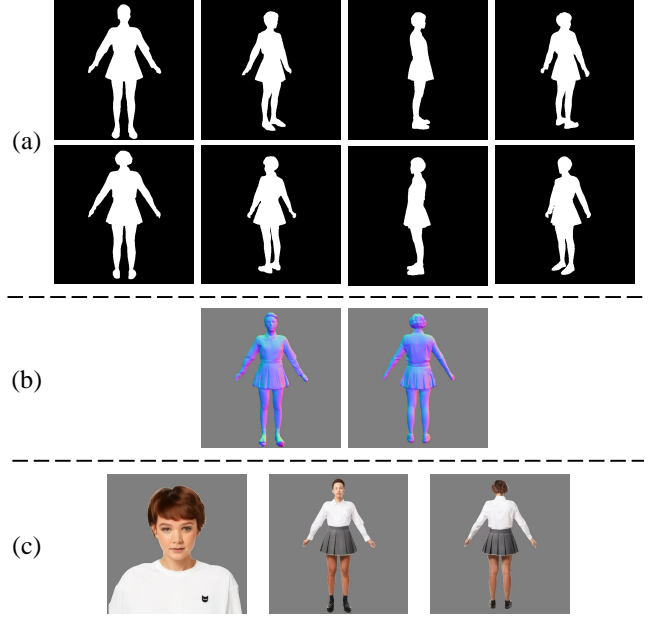


Figure 1: (a) Examples of mask templates. (b) Examples of normal templates. (c) Examples of RGB templates.

3 ADDITIONAL EXPERIMENT RESULTS

We provide additional visual comparisons among our proposed DreamVTON and the existing 3D human generation methods (i.e., DreamWaltz [2], TEXTure [4], and TeCH [3]) in Figure 3 and Figure 4. Furthermore, we also provide the rotated views of DreamVTON’s 3D try-on results in the accompanying video (named **DreamVTON-video.mp4**). Please refer to video for more details.

4 ARCHITECTURE DETAILS OF DREAMVTON

For the network architecture, our DreamVTON follows Fantasia3D [1], the geometry network Ψ_g is implemented as a three-layer MLP, while the texture network Ψ_t is implemented as a two-layer MLP. We illustrate the network architecture in Table 1.

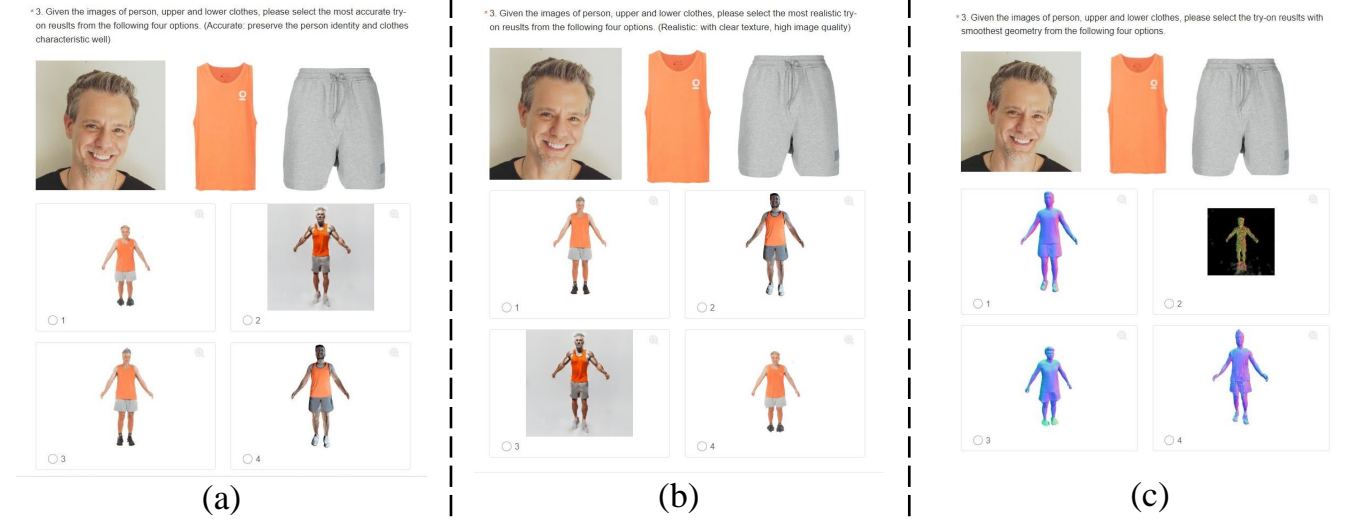


Figure 2: Interface of the questionnaires used to evaluate Accuracy, Realism, and Geometry Smoothness of the generated 3D try-on results.

Network	Architecture
Geometry Network Ψ_g	(Encoder): Linear(in_feat=3, out_feat=32, bias=False) (MLP): Linear(in_feat=32, out_feat=32, bias=False) ReLU() Linear(in_feat=32, out_feat=32, bias=False) ReLU() Linear(in_feat=32, out_feat=4, bias=False)
Texture Network Ψ_t	(Encoder): Linear(in_feat=3, out_feat=32, bias=False) (MLP): Linear(in_feat=32, out_feat=32, bias=False) ReLU() Linear(in_feat=32, out_feat=9, bias=False)

Table 1: Architecture of Geometry Network Ψ_g and Texture Network Ψ_t .



Figure 3: Qualitative Comparisons. Using the same clothes, person image, and prompt as inputs, our method achieves superior results.



Figure 4: Qualitative Comparisons. Using the same clothes, person image, and prompt as inputs, our method achieves superior results.

REFERENCES

[1] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. 2023. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873* (2023).

[2] Yukun Huang, Jianan Wang, Ailing Zeng, He Cao, Xianbiao Qi, Yukai Shi, Zheng-Jun Zha, and Lei Zhang. 2023. DreamWaltz: Make a Scene with Complex 3D Animatable Avatars. *arXiv preprint arXiv:2305.12529* (2023).

[3] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. 2024. TeCH: Text-guided Reconstruction of Lifelike Clothed Humans. In *International Conference on 3D Vision (3DV)*.

[4] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. 2023. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721* (2023).

465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522

523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580