DreamCoser: Controllable Layered 3D Character Generation and Editing Supplementary Material

Abstract

In this document, we provide the following supplementary contents:

- Implementation Details.
- Application.
- Ablation Study.
- Qualitative Results.

CCS Concepts

• **Computing methodologies → Artificial intelligence**; **Shape modeling**; **Image manipulation**.

ACM Reference Format:

. 2025. DreamCoser: Controllable Layered 3D Character Generation and Editing Supplementary Material. In *SIGGRAPH Asia 2025 Technical Communications (SA Technical Communications '25), December 15–18, 2025, Hong Kong, Hong Kong.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3757376.3771404

1 Implementation Details

Training Details. For our SDG network and PUP module, we adopt the Stable Diffusion 2.1 model as the base architecture. A normal prediction diffusion model is trained based on this stable diffusion image variant [Rombach et al. 2022]. A key modification in our approach involved the introduction of a reference U-Net [Ronneberger et al. 2015], which mirrors the network structure and initialization of the original model. This reference U-Net provides pixel-level reference attention exclusively to the newly incorporated attention layers of the main network. The normal map prediction is trained for 15,000 iterations with a batch size of 128.

Hyperparameters. (1) For generation based on character images using the PUP module, in the coarse stage, we optimize the DMTet [Shen et al. 2021] representation with 1000 steps, with $\lambda_1 = 1$, $\lambda_2 = 0.2$, $\lambda_3 = 0.5$. In the generation refinement stage, the DMTet representation is optimized for 1500 steps, with $\lambda_1 = 0.1$, $\lambda_2 = 0.1$, $\lambda_3 = 0.25$, $\lambda_4 = 1$. (2) For sketch-based editing, in the layered stage, we optimize the geometry for 2500 steps, with $\alpha_1 = 1$, $\alpha_2 = 1$, $\alpha_3 = 1e-3$. Specifically, alternate training is used in the layered refinement stage, and the training ratio of the nth layer to the combination of the previous n layers is $1:5$. (3) In the texture completion stage of editing, the texture of the edited object is optimized for 2000 steps, with $\beta_1 = 1$, $\beta_2 = 1$, $\beta_3 = 1e-3$, $\beta_4 = 0.1$. In particular, alternate training is used in the texture completion stage, and the training ratio of the nth layer to the combination of the previous n layers is $5:1$. (4) In the vertex anti-penetration stage of editing, multi-layer geometry is co-optimized for 500 steps, with $\mu_1 = 1.0$, $\mu_2 = 1.0$, $\mu_3 = 0.1$. The generation case takes 8 minutes to optimize, while the editing case takes about 10 minutes to optimize. Additionally, our method can improve the generation speed by adjusting the parameters of the PUP module.

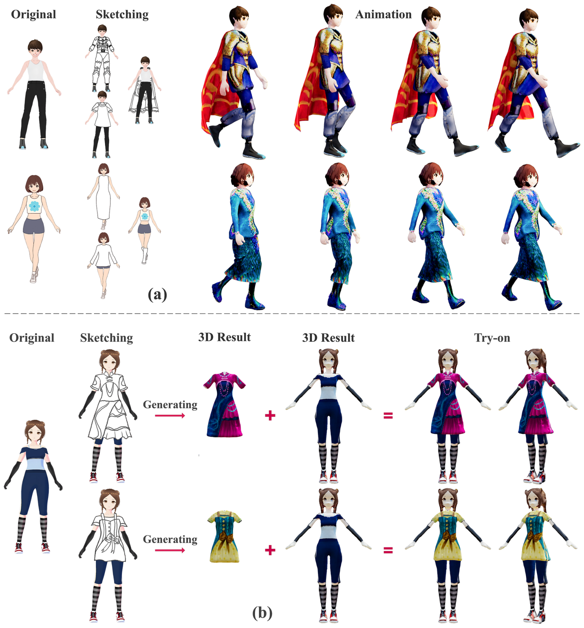

Figure 1: Application results. (a) Thanks to geometric disentanglement, our multi-layer dressed characters can be rigged for animation and simulate physical collisions between clothing layers. (b) Our method enables diverse clothing design for characters while generating 3D-compatible garments for virtual dressing.

2 Application

Benefiting from layered generation and local editing capabilities, our method can: (1) simulate physical collisions in multi-layer clothing (Fig. 1a), (2) enable virtual try-on for 3D characters (Fig. 1b), and (3) perform localized modifications on 3D characters (Fig. 5). These functionalities are achieved through free-hand sketch editing alone.

3 Ablation Study

Effectiveness of the Sketch-to-3D Decoupled Generation (SDG) Network. As shown in Fig. 2(c)-(d), the decoupled MVD ensures semantic consistency between edited content and input images while

SA Technical Communications '25, December 15–18, 2025, Hong Kong, Hong Kong

ACM ISBN 979-8-4007-2136-6/2025/12
https://doi.org/10.1145/3757376.3771404

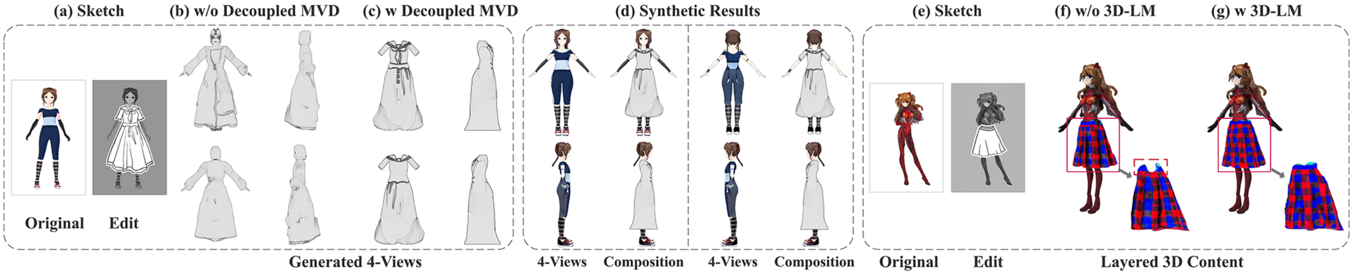

Figure 2: Ablation study of the Sketch-to-3D Decoupled Generation (SDG) Network: without decoupled MVD (Multi-view diffusion), multi-view outputs (b) exhibit semantic inconsistencies and tangled body parts; (c) adding MVD enables semantically consistent clothing generation that properly matches the body shape of character (d). Furthermore, (f) without 3D layered module (3D-LM), clothing layers appear incomplete and mismatched, while (g) the complete 3D-LM produces fully coherent clothing layers that semantically align with the 3D character model.

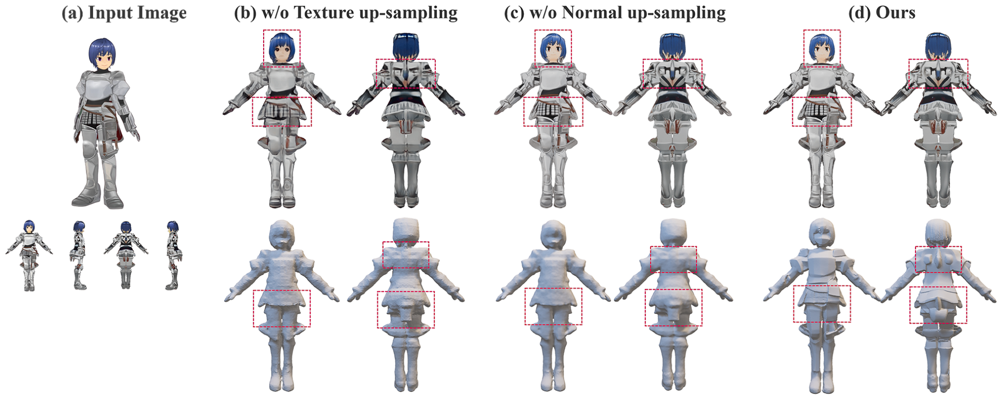

Figure 3: Ablation on Progressive Upsampling Module (PUP): (b) Without PUP: degraded textures and rough geometry; (c) Texture upsampling only: improves visuals but geometry remains coarse; (d) Full PUP: achieves both high-fidelity textures and refined geometry.

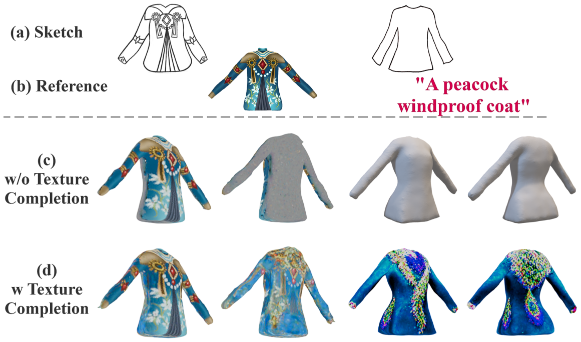

Figure 4: Ablation on dual-mode texture completion module: (c) without this module, textures are incomplete/unnatural; (d) with this module, textures become complete and semantically/tonally consistent with reference.

maintaining precise multi-view alignment. Fig. 2(f)-(g) demonstrates our 3D layered module effectively resolves layer incompleteness and semantic inconsistencies caused by sparse multi-view inputs.

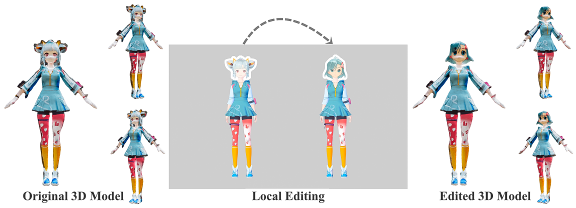

Figure 5: Results of local 3D content modification.

Effectiveness of the Progressive Upsampling (PUP) Module. Fig. 3 demonstrates our PUP module effectively enhances the resolution of both generated RGB images and corresponding normal images, ultimately producing high-quality 3D characters with texture and geometric details that semantically match the input image. **Effectiveness of the Dual-Mode Texture Completion Module.** Fig. 4(d) demonstrates that our texture completion module can utilize either a single image or text as reference to perform detailed texture completion on 3D models, while maintaining both semantic and tonal consistency with the texture reference input.

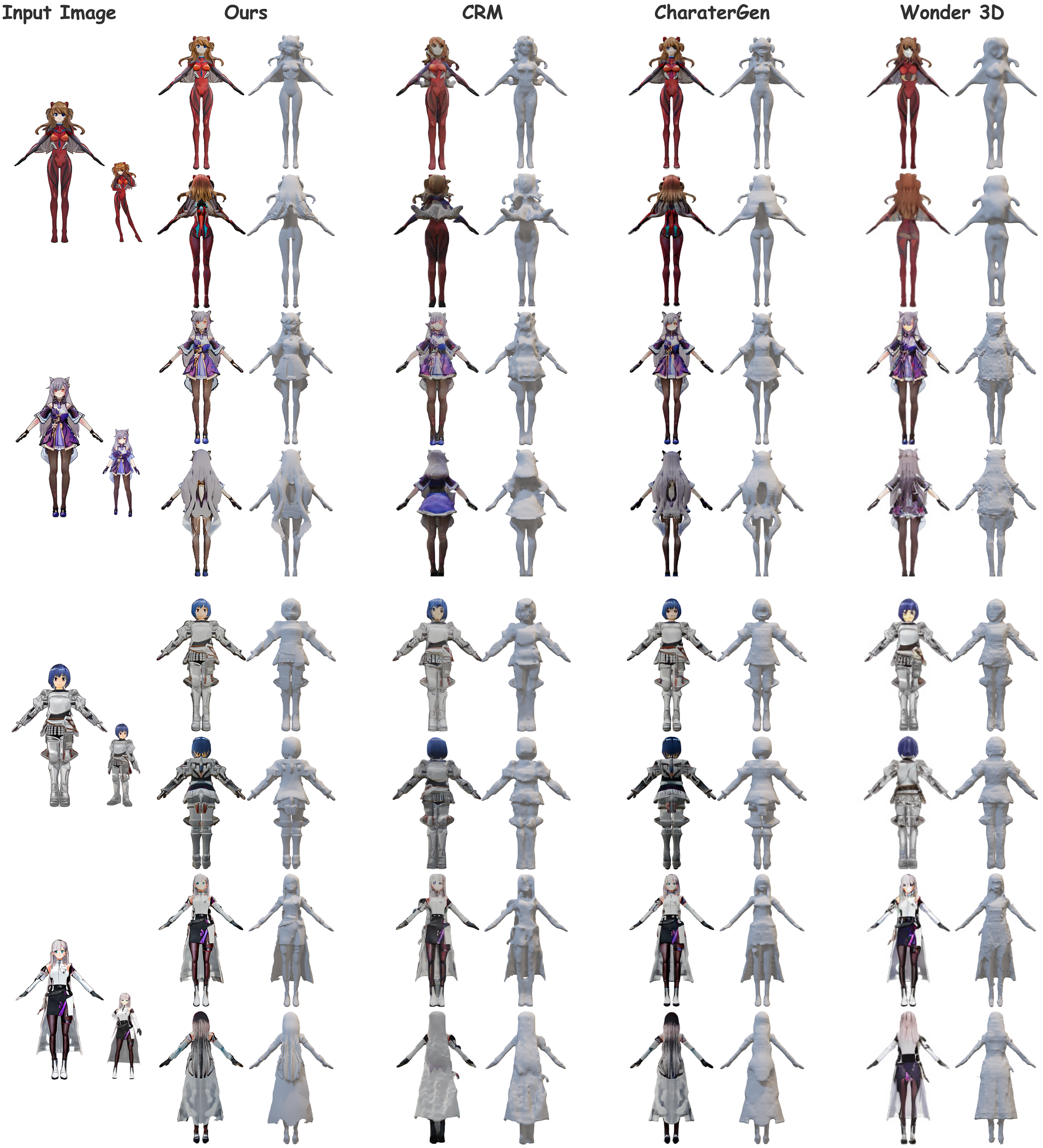

Figure 6: Qualitative comparison of single-image-based methods. We use A-pose character image as input for 3D character generation.

4 Qualitative Results

To compare under a unified posture, we use a single character image in A-pose as input for qualitative comparison between our method and SoTA methods [Long et al. 2024; Peng et al. 2024; Wang et al. 2025]. Fig. 6 shows that our results visually outperform those from SoTA methods. CRM [Wang et al. 2025] fails to represent complex structures and high-frequency features due to the limitations of convolutional layers in capturing global contextual information and complex topologies. CharacterGen [Peng et al. 2024] loses local geometry such as hair or clothing, although it introduces multi-view pose normalization to improve the handling of complex poses. Although Wonder3D [Long et al. 2024] includes cross-domain alignment for global feature capture, it falls short in texture detail fidelity, especially in reconstructing high-resolution textures and fine details. In contrast, our method generates high-quality textured 3D content, which we attribute to our proposed PUP module. Moreover, complex local geometric details, such as the hair and clothing details shown in Fig. 6, are captured through the multi-view consistent normal upsampling of the PUP module. Furthermore, the compared methods [Long et al. 2024; Peng et al. 2024; Wang et al. 2025] cannot perform layered generation and editing of 3D content, whereas our method ensures high-quality generation while enabling fine-grained local and layered editing.

References

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. 2024. Wonder3D: Single image to 3D using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9970–9980.

Hao-Yang Peng, Jia-Peng Zhang, Meng-Hao Guo, Yan-Pei Cao, and Shi-Min Hu. 2024. Charactergen: Efficient 3D character generation from single images with multi-view pose canonicalization. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–13.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. 2021. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems* 34 (2021), 6087–6101.

Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. 2025. Crm: Single image to 3D textured mesh with convolutional reconstruction model. In *European Conference on Computer Vision*. Springer, 57–74.