1. Additional Details

In this section, we provide additional details to clarify our method, including metric computation, data details, and skeleton definition.

1.1. Metric Computation

Given the sensitivity of Chamfer Distance (CD) to shape alignment and point distribution—particularly the topological discrepancies between watertight and non-watertight meshes—we adopt Mean Angular Error (MAE) and Cosine Similarity (SIM) of rendered normal maps to assess geometric accuracy. Specifically, our evaluation begins by rendering the front-view normal map of the predicted mesh. To account for potential azimuthal misalignments between the predicted meshes (including those generated by baseline methods such as Hunyuan3D 2.1 [8] and CraftsMan [11]) and the ground truth (GT) models, we render the GT normals across 36 azimuths uniformly sampled at 10° intervals (from 0° to 350°). We then compute the MAE and SIM between the predicted front-view normals and the GT normals under each viewpoint. The best-matching score among all views is reported as the final metric, ensuring a robust evaluation of geometric fidelity.

1.2. Data details.

Our constructed dataset comprises animatable dynamic assets and static assets in an approximate 3:2 ratio. For static meshes lacking skeletal structures, we employ an auto-regressive model fine-tuned from UniRig [20] for automatic rig generation. For the dynamic assets, the characters from ReadyPlayerMe [1] and VRoid [4] are driven by Mixamo motion sequences, whereas the assets from Playbox [3] are animated by VMD motion sequences compatible with the MMD data format.

Image rendering. For dynamic characters, we fix the camera azimuth, allowing the character's articulated motions to naturally induce viewpoint variations. Furthermore, we randomize the camera elevation between 0° and 10° and dynamically adjust the camera distance to ensure the animated character remains centered in the viewpoint. For static meshes, we render four canonical views (front, back, left, and right) alongside 32 uniformly sampled random views. All renderings are generated using perspective projection.

1.3. Skeleton Definition

Following previous methods [15, 16] that utilize Openpose [5] for pose conditioning, we select the body's bones and hand's bones in the 3D skeleton as our skeleton system. This selection excludes hair and skirt bones, yet remains sufficient to define a humanoid pose effectively.

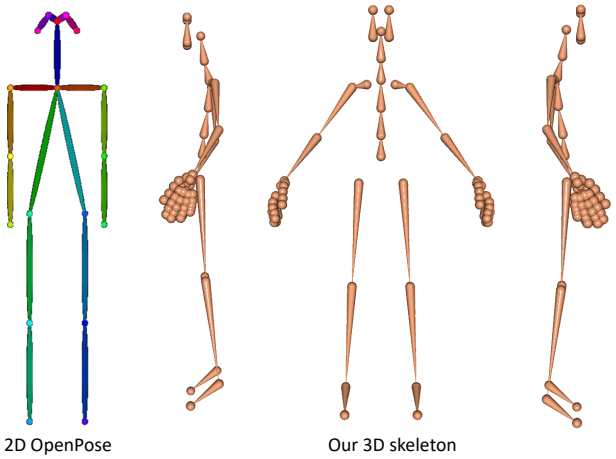

Figure 1. The visualization of the standard skeletons of Openpose [5] and ours (here, we take the skeleton from Readyplayeme's [1] data as the example).

Table 1. The quantitative comparison for arbitrary pose stylization by using the target-pose images edited from Qwen-Image [17] as the baselines' input.

Method	MAE ↓	SIM ↑	Uni3D-I ↑	ULIP-I ↑
Trellis [19]	8.50	0.874	0.293	0.160
CraftsMan [11]	8.55	0.876	0.302	0.148
Hunyuan3D 2.1 [8]	8.15	0.874	0.293	0.160
PoseMaster (Ours)	**5.28**	**0.935**	**0.313**	**0.172**

2. Additional Experimental Results

In this section, we present extended comparative evaluations, featuring additional qualitative visualizations for both arbitrary-pose stylization and pose canonicalization. Following these comparisons, we provide a comprehensive analysis of the computational efficiency and structural robustness of our proposed framework. Finally, we report supplementary texture synthesis results and demonstrate a practical downstream application.

2.1. More results on Arbitrary-pose Stylization

To isolate the effects of 2D image distortion inherent in decoupled pose stylization pipelines, we utilized ground-truth target-pose images as inputs for existing 3D native methods (i.e., Trellis [19], CraftsMan [11], and Hunyuan3D 2.1 [8]) in the arbitrary-pose stylization comparisons presented in Sec. 5.2.2. Therefore, to explicitly expose the limitations of the decoupled paradigm—where 2D pose transformation and 3D generation are treated as independent stages—we construct a comprehensive baseline pipeline utilizing Qwen-Image [17] to synthesize target-pose inputs. Specifically, we project 3D skeletons into Openpose-style

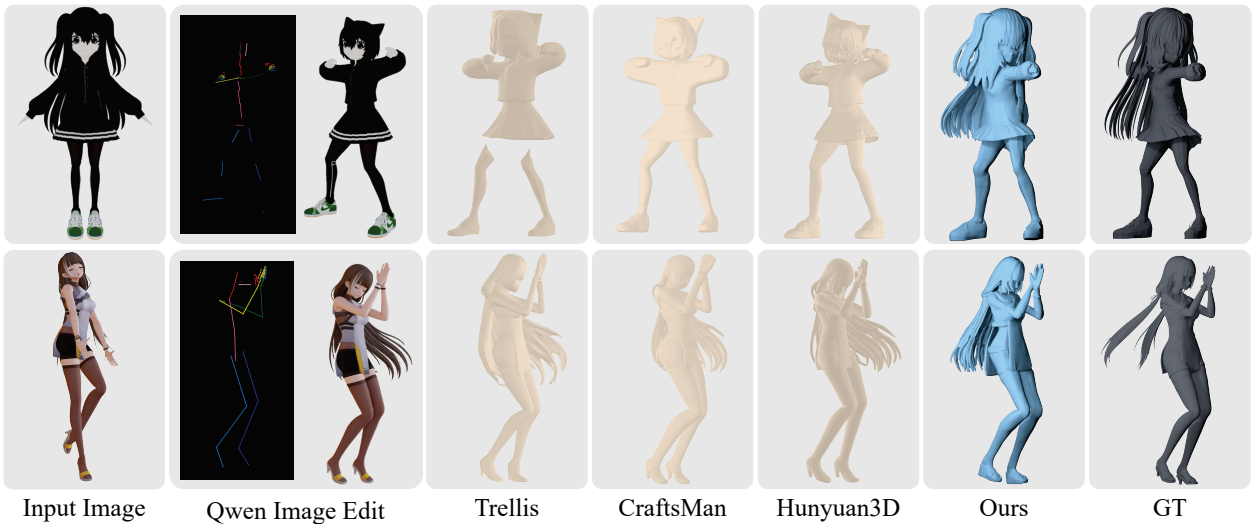

Figure 2. The qualitative comparison for arbitrary-pose stylization by using the target-pose images edited from Qwen-Image [17] as the baselines' input.

Table 2. Time comparison of generation speed.

Method	CharacterGen [15]	StdGen [7]	Ours
Inference time	$\sim$ 32.98 (s)	$\sim$ 61.54 (s)	$\sim$ 23.48 (s)

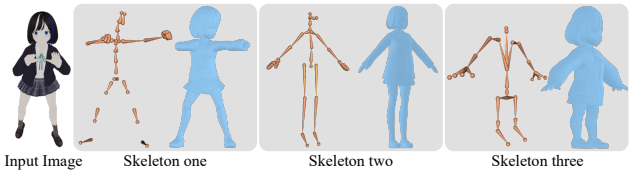

Figure 3. The visualization for robustness analysis.

2D skeleton maps to guide Qwen-Image in generating pose-edited images, which are subsequently fed into 3D generation models to produce meshes.

As shown in Table 1 and Figure 2, our method presents significant quantitative and qualitative superiority over both baseline methods. Visual inspection reveals that Qwen-Image-based editing frequently compromises the preservation of original identity features. Moreover, the 2D interpretation of skeletons suffers from viewpoint-dependent ambiguity, leading to geometric inaccuracies in the generated poses. Consequently, the performance of the Qwen-Image-based pipeline falls short of the setting that uses ground-truth inputs. This performance gap strongly underscores the critical necessity of a unified framework and explicit 3D skeleton guidance for robust 3D pose stylization.

Additional qualitative results for arbitrary-pose stylization are presented in Figure 9 and Figure 10. These visualizations further demonstrate the robustness of PoseMaster. By seamlessly integrating pose stylization and 3D generation into a unified generative framework, we achieve precise pose control directly in the 3D domain. Moreover, the incorporation of 3D skeletons facilitates accurate pose recovery and manipulation, substantially enhancing the model's practical usability.

2.2. Efficiency Analysis

We evaluate the computational efficiency of our method against previous multi-stage pose stylization approaches, including CharacterGen [15] and StdGen [7]. All inference timings are measured on a single NVIDIA H20 GPU. As detailed in Table 2, PoseMaster achieves faster generation speeds. This acceleration is primarily attributed to our integrated, single-stage generative framework, which streamlines the conventional multi-stage workflow and eliminates intermediate computational bottlenecks.

2.3. Robustness Analysis

In this section, we introduce various skeleton conditions to validate the robustness of our model. As illustrated in Figure 3, we design test cases featuring misaligned image-skeleton pairs with significant discrepancies in body proportions and topology. Remarkably, PoseMaster exhibits exceptional structural adherence. Even when conditioned on heavily mismatched skeletons, the model consistently synthesizes meshes that strictly conform to the target skeletal topology while faithfully preserving the appearance of the source image. This flexibility is particularly valuable for downstream game asset creation, such as character retarget-

Figure 4. The visualized results for textured mesh. We employ 3D native texturing approaches to synthesize textures for the meshes generated by PoseMaster. The final textured assets are rendered from four canonical viewpoints.

ing and body proportion customization

2.4. Texture Generation

In this work, geometry and texture synthesis are formulated as decoupled tasks; hence, explicit texture generation remains orthogonal to our main contributions. To demonstrate the practical utility of our approach, we provide texturing results leveraging state-of-the-art models. While the inherent pose inconsistency between the input image and the generated geometry poses a substantial challenge for texture mapping, we mitigate this by integrating advanced image-editing models (e.g., Qwen-Image [17]) and 3D native texturing approaches (e.g., UniTEX [13], NaTex [10], LaFiTe [6]). As depicted in Figure. 4, these frameworks demonstrate strong robustness against pose misalignment. For minor persistent artifacts, a promising future direction is to fine-tune these models specifically on misaligned image-geometry datasets. Note that we employ Hunyuan3D 2.5 [9] as an intermediate geometric refinement step prior to the texturing phase.

2.5. More Results on Pose Canonicalization

Results on in-the-wild images. To assess the zero-shot generalization capability of our model, we compile a diverse in-the-wild test set comprising both virtual avatars and real-world photographs. We evaluate our method against

CharacterGen [15], StdGEN [7], and Hunyuan3D 2.1 [8]. For the Hunyuan3D 2.1, we utilized Qwen-Image [17] to transform the pose of the images to form its inputs. To circumvent the instability of Qwen-Image in generating A-pose outputs, we prompted the model to synthesize T-pose images instead. As shown in the qualitative comparisons, our method exhibits superior pose customization and geometric fidelity. CharacterGen and StdGEN, which are predominantly trained on single-style VRoid data, struggle to generalize to diverse artistic styles; their generated A-pose images often suffer from severe artifacts and distortions, fundamentally bottlenecking the subsequent 3D reconstruction. While Qwen-Image can alleviate some 2D distortion, controlling poses purely via text descriptions introduces significant randomness. For instance, in the inputs generated for Hunyuan3D, although the overall posture approaches a T-pose, critical details such as arm elevation, leg spacing, and perspective lack precise control. In contrast, our end-to-end framework bypasses the error-prone 2D editing stage. By employing explicit 3D skeletons as conditioning, we achieve strictly standardized pose canonicalization. Furthermore, inheriting the powerful priors of native 3D generation models ensures high-fidelity mesh topologies.

Additionally, as illustrated in Figure 6, our method robustly handles real-world human images from the DeepFashion dataset [14]. The incorporation of a large-scale, diverse training dataset empowers our model to bridge the domain gap, significantly broadening its real-world applicability.

Results on AI-synthesized images. We further evaluate generalization on synthetic images produced by text-to-image models. As illustrated in Figure 7, PoseMaster exhibits superior stability in pose canonicalization tasks compared to pipeline-dependent approaches. Notably, for the Hunyuan3D baseline, the required T-pose inputs were synthesized via text-prompted Qwen-Image, which often struggles with complex pose transformations.

Results on CharacterGen's testing set. For a rigorous and fair comparison, we evaluate our method on the official test set provided by CharacterGen [15]. As shown in Figure 8, PoseMaster significantly outperforms both CharacterGen and StdGen [7] in terms of geometric quality and pose accuracy, further validating that native pose stylization yields optimal results for pose canonicalization.

2.6. Application in 3D Printing

Leveraging our arbitrary-pose stylization capabilities, we demonstrate a practical downstream application: a customized 3D printing pipeline for stylized anime characters. As depicted in Figure 5, the workflow begins with a 2D anime character generated via text-to-image models. Users can then author desired poses using standard tools like Blender [2] or Openpose Editor. By extracting the 3D skeleton from the authored pose, PoseMaster directly reconstructs a 3D character mesh matching both the input identity and the target pose. Following standard post-processing, the stylized mesh is readily 3D-printed.

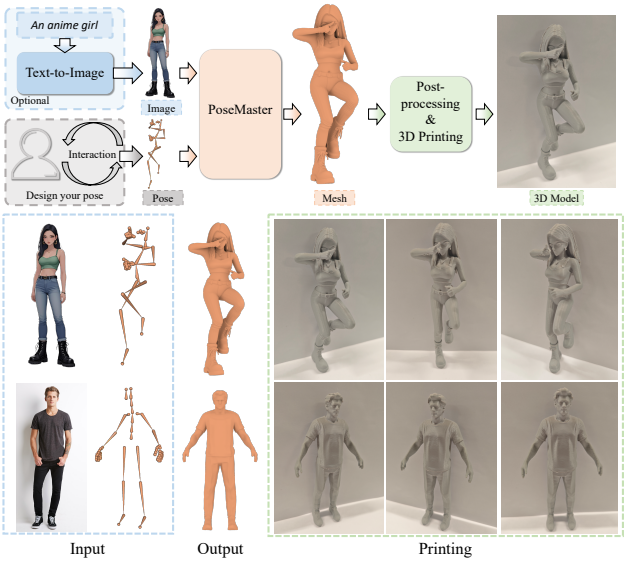

Figure 5. The system of 3D printing based on our PoseMaster. PoseMaster can customize the pose of a character from a single image for figure printing.

3. Limitation and Discussion

First, we acknowledge that native 3D pose stylization remains a highly challenging task. While our framework demonstrates strong overall pose controllability, synthesizing fine-grained geometric details—such as intricate hand gestures, flowing skirts, and complex hairstyles—requires further exploration. Furthermore, our current implementation utilizes a single-stage generation paradigm operating at a spatial resolution of 512. Consequently, the high-frequency geometric fidelity of our outputs may lag behind recent state-of-the-art multi-stage refinement methods (e.g., Direct3D-S2 [18], Hunyuan3D 2.5 [9], and Sparc3D [12]). A promising avenue for future work is to integrate a high-resolution geometric refinement module to upscale and enhance the coarse meshes generated in the first stage, thereby achieving photorealistic and highly detailed pose stylization.

References

[1] Readyplayerme, 2025. https://readyplayer.me/.
[2] Blender, 2025. https://www.blender.org/.
[3] Playbox, 2025. https://www.aplaybox.com.
[4] Vroid-hub, 2025. https://hub.vroid.com/.
[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose

estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.

[6] Chia-Hao Chen, Zi-Xin Zou, Yan-Pei Cao, Ze Yuan, Guan Luo, Xiaojuan Qi, Ding Liang, Song-Hai Zhang, and Yuan-Chen Guo. Lafite: A generative latent field for 3d native texturing. *arXiv preprint arXiv:2512.04786*, 2025.

[7] Yuze He, Yanning Zhou, Wang Zhao, Zhongkai Wu, Kaiwen Xiao, Wei Yang, Yong-Jin Liu, and Xiao Han. Stdgen: Semantic-decomposed 3d character generation from single images. *arXiv preprint arXiv:2411.05738*, 2024.

[8] Team Hunyuan3D, Shuhui Yang, Mingxin Yang, Yifei Feng, Xin Huang, Sheng Zhang, Zebin He, Di Luo, Haolin Liu, Yunfei Zhao, et al. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material. *arXiv preprint arXiv:2506.15442*, 2025.

[9] Zeqiang Lai, Yunfei Zhao, Haolin Liu, Zibo Zhao, Qingxiang Lin, Huiwen Shi, Xianghui Yang, Mingxin Yang, Shuhui Yang, Yifei Feng, et al. Hunyuan3d 2.5: Towards high-fidelity 3d assets generation with ultimate details. *arXiv preprint arXiv:2506.16504*, 2025.

[10] Zeqiang Lai, Yunfei Zhao, Zibo Zhao, Xin Yang, Xin Huang, Jingwei Huang, Xiangyu Yue, and Chunchao Guo. Natex: Seamless texture generation as latent color diffusion. *arXiv preprint arXiv:2511.16317*, 2025.

[11] Weiyu Li, Jiarui Liu, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. *arXiv preprint arXiv:2405.14979*, 2024.

[12] Zhihao Li, Yufei Wang, Heliang Zheng, Yihao Luo, and Bihan Wen. Sparc3d: Sparse representation and construction for high-resolution 3d shapes modeling. *arXiv preprint arXiv:2505.14521*, 2025.

[13] Yixun Liang, Kunming Luo, Xiao Chen, Rui Chen, Hongyu Yan, Weiyu Li, Jiarui Liu, and Ping Tan. Unitex: Universal high fidelity generative texturing for 3d shapes. *arXiv preprint arXiv:2505.23253*, 2025.

[14] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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

[16] Shuai Tan, Biao Gong, Xiang Wang, Shiwei Zhang, Dandan Zheng, Ruobing Zheng, Kecheng Zheng, Jingdong Chen, and Ming Yang. Animate-x: Universal character image animation with enhanced motion representation. *arXiv preprint arXiv:2410.10306*, 2024.

[17] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

[18] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Yikang Yang, Yajie Bao, Jiachen Qian, Siyu Zhu, Xun Cao, Philip Torr, and Yao Yao. Direct3d-s2: Gigascale 3d generation made easy with spatial sparse attention. *CoRR*, abs/2505.17412, 2025.

[19] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21469–21480, 2025.

[20] Jia-Peng Zhang, Cheng-Feng Pu, Meng-Hao Guo, Yan-Pei Cao, and Shi-Min Hu. One model to rig them all: Diverse skeleton rigging with unirig. *ACM Trans. Graph.*, 44(4), 2025.

Figure 6. The qualitative comparison for pose canonicalization on real-world images from the DeepFashion dataset [14].

Figure 7. The qualitative comparison for pose canonicalization on AI-synthesized images.

Figure 8. The qualitative comparison for pose canonicalization on CharacterGen's [15] testing dataset.

Figure 9. The qualitative results for arbitrary-pose stylization.

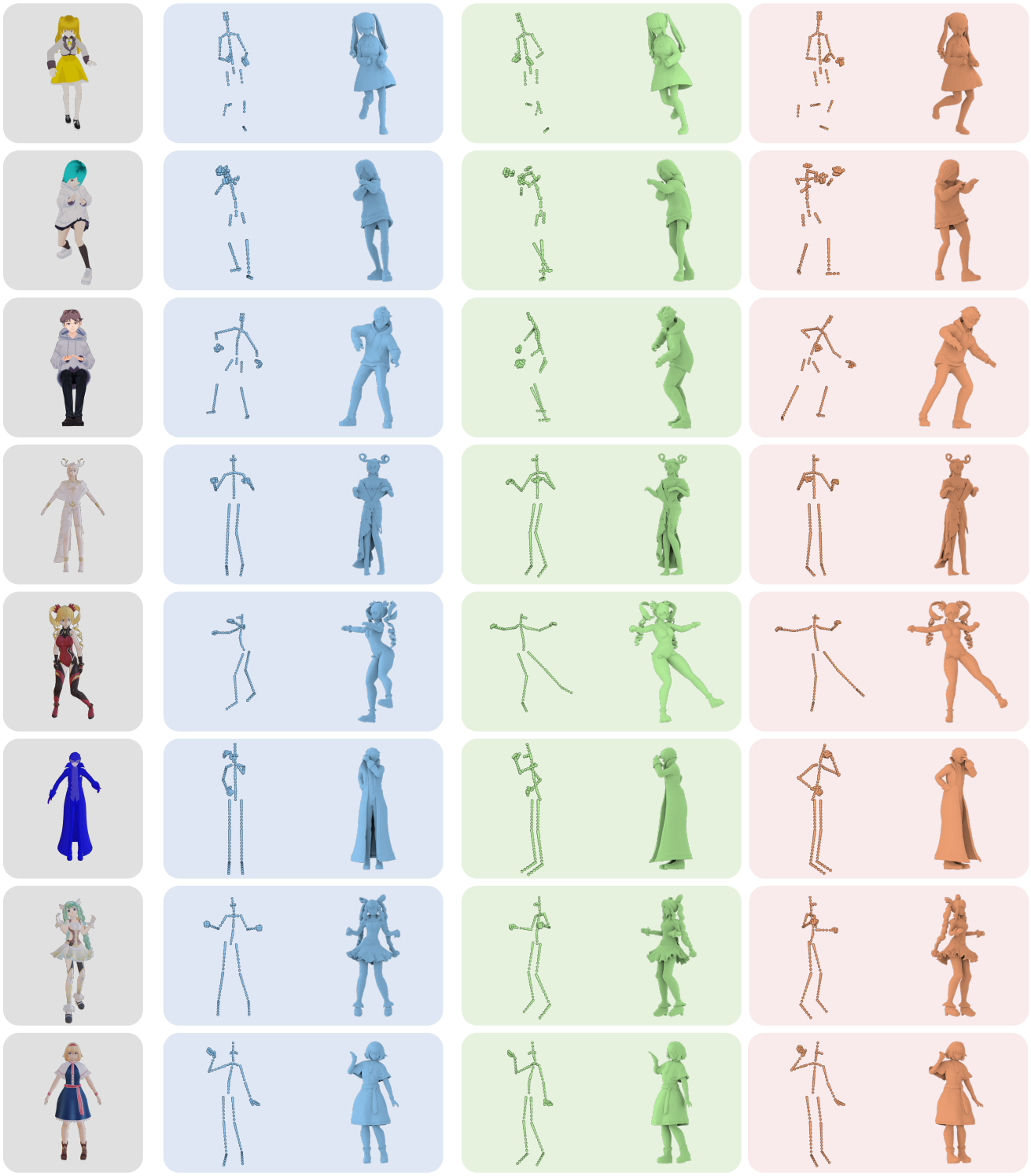

Figure 10. The qualitative results for arbitrary-pose stylization.