SerialGen: Personalized Image Generation by First Standardization Then Personalization

Supplementary Material

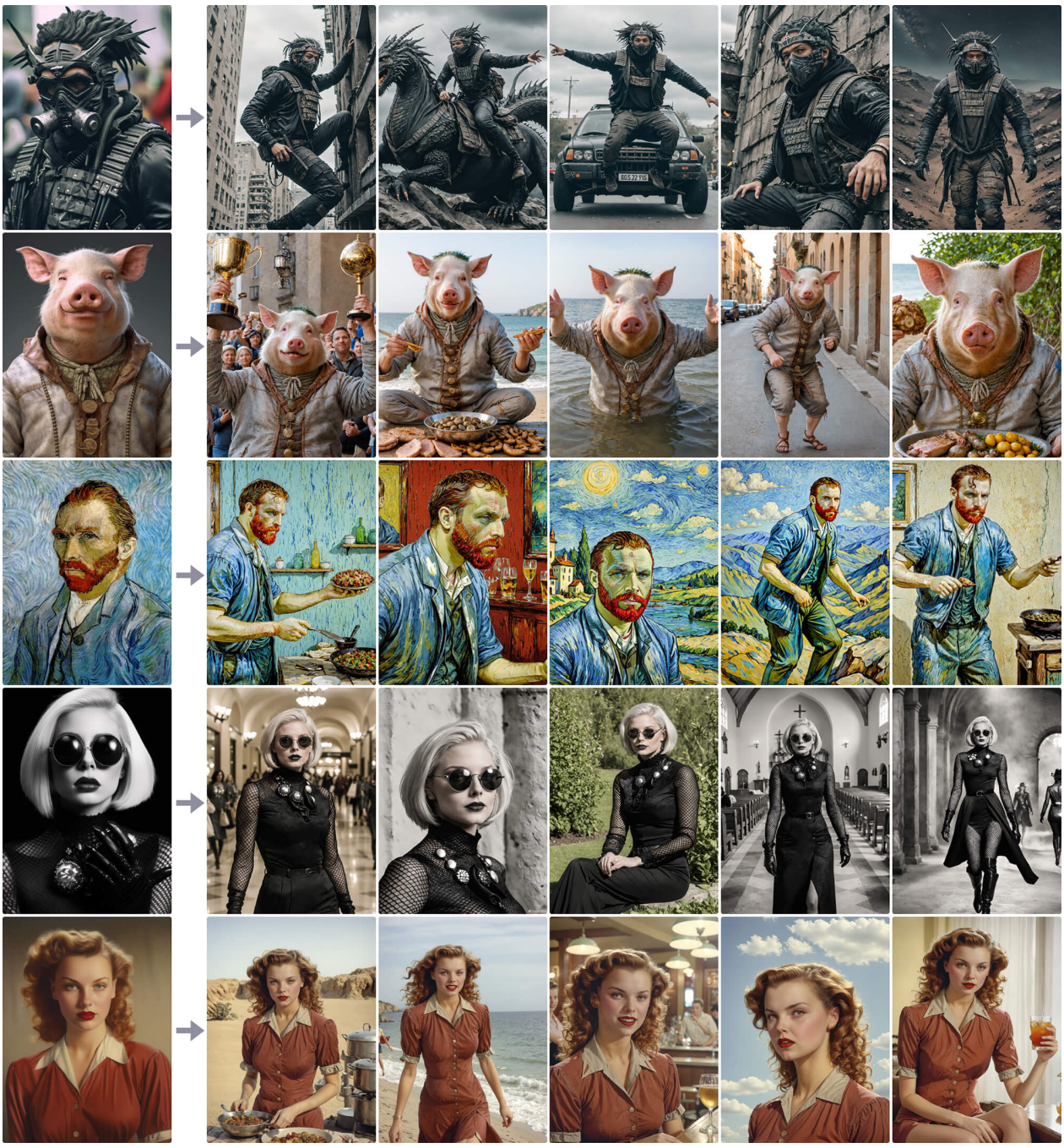

Figure 8. More serial images generated by SerialGen, showcasing its outstanding ability to maintain whole-body appearance consistency across different types of characters, including non-human subjects.

Figure 9. Extension of Figure 8.

6. Details of the Reference Pose Injection Module

We utilize a light convolutional network to extract pose feature maps from pose images. The architectural setup is depicted in Figure 10, where 3×3 conv, 32, $\downarrow 2$ indicates a convolutional layer with a kernel size of 3×3, a channel number of 32, and a stride size of 2. The term silu refers to a SiLU activation layer. The network processes the input through several convolutional stages with channel counts of $16, 32, 96, 256$, and 320, progressively reducing the spatial resolution by a factor of 8. Before being added to the $\tilde{\mathbf{f}}_r$ feature in each self-attention module, an additional convolutional layer is introduced, followed by interpolation-based downsampling to align the dimensions of pose feature with the $\tilde{\mathbf{f}}_r$ feature.

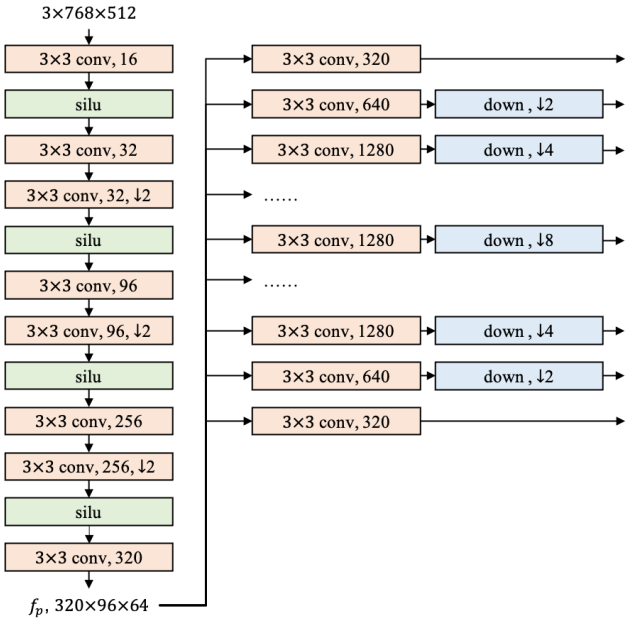

Figure 10. Details of the Reference Pose Injection Module.

7. Impact of 3D Style Bias

As depicted in the second paragraph of Section 3.4, we demonstrate the impact of 3D style bias introduced by the synthetic data. As shown in Figure 11, the standardization stage introduces a slight 3D style bias when standardizing images. This bias is effectively mitigated during the personalization stage. Specifically, as shown in the last row of Figure 11, given a head-only image, the standardization stage generates clothing with a noticeable 3D style. However, the personalization stage subsequently recovers realistic clothing appearances.

8. More Comparison Results and Analysis

This part gives supplementary comparisons and analysis in Section 4.2. We give more comparison results with FastComposer, IP-Adapter and StoryMaker. As shown in the Figure 12, we selected four different characters for analysis, which include two anime characters and two real-life humans. The evaluation prompts are categorized into four descriptive types: action, background, viewpoint, and ex-

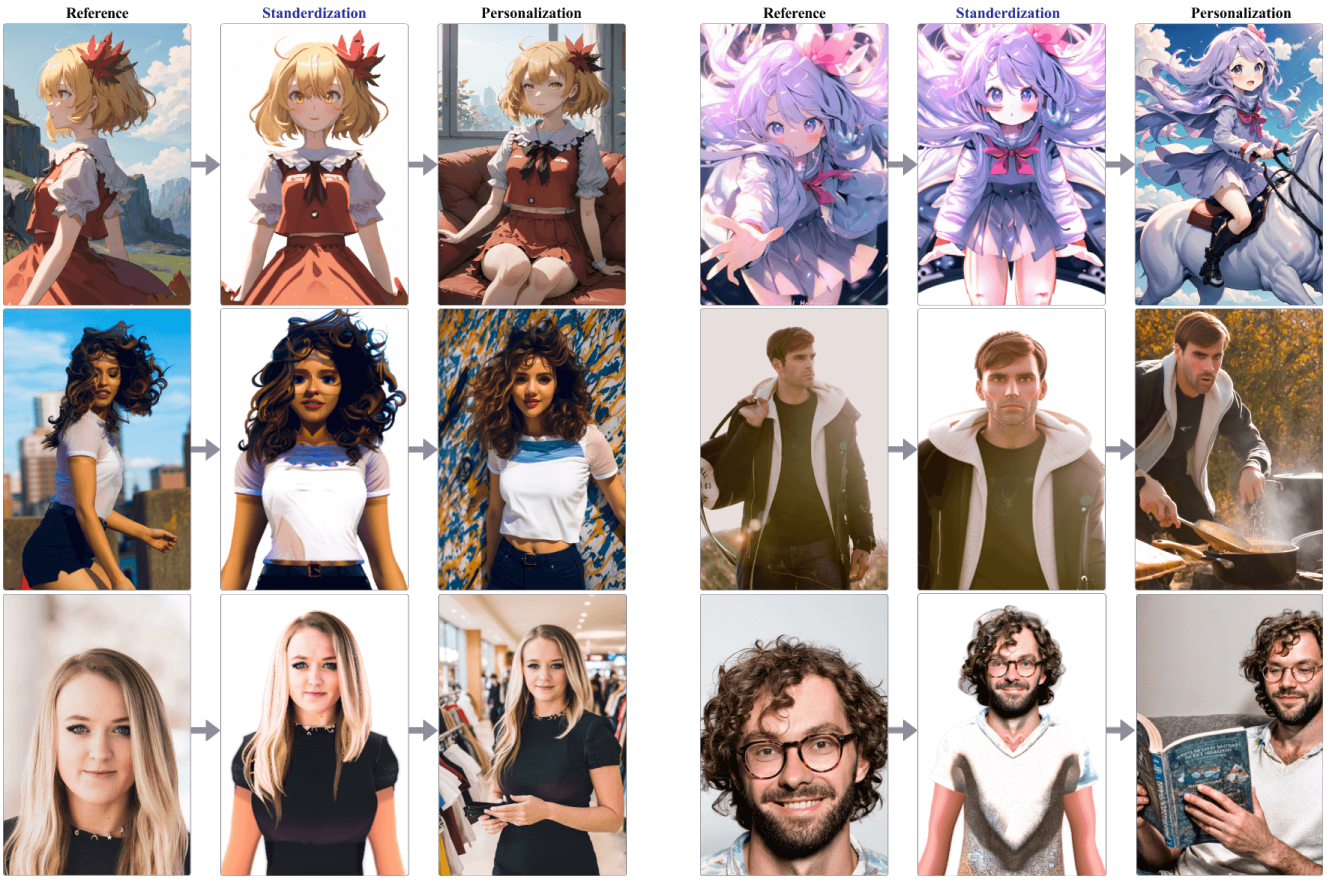

Figure 11. The standardization introduces a slight 3D style bias, particularly evident in head-only inputs (last row), resulting in clothing with a 3D appearance. This bias is effectively mitigated during the personalization stage.

pression, arranged from the first row to the fourth row, respectively. In the first row, both StoryMaker and our method successfully alter the character's action, with our method maintaining a more accurate hairstyle. In the second row, both FastComposer and our method produce images featuring a clear jungle background; however, FastComposer does not accurately depict the character's clothes. In the third and fourth rows, while IP-Adapter manages to capture the anime character's appearance, it struggles to modify the viewpoint and expression, a limitation attributed to its unpaired one-stage training strategy. Conversely, our method effectively generates images that match the descriptions of expressions and viewpoints accurately. Our approach demonstrates superior performance in maintaining appearance consistency and textual controllability compared to other leading-edge methods.

9. User Study

As shown in Table 5, we design three criteria for comparison, where each criterion receives 600 valid votes (30 participant $\times$ 20 text-image pairs). The detailed questions are as follows: 1) Whole-body Appearance Consistency: Which method best preserves the input character's whole-body appearance? 2) Text Controllability: Which method generates images that best align with the input text prompt? 3) Visual Appeal: Which method produces the most visually appealing image? To ensure objectivity, the names of all methods are anonymized, and the methods are presented in a randomized order for each question.

Method	WAC	TC	VA
IP-Adapter [33]	20.00%	4.33%	5.67%
FastComposer [30]	4.67%	3.67%	0.67%
StoryMaker [36]	20.33 %	37.67%	22.67%
Ours	**55.00%**	**54.33%**	**71.00%**

Table 5. User preference in personalized image generation, evaluated across three criteria: whole-body appearance consistency (WAC), text controllability (TC), and visual appeal (VA).

Figure 12. More comparison with other methods.

10. Limitations of Unpaired Training

In these experiments, we train models on unpaired image data, using identical images as both reference and target. For the reference encoder, we employ IP-Adapter [33], while SDXL is utilized as the diffusion model. The feature size extracted from the reference image can be adjusted by modifying a setup parameter in IP-Adapter, known as the number of token features. An increase in the number of token features corresponds to a more powerful reference encoder.

We systematically train a series of IP-Adapter reference encoders, including both powerful and weak configurations,

Figure 13. Visual comparison of different numbers of token features. Leftmost is the reference image. Token-i indicates the model trained with a token feature number of i.

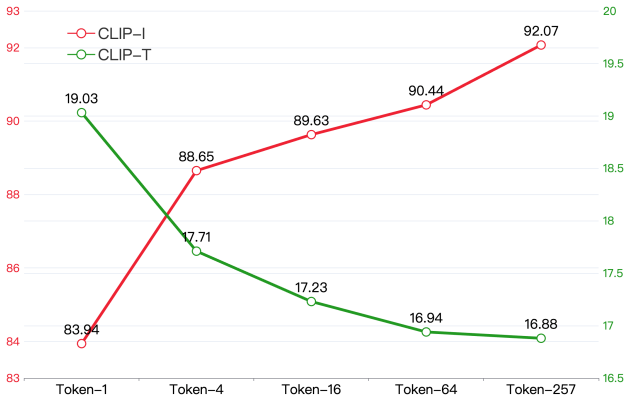

Figure 14. Comparison of different numbers of token features .

by varying the number of token features. Figure 13 presents some qualitative results with varying numbers of tokens. From these results, it is evident that unpaired training struggles to meet the objectives of personalized generation tasks: the models either compromise text controllability to maintain high appearance consistency or sacrifice appearance consistency to enhance text controllability. In the settings of powerful encoders—those equipped with a larger number of tokens—the models can easily replicate the reference image, achieving high appearance consistency but showing inadequate responsiveness to text prompts. Conversely, in the settings of weak encoders, there is a better alignment with text prompts, albeit at the expense of compromised appearance consistency. The quantitative results depicted in Figure 14 also align with these visual observations. As the number of tokens increases, indicating more powerful encoders, there is an observed rise in the CLIP-I score, from 83.94 to 92.07, while the CLIP-T score decreases, moving from 19.03 to 16.88.

Method	FVD↓	FID-VID↓
DisCo [28]	292.8	59.9
MagicPose [2]	-	46.3
MagicAnimate [31]	179.07	21.75
Animate Anyone [15]	171.9	-
Champ [37]	160.82	21.07
TCAN [16]	154.84	19.42
Ours	**149.95**	**14.75**

Table 6. Quantitative comparison on TikTok dataset.

11. Comparison to Human Image Animation Models

As discussed in Section 4.3.2, we compare the architecture of our standardization model with other leading human image animation models, including DisCo [28], MagicPose [2], MagicAnimate [31], Animate Anyone [15], Champ [37], and TCAN [16]. Experiments are conducted using the benchmark dataset TikTok [28]. No additional training data was utilized to ensure a fair comparison. To enable training on video datasets, a temporal layer [15] is incorporated into the architecture described in Section 3.3. The results presented in Table 6 demonstrate that our method significantly outperforms existing state-of-the-art approaches, achieving superior performance in both FVD and FID-VID metrics. These results justify our architec-

Module	CLIP-I ↑	Face Sim.↑
after standardization	89.47	0.69
after personalization	85.49	0.53

Table 7. Ablation study on identity loss at each stage.

tural choices.

12. Ablation Study on Identity Loss

We conduct an ablation study to evaluate the impact of each stage on identity preservation. As shown in Table 7, after the standardization stage, CLIP-I is 89.47, and Face Sim. is 0.69. Following the personalization stage, CLIP-I decreases to 85.49, while Face Sim. drops to 0.53. These results indicate that identity consistency remains relatively high after the standardization stage.

13. More Quantitative Comparisons

We also made a quantitative comparison between our method and the recent face-oriented approach LCM-Lookahead [9], which achieved a Face Sim. score of 0.46, CLIP-I score of 74.56, and CLIP-T score of 24.63 on the test dataset. Our method outperforms LCM-Lookahead in both CLIP-I and Face Sim. metrics, with only a disadvantage in CLIP-T. Notably, LCM-Lookahead achieves good text controllability at the cost of consistency.