CompleteMe: Reference-based Human Image Completion

Supplementary Material

6. Overview

This supplementary material presents additional results to complement the main manuscript. We first include more diverse applications with our method in Sec. 7. We then introduce more implementation details in Sec. 8 and compare the training dataset with other methods in Sec. 9. We further provide more benchmark settings and examples in Sec. 10. In Sec. 11, we provide a more detailed explanation for our ablation study on different training strategies. In Sec. 12, we provide a visual comparison of different inputs for our model during inference, which shows the robustness and flexibility of our model. In Sec. 13, we provide more qualitative comparisons with the reference-based methods. Finally, we discuss the limitations and provide future works in Sec. 14.

7. More Diverse Applications

Our method, *CompleteMe*, demonstrates versatile applicability beyond basic completion, effectively supporting realistic virtual try-on and advanced image editing tasks, as shown in Fig. 8. By leveraging detailed reference guidance and the Region-focused Attention mechanism, *CompleteMe* accurately transfers complex clothing patterns and accessories, enabling high-quality content generation suitable for fashion, e-commerce, and creative image editing applications.

8. More Implementation Detail

All experiments are conducted with the resolution of 512 $\times$ 512 and resized back to the original resolution to show the visual results. For our masking strategy during training, the mask grid size is between 3% to 25% of the image resolution, and we randomly apply these mask grids from 1 to 30 times in random positions.

For inference, we use only one reference image and text prompt for our method and apply reference masks for our model. We want to note that text prompts and reference masks are optional inputs for our model.

9. Training Dataset Comparison

Compared methods are trained on significantly larger or broader datasets as follows:

- LOHC [41]: 57K images from the AHP dataset, specifically focused on humans.
- BrushNet [12]: 1.2 billion images from Laion-Aesthetic.
- Paint-by-Example [35]: 1.9 million images from OpenImage.
- AnyDoor [4]: 410K images from various video datasets.
- LeftRefill [1]: 820K image pairs from MegaDepth.
- MimicBrush [3]: 100K video frames and 10 million images from SAM.

Despite having a smaller training dataset than these methods, our model achieves superior results by leveraging human-specific priors and a carefully curated benchmark. This demonstrates the efficiency of our approach in using targeted human data rather than vast generic datasets.

10. Benchmark Detail

To better evaluate the performance of different methods, we construct our benchmark from the Wpose dataset in UniHuman [16]. The Wpose dataset contains 872 distinct person IDs, and some IDs have more than one input-reference pair. We mainly use one input-reference pair for each person's ID. We crop a rectangle centering the subject in the image and resize its longer side to 1,024 pixels. We show more visual examples for our benchmark in Fig. 9.

We use LLaVA [17, 18] to generate text prompts describing the source image. We provide some text prompt examples here:

- A woman wearing a white shirt and white pants sits on a brick staircase.
- A woman wearing a black dress with red roses on it is standing in front of a door.
- A woman wearing a striped sweater and tan pants sits on a wooden post by the water.
- A man wearing a white shirt and black shorts with white socks.
- A man wearing a blue jean jacket and a red jersey with the number 23 on it.
- A man wearing a white shirt and khaki pants is leaning against a wall.

11. Detail Explanation for Ablation on Different Training Strategies.

As shown in Fig. 7, in Exp. (a) Freeze U-Net, the model generates plausible results but still lacks some specific details, such as the missing hand in the top image. In Exp. (b) Freeze U-Net+Prompt, after incorporating an additional text prompt, the model improves by recovering the hand pose in the top image and adding detailed texture to the pants in the bottom image. Furthermore, in Exp. (c) Freeze U-Net+Prompt+Ref Mask, we introduce a reference mask that contains only the human regions for our Region-focused Attention Block, allowing the model to better focus on the human body, identify correct correspondences, and

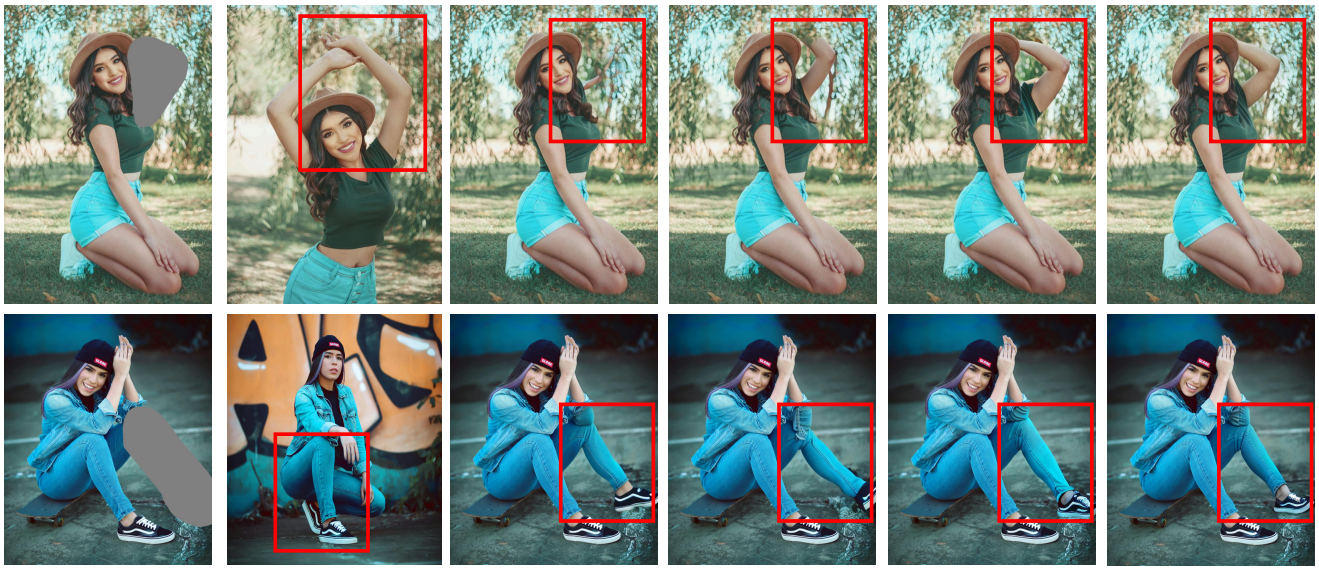

Figure 7. **Qualitative Comparison on Different Training Strategies.** The experimental index follows configurations in Table 5. The Red box highlights the finely detailed regions where different models exhibit varying performance based on distinct training strategies.

generate accurate details, such as the shape of the arm and the texture of the shoes.

Finally, with *CompleteMe*, we train the Reference U-Net to better align its feature space with that of the Complete U-Net. This alignment allows us to preserve fine details from the reference image, enabling the completion of missing regions with realistic content.

12. Different Inference Inputs

During inference, *CompleteMe* demonstrates the flexibility to accept various inputs, including optional text prompts and reference masks. As shown in Fig. 10, we compare the visual results generated with different inference inputs, highlighting the robustness and adaptability of our method in handling diverse conditions while maintaining high-quality completions.

13. More Visual Comparison

We provide more visual comparisons with reference-based methods: Paint-by-Example [35], AnyDoor [4], LeftRefill [1], and MimicBrush [3]. As shown in Fig. 11 to Fig. 20, *CompleteMe* effectively completes the masked region by accurately preserving identical information and correctly mapping corresponding parts of the human body from the reference image.

14. Limitation and Future Work

While our *CompleteMe* model demonstrates strong performance in human image completion, it faces limitations that highlight avenues for future improvement. Our model depends on the quality and availability of reference images; when these references fail to capture specific details or perspectives, the completion results may lack fidelity. Additionally, the reliance on pre-trained models such as Stable Diffusion [24] and CLIP [23] embeddings restricts adaptability to domains where these pre-trained backbones perform suboptimally.

To address these challenges, our future work focuses on adapting the model to leverage new and more versatile backbones, like Stable Diffusion 2, enhancing its applicability across diverse scenarios. Moreover, expanding our benchmark datasets to include a wider variety of tasks, poses, and object types enables a more comprehensive evaluation of the model's robustness and versatility, driving progress in both human-centric and generalized image completion tasks.

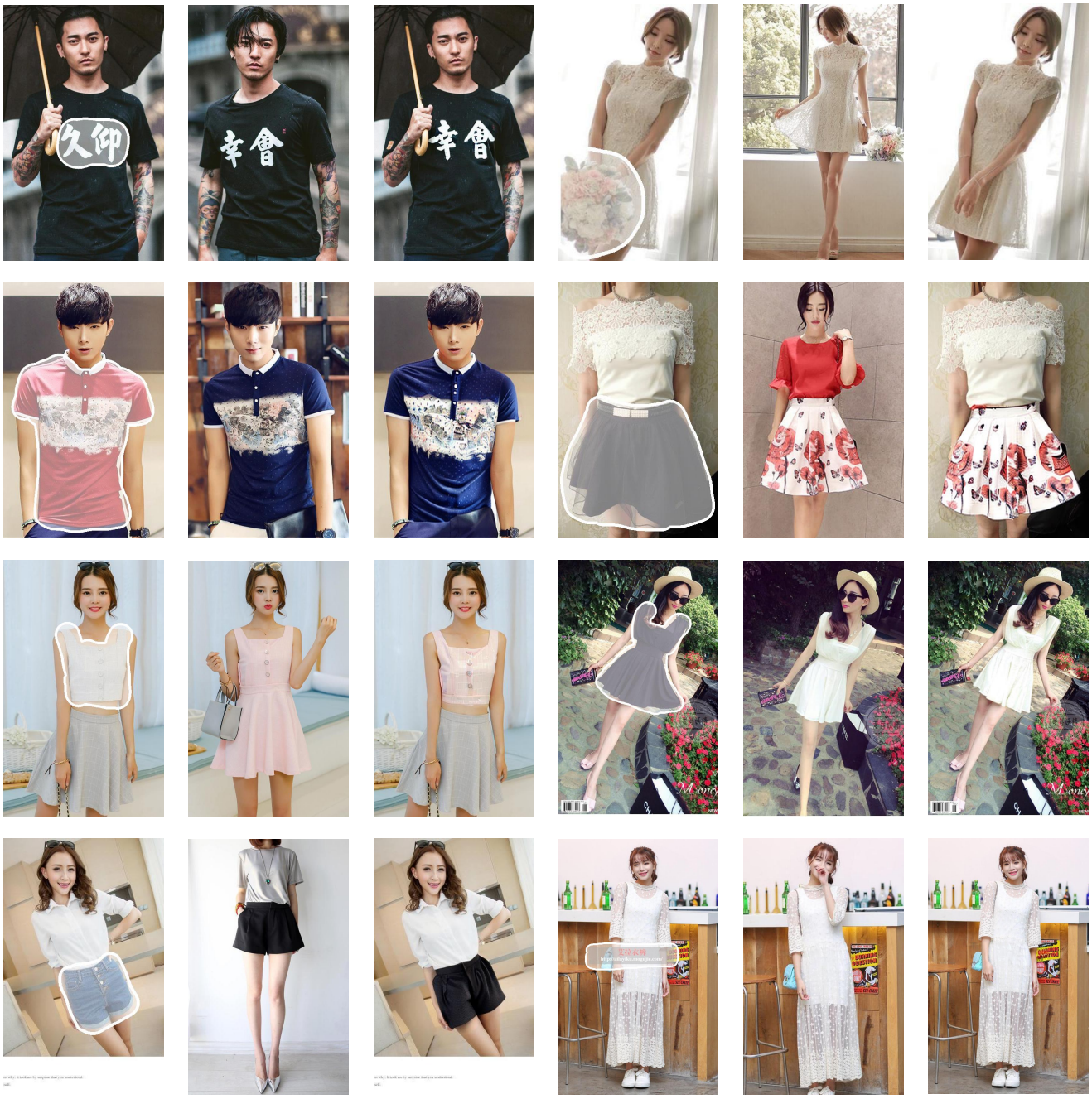

Figure 8. **More Diverse Applications.** We provide more diverse applications with our method on virtual try-on and image editing tasks.

Figure 9. **More Benchmark Examples.** We provide more examples from our benchmark, including the source image, inpainting area, and reference image.

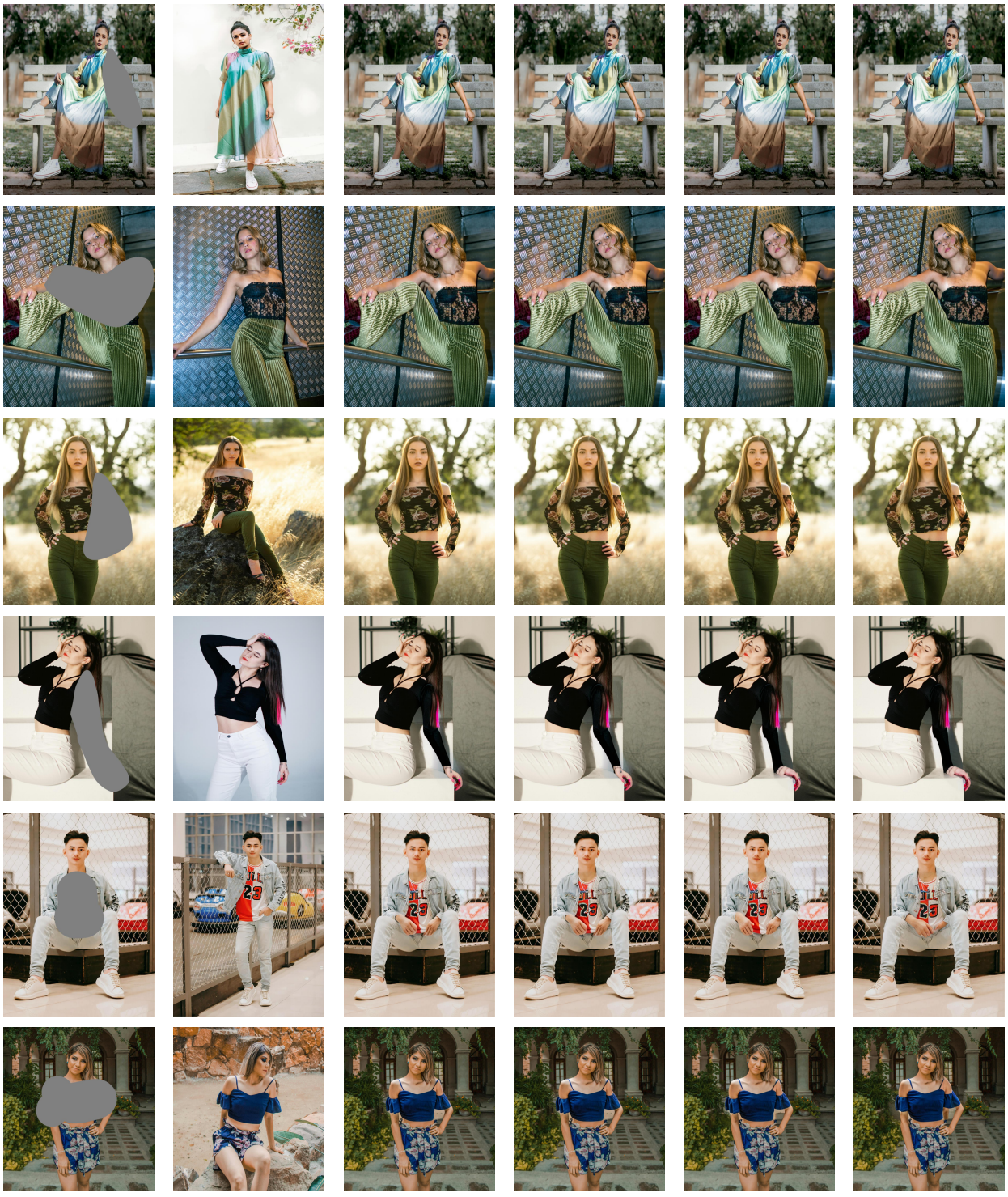

Figure 10. **Different Inference Inputs.** We provide more examples of different inputs for our model during inference time, in which text prompts and reference masks are optional inputs for our model. *CompleteMe* use the inputs with text prompt and reference mask for best performance.

Figure 11. **Qualitative Comparison with Reference-based Methods on Our Benchmark (Sec. 3.4).** Our *CompleteMe* can generate more realistic and preserve identical information from the reference image. Please zoom in for a better comparison.

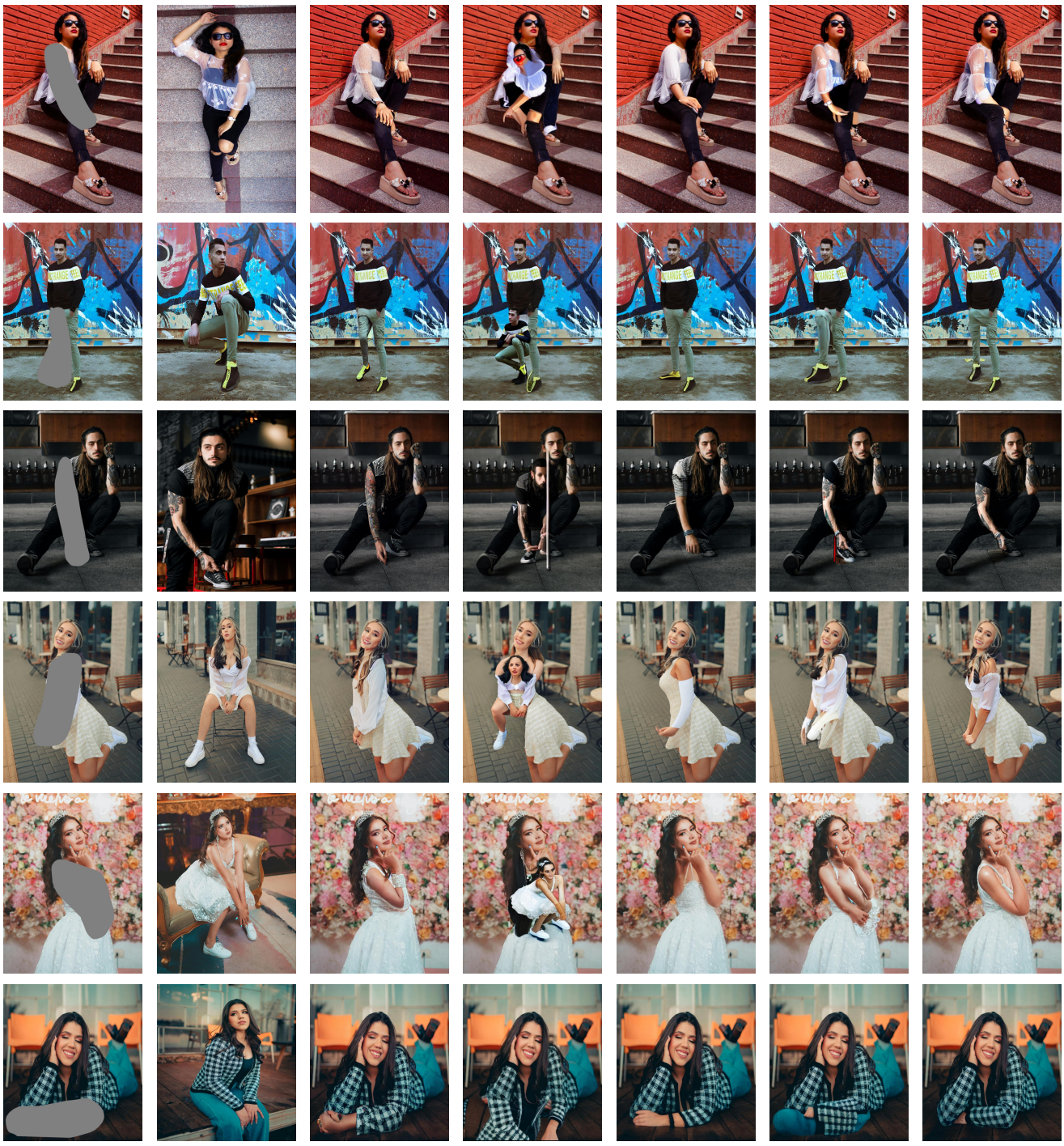

Figure 12. **Qualitative Comparison with Reference-based Methods on Our Benchmark (Sec. 3.4).** Our *CompleteMe* can generate more realistic and preserve identical information from the reference image. Please zoom in for a better comparison.

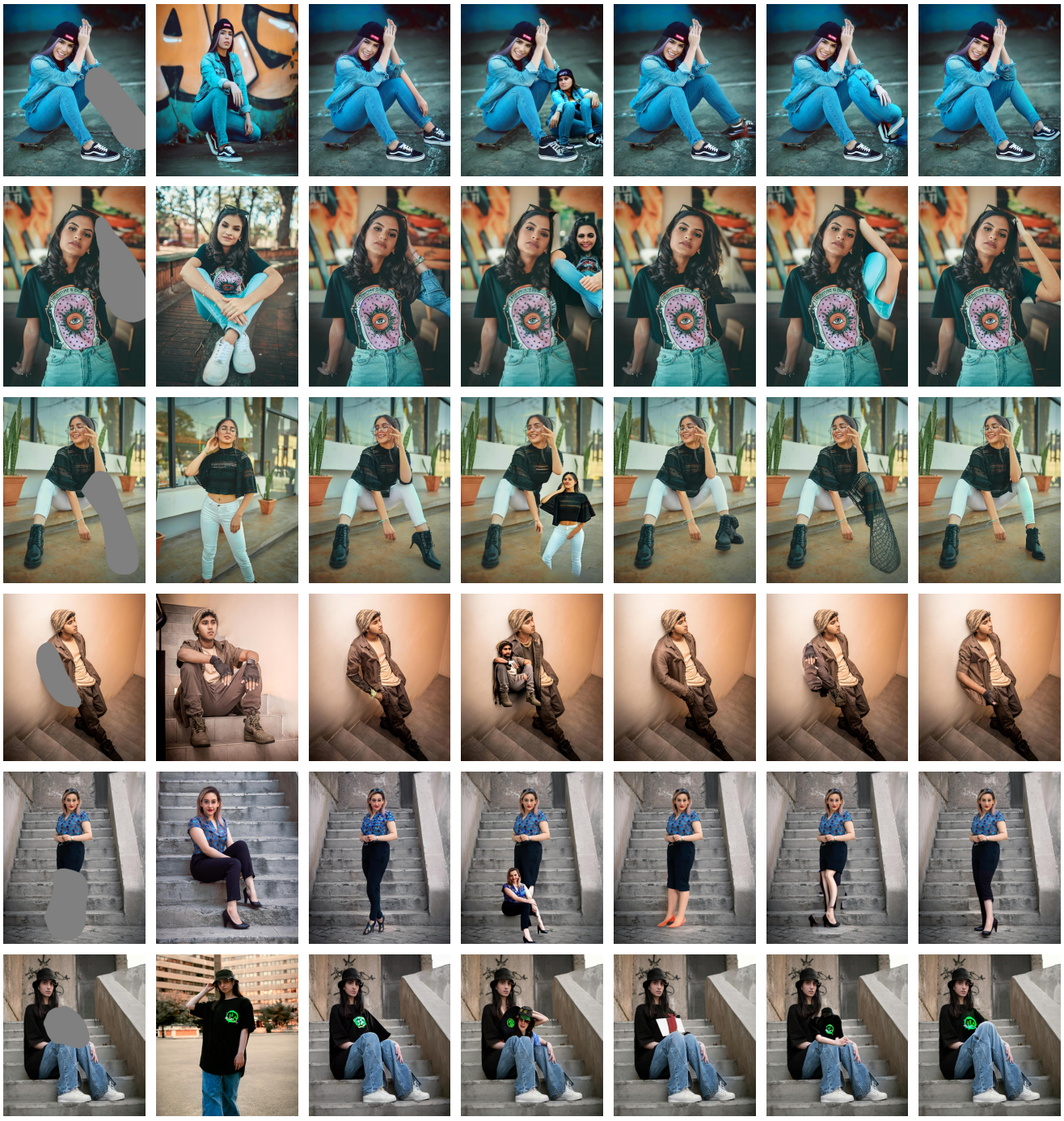

Figure 13. **Qualitative Comparison with Reference-based Methods on Our Benchmark (Sec. 3.4).** Our *CompleteMe* can generate more realistic and preserve identical information from the reference image. Please zoom in for a better comparison.

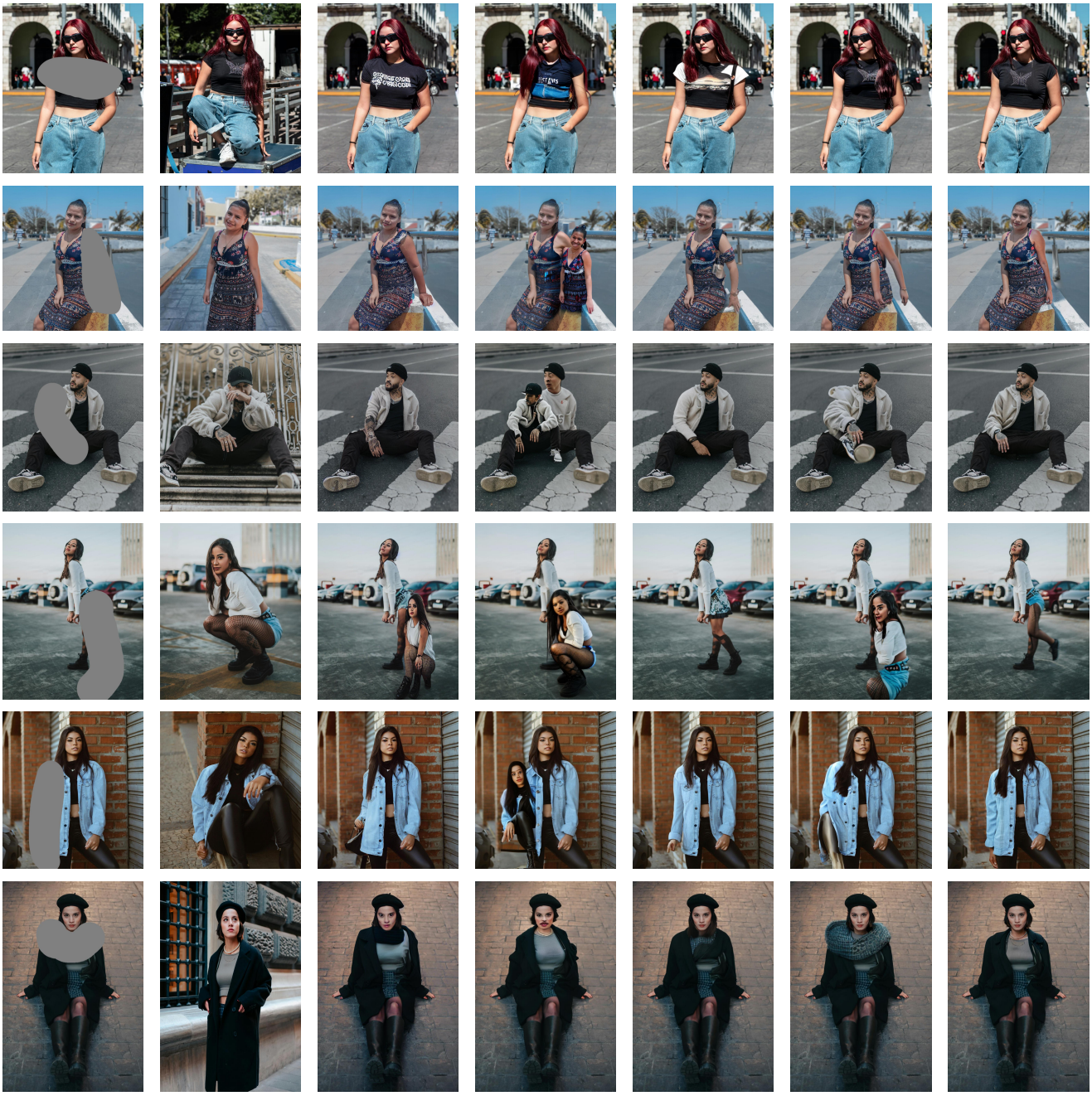

Figure 14. **Qualitative Comparison with Reference-based Methods on Our Benchmark (Sec. 3.4).** Our *CompleteMe* can generate more realistic and preserve identical information from the reference image. Please zoom in for a better comparison.

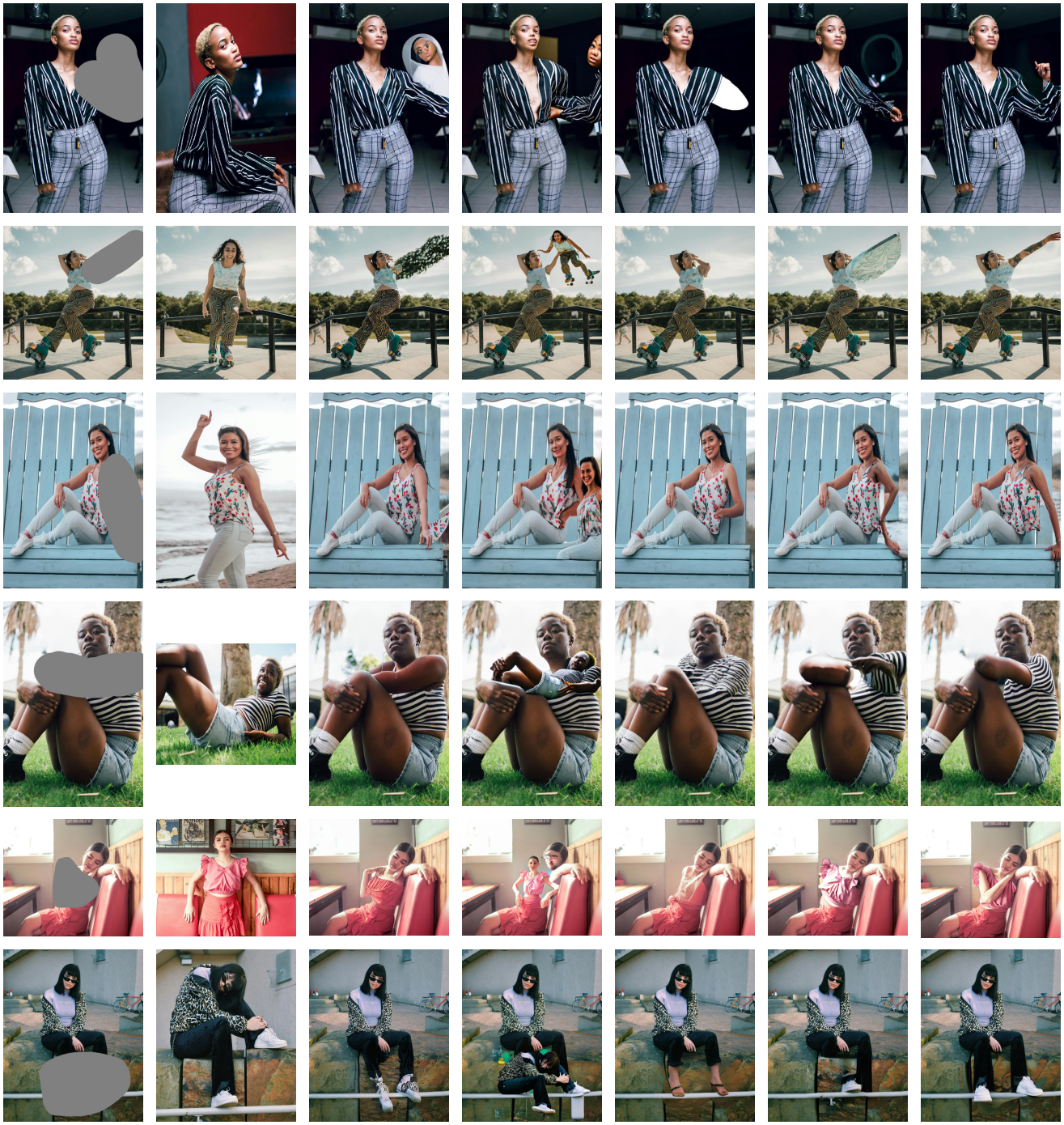

Figure 15. **Qualitative Comparison with Reference-based Methods on Our Benchmark (Sec. 3.4).** Our *CompleteMe* can generate more realistic and preserve identical information from the reference image. Please zoom in for a better comparison.

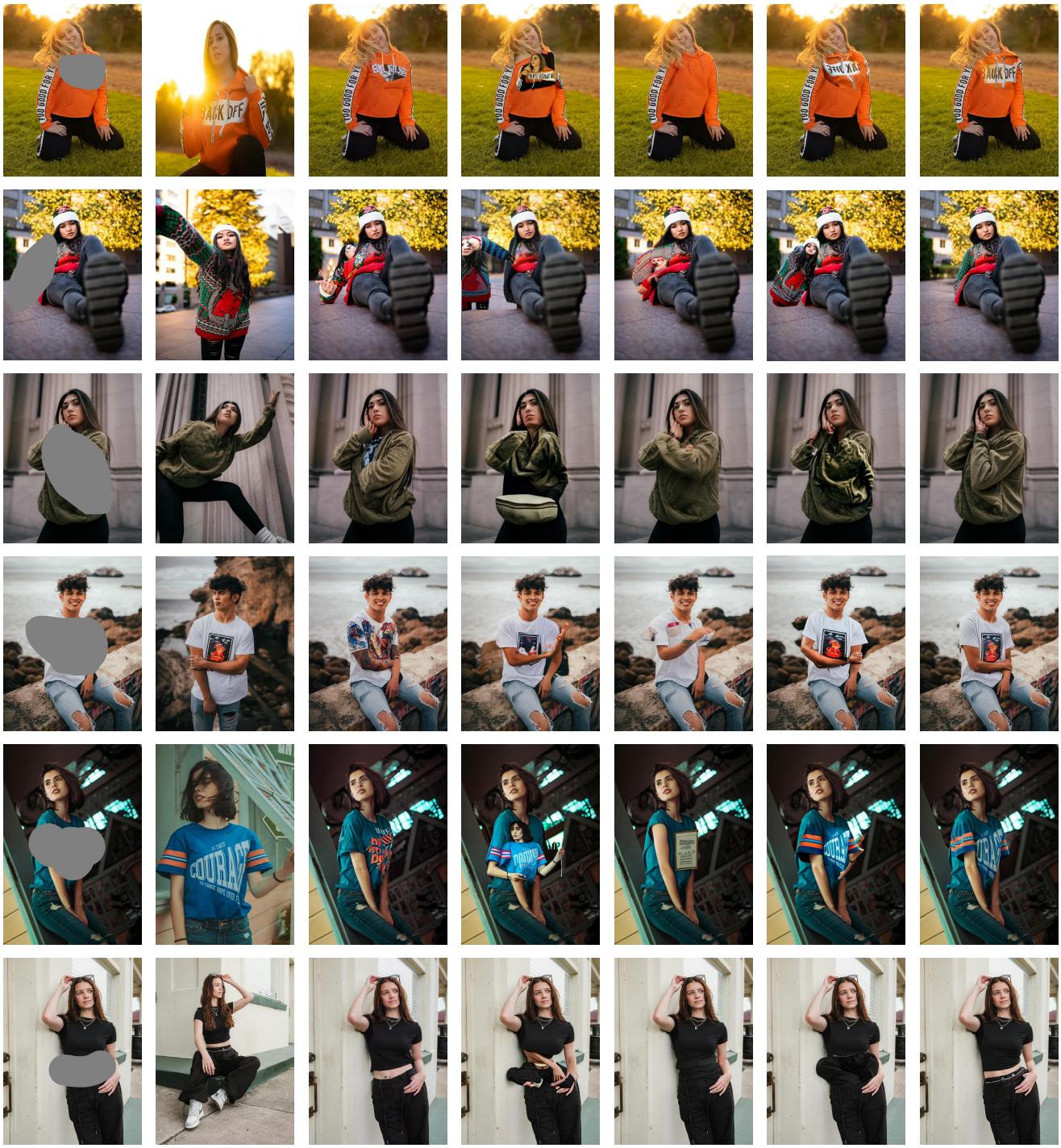

Figure 16. **Qualitative Comparison with Reference-based Methods on Our Benchmark (Sec. 3.4).** Our *CompleteMe* can generate more realistic and preserve identical information from the reference image. Please zoom in for a better comparison.

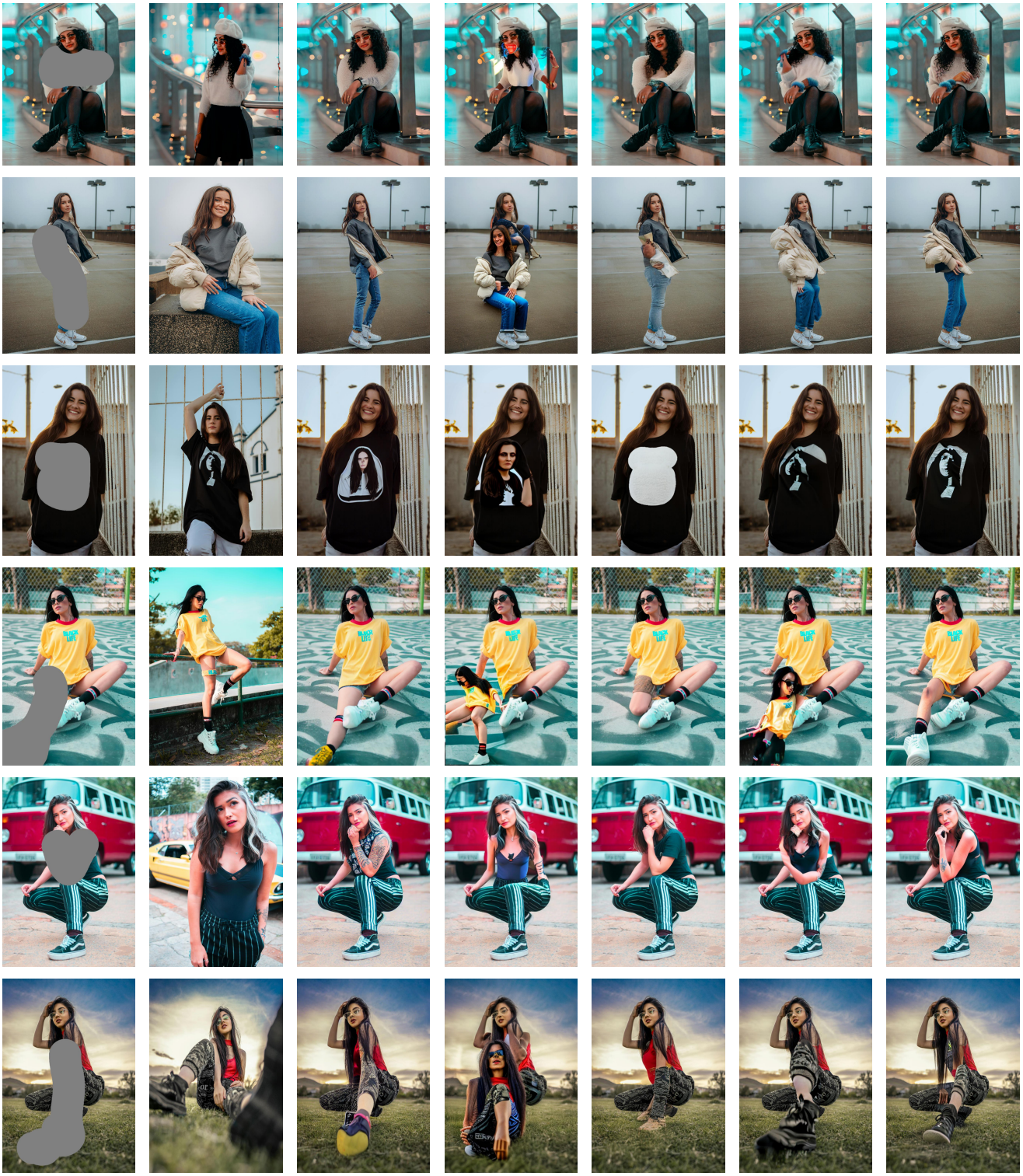

Figure 17. **Qualitative Comparison with Reference-based Methods on Our Benchmark (Sec. 3.4).** Our *CompleteMe* can generate more realistic and preserve identical information from the reference image. Please zoom in for a better comparison.

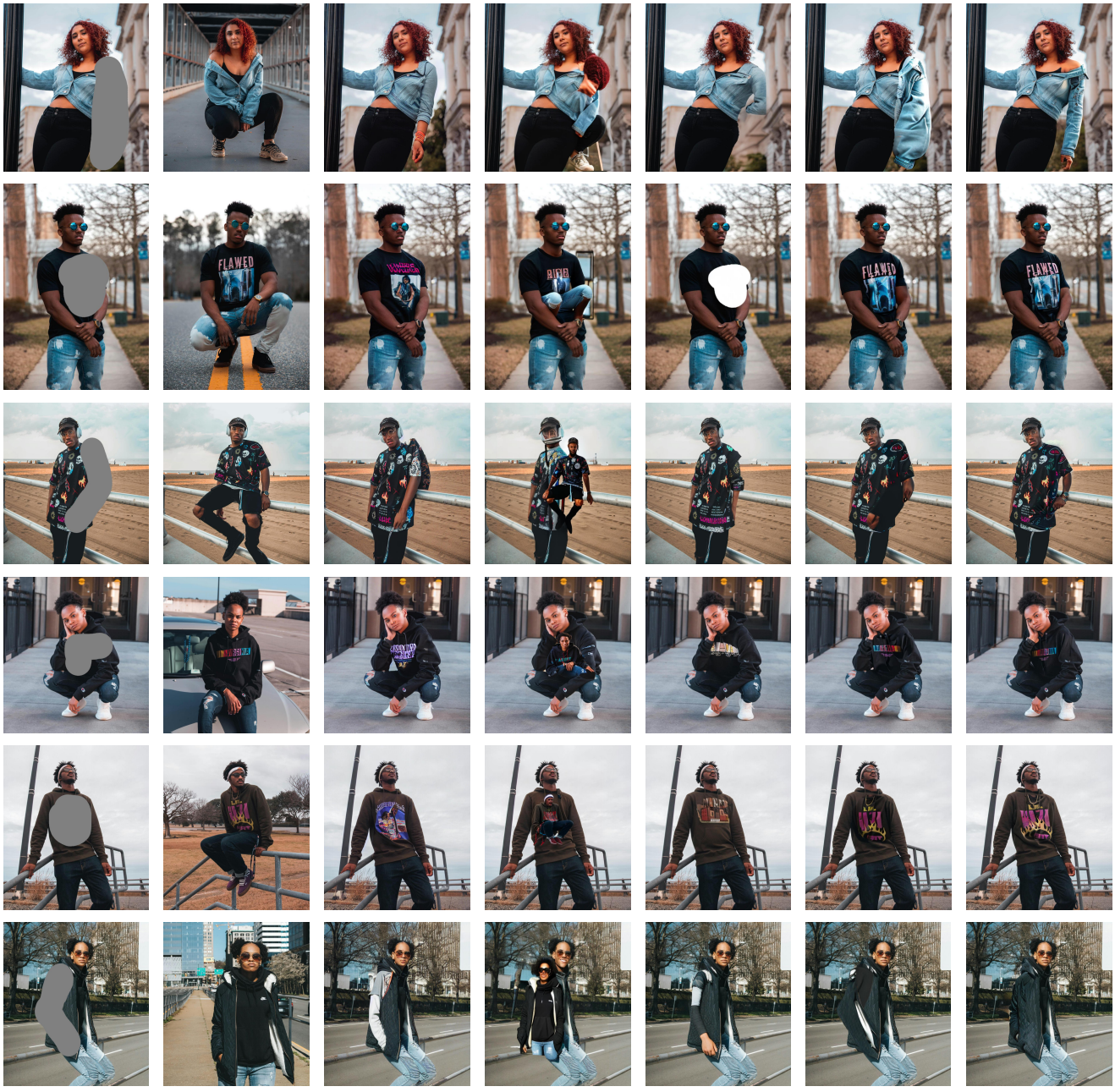

Figure 18. **Qualitative Comparison with Reference-based Methods on Our Benchmark (Sec. 3.4).** Our *CompleteMe* can generate more realistic and preserve identical information from the reference image. Please zoom in for a better comparison.

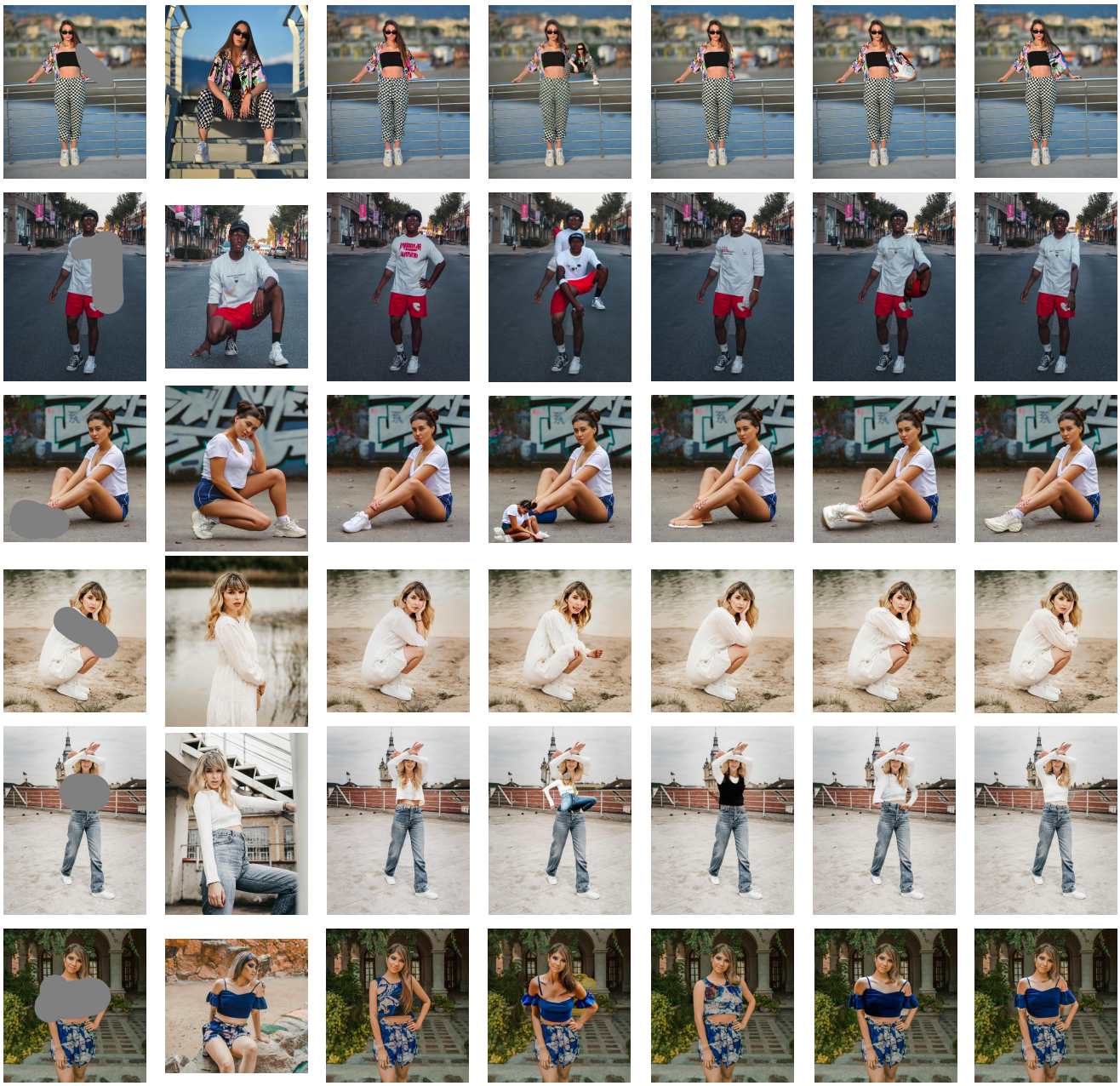

Figure 19. **Qualitative Comparison with Reference-based Methods on Our Benchmark (Sec. 3.4).** Our *CompleteMe* can generate more realistic and preserve identical information from the reference image. Please zoom in for a better comparison.

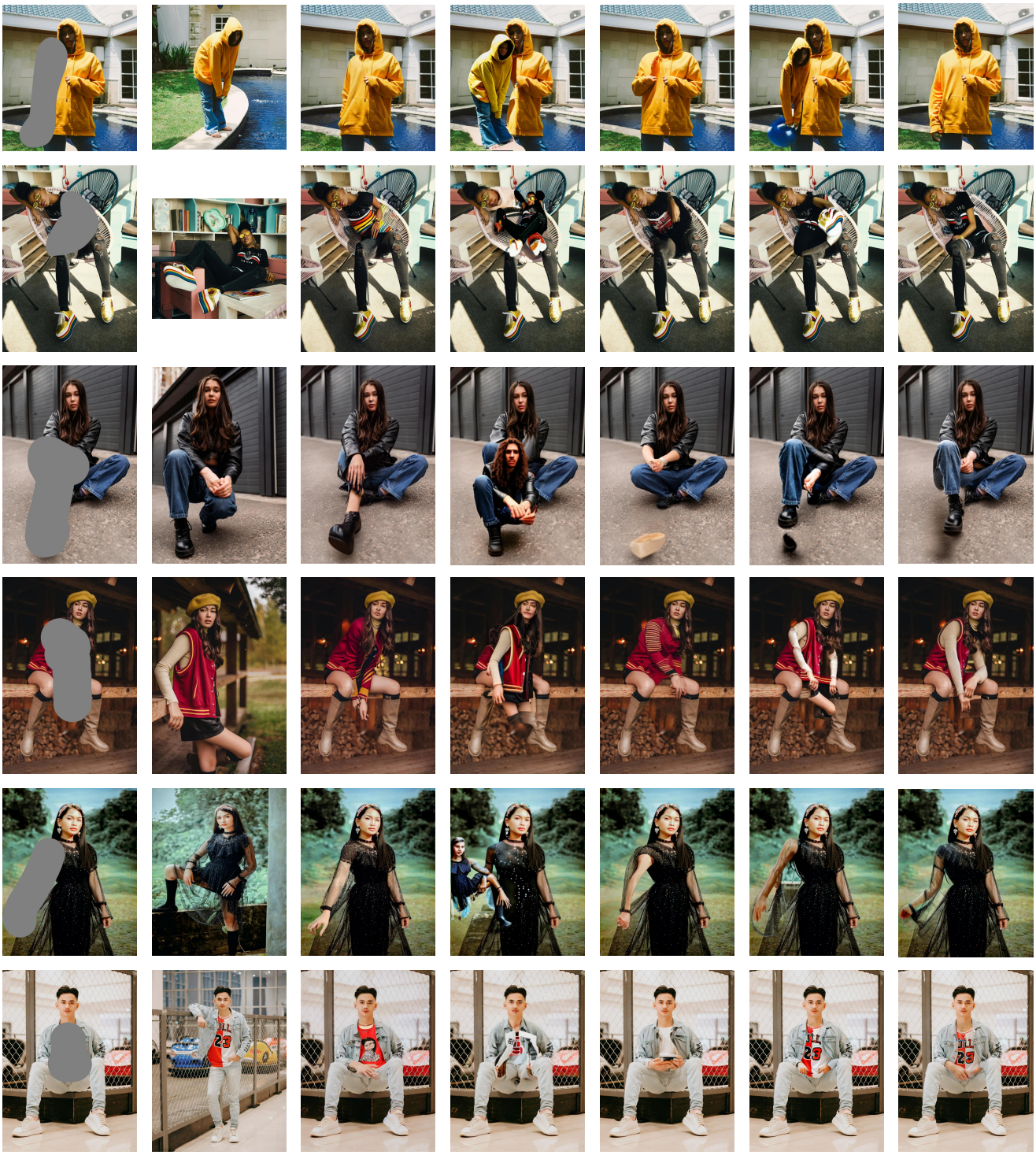

Figure 20. **Qualitative Comparison with Reference-based Methods on Our Benchmark (Sec. 3.4).** Our *CompleteMe* can generate more realistic and preserve identical information from the reference image. Please zoom in for a better comparison.