CompleteMe: Reference-based Human Image Completion

Supplementary Material

7. Overview

This supplementary material presents additional results to complement the main manuscript. We first include more diverse applications with our method in Sec. 8. We then introduce more implementation details in Sec. 9 and compare the training dataset with other methods in Sec. 10. We provide more quantitative comparisons in Sec. 11. We further provide more benchmark settings and examples in Sec. 12. In Sec. 13, we discuss the discrepancy between quantitative evaluation and user study. In Sec. 14, we provide a more detailed explanation for our ablation study on different training strategies. In Sec. 15, we conduct an ablation with multiple reference images with our method. In Sec. 16, we provide a visual comparison of different inputs for our model during inference, which shows the robustness and flexibility of our model. In Sec. 17, we provide more qualitative comparisons with the reference-based methods. Finally, we discuss the limitations and provide future works in Sec. 18.

8. 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. 9. 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.

9. More Implementation Detail

All experiments are conducted with the resolution of 512 × 512 and resized back to the original resolution to show the visual results.

We use the unchanged VAE from the SD v1.5 checkpoint for our pipeline. For reference feature extraction, each reference image is encoded once to a latent tensor 64×64×4 by the SD v1.5 VAE encoder. For training and inference, every masked source image is encoded in the same latent space. During training, we use the 42-class human parsing masks provided by the DeepFashion-MultiModal dataset (no manual work).

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.

During inference, we use a pretrained SegFormer-B2-Clothes [32] model to generate the binary mask from each reference image. Users may optionally override it with their mask.

Additionally, 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.

10. Training Dataset Comparison

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

- LOHC [43]: 57K images from the AHP dataset, specifically focused on humans.
- BrushNet [12]: 1.2 billion images from Laion-Aesthetic.
- Paint-by-Example [37]: 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.

11. More Quantitative Comparison

In Table 6, each method with official model weight receives the same test group: (1) the occluded image, (2) the full body reference image, and (3) a binary human mask produced by our SegFormer [32] parser. We want to emphasize that Paint-by-Example [37] and AnyDoor [4] already accept reference masks in our main paper. We retrain the LeftRefill [1] with the DeepFashion-MultiModal [11, 19] training dataset with experimental settings from their paper. We provide the evaluation results in the following table. The results show that all methods are evaluated with the same inputs, where LeftRefill(Retrain) is retrained on the human data, and our *CompleteMe* still achieves the best scores. We also want to emphasize that the results are comparable between the original LeftRefill and the Retrained LeftRefill. Therefore, our advantage comes from the proposed RFA mechanism rather than from special inputs or a different training set.

Table 6. **More Quantitative Comparison on Our Benchmark**

Method	LeftRefill (Retrain)	LeftRefill	MimicBrush	*CompleteMe*
DINO ↑	96.23	96.17	93.15	**96.29**
DreamSim ↓	0.0461	0.0462	0.0839	**0.0419**
LPIPS ↓	0.0611	0.0606	0.0722	**0.0588**

12. 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. 10.

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.

13. Discrepancy between Quantitative Evaluation and User Study

Existing perceptual metrics (LPIPS, DINO, DreamSim) average over the full image, which tends to ignore the fine-detailed area where success or failure is most visible to human perception. We provide the two visual examples in Fig. 7 and the evaluation score in Table 7. LeftRefill scores better numerically, while it clearly misses specific details from the reference image (Red Boxes). For our user study, we want to clarify that pairs are randomly sampled from our benchmark, and the method order is also random. Hence, the user study is reliable, and the mismatch stemmed from whole-image perceptual metrics that do not capture fine inpainting quality, not from a bias in sampling. We provide many additional qualitative comparisons in this supplementary material to further demonstrate our strengths.

Table 7. **Quantitative Comparison for Fig. 7**

Left	DINO ↑	DreamSim ↓	LPIPS ↓	Right	DINO ↑	DreamSim ↓	LPIPS ↓
LeftRefill	**98.31**	**0.0372**	0.0516		**97.00**	**0.0287**	**0.0674**
Ours	96.13	0.0521	**0.0493**		94.61	0.0590	0.0815

14. Detail Explanation for Ablation on Different Training Strategies.

As shown in Fig. 8, 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 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.

15. Ablation Study with Multiple Reference Images

We conduct an ablation with multiple reference images in the following Table 8. When the reference images increase, the scores do not change significantly, which shows our model is robust to the number of extra references. In practice, users pass a list of images and reference pairs to our model. The common case is a single full-body shot, but additional zoom-ins or different photos of the same outfit are accepted without further tuning.

Table 8. **Ablation Study with Multiple Reference Images**

# Reference Images	1	2	3	4	5	6
DINO ↑	**96.29**	96.09	95.92	95.83	95.61	95.43
DreamSim ↓	**0.0419**	0.0439	0.0449	0.0455	0.0470	0.0481
CLIP-T ↑	29.83	29.77	29.83	29.82	**29.85**	29.83

16. Different Inference Inputs

During inference, *CompleteMe* demonstrates the flexibility to accept various inputs, including optional text prompts and reference masks. As shown in Fig. 11, 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.

17. More Visual Comparison

We provide more visual comparisons with reference-based methods: Paint-by-Example [37], AnyDoor [4], LeftRefill [1], and MimicBrush [3]. As shown in Fig. 12 to

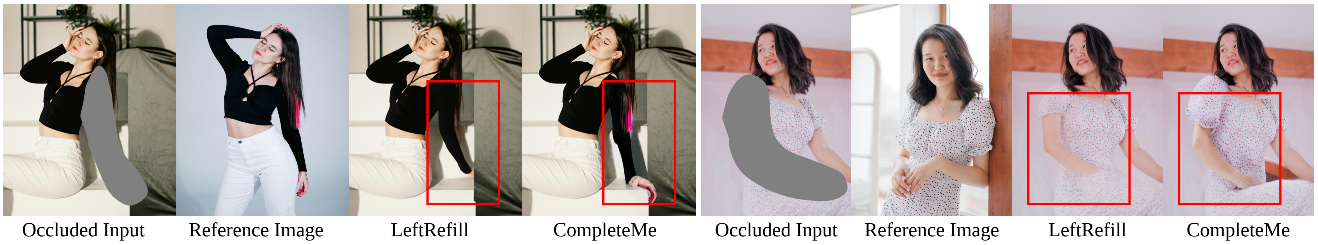


Figure 7. **Example for Discrepancy between Quantitative Evaluation and User Study.** Please refer to Table 7 for the quantitative evaluation.

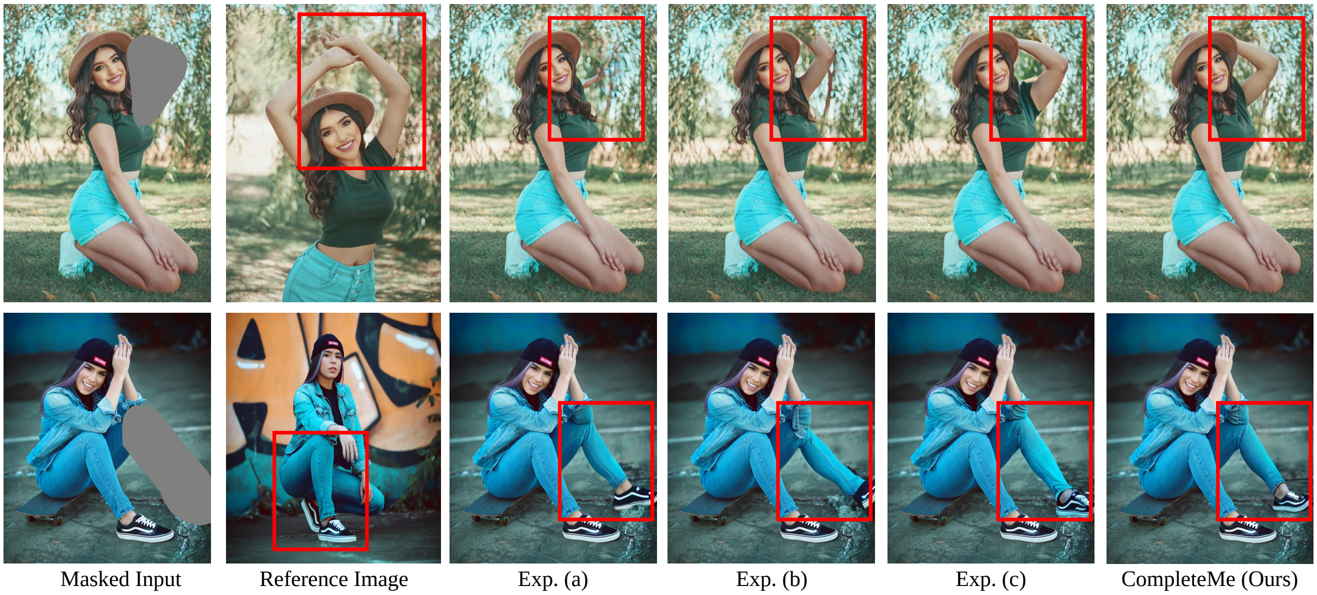


Figure 8. **Qualitative Comparison on Different Training Strategies.** The experimental index follows configurations in Table 5. The Red box highlights the finely detailed regions where different models exhibit varying performance based on distinct training strategies.

Fig. 21, *CompleteMe* effectively completes the masked region by accurately preserving identical information and correctly mapping corresponding parts of the human body from the reference image.

18. 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 [25] and CLIP [24] 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.

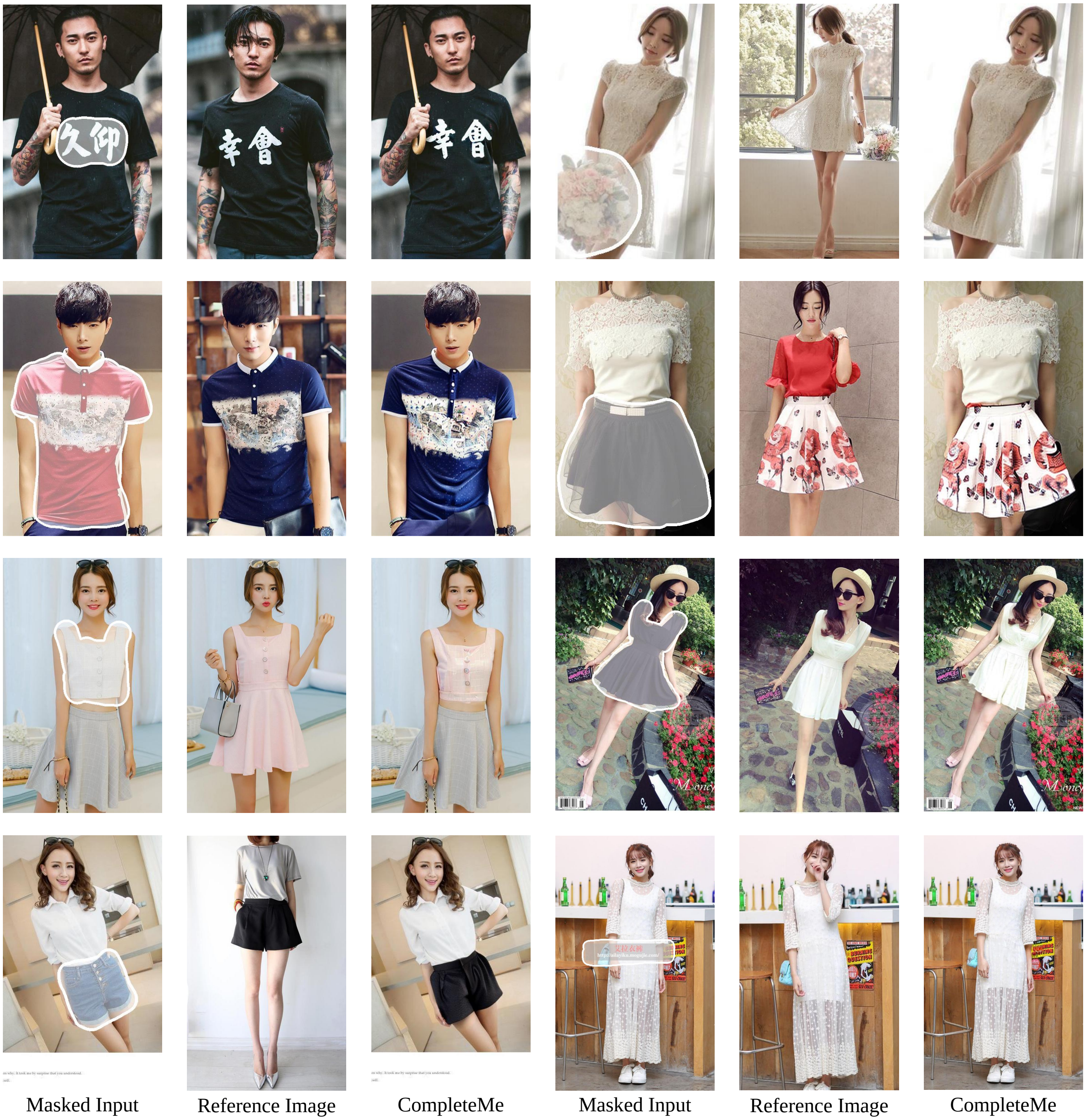


Figure 9. **More Diverse Applications.** We provide more diverse applications with our method on virtual try-on and image editing tasks.

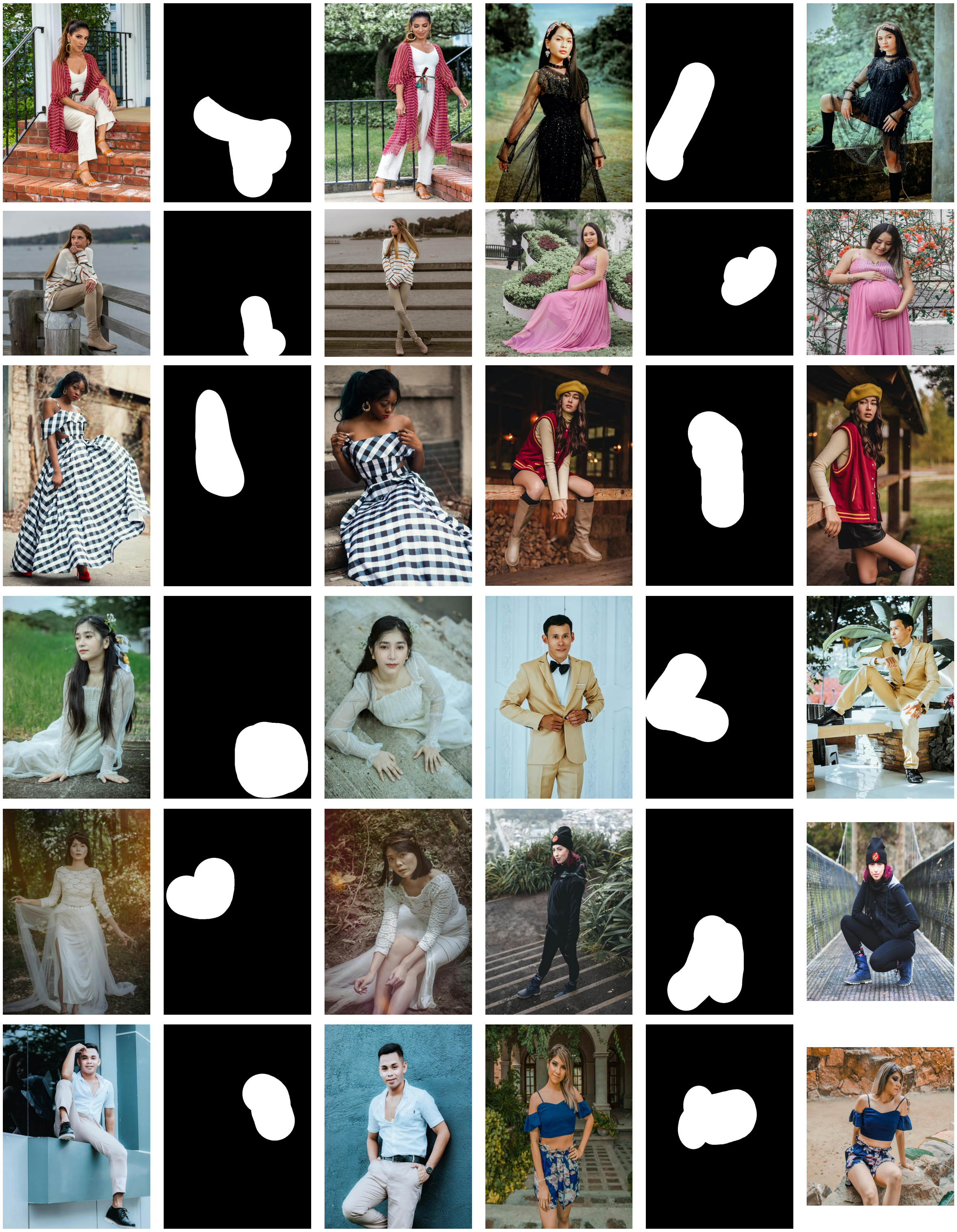


Figure 10. **More Benchmark Examples.** We provide more examples from our benchmark, including the source image, inpainting area, and reference image.

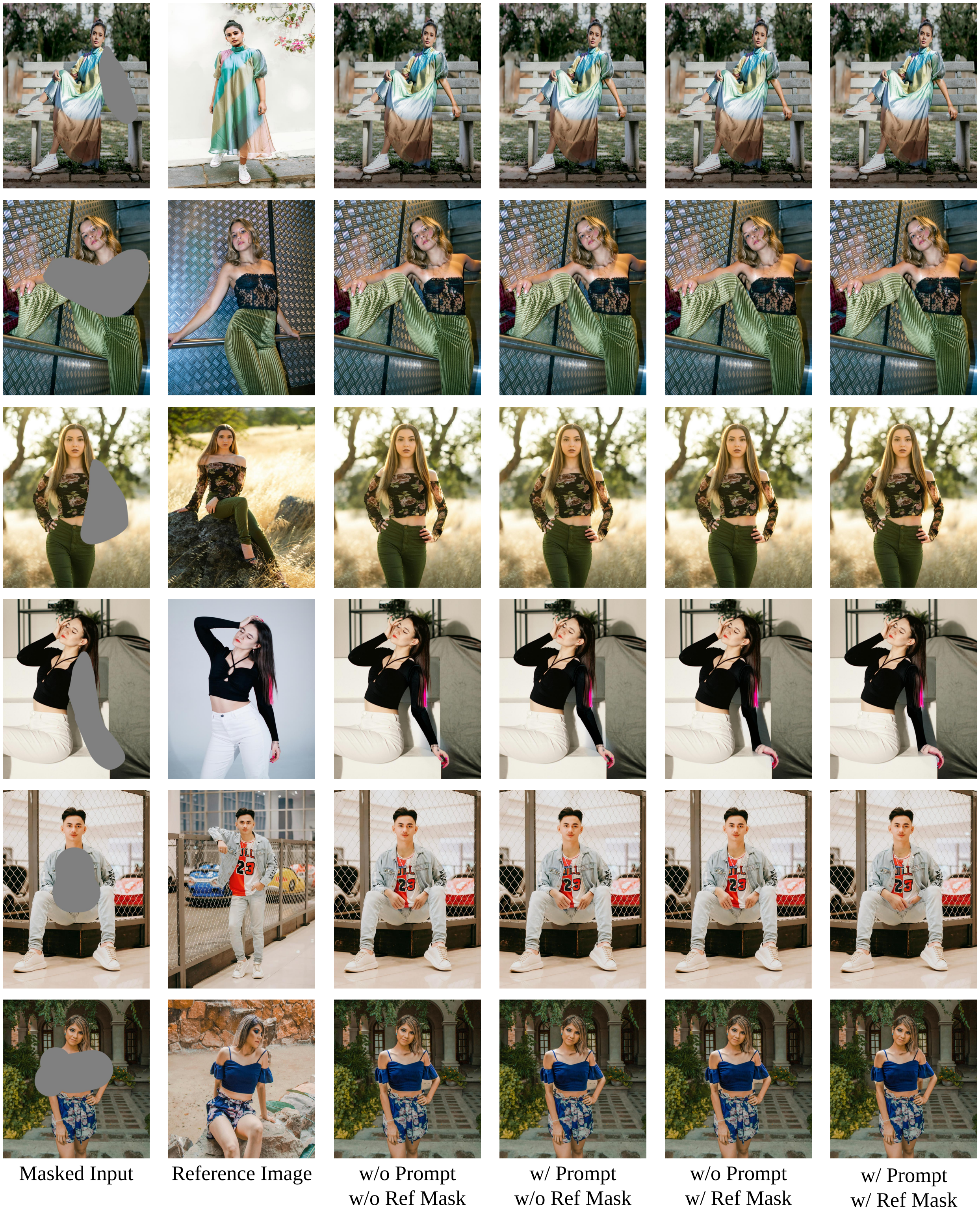


Figure 11. **Different Inference Inputs.** We provide more examples of different inputs for our model during inference time, in which text prompts and reference masks are optional inputs for our model. *CompleteMe* use the inputs with text prompt and reference mask for best performance.

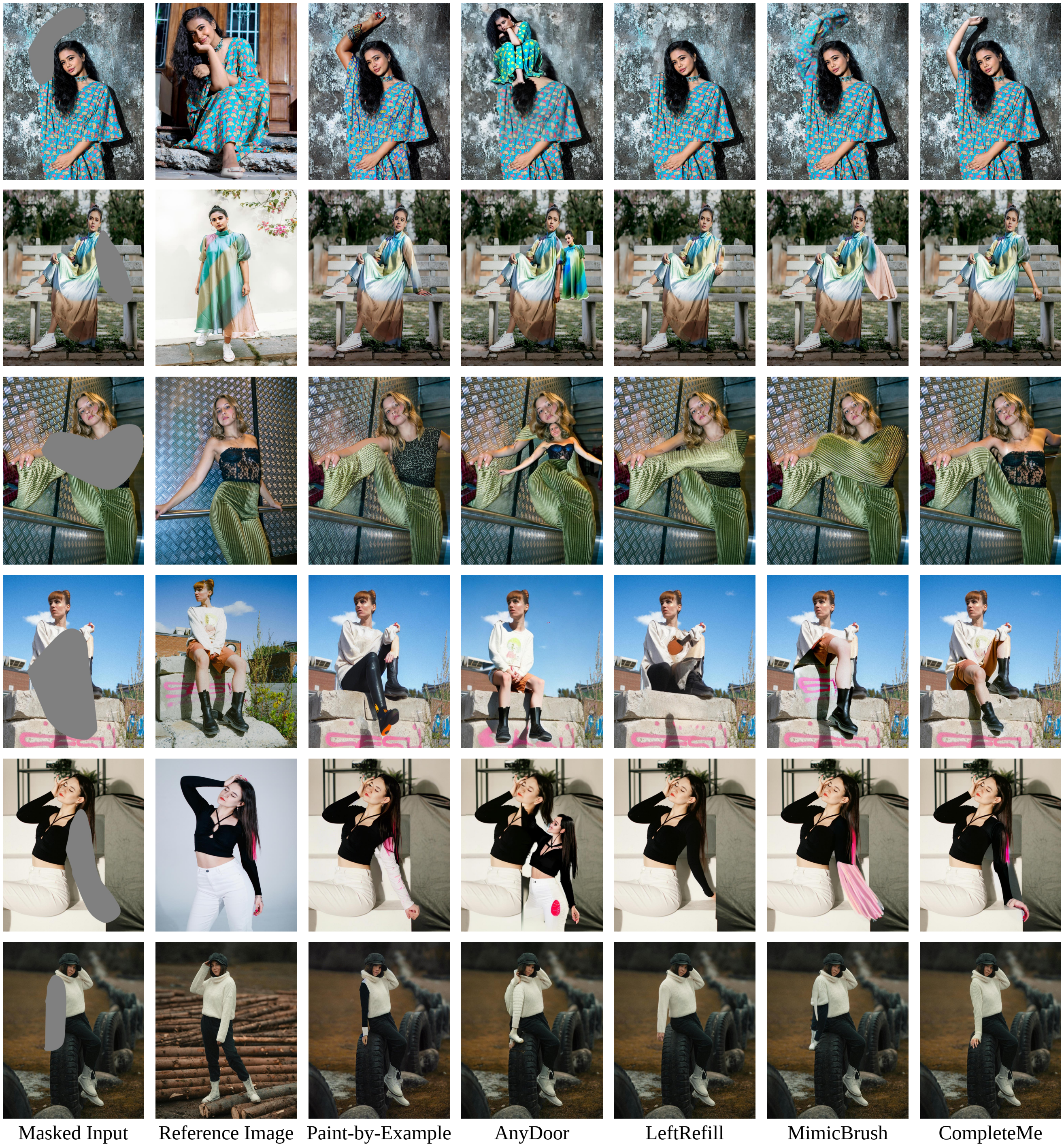


Figure 12. **Qualitative Comparison with Reference-based Methods on Our Benchmark (Sec. 3.4).** Our *CompleteMe* can generate more realistic and preserve identical information from the reference image. Please zoom in for a better comparison.

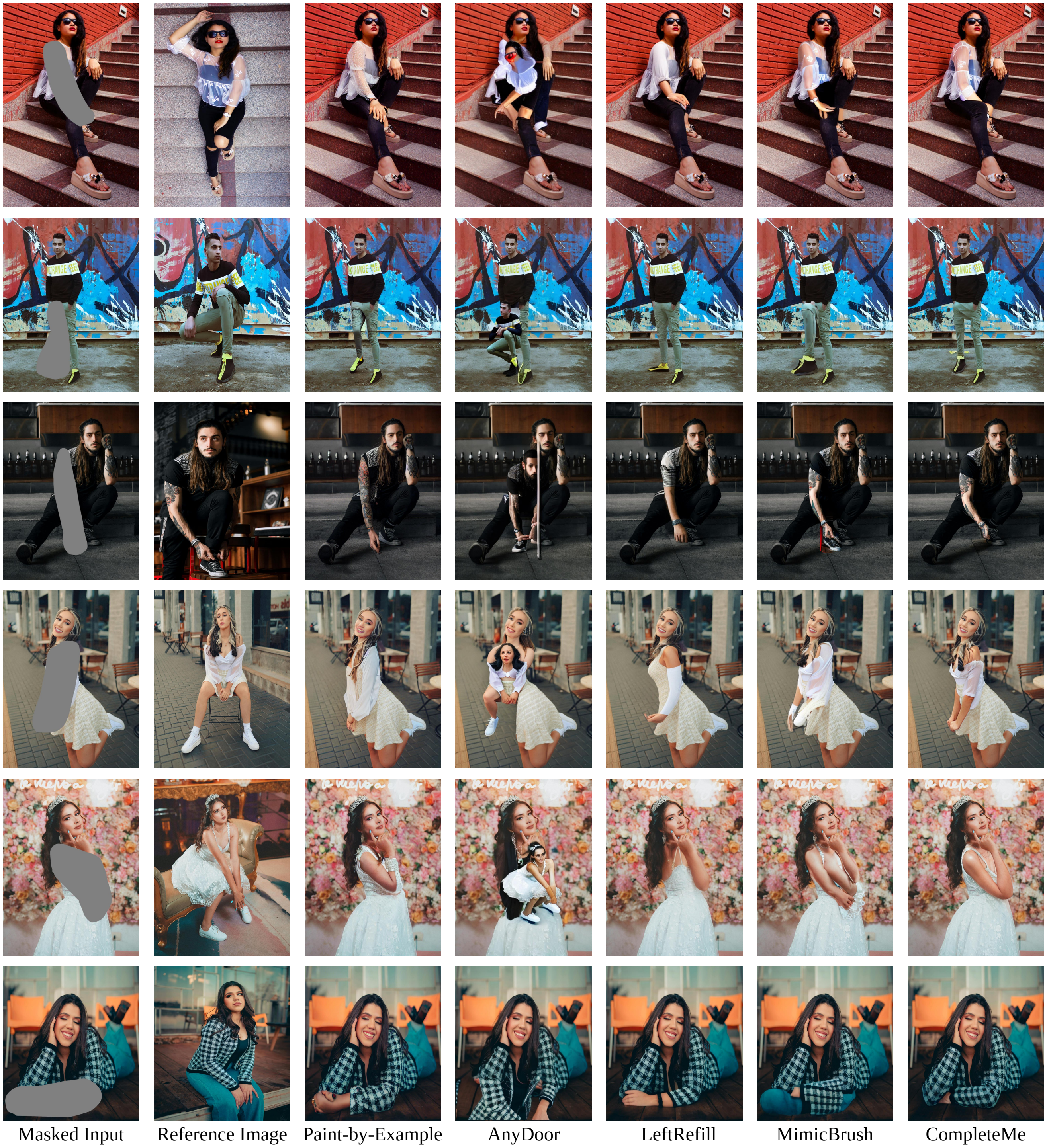


Figure 13. **Qualitative Comparison with Reference-based Methods on Our Benchmark (Sec. 3.4).** Our *CompleteMe* can generate more realistic and preserve identical information from the reference image. Please zoom in for a better comparison.

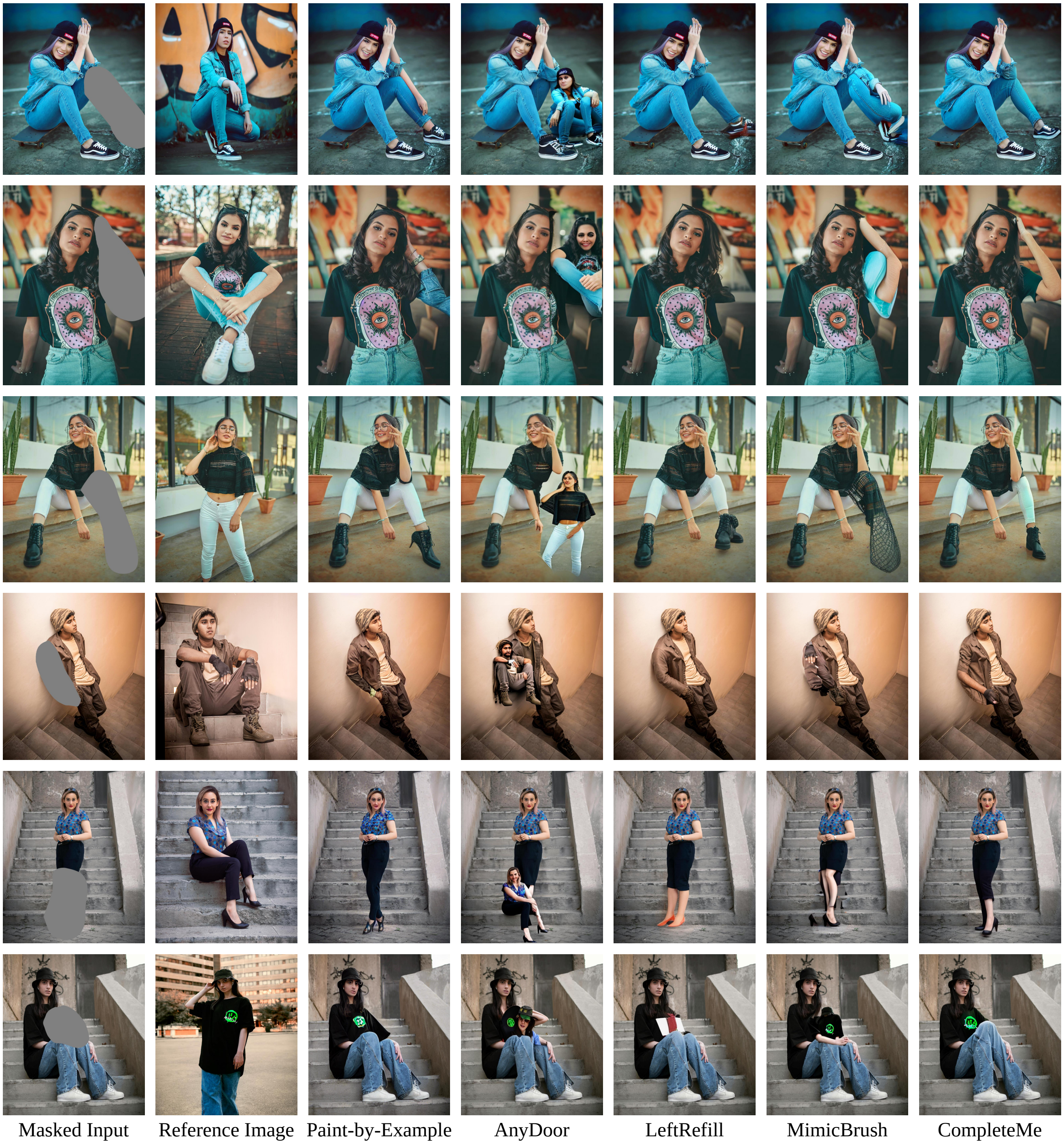


Figure 14. **Qualitative Comparison with Reference-based Methods on Our Benchmark (Sec. 3.4).** Our *CompleteMe* can generate more realistic and preserve identical information from the reference image. Please zoom in for a better comparison.

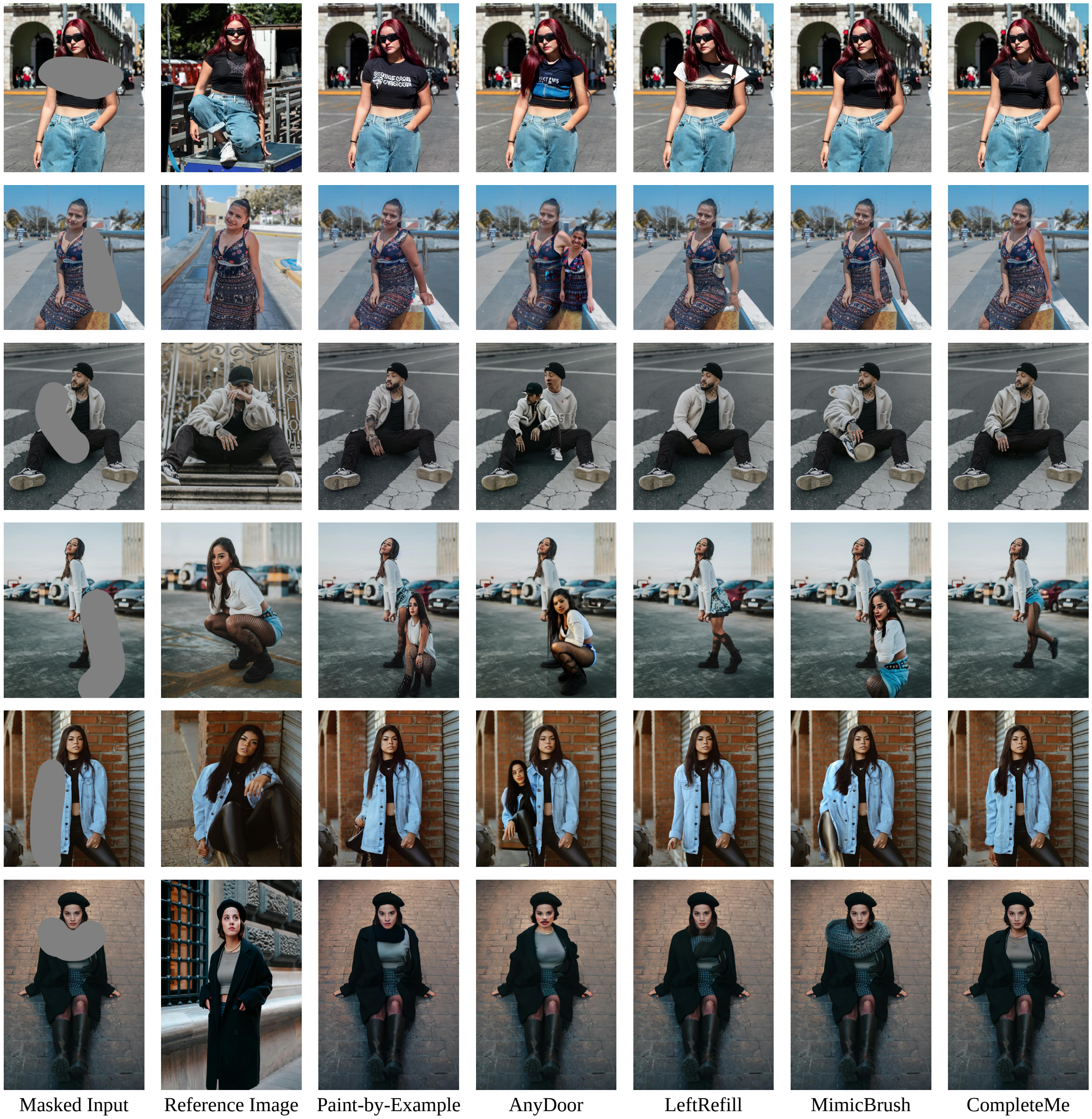


Figure 15. **Qualitative Comparison with Reference-based Methods on Our Benchmark (Sec. 3.4).** Our *CompleteMe* can generate more realistic and preserve identical information from the reference image. Please zoom in for a better comparison.

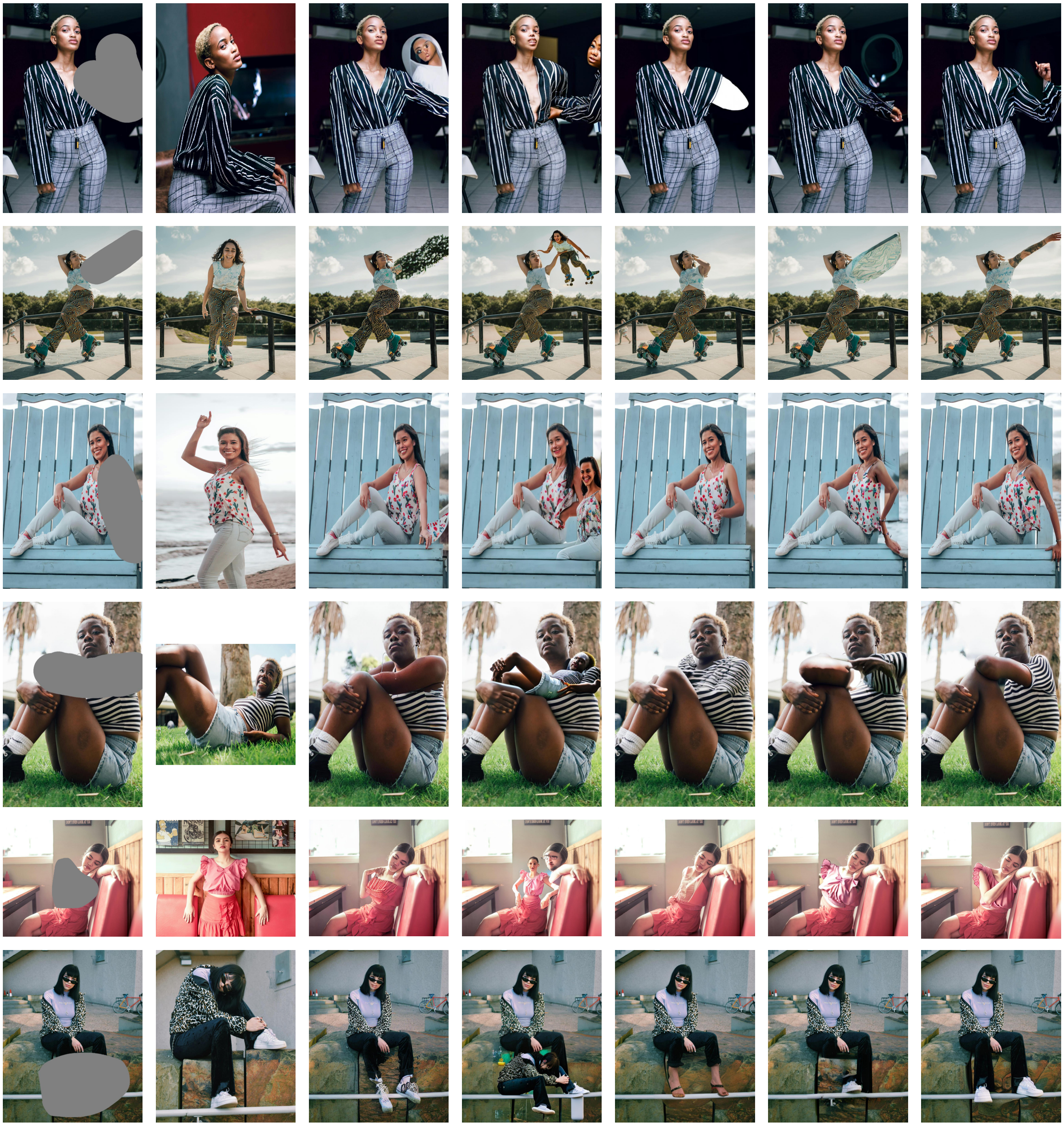


Figure 16. **Qualitative Comparison with Reference-based Methods on Our Benchmark (Sec. 3.4).** Our *CompleteMe* can generate more realistic and preserve identical information from the reference image. Please zoom in for a better comparison.

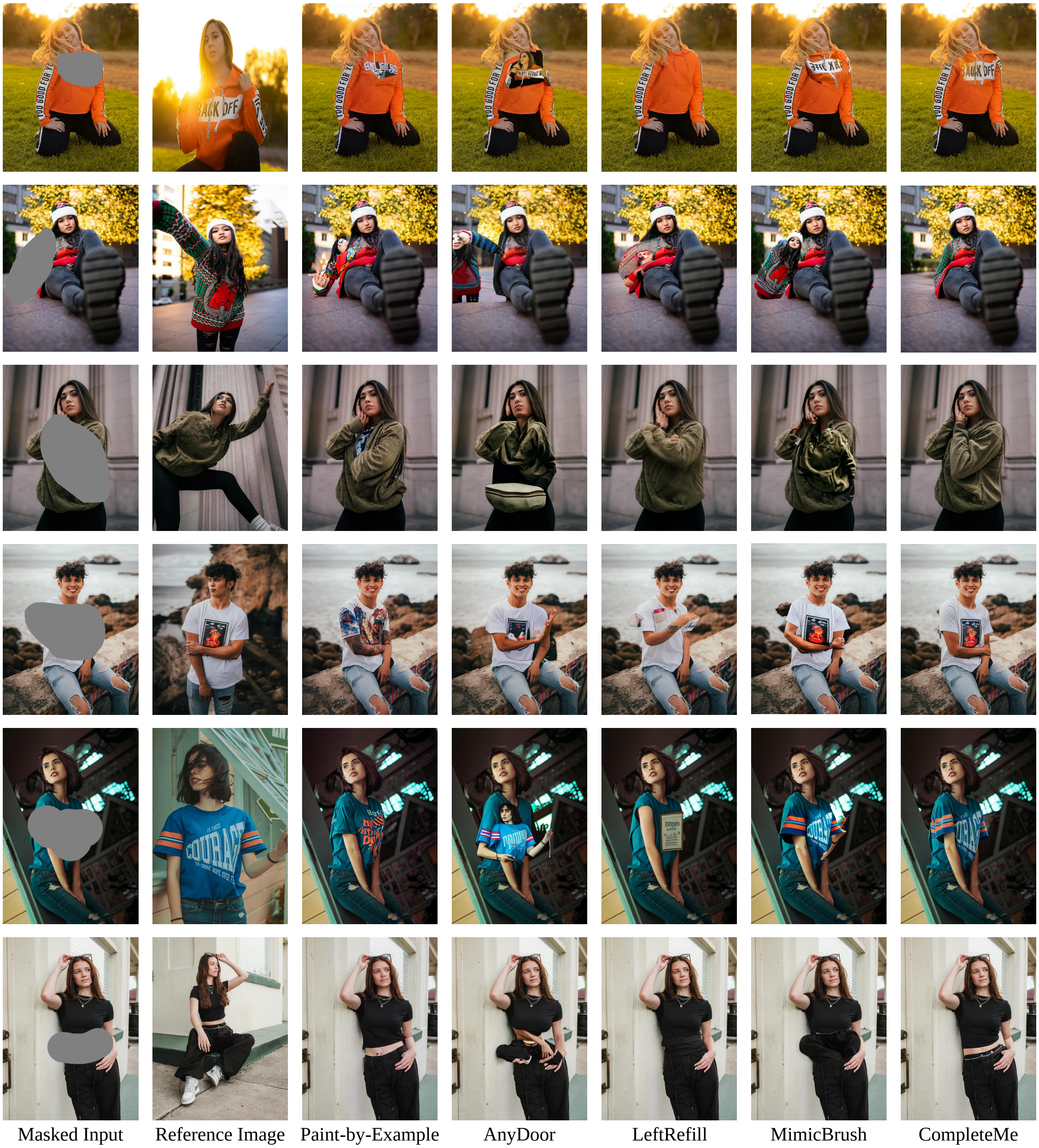


Figure 17. **Qualitative Comparison with Reference-based Methods on Our Benchmark (Sec. 3.4).** Our *CompleteMe* can generate more realistic and preserve identical information from the reference image. Please zoom in for a better comparison.

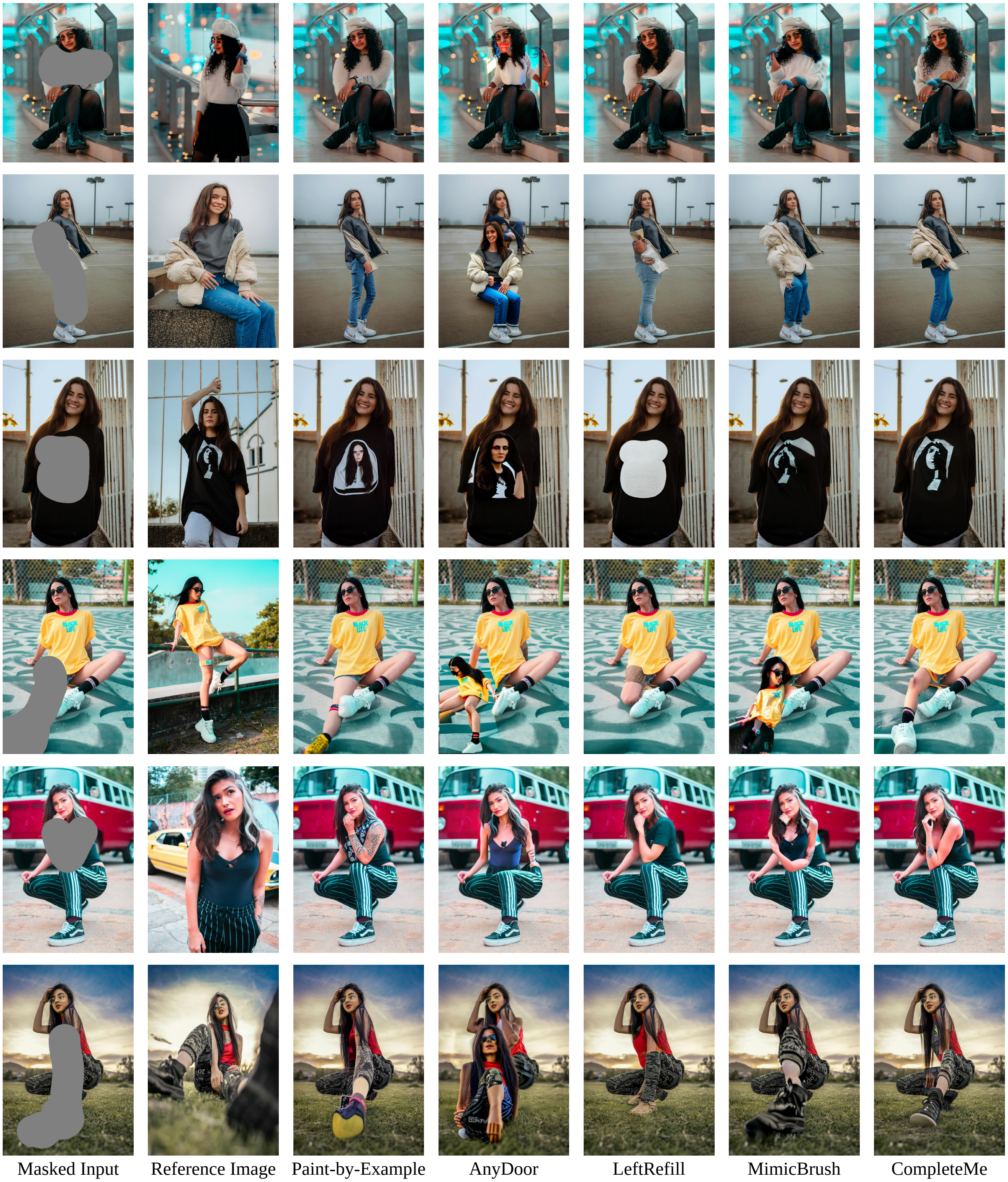
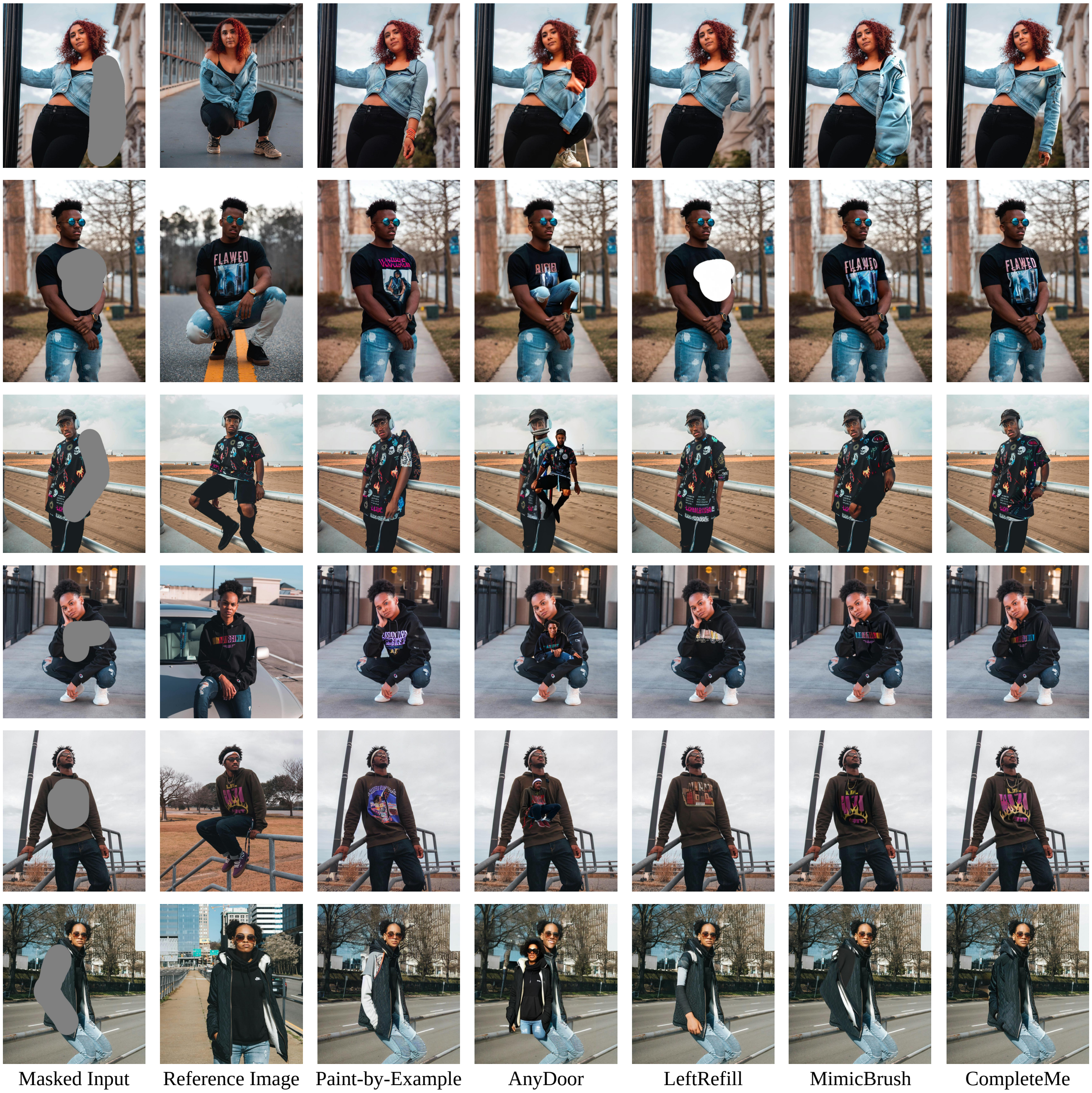


Figure 18. **Qualitative Comparison with Reference-based Methods on Our Benchmark (Sec. 3.4).** Our *CompleteMe* can generate more realistic and preserve identical information from the reference image. Please zoom in for a better comparison.



Masked Input Reference Image Paint-by-Example AnyDoor LeftRefill MimicBrush CompleteMe

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.

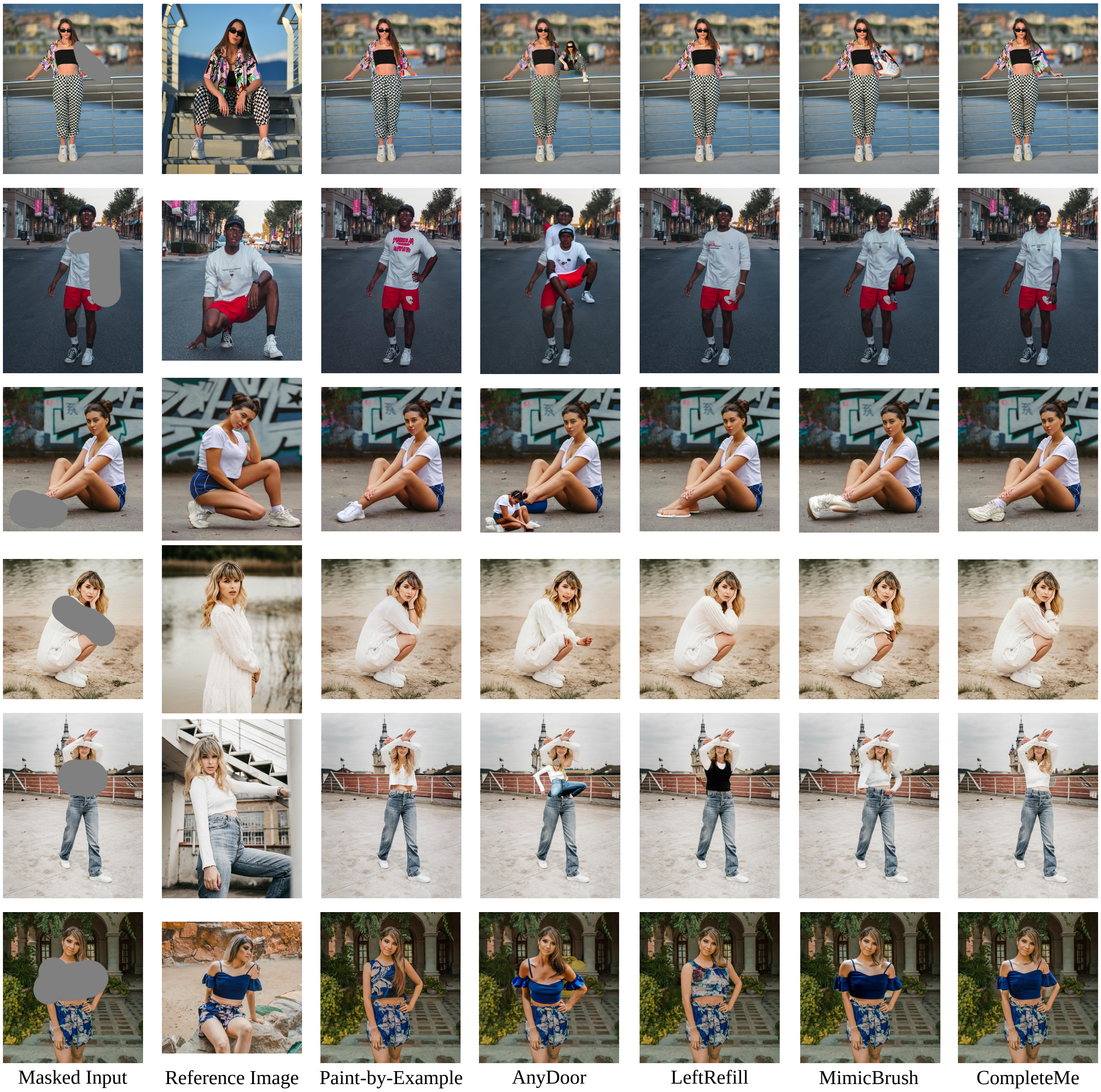


Figure 20. **Qualitative Comparison with Reference-based Methods on Our Benchmark (Sec. 3.4).** Our *CompleteMe* can generate more realistic and preserve identical information from the reference image. Please zoom in for a better comparison.

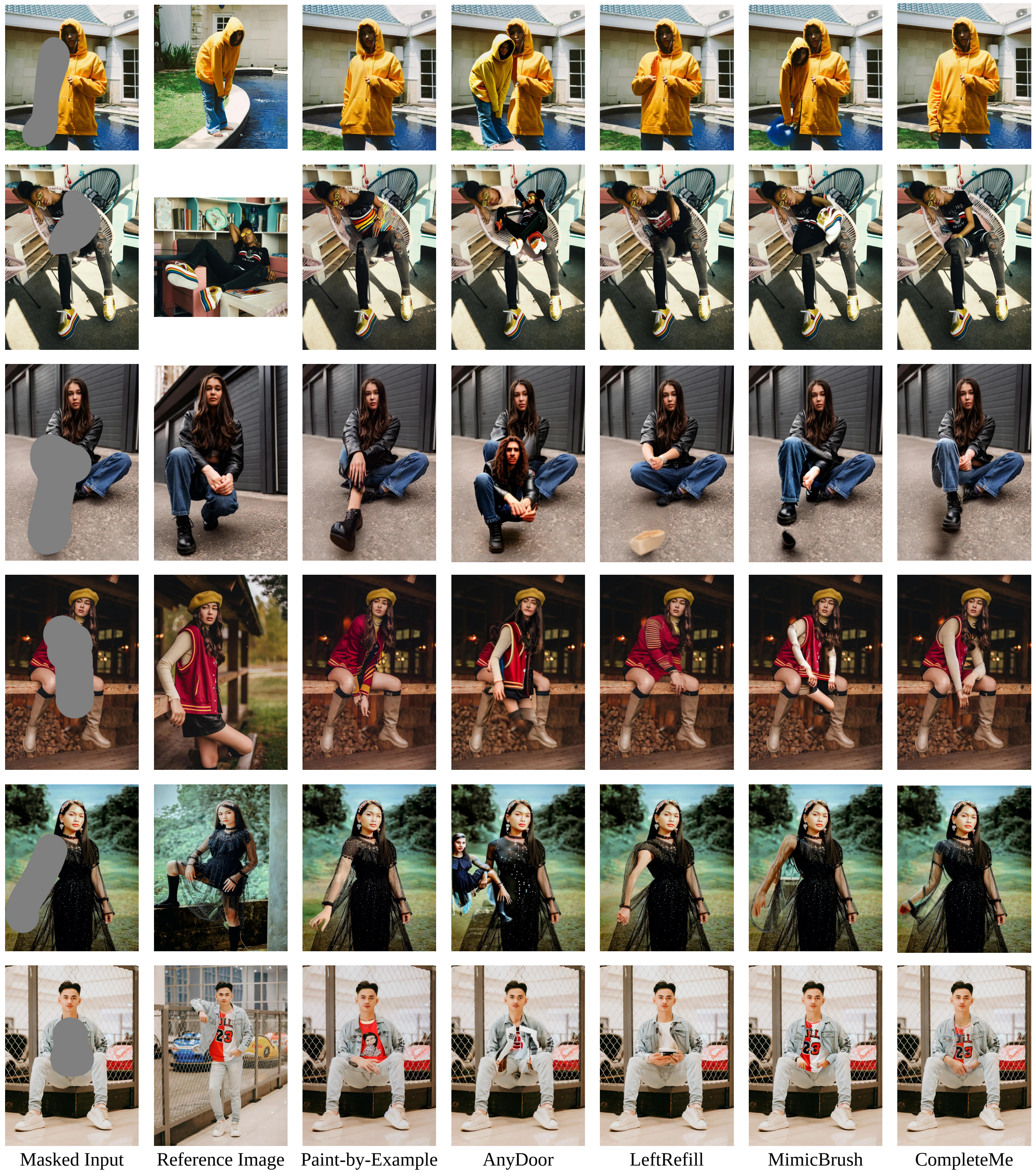


Figure 21. **Qualitative Comparison with Reference-based Methods on Our Benchmark (Sec. 3.4).** Our *CompleteMe* can generate more realistic and preserve identical information from the reference image. Please zoom in for a better comparison.