MultiCrafter: High-Fidelity Multi-Subject Generation via Disentangled Attention and Identity-Aware Preference Alignment

Supplementary Material

A. More Details for Identity-Preserving Preference Optimization

We provide a detailed algorithm execution flow for Identity-Preserving Preference Optimization Sec. 4.3, as shown in Algorithm 1. We conducted detailed ablation experiments on each reward model. As shown in Tab. S-1, individual rewards significantly improve specific capabilities. By combining multiple rewards, we balance the advantages of each reward and achieve improvements across all metrics compared to not using IPPO.

CLIP	HPS	Face Reward	CLIP-T	Face-Sim	DINO-I	CLIP-I	AES
✗	✗	✗	0.2674	0.5154	0.8107	0.848	0.2661
✓	✗	✗	**0.2793**	0.4745	0.7892	0.7967	0.2631
✗	✓	✗	0.2716	0.4979	0.8043	0.8144	**0.3037**
✗	✗	✓	0.2638	**0.5822**	**0.8362**	**0.8613**	0.2613
✓	✓	✓	0.2753	0.5284	0.8294	0.8524	0.2915

Table S-1. Ablation of each reward on Identity-Preserving Preference Optimization.

B. More Implementation Details.

In this section, we provide more details on the hyperparameter settings and specific training details. As mentioned in the main text, we decouple the training process into two parts, and the details of each phase are as follows:

Subject Fidelity Focused Pre-training. As described in Sec. 4.1, in this stage, we train the model to primarily ensure it can generate images with high subject fidelity. Following [3, 68], we adopt a progressive training approach, first pre-training on single-subject data, and then using this as a foundation to train on multi-subject data. *1. Single-Subject Pre-training.* We first pre-train the model for 40,000 steps on an internal single-subject dataset to equip it with foundational subject customization capabilities. In this stage, only $\mathcal{L}_{diff}$ is used for supervision. We use the AdamW optimizer with a 3×10^{-5} learning rate and a weight decay of 1×10^{-2}. We use 8 cards for training and set the batch size of each card to 6. *2. Multi-Subject Customization Training.* Then we train two models for customized human generation and object generation, respectively. For multi-human customized generation, we decrease the learning rate to 1×10^{-5}, set the loss weight $\lambda = 0.3$, and introduce the attention loss $\mathcal{L}_{attn}$. This stage runs for 25,000 steps. We use 8 cards at this stage and set the batch size to 4. For multi-object customized generation, since the size of the dataset is smaller than multi-human datasets, we only train for 15,000 steps.

Multi-Dimensional Preference Alignment Post-training. As described in Sec. 4.1, after completing the Subject Fidelity Focused Pre-training, in the second stage, we primarily use post-training to align it with multi-dimensional human preferences. We fine-tune the model using our proposed Identity-Preserving Preference Optimization. Following [32], we configure with a sampling step of 16, a window size of $w = 2$, a shift interval of $\tau = 50$, and a window stride of $s = 1$. This stage consists of 300 steps. For multi-human customization, the reward weights are set to $w_{id} = 0.5$, $w_{text} = 1.4$, and $w_{aes} = 0.7$. For multi-object customization, we adjust the subject fidelity weight to $w_{id} = 1.0$, while keeping $w_{text} = 1.4$ and $w_{aes} = 0.7$. In this stage, we used 16 cards for training and set the batch size of each card to 1.

C. Training Dataset Construction Pipeline.

Due to the scarcity of public multi-human customization datasets with adequate annotations, we designed a comprehensive and automated data preparation process to extract training data from OpenHumanVid [31] video datasets, as shown in Fig. S-1. This pipeline processes raw video clips to generate structured training samples, each containing a target image with multiple subjects, corresponding identity reference images, segmentation masks, and a detailed textual description. The entire workflow ensures subject fidelity, high image quality, and rich annotation. Specifically, we sample video clips featuring two individuals from a large database. The process begins with frame selection and subject localization. For each input video, we sample an initial frame as the source for reference images and a middle frame as the target scene. We first employ a YOLO-Pose [48] to obtain initial bounding boxes and keypoint information for each person. Following localization, we leverage the Segment Anything Model (SAM) [55] to generate high-fidelity segmentation masks for each individual, effectively isolating them from the background. To refine this output, only the largest connected component of the mask is retained. A critical subsequent step is ensuring subject fidelity across frames. We apply a face detection model [7] to the segmented portraits to locate facial regions, and then use a face recognition model [7] to extract a normalized feature embedding for each face. By computing the cosine similarity between embeddings from the reference and target frames, we enforce a stringent threshold to discard pairs where identity cannot be confidently verified. Once a pair of frames

Algorithm 1 Identity-Preserving Preference Optimization Training Process

Require: initial policy model π_θ; composite reward model R; prompt dataset $\mathcal{C}$; reference subjects dataset $\mathcal{Z}_{\text{data}}$; total sampling steps T; number of samples per prompt N; sliding window $W(l)$, window size w, shift interval τ, window stride s

Init left boundary of $W(l)$: $l \leftarrow 0$
for training iteration $m = 1$ **to** M **do**
  Sample batch prompts $\mathcal{C}_b \sim \mathcal{C}$ and corresponding subjects $\mathcal{Z}_b \sim \mathcal{Z}_{\text{data}}$
  Update old policy model: $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
  for each prompt $c \in \mathcal{C}_b$ and subject $\mathcal{Z} \in \mathcal{Z}_b$ **do**
    Init the same noise $s_0 \sim \mathcal{N}(0, \mathbf{I})$
    for generate i-th image from $i = 1$ **to** N **do**
      for sampling timestep $t = 0$ **to** $T - 1$ **do** ▷ $\pi_{\theta_{\text{old}}}$ mixed sampling loop
        if $t \in W(l)$ **then**
          Use SDE Sampling to get s^i_{t+1}
        else
          Use ODE Sampling to get s^i_{t+1}
        end if
      end for
    end for
    Calculate advantage: $A_i \leftarrow \frac{R(s^i_T, c, \mathcal{Z}) - \text{mean}(\{R(s^j_T, c, \mathcal{Z})\}_{j=1}^N)}{\text{std}(\{R(s^j_T, c, \mathcal{Z})\}_{j=1}^N)}$
    for optimization timestep $t \in W(l)$ **do** ▷ optimize policy model π_θ
      Update policy model via gradient ascent: $\theta \leftarrow \theta + \eta \nabla_\theta \mathcal{J}_{GSPO}$
    end for
  end for
  if $m \bmod \tau = 0$ **then** ▷ move sliding window
    $l \leftarrow \min(l + s,\ T - w)$
  end if
end for

Table S-2. Analysis of Data Quality and Preference Alignment. We compare the original UNO with a version retrained on our dataset (UNO†). Training with UNO on our data also caused the aesthetic metrics to drop to levels similar to those before we used IPPO.

Methods	CLIP-T ↑	Face-Sim ↑	DINO-I ↑	CLIP-I ↑	AES ↑
UNO (Official)	0.2645	0.1474	0.5972	0.6489	**0.2954**
UNO (Retrained)†	0.2658	0.2879	0.7173	0.7528	0.2656
Ours (w/o IPPO)	0.2674	0.5154	0.8107	0.8480	0.2661
Ours (Full)	**0.2753**	**0.5284**	**0.8294**	**0.8524**	0.2915

passes this verification, the pipeline generates the final training sample. The segmented portraits and cropped faces from the initial frame are saved as the reference =images. The target frame is cropped into a square as target image, with its corresponding body and face masks preserved. Finally, a powerful vision-language model, Qwen2.5-VL [1], is prompted with the reference images and the target image to produce a rich text prompt of the entire scene, ensuring descriptive consistency for each subject.

D. Impact Analysis of Dataset Quality

To investigate whether the variance in aesthetic metrics between our method and certain baselines for multi-human generation task (e.g., UNO [68]) stems from algorithmic limitations or discrepancies in training data quality, we conducted a detailed statistical analysis and ablation study. It is important to note that many existing methods utilize private, high-quality datasets, whereas our model is trained on a subset curated from open-source video data, which inevitably contains frames with motion blur or lower aesthetic appeal.

Data Aesthetic Distribution. We first analyzed the aesthetic quality of our training dataset using the HPS v2 scoring model. As illustrated in Fig. S-3, the aesthetic scores of our training samples follow a normal distribution with a mean of 0.2552 and a median of 0.2556. This relatively low baseline suggests that the model's "upper bound" for aesthetics is naturally constrained by the training data when using standard supervised learning.

Baseline Retraining and Comparison. To verify this hypothesis, we conducted ablation study using UNO [68] as a baseline. We retrained the UNO model on our dataset (denoted as UNO†) and compared it with the official UNO release and our method. As shown in Tab. S-2, when UNO is retrained on our data, its aesthetic score (AES) drops significantly from 0.2954 (Official) to 0.2656. Crucially, this result (0.2656) is highly consistent with the performance of our method before the post-training stage (Ours w/o IPPO, 0.2661). This observation supports two key conclusions:

- The decline in aesthetic quality is primarily attributable

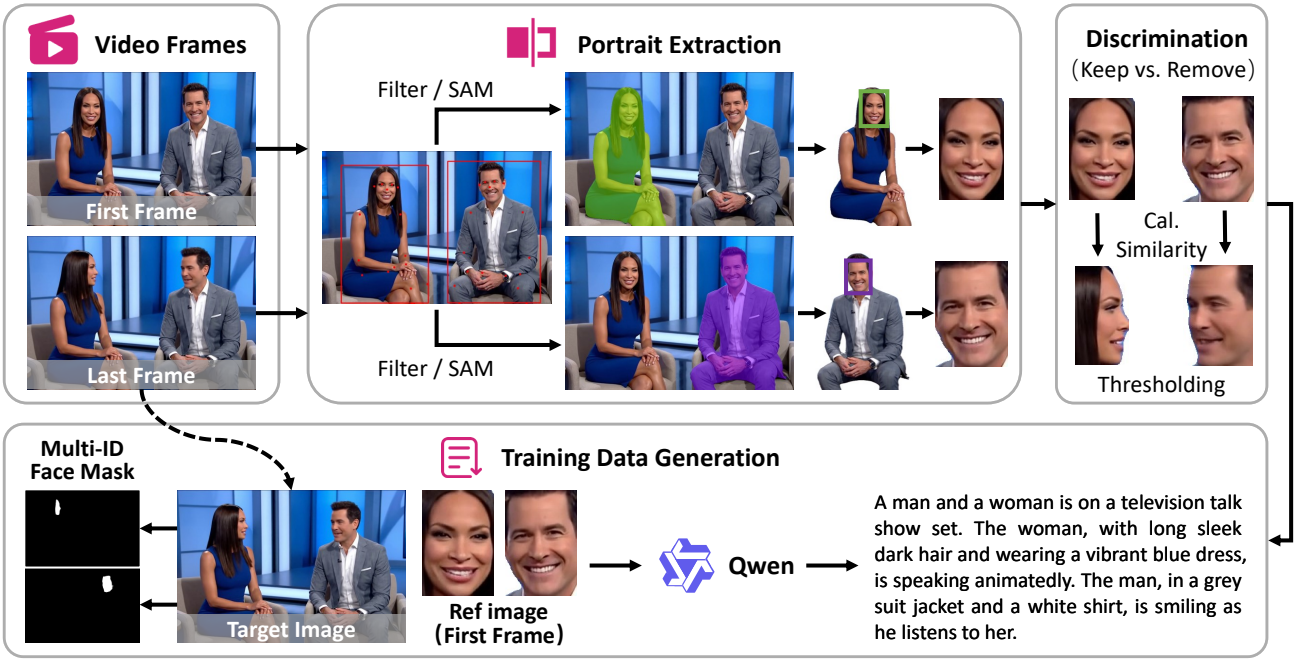

Figure S-1. Data processing pipeline for customized multi-human image generation.

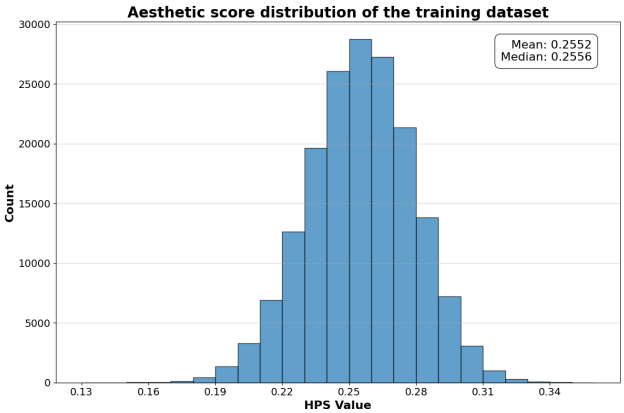

Figure S-2. Aesthetic Score Distribution of Training Data. The histogram illustrates the frequency of HPS v2 scores within our training dataset. The distribution is centered around a mean of 0.2552.

to the training data rather than the model architecture.

- Even with suboptimal data, our full method (Ours Full) successfully recovers the aesthetic quality to 0.2915 via the proposed Identity-Preserving Preference Optimization (IPPO).

This demonstrates that our two-stage decoupled framework effectively aligns the model with human aesthetic preferences, overcoming the limitations of the underlying training data quality.

E. More Details for Our Benchmark.

To advance research on high-fidelity multi-subject generation, we constructed a benchmark dataset by collecting images from publicly available sources and extracting the corresponding facial regions as reference images. The dataset comprises 80 celebrities and 80 non-celebrities, covering diverse attributes in terms of gender, age, and ethnicity (male/female; young/elderly; Caucasian, Black, and Asian). We used this face collection as a reference pool and paired faces within it. For each pair, we employed Qwen2.5-VL to generate distinctive natural-language prompts to provide diverse textual descriptions. Representative samples are shown in Fig. S-3.

F. Limitation.

Although MultiCrafter has achieved excellent performance in multi-subject driven image generation tasks, our work still has certain limitations, which also point the way for future research.

First, the scale and quality of the training data are the primary limiting factors. We discussed in detail the limitations imposed by the quality of our data in Sec. D. Currently, high-quality, publicly available datasets for multi-subject driven generation remain scarce [3, 68]. Although we have designed a complete automated data processing pipeline to extract training samples from videos, our dataset is still limited in scale and diversity due to the quantity and quality of open-source video data.

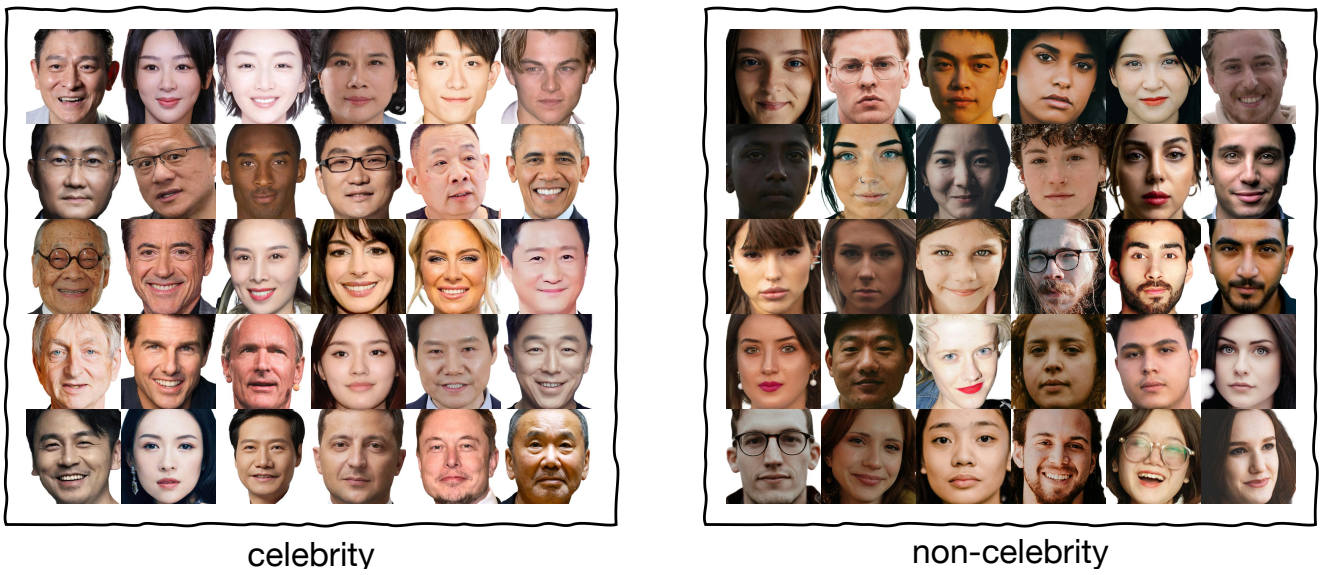

Figure S-3. Visualization for part of our multi-human evaluation benchmarks.

Second, the effectiveness of our method has been primarily validated in two-subject scenarios. Since our multi-person dataset and the public MUSAR dataset [19] mainly contain two subjects, the experiments in this paper were centered around this setting. Although we have supplemented some results from a small number of customized generation experiments with 3 person in Sec. I, the model's generation capabilities when dealing with more objects have not yet been fully validated. It is worth noting that our framework was designed with scalability in mind; both the attention regularization mechanism and the Multi-ID Alignment Reward (based on the Hungarian algorithm) in the reinforcement learning framework can be directly extended to scenarios with more subjects.

For future work, we plan to explore improvements from both data and model perspectives. On one hand, we will attempt to construct larger, higher-quality datasets containing a more diverse number of subjects by combining synthetic data with image editing [66] techniques. On the other hand, we will train and evaluate the model in scenarios with more subjects to further enhance the generalization and robustness of MultiCrafter, enabling it to handle more complex personalized image generation.

G. User Study

To complement our quantitative analysis and assess human subjective preference, we conducted a user study comparing MultiCrafter against five state-of-the-art baselines: UNO [68], OmniGen [72], OmniGen2 [67], DreamO [51], and XVerse [3]. We evaluated the generated results across five key dimensions: text alignment, subject fidelity, aesthetics, realism, and overall preference. As illustrated in Fig. S-4, our method demonstrates a commanding lead in subject fidelity, significantly outperforming the second-best method. This result strongly validates the effectiveness of our Identity-Disentangled Attention Regularization in resolving attribute leakage and preserving intricate identity details. Furthermore, MultiCrafter achieves the highest scores in realism and overall quality, while maintaining a top-tier performance in text alignment and competitive aesthetics. These findings confirm that our decoupled training framework successfully resolves the trade-off between fidelity and preference, producing images that are not only faithful to the user-provided subjects but also aesthetically pleasing and semantically accurate. Furthermore, we introduced Gemini-3-Pro and GPT-5.1 to score the generated images, and as shown in the figure, our method shows a significant improvement over existing methods.

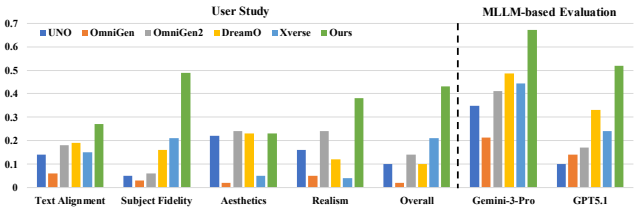

Figure S-4. User Study. Our method achieved the best results in both subject fidelity and overall quality, demonstrating its effectiveness.

H. More Discussion about MoE-LoRA.

To further validate the effectiveness of our MoE-LoRA architecture in expanding model capacity and handling di-

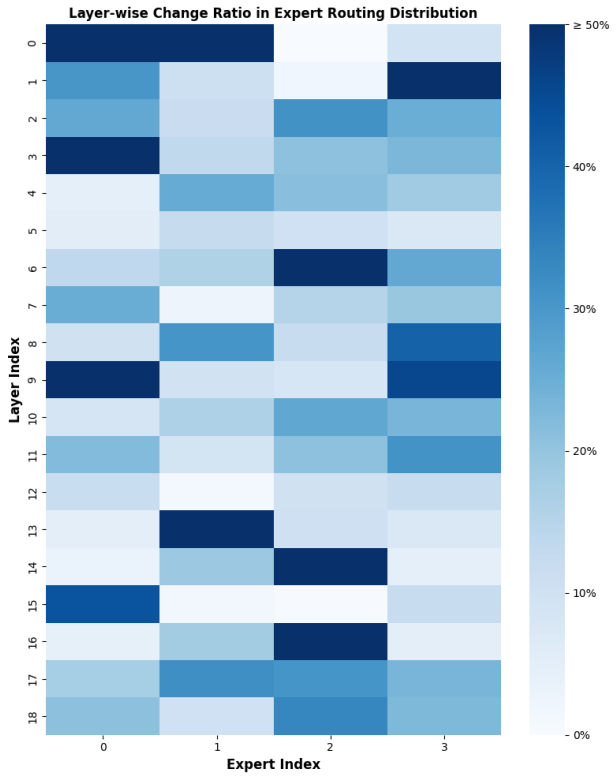

Figure S-5. Layer-wise Change Ratio in Expert Routing Distribution. The heatmap illustrates the difference in expert activation between "Left-Right" and "Top-Down" layouts.

verse spatial layouts, we analyzed the expert routing behavior under distinct spatial instructions. If the MoE mechanism truly functions as intended, the routing network should dynamically allocate different experts to handle different spatial layouts. We designed a controlled experiment using two distinct spatial prompts: a 'side-by-side' (left-right) layout and a 'top-down' (up-down) layout. To isolate the impact of spatial configuration, we kept all other variables constant, including the subject identities, background descriptions, and random seeds. We conducted inference across 4 different seeds and calculated the average change ratio of the expert routing distribution within the Double Blocks. The change ratio for an expert e at layer l is defined as $|P_A^{(l,e)} - P_B^{(l,e)}|/(P_A^{(l,e)} + \epsilon)$, where P_A and P_B represent the routing probabilities under the two different layouts.

The results are visualized in Fig. S-5. As shown in the heatmap, significant variations in expert activation (indicated by deep blue regions, where the change ratio exceeds $40\% - 50\%$) are observed across many layers. This phenomenon proves that the routing network effectively perceives the change in spatial prompts. The distinct activation patterns confirm that the model utilizes different experts to construct different layouts. This experiment strongly supports our claim that MoE-LoRA successfully decouples the learning of diverse spatial scenarios. By assigning specialized experts to different spatial layouts, the framework avoids the "attribute averaging" issue inherent in single-LoRA approaches, thereby achieving high fidelity across complex and varied compositions. In addition, we have added more visualizations of the results for multi-human and multi-subject generation. As shown in Fig. S-6, after adding MoE-LoRA, many scenes that could not be generated before adding MoE-LoRA can be generated accurately.

Besides, we ablate the number of experts of MoE-LoRA. Tab. S-3 shows that increasing the number of experts improves text alignment across diverse layouts while maintaining subject fidelity.

Expert Nums	CLIP-T	Face-Sim	DINO-I	CLIP-I	AES
1	0.2637	0.4983	0.7953	0.8032	0.2653
2	0.2646	0.5016	0.8021	0.8137	0.2657
4	0.2674	0.5154	0.8107	0.8480	0.2661

Table S-3. Ablation study of the expert number of MoE-LoRA.

I. More Results of Multi-Human Generation.

To further demonstrate our method's performance in multi-human personalization, we present qualitative comparisons in Fig. S-7 and Fig. S-8. The results show that our model effectively preserves the identity of each subject and avoids the "attribute leakage" common in other methods. This outcome validates the efficacy of our Identity-Disentangled Attention Regularization (IDAR). While some baselines produce more stylized outputs that may yield higher HPSv2 scores, this is often at the expense of subject fidelity. Our method prioritizes photorealism and faithful subject appearance consistency, which leads to more reliable results in multi-subject customization. Furthermore, we extended our method to scenarios involving more subjects to validate the effectiveness of our framework. We curated 10,000 samples of three-person customized generation data from video data to train with our method. As shown in Fig. S-9, our method continues to significantly improve subject fidelity compared to existing methods, while maintaining text alignment and image aesthetics, effectively aligning with multi-dimensional human preferences. This demonstrates the effectiveness of our framework.

J. More Results of Multi-Object Generation.

To evaluate the generalization of our framework, we showcase multi-object customization comparisons in Fig. S-10. Our method demonstrates high object fidelity, accurately preserving core visual attributes such as a toy's texture or a glass's geometry. In contrast, competing approaches often introduce artifacts like deformation and detail loss. This

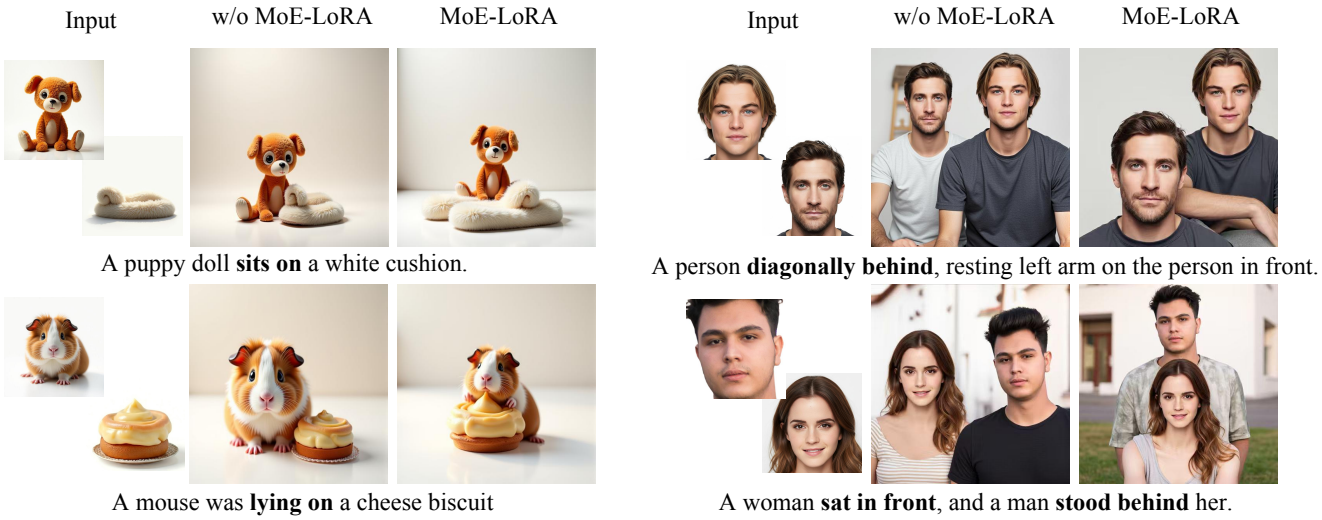

Figure S-6. More Visualization of the effectiveness of MoE-LoRA.

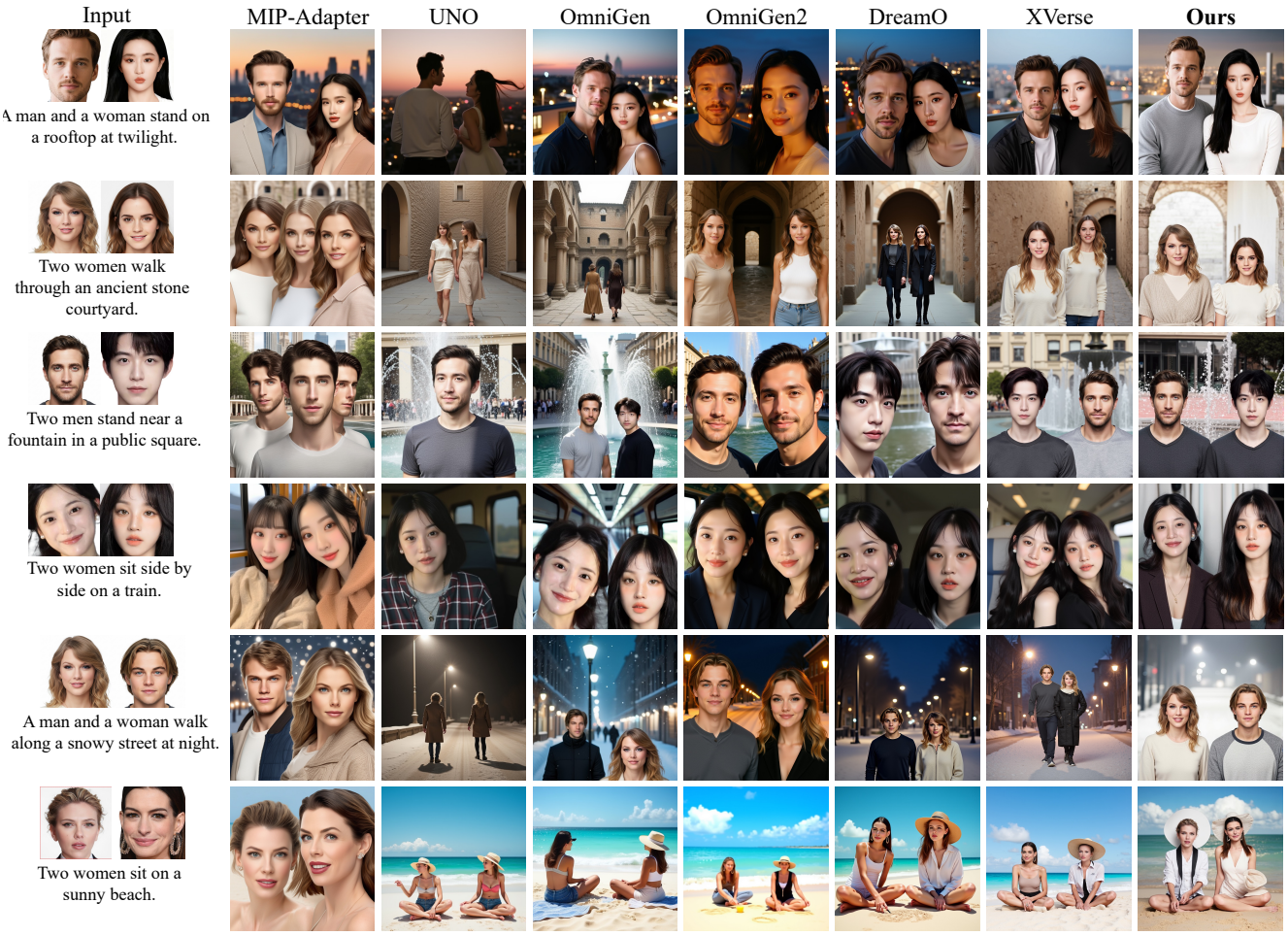

Figure S-7. More Visualization of our method in Multi-human Generation.

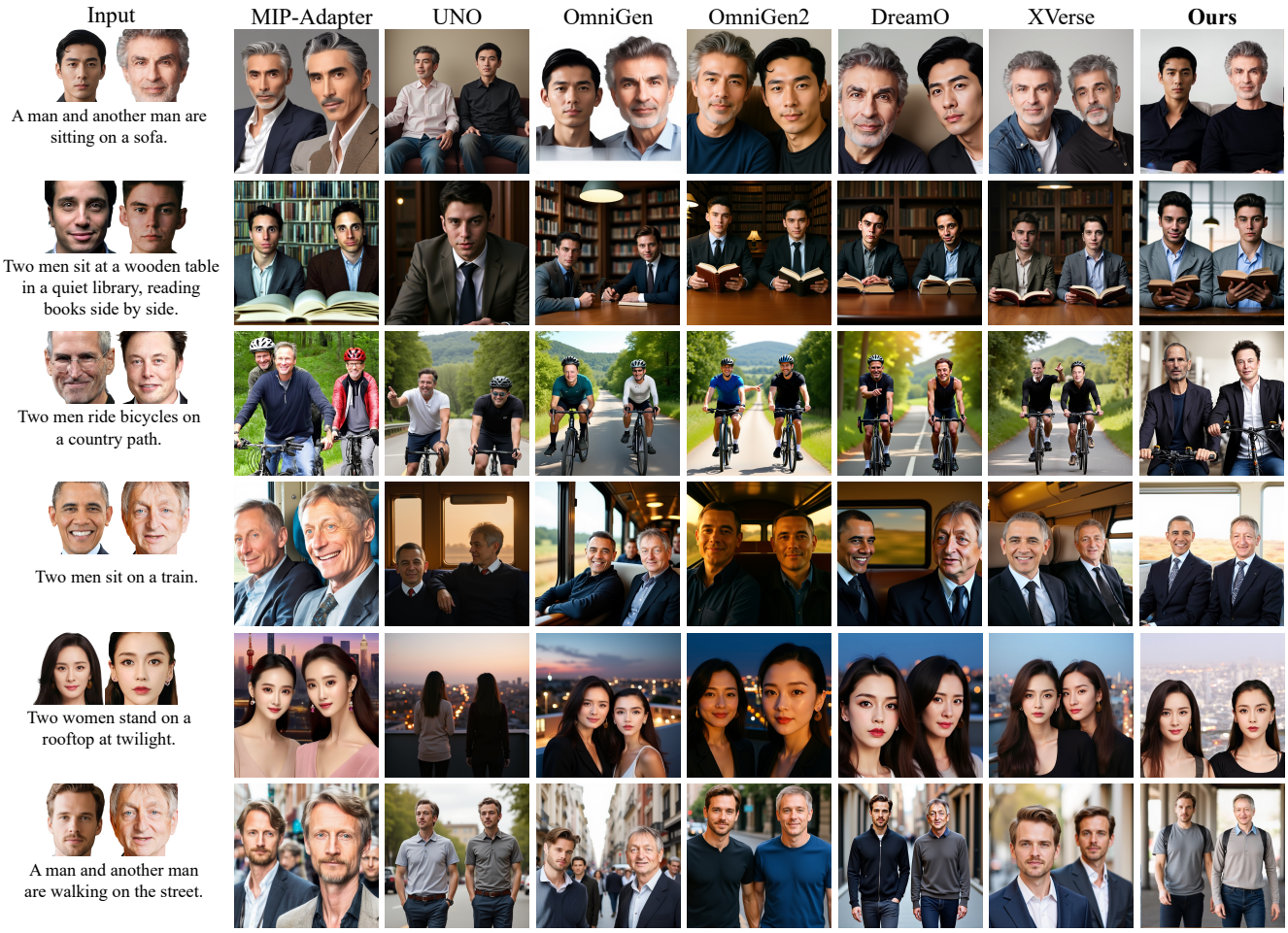

Figure S-8. More Visualization of our method in Multi-human Generation.

highlights our model's strength in precise subject representation rather than hyper-stylization, a crucial capability for practical applications that require accuracy.

K. More Results of Single-Subject Generation.

Effective multi-subject generation builds on strong single-subject performance. We validate this capability in Fig. S-11 and Fig. S-12, showing six diverse samples for each of four individuals and six frontal single-subject comparisons against SOTA models. Our method consistently preserves identity across varying styles, poses, and scenes, and even improves fidelity to the reference over baselines. These results confirm that the proposed framework not only enables reliable multi-subject generation but also enhances single-subject identity fidelity.

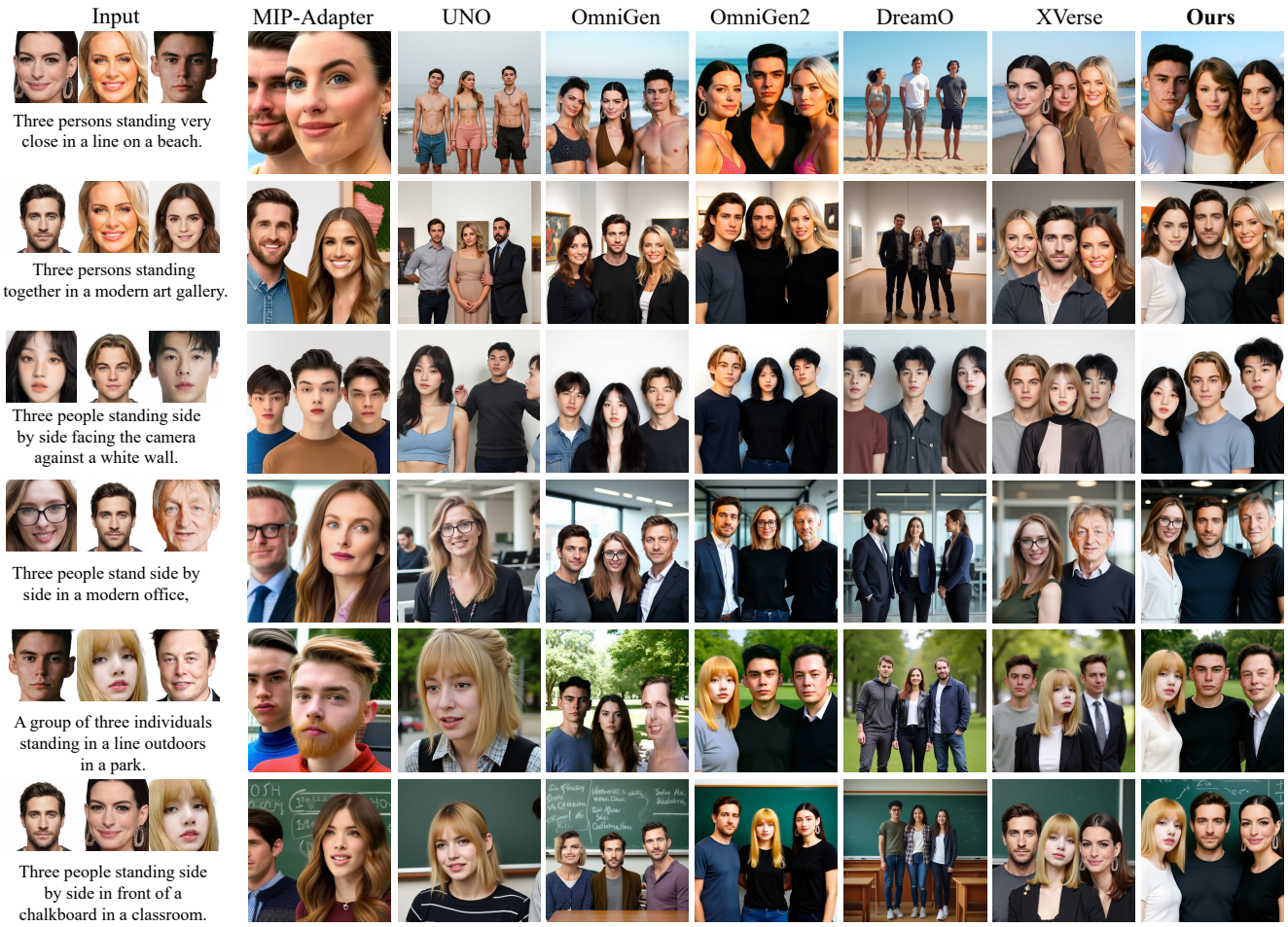

Figure S-9. More Visualization of our method in three-human Generation.

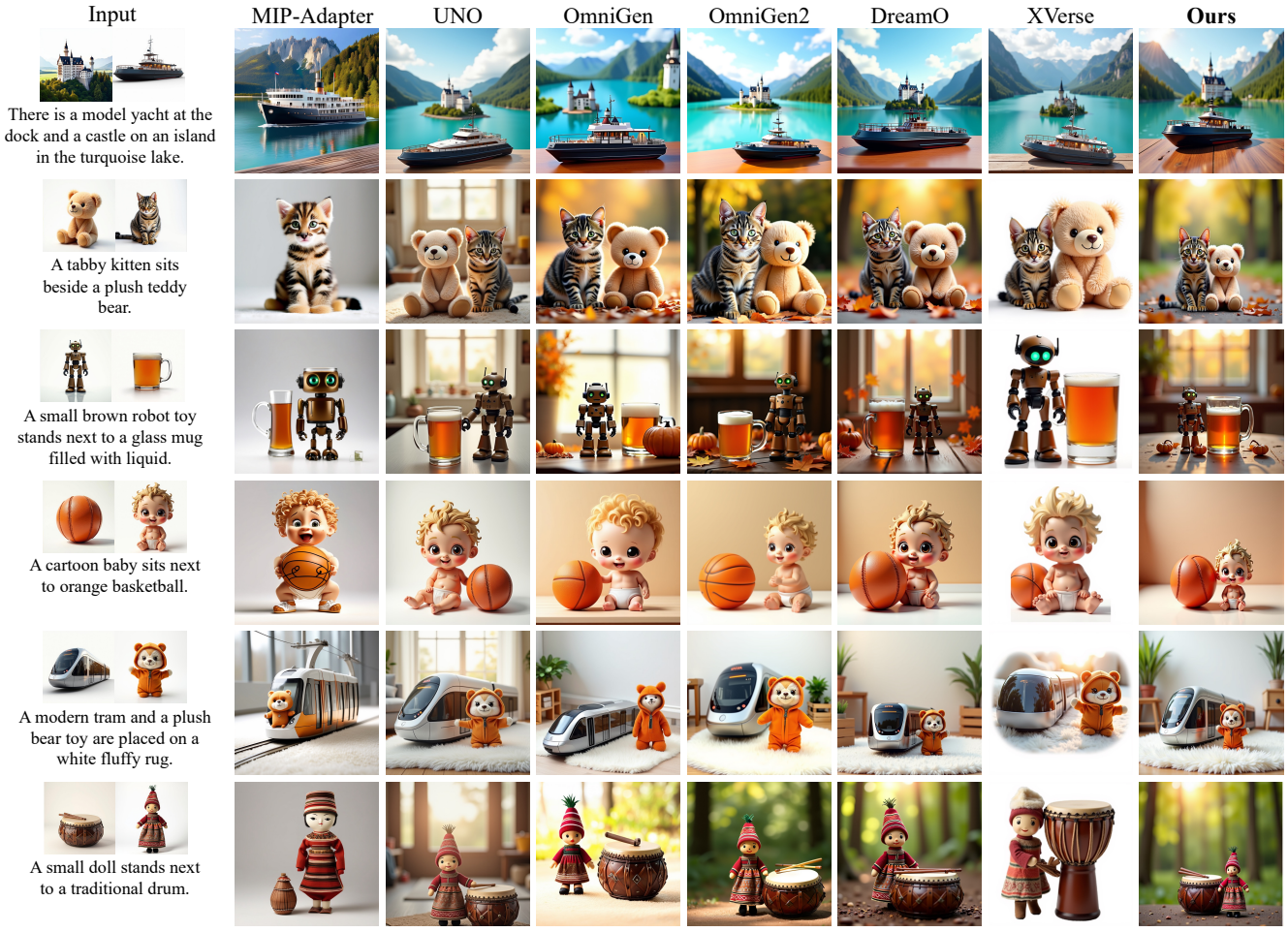

Figure S-10. Visualization of our method in Multi-object Generation.

(a) A young woman waiting at a modern, minimalist bus stop, looking towards the camera. (b) A portrait of a young woman sitting on a grassy coastal hill, looking at the camera. The bright sky and a sliver of the distant sea are visible in the background. (c) A modern, stylish portrait of a woman standing against a solid, textured concrete wall. She looks confidently at the camera. (d) A stunning portrait of a woman standing under a blooming cherry blossom tree, looking at the camera. (e) A friendly women in a simple apron, standing in front of a clean, minimalist cafe counter. He is smiling and looking at the camera as if about to take an order. Warm, inviting indoor lighting. professional portrait. (f) A clean, bright portrait of a young woman in a simple white linen shirt, walking on a white sand beach. She is looking at the camera with a soft smile.

(a) A women with a gentle smile, looking directly at the camera. She is holding a ceramic coffee mug, sitting by a large window with soft morning light. (b) A modern, stylish portrait of a woman standing against a solid, textured concrete wall. (c) Medium shot of a woman in a casual jacket, standing on a quiet European-style cobblestone street corner, looking at the camera. (d) A stunning portrait of a woman standing under a blooming cherry blossom tree, looking at the camera. (e) A clean, bright portrait of a young woman in a simple white linen shirt, walking on a white sand beach. (f) A gentle close-up of a woman holding a single white daisy, looking at the camera with a soft smile.

(a) A man in a fashionable coat, standing in the middle of a charming European cobblestone alley. (b) A man in a sharp, dark suit standing in front of a modern glass skyscraper. (c) A man in a casual shirt sitting on a weathered wooden bench by the sea at sunset. (d)A cheerful man in a vibrant, stylish tropical-print shirt, leaning against a palm tree on a sunny beach. (e) A man in a smart-casual look with a blazer and white trousers, sitting at an elegant outdoor lounge at a luxury beach resort. (f) A relaxed man in an open white linen shirt and shorts, walking along the shoreline.

(a) A man in a fashionable coat, standing in the middle of a charming European cobblestone alley. (b) A man in a sharp, dark suit standing in front of a modern glass skyscraper. (c) A man walking through a bright, airy modern art gallery with high ceilings and white walls. He pauses and turns to look at the camera. (d)A cheerful man in a vibrant, stylish tropical-print shirt, leaning against a palm tree on a sunny beach. (e) A stylish man standing on a bustling Tokyo street at night, looking directly at the camera. (f) A relaxed man in an open white linen shirt and shorts, walking along the shoreline.

Figure S-11. Visualization of our method in Single-subject generation.

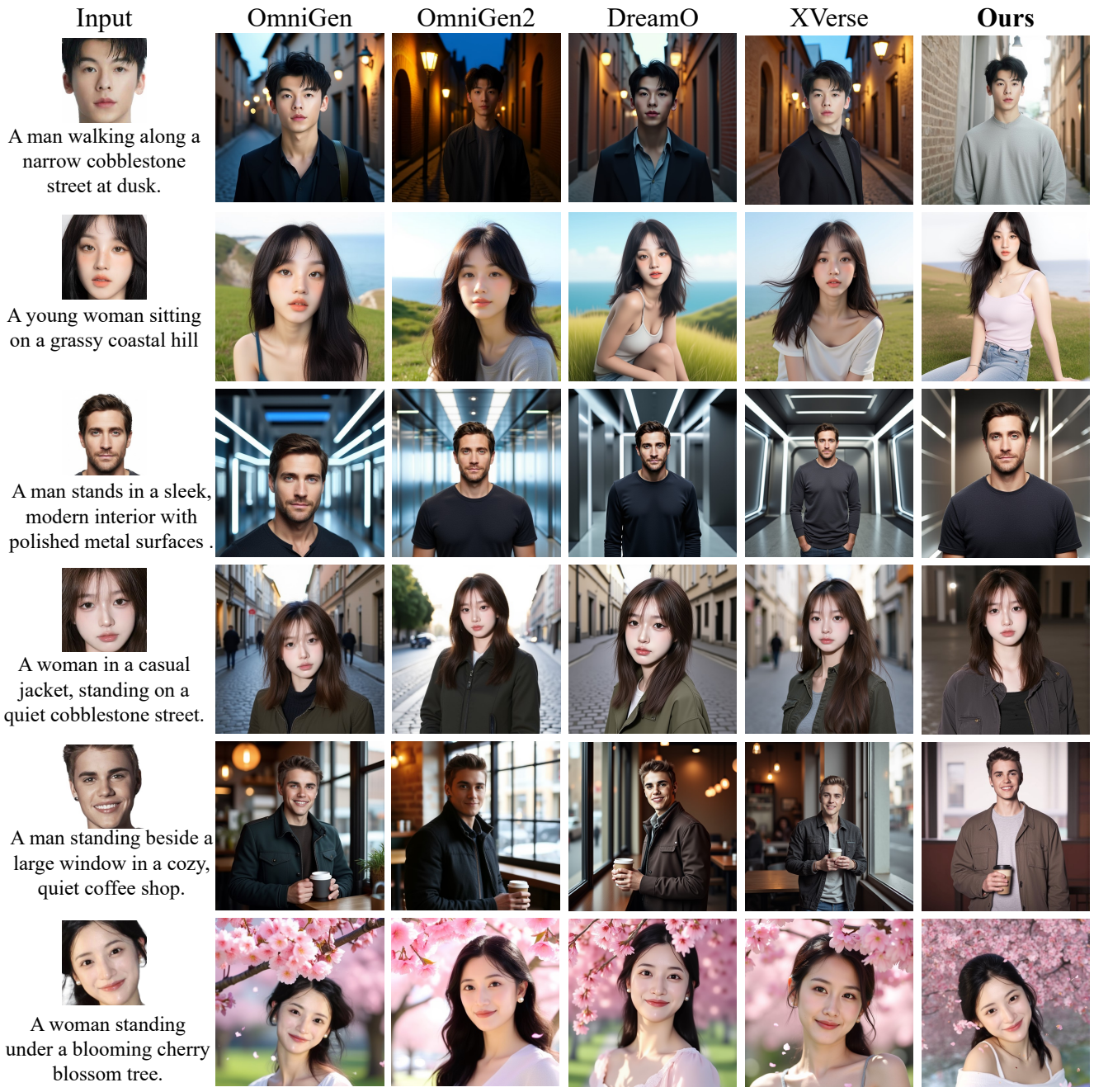

Figure S-12. Qualitative comparison with existing methods on the single human generation.