FairHuman: Boosting Hand and Face Quality in Human Image Generation with Minimum Potential Delay Fairness in Diffusion Models

Supplementary Material

1. Proofs

1.1. Derivation of Eq.(5)

Assuming there exists a constant M such that $\|\nabla l_i(\theta)\| \leq M$, we have:

$$\begin{aligned}\mathcal{F}(d) &= \sum_{i=1}^{3} \frac{1}{\|\text{proj}_{\nabla l_i(\theta)}(d)\|} \\ &= \sum_{i=1}^{3} \frac{1}{\nabla l_i(\theta)^\top d / \|d\|} \\ &= \sum_{i=1}^{3} \frac{\|d\|}{\nabla l_i(\theta)^\top d} \\ &\leq \sum_{i=1}^{3} \frac{M}{\nabla l_i(\theta)^\top d} \\ &= M \sum_{i=1}^{3} \frac{1}{\nabla l_i(\theta)^\top d}\end{aligned}$$

Since $\mathcal{F}'(d) = \sum_{i=1}^{3} \frac{1}{\nabla l_i(\theta)^\top d}$, we can get that:

$$\mathcal{F}(d) \leq M\mathcal{F}'(d)$$

Therefore, $\mathcal{F}'(d)$ is proportional to an upper bound of $\mathcal{F}(d)$, with the proportionality constant being M.

1.2. Derivation of Eq.(6)

Now, we have the following optimization objective:

$$\min_d \mathcal{F}'(d) = \sum_{i=1}^{3} \frac{1}{\nabla l_i(\theta)^\top d}\ ,\ \text{s.t. } \|d\|^2 \leq r^2 \qquad (1)$$

where $\nabla l_i(\theta)$ represents the gradient of each objective function $l_i(\theta)$ and d is the update direction for all objectives. r is the radius of the boundary space. Notably, the objective function above is non-decreasing for any feasible d. Therefore, for any d in the boundary space, there must exist a point in the same direction but on the boundary. For the maximization of utility, we can conclude that the optimal d^* must lie on the boundary, *i.e.*, $\|d^*\| = r$. To solve the extremum problem above, we apply Lagrange multiplier method in the following steps:

- **Construct Lagrange function.** By introducing the Lagrange multiplier λ, we can obtain the Lagrangian function:

$$\mathcal{L}(d, \lambda) = \sum_{i=1}^{3} \frac{1}{\nabla l_i(\theta)^\top d} + \lambda \left(\|d\|^2 - r^2\right) \qquad (2)$$

- **Take the derivative and set it equal to zero.** Take the partial derivatives of λ and d, set them equal to zero:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial d} &= -\sum_{i=1}^{3} \frac{\nabla l_i(\theta)}{(\nabla l_i(\theta)^\top d)^2} + 2\lambda d = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda} &= \|d\|^2 - r^2 = 0\end{aligned} \qquad (3)$$

- **Solve the equation.** From the first formula in Eq.(3), we can obtain:

$$d = \frac{1}{2\lambda} \sum_{i=1}^{3} \frac{\nabla l_i(\theta)}{(\nabla l_i(\theta)^\top d)^2} \qquad (4)$$

Here, λ is set to $\frac{1}{2}$ for simplicity following Nash-MTL, which has no impact on the final results. It indicates that the update direction d can be regarded as a weighted sum of gradient $\nabla l_i(\theta)$. The weight is allocated depending on the dot product of the gradient and the current update direction. Through this method, we can find the optimal update direction within the boundary space.

2. Dataset Details

2.1. Construction of Training Set

About the two public datasets we use, CosmicMan is specifically designed for generating highly realistic and photorealistic human images, containing 6 million high-resolution real-world human images and detailed descriptions of 115 million diverse attributes, which helps the model learn a wide range of human details and scene information. BodyHands focuses on human poses, particularly the torso and hands, providing additional information on the structure of the human body. To enable effective training, we utilize state-of-the-art detectors to filter low-resolution images (smaller than 1024×1024) or unrelated images (containing little information about humans). Furthermore, we apply visual language models (VLMs) to achieve accurate and concise image caption generation. For each sample annotation, in addition to the text description, we add extra positional information based on the detection results for our target optimization regions, namely the faces and hands, while also providing pose and depth information for the controllable generation.

2.2. Construction of Validation Set

To provide a more comprehensive evaluation of the model's performance in human image generation, we construct our validation set based on the large-scale human-centric image dataset proposed by MoLE, which mainly includes the following characteristics:

- **Brief and clear text prompts:** All text prompts are processed through VLMs and manual review to remove information unrelated to the image content, such as complex adjectives and clauses, retaining only the key information related to humans. For example: "A woman in dress standing in front of a building", "A man in a suit and tie standing with his hands on his hips". This ensures that the model does not generate content inconsistent with the target, thereby avoiding a negative impact on the final evaluation.
- **Diverse content and scenarios:** To cover human images in various scenarios and content as comprehensively as possible, we randomly sample 5k images and their corresponding text descriptions from the 23411 selected human-in-the-scene images for each batch testing. This includes different human races, genders, and scenarios.
- **High-quality ground truth images:** To ensure the quality of the image data, we filter the existing dataset based on resolution, clarity, and content relevance.

3. Implementation Details

We choose Stable Diffusion XL as the base model since it already possesses prior knowledge about human image generation through pre-training but has major limitations in generating local hand and face details. Regarding the generation of local masks for the faces and hands, we first generate mask images using the pre-annotated bounding box information from the dataset. Then, we apply the same preprocessing methods as used for the original image, including resizing, cropping, and random flipping. Our code is developed based on Diffusers. For the LoRA fine-tuning, we set the rank to 256 and train it for 50k steps with a learning rate at 1e-5. The Adam optimizer is deployed. Compared to the default settings, we mainly increase the rank size while reducing the learning rate, as we expect the model to learn detailed content better. For the ControlNet fine-tuning, we train it 60k steps with a learning rate set to 1e-5. It is worth mentioning that we apply a dropout probability of 0.3 to the input control conditions to enhance robustness and generalization. The overall image resolution is set as 1024×1024. Our resource-friendly training and evaluation processes can be implemented on a single 80G NVIDIA A800 GPU.

4. More Evaluation Details

4.1. More Quantitative Comparisons

In addition to comparing the image quality generated by the models, we also conduct experiments on memory usage and inference speed. Specifically, both our LoRA-based method and MoLE are deployed on SDXL as the backbone model. Therefore, here we primarily present the extra memory consumption and inference time required per image (set inference steps to 30 and employ UniPCMultistepScheduler as the sampler). The results are demonstrated in Tab.1. In particular, MoLE uses two low-rank modules and a gate network to achieve adaptive generation. Our method, on the other hand, achieves joint optimization of multiple objectives through a single low-rank module, thereby reducing memory requirements by nearly half and also significantly reducing the time required for inference.

Methods	Extra Memory Usage	Inference Time
MoLE	2611.75MB	×3
Ours	1413.12MB	×1.4

Table 1. Comparisons between MoLE and our method in terms of memory usage and inference speed.

4.2. User Study Details

The user study involves 50 participants to evaluate 100 pairs of images in total with corresponding annotations generated by different methods. Images with irrelevant content are pre-filtered and removed. In complying with the quantitative analysis, participants are asked to rate them according to the following two criteria, respectively:

- **Overall quality:** assessing the general appearance, realism, and coherence of the entire image.
- **Regional quality:** evaluating specific regions of interest (faces and hands) for detail, plausibility, and naturalness.

We note that all the participants are unaware of which image corresponds to which method and rank the images based on their preferences. For the highest rank in each group, we record its score as 1 and the rest as 0. In addition, for tied rankings, we assign a score of 0.5 to each. Finally, we separately calculate the scores for each method based on the two aforementioned criteria and visualize them using the bar chart, which provides a more comprehensive and intuitive reflection of the image generation quality.

4.3. More Ablation Studies

For our LoRA-based method, we conduct additional ablation studies on the choice of rank and multi-objective optimization strategy. The results are demonstrated in Tab.3 of the main article and in Tab.2, respectively.

LoRA Rank. As for the choice of LoRA rank, we experiment with three settings: 64, 128, and 256. Generally, a larger rank means more trainable parameters are introduced, enhancing the model's adaptability to new data. However, this also increases computational and memory demands, potentially leading to over-fitting. Therefore, we aim to identify the most suitable parameter choice for human image generation through comparative experiments.

Multi-Objective Optimization Strategy. We further deploy several classic multi-task learning strategies during the training process and conduct related comparative experiments. Specifically, Linear Scalarization (LS), Dy-

namic Weigth Average (DWA), Uncertainty Weighting (UW), Random Loss Weighting (RLW), Scale-Invariant (SI), Nash-MTL, and Minimum Potential Delay Fairness Grad (MPD-FairGrad) are implemented here. Notably, SI utilizes the proportional fairness principle. Results in Tab.2 demonstrate that utilizing the MPD fairness principle to the multi-objective optimization for human image generation can achieve a more balanced performance.

Strategy	Image Quality		Regional Quality		
	HPS(%) ↑	IR(%) ↑	FID ↓	Hand Confi. ↑	Face Confi. ↑
LS	32.60	153.23	47.22	93.70	84.82
DWA	32.53	152.58	46.34	93.40	84.86
UW	32.61	152.82	47.42	93.76	84.15
RLW	32.58	154.56	47.90	93.80	84.24
SI	32.52	154.58	**46.01**	93.99	85.31
Nash-MTL	32.63	153.11	47.50	93.68	84.33
MPD-FairGrad	**32.73**	**154.60**	46.11	**94.03**	**85.34**

Table 2. Ablation results for the choice of multi-objective optimization strategy.

5. More Visualizations

More visualization results compared to existing methods are demonstrated in Fig.1, Fig.2, Fig.3, and Fig.4. In addition to the baseline methods mentioned in our paper, we also compare two close-source methods for human image generation called HanDiffuser and HyperHuman. Here, we extract example images from their paper for comparison.

Figure 1. Comparison with HanDiffuser.

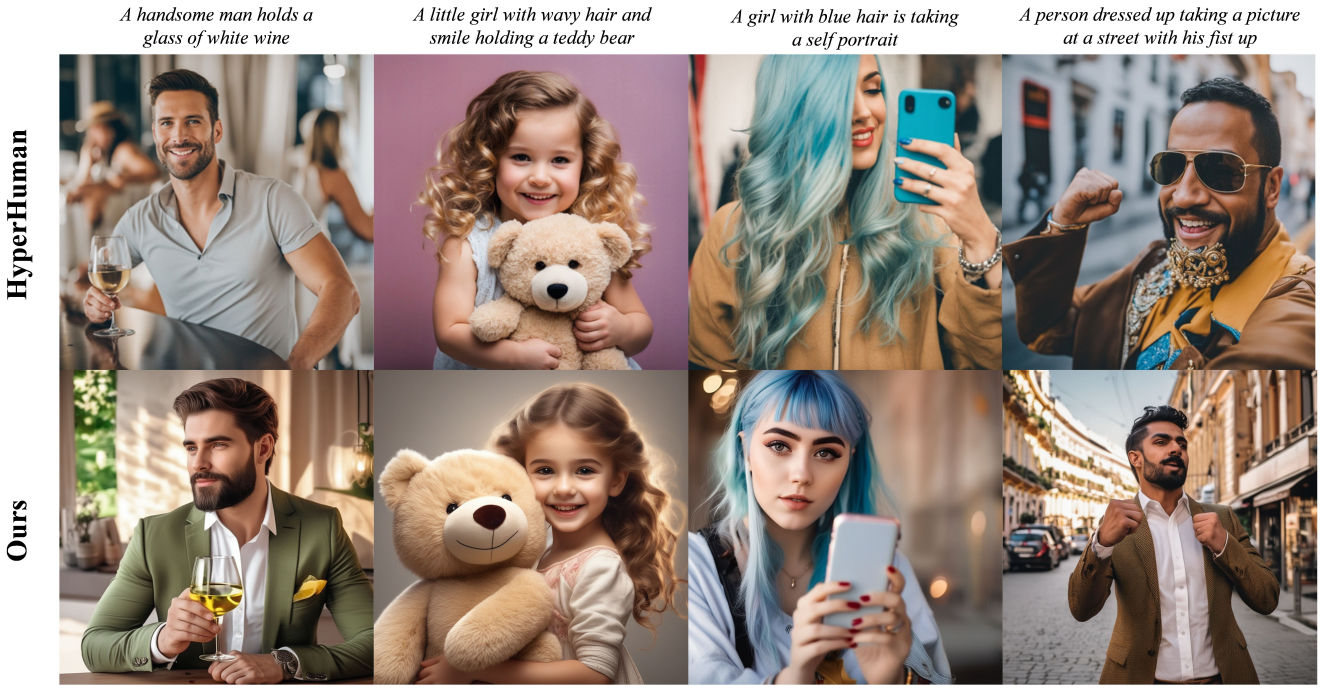

Figure 2. Comparison with HyperHuman.

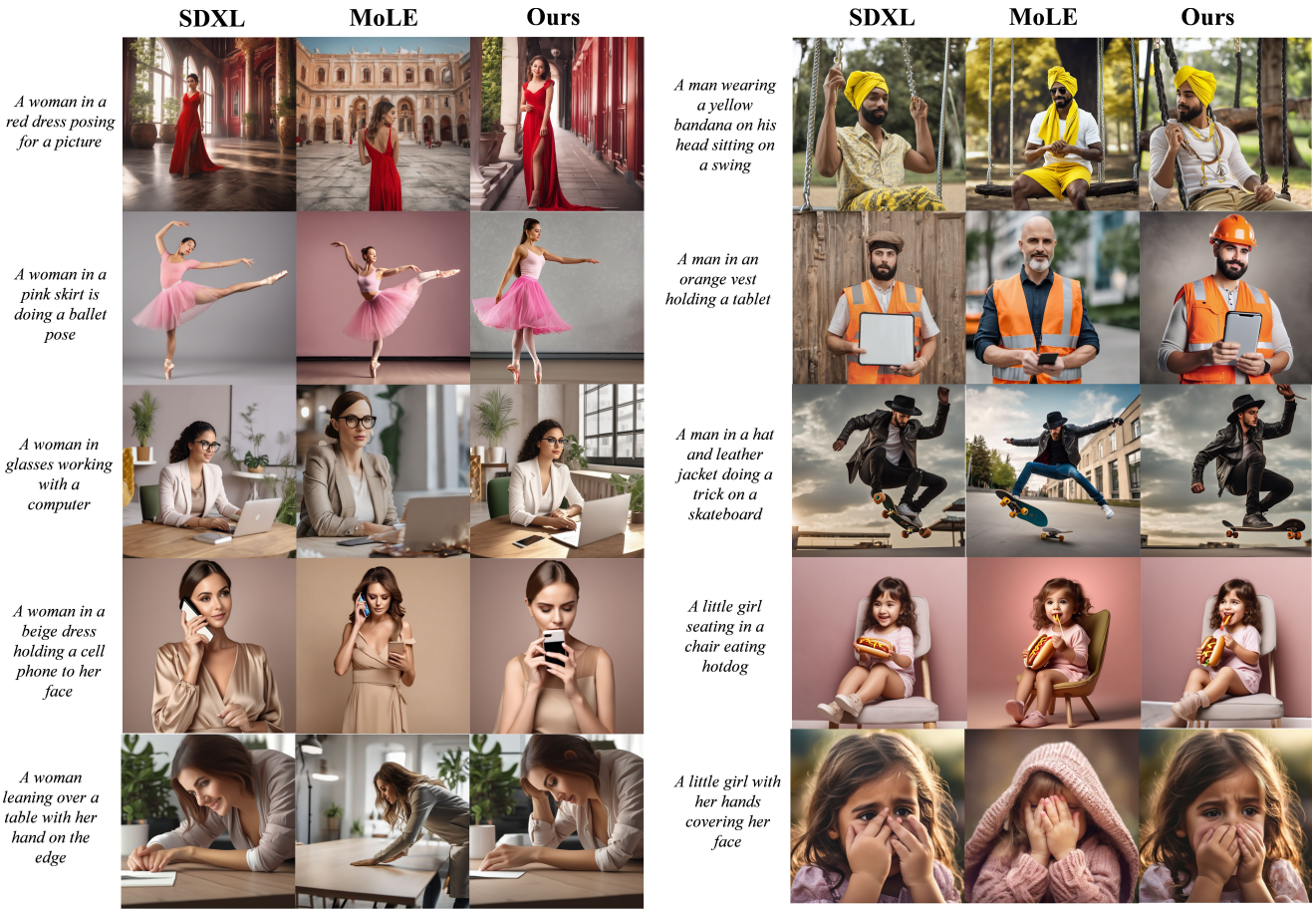

Figure 3. Comparison with general T2I methods.

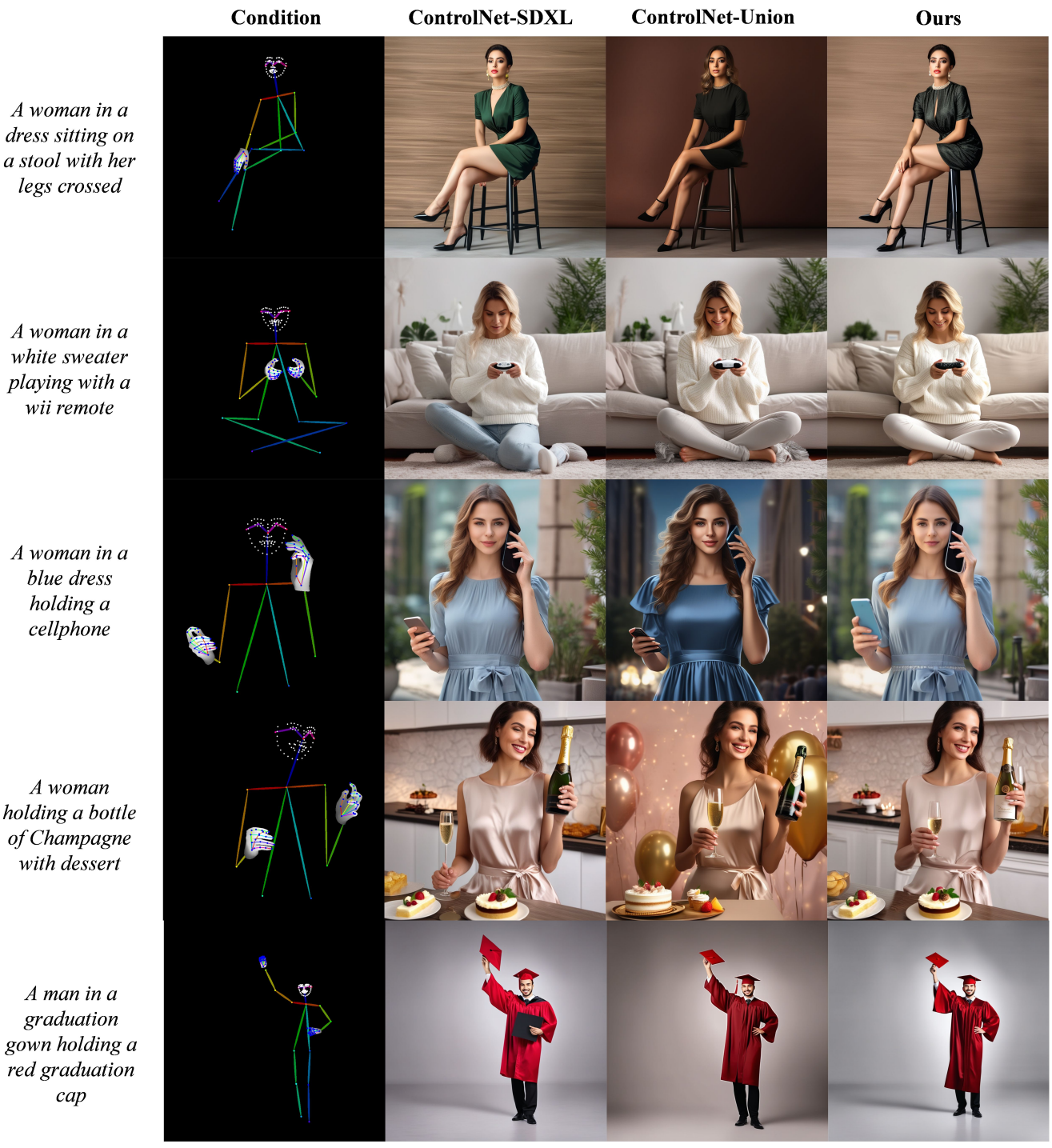

Figure 4. Comparison with controllable methods.