Supplementary Material of Human-MoE: Multimodal Full-Body Human Image Synthesis with Component-driven Mixture of Experts

Yu-Jiu Huang
College of Computer Science
National Yang Ming Chiao Tung University
Hsinchu City, Taiwan
joehuang1999.cs11@nycu.edu.tw

I-Chen Lin
College of Computer Science
National Yang Ming Chiao Tung University
Hsinchu City, Taiwan
ichenlin@cs.nycu.edu.tw

We provide additional details that support the findings presented in the main paper, including preliminaries, implementation details, evaluation metrics, ablation studies, overheads analysis, limitations, and more visual results.

A. Preliminaries

We briefly introduce the concept and notation of LDM and VQGAN. Please refer to [1] and [2] for details.

A. LDM

To achieve high computational efficiency and maintain generation quality, we adopt LDMs [1]. Content of LDM is mainly operated in the latent space. It requires an autoencoder, comprising an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. Given an image $x \in \mathbb{R}^{H\times W\times 3}$, it can encoded as a latent representation $\mathcal{E}(x)$. A LDM consists of a noise predictor ϵ_θ and can be expanded with condition encoders for various modalities. ϵ_θ is responsible for estimating noise based on the noisy images during the reverse diffusion process. Input conditions can be incorporated through concatenation with random noise and cross-attention. The spatial input is denoted by γ, and the attention conditioning input by ψ. The training objective for ϵ_θ is shown in the following equation:

$$L = \mathbb{E}_{\mathcal{E}(x),Y,\epsilon\sim N(0,1),t}\left[\|\epsilon - \epsilon_\theta(\gamma,\psi)\|_2^2\right], \tag{1}$$

where Y is the control condition, $\epsilon \sim N(0,1)$ is the Gaussian noise added to $\mathcal{E}(x)$, and $t = 1, ..., T$ signifies the time step. For the unconditional model, we obtain $\gamma = z_t$ and $\psi = t$, where z_t is a noisy image at t.

B. VQGAN

We use the VQGAN [2] as the autoencoder, which combines vector quantization to efficiently represent data in latent space. A VQGAN consists of an encoder $\mathcal{E}$, a decoder $\mathcal{D}$, a codebook $\mathcal{C}$, and a discriminator $\mathcal{D}_{disc}$. Given an image $x \in \mathbb{R}^{H\times W\times 3}$ in pixel space, $\mathcal{E}$ encodes x into a latent code $z = \mathcal{E}(x)$, where $z \in \mathbb{R}^{h\times w\times c}$. To represent the key features in the distribution, the vectors in z are replaced with the closest codewords from $\mathcal{C}$, which results in $z_q = \mathcal{Q}(z)$. $\mathcal{D}$ performs the reverse operation, and reconstructs the image $\hat{x} = \mathcal{D}(z_q)$. To further improve model performance, $\mathcal{D}_{disc}$ is introduced to assess the authenticity of generated results and to enhance the learning process. The objective of VQGAN is to keep the original image x to be as similar as possible to the generated image $\hat{x}$ with a limited codebook. The loss function we used is similar to that of [2].

B. Implementation Details

The following discussion focuses on the basic generative model. The editing model and experts in the refinement network follow similar configurations to the basic model.

A. VQGAN

The codebook size of the VQGAN is 8192, and the downsampling factor is 8. The encoder, decoder, codebook, and discriminator in VQGAN were trained from scratch. During training, an Adam optimizer is used with a batch size of 4 and an initial learning rate of 1×10^{-5}. The loss function is shown below:

$$\mathcal{L}_{VQ} = s_{rec}\|x - \hat{x}\|^2 + s_{cod}\|sg(\mathcal{E}(x)) - z_q\|^2 + s_{com}\|sg(z_q) - \mathcal{E}(x)\|^2, \tag{2}$$

$$\mathcal{L}_G = \frac{1}{2}(\mathcal{D}_{disc}(\hat{x}) - 1)^2, \mathcal{L}_D = \frac{1}{2}((\mathcal{D}_{disc}(x) - 1)^2 + (\mathcal{D}_{disc}(\hat{x}) - 0)^2), \mathcal{L}_{GAN} = \mathcal{L}_G + \mathcal{L}_D, \tag{3}$$

$$\mathcal{L} = \mathcal{L}_{VQ} + \mathcal{L}_{GAN} + \mathcal{L}_{lpips}, \tag{4}$$

where sg denotes gradient stopping, and s_{rec}, s_{cod}, and s_{com} represent the scales of reconstruction loss, codebook loss, and commitment loss, respectively. We set $s_{rec} = 1$, $s_{cod} = 1$, and $s_{com} = 0.2$. The key difference from the original VQGAN is the additional LPIPS [3] loss, which encourages the model to reconstruct results that are more aligned with human perception.

B. LDM

The noise step T is 1000. The text encoder is initialized with the pre-trained CLIP [4] model: CLIP-ViT-H-14-laion2B-s32B-b79K, while all other modules of the LDM are trained from scratch. During training, an Adam optimizer is used, along with a batch size of 8 and an initial learning rate of 5×10^{-6}. We set the dropout ratio for each condition to 0.1. In the first 20 epochs of training, we did not use the parsing map as a spatial condition in order to encourage the model to prioritize learning the basic pose and structure provided by the pose map. Sampling utilizes DDIM [5] with 20 steps. For hybrid multi-condition CFG, the guidance scales, ω_{joing} and ω_{indep}, are both set to 3. For the independently gradient part, we only consider the text condition.

C. Composition Pipeline

We perform composition in the order of the face, hands, upper garment, and lower garment. Two methods are provided. The first method uses multiple predefined mask sizes and calculates the gradient at the mask boundaries using a Laplacian filter. If the gradient of a mask exceeds a given threshold of 3, the mask configuration is discarded. Then, the mask setting closest to the original mask is selected for Poisson blending [6]. Next, we calculate the SSIM [7] for the component areas before and after blending. If the similarity is below 0.85, we switch to the second method, which directly pastes the source image onto the target image. Finally, the edges are repaired by the fine-tuned SD [1] inpainting model. A kernel of size 5 is applied to dilate the edges and obtain the repair mask.

D. Compared Methods

We applied the pre-trained Text2Human [8] model provided by the authors. For ControlNet [9], we took SD v1.5 as its pre-trained model and fine-tuned a copy branch with parsing and pose maps as spatial conditions. For Pix2Pix-HD [10] and SPADE [11], we followed the initial settings and trained each corresponding model from scratch.

C. Evaluation Metrics

We follow previous methods for calculation of FID [12] and SSIM. How we evaluate the scores for CLIP-Score [13], PD, and SI are introduced as follows.

A. CLIP Score

To ensure the stability of measuring text-image alignment, we introduced five pre-trained CLIP models: clip-vit-base-patch16, clip-vit-base-patch32, clip-vit-large-patch14, clip-vit-large-patch14-336, and CLIP-ViT-H-14-laion2B-s32B-b79K. The final CLIP-Score is calculated by averaging the outputs from these models.

B. Pose Distance

We use OpenPose [14] to predict the keypoints of both the ground truth and the generated human. PD is calculated using MPJPE [15], which measures the distance between the keypoints of the generated human and the ground truth. The final score is weighted by the confidence scores of the ground truth to reduce calculation errors.

C. Semantic Intersection

We fine-tuned SAM [16] to obtain 23 binary segmentation models. The output of each model is a binary mask that delineates the region of object distribution. A human image corresponds to 23 masks. The SI score is calculated by averaging the mIoU between corresponding masks and then computing the overall average.

D. Additional Ablation Study

A. Varying Input Modalities

We investigated the impact on our model when specific input modalities are omitted, as shown in Table A. The reference model uses semantic maps, text descriptions, and pose maps as input conditions without MoE. We sequentially omitted text, pose, and both text and pose, and replaced them with zero-valued tensors to maintain dimensional alignment. Omitting text conditions reduced the CLIP-Score, as the absence of text weakened the semantic understanding of the model. Masking pose conditions resulted in a slight decrease in PD, as the model lacked precise pose information and relied solely on parsing maps. Further omitting both text and pose slightly reduced other quality and similarity metrics. These results confirm that our model can effectively handle various conditions in a decoupled manner. Additionally, we observed that the addition of pose

conditions helps the model achieve higher precision in body posture control. For example, due to the ambiguity in semantic representations, artifacts often appear when the legs are crossed, as shown in Fig. A. Our framework mitigates this issue by incorporating pose maps that provide limb crossing information.

TABLE A: Quantitative comparison of the impact of modal condition inputs on generation results. T stands for text, P stands for pose, and TP refers to both text and pose. The scores highlighted in bold represent the best performance.

Methods	FID↓	CLIP-S↑	SSIM↑	PD↓	SI↑
Ours (w/o MoE)	**21.64**	**0.261**	**0.826**	1.50	0.943
w/o P	22.71	0.255	0.817	1.62	0.943
w/o T	26.30	0.231	0.811	**1.46**	**0.944**
w/o TP	25.63	0.230	0.809	1.52	0.943

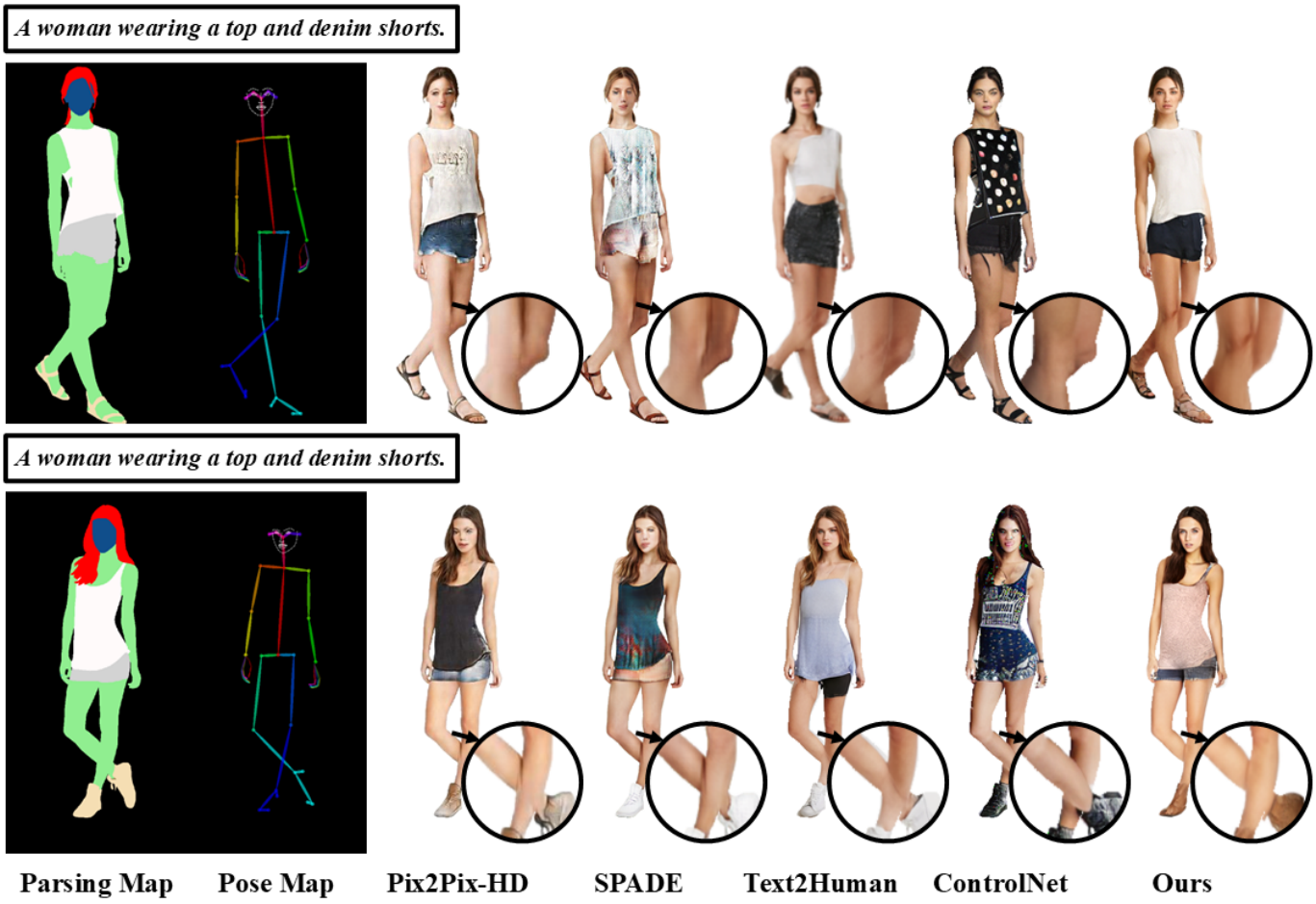

Fig. A: Illustration of the results of leg-crossing poses. Obvious artifacts occurs in results of related methods.

B. Impact of CFG Scale

This study aims to analyze the impact of scale strength in the proposed hybrid multi-conditioned CFG. The ablation experiment in this sub-section does not consider the inclusion of the refinement network. From Table B, it can be observed that when only joint CFG guidance is used, the performance in terms of FID, SSIM, and LPIPS is optimal when ω_{joint} is set to 3. As ω_{joint} weight increases, although the other metrics remain almost unchanged, the FID significantly worsens. Therefore, we chose ω_{joint} to be 3 as the starting point. After introducing Independent CFG guidance, the lowest FID and highest CLIP-Score are achieved when ω_{indep} is set to 7, but the SSIM and LPIPS performances worsen. When ω_{indep} is set to 3, FID and CLIP Score remain acceptable, while SSIM and LPIPS performances are the best. Considering the balance of these metrics, we eventually chose ω_{indep} to be 3 as the experimental setting.

TABLE B: Ablation comparison of CFG scale configurations. The scores highlighted in bold represent the best performance.

CFG settings	FID↓	CLIP-S↑	SSIM↑	LPIPS↓
$\omega_{joint} = 1$	26.47	0.228	0.803	0.075
$\omega_{joint} = 3$	24.06	0.230	0.808	0.074
$\omega_{joint} = 5$	24.50	0.231	0.808	0.074
$\omega_{joint} = 7$	24.91	0.231	0.808	0.074
$\omega_{joint} = 3$, $\omega_{indep} = 1$	23.79	0.239	0.813	0.072
$\omega_{joint} = 3$, $\omega_{indep} = 3$	21.64	0.261	**0.826**	**0.067**
$\omega_{joint} = 3$, $\omega_{indep} = 5$	21.27	0.260	0.820	0.068
$\omega_{joint} = 3$, $\omega_{indep} = 7$	**21.04**	**0.262**	0.818	0.069

E. Overheads Analysis

To ease the framework scalability and limit the VRAM usage, in our proof-of-concept system, we streamlined the VAE and U-Net architectures in both the synthesis model and experts to make them more lightweight compared to the original LDM. A more sophisticated pre-trained text encoder is employed to enhance the cross-modal alignment capability of the synthesis model and experts. We also use a fine-tuned SD inpainting model for edge refinement to address boundary imperfections. Table C lists the parameter counts (M), model size in VRAM (MB), and inference time (ms) of individual stages of our method and those of ControlNet with additional inputs, such as parsing maps, poses, during the inference phase. The model size includes both parameters and buffer sizes in float32 precision. We used the DDIM sampler with 20 inference steps and performed inference on an RTX 3070 GPU. The five experts operate based on the output of the main model and do not rely on each other. Alternatively, they can perform in parallel to shorten the inference time. Regardless of whether the experts operate in a sequential or parallel manner, the inference time of our system is lower than that of ControlNet.

TABLE C: Parameter counts (M), model size in VRAM (MB), and inference time (ms) of our method and ControlNet.

Method/Stages	Parameter Counts (M)	Model Size (MB)	Inference Time (ms)
ControlNet	1428	5446	94450
Ours (Synthesis Stage)	594	2265	3779
Ours (Sequential Expert Stages)	584	2229	3731 × 5
Ours (Alternative Parallel Expert Stage)	584 × 5	2229 × 5	3731
Ours (Edge Refinement Stage)	1067	3867	46981

F. Limitations

A. Computational Cost

Our current proof-of-concept expert models adopt the same design as the generative model, which may introduce unnecessary overhead. We believe this issue can be effectively mitigated through model compaction techniques such as knowledge distillation.

B. Controllability

Our models tend to overlook unexpected control conditions, such as inputting a text description "three hands" or arbitrarily scribbling on the parsing map. This limitation arises because the model is trained from scratch, it captures only the characteristics of the dataset without incorporating broader prior knowledge. We think the generalization ability can be enhanced by incorporating a pre-trained generative model, where richer priors are provided.

G. More Visual Results

A. Image Generation

Results generated by Human-MoE are of high fidelty and meet the given conditions, as shown in Fig. B. Please refer to the supplementary video for more results.

B. Image Editing

As shown in Fig. C, our method offers excellent flexibility in editing capabilities. In semantic-driven editing, it allows control over object layouts, such as fabrics, accessories, hairstyles, and body types. In text-driven editing, it enables the manipulation of object appearances, including clothing materials and colors, as well as human hair and skin tones. Component-driven editing provides the freedom to creatively adjust elements within the same geometry as the parsing map.

Fig. B: Additional visualization of humans wearing dresses.

C. User Interface

Fig. D shows the user interface, which allows users to express their preferences of multiple modalities in an intuitive way.

REFERENCES

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.

[2] Patrick Esser, Robin Rombach, and Bjorn Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12873–12883.

[3] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[5] Jiaming Song, Chenlin Meng, and Stefano Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[6] Patrick Pérez, Michel Gangnet, and Andrew Blake, "Poisson image editing," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 577–582. 2023.

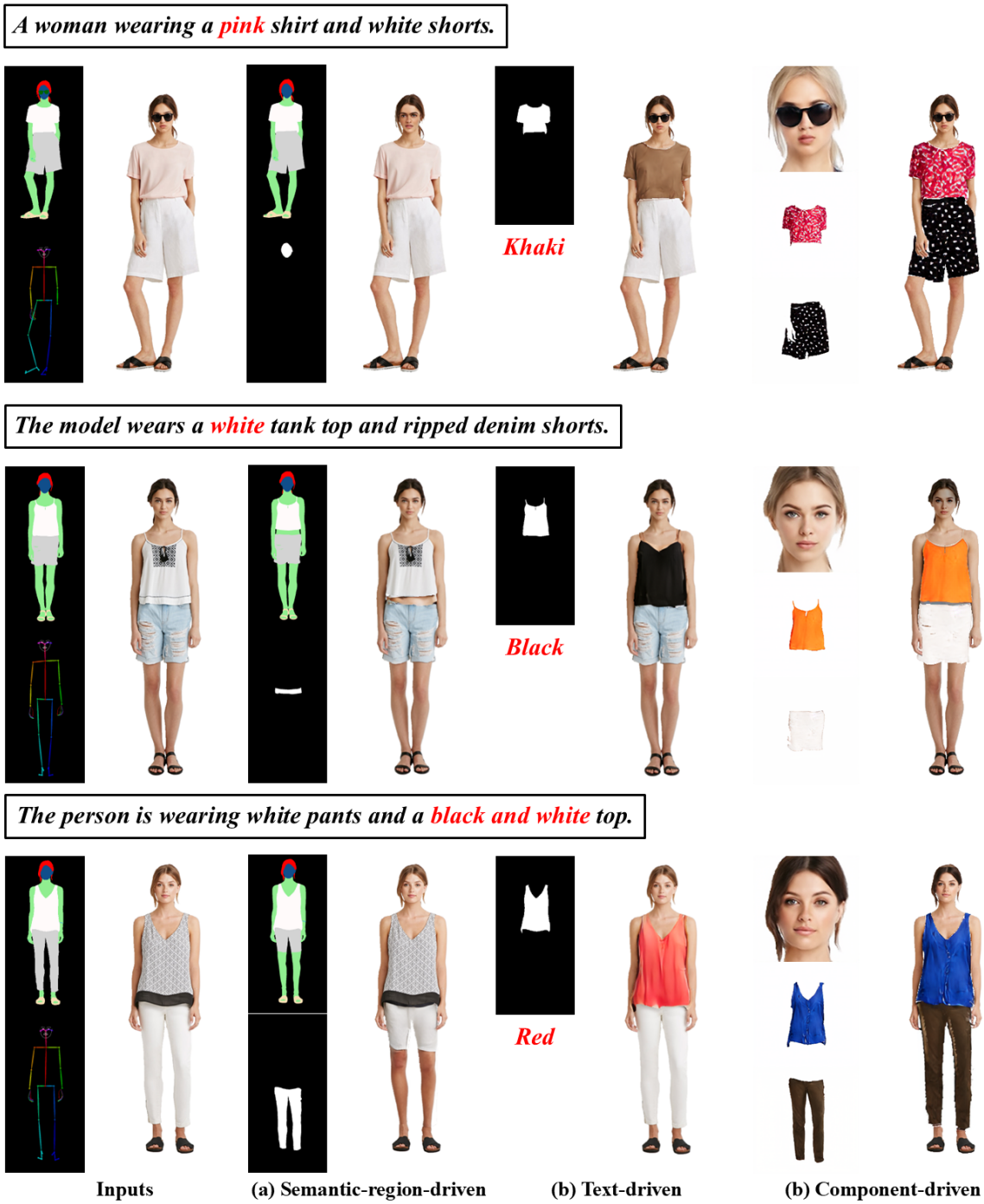

Fig. C: Visualization of flexibility in our editing technique. In (b), the text-driven editing change the color of a region according to the text.

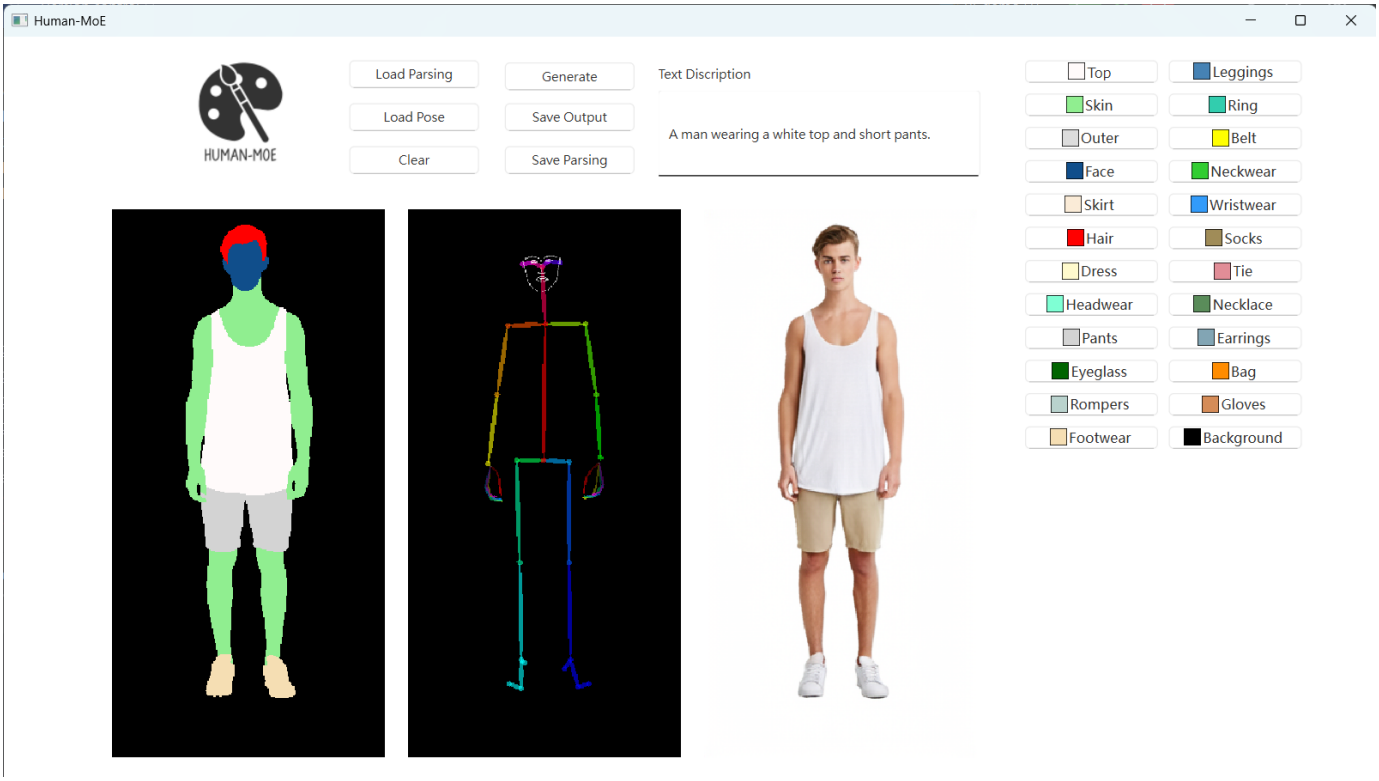

Fig. D: User interface of our human image synthesis system. Please refer to the supplementary video for the process of image generation and editing.

[7] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[8] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu, "Text2human: Text-driven controllable human image generation," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–11, 2022.

[9] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[10] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.

[11] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2337–2346.

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[13] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi, "Clipscore: A reference-free evaluation metric for image captioning," *arXiv preprint arXiv:2104.08718*, 2021.

[14] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[15] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7753–7762.

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al., "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.