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

***Abstract*—Conditional full-body human synthesis is to generate and edit realistic images based on given conditions. Previous methods lay a solid foundation but may have limitations in adjusting human poses and appearance. They usually adopt a monolithic design, and the details of generated images are prone to be indistinct or distorted due to the high variation of human appearances. To tackle the above-mentioned challenges, we propose Human-MoE for multi-modal full-body human synthesis. Users can control the image generation through three types of input representations: parsing maps for geometry, text descriptions for appearance attributes, and pose maps to distinguish postures. Our framework specifically designs a mixture-of-experts module to capture and synthesize details in specific regions with high fidelity. These synthesized details are then applied to refine the appearance. Our method achieves top scores in experiments by multiple metrics, especially FID and SSIM, demonstrating its advance in visual quality and controllability.**

***Index Terms*—Human body image synthesis, mixture of experts, multimodal control, latent diffusion model**

I. Introduction

Full-body Human Image Synthesis (FHIS) aims at generating complete and coherent representations of the entire human body, where the head, torso, limbs, and clothing details are revealed. It has become attractive and employed in various applications, such as virtual try-ons [1] and pose transfer [2]. To make the generated results meet the expectation of users, recent related methods [3], [4] use multi-modal frameworks to conditionally synthesize images according to various inputs, such as poses, text descriptions, or reference images.

While previous methods generated impressive results, we found that they usually encountered two challenges, realism and controllability. Regarding realism, they often focus on global features, and have difficulty in capturing intricate facial and hand details, leading to blurriness in these areas. When the legs of a target cross, artifacts frequently occur (as shown in Fig.A of the supplementary document). Regarding controllability, we found that there are limitations and it is not always easy to specify the desired appearance conditions in related framework. For instance, in Text2Human [3], the text control component relies solely on one-hot encoding, which restricts users to selecting specific tokens.

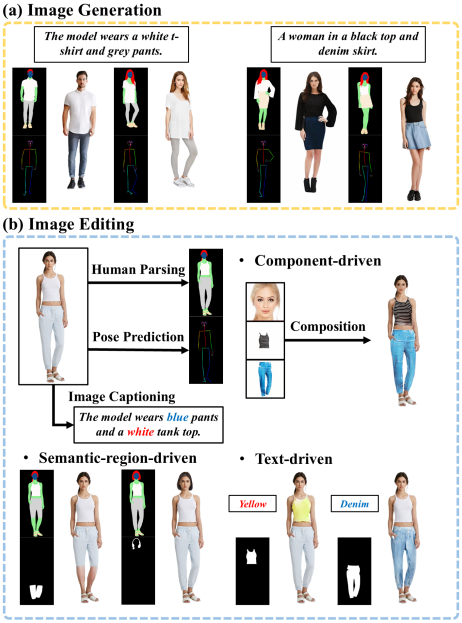

Fig. 1: Overview of the functions of our framework: (a) image generation according to parsing maps, pose maps, and text descriptions, (b) image editing according to multimodal and multi-component conditions. Its provides users with text-driven, semantic-region-driven, and component-driven editing. The text-driven editing changes the color of a region according to the replaced text.

The objective of our work is to achieve high-fidelity human image synthesis and keep the generation process controllable. We utilize a Latent Diffusion Model (LDM) [5] as the foundational architecture and a Vector-Quantized Generative Adversarial Network (VQGAN) [6] as the autoencoder. To more precisely and intuitively control the generation process, our framework takes parsing maps, text descriptions, and pose maps as input. Parsing maps provide multi-class labels that enable intuitive control over the shape and structure of

objects, and allow for precise geometric management. Text descriptions indicate appearance attributes such as gender, skin tone, as well as clothing color and texture. Pose maps mark human keypoints to disambiguate the limb postures and remove related artifacts. To address the detail missing issue, we introduce a component-based Mixture of Experts (MoE) module, in which specialized autoencoders and LDMs are trained as regional experts for the face, hands, and both upper and lower clothing. A gating mechanism allocates regions to the corresponding experts, and the generated regional details are fused with the target image using composition techniques.

Our work can not only generate human images but also be applied for editing. For the generative model, the LDM operates in latent space and uses a noise predictor to estimate noise in a noisy image based on given conditions. The decoder then maps the generated embeddings back into pixel space to produce the final result. For the editing model, additional information from a binary mask is incorporated through concatenation, while the rest remains similar to the generative model. We compare our method with related state of the arts through multiple quantitative metrics, including FID, CLIP-Score, SSIM, pose distance, and semantic intersection. They demonstrates the diversity, realism, and controllability of the generated results by the proposed method. Fig. 1 demonstrates the functionality of our work. Our contributions are summarized as follows:

- Human-MoE framework is proposed, which integrates multiple conditions (parsing maps, pose maps, and text descriptions) to achieve highly controllable human image generation and editing.
- The component-based MoE module is introduced to refine the quality of the face, hands, and clothing, and it can be extended to other body parts.

II. Related Work

The emergence of Diffusion Models (DMs) [7] has significantly advanced image generation tasks. DMs generate images by iteratively denoising a variable starting from noise. LDMs [5] apply this process in latent space, reducing computational complexity while maintaining high quality. Classifier Free Guidance (CFG) is the main method to control LDMs [8]. CFG controls the generation process without an external classifier by leveraging a single DM trained on both conditional and unconditional data. SpaText [9] further proposes multi-condition CFG for finer control. The advent of LDMs has driven significant advancements in various applications. Large cross-modal models like Stable Diffusion (SD) [5] pave the way for cutting-edge developments.

Palette [10] introduces a general framework for image-to-image tasks based on conditional DMs. It demonstrates that concatenating condition images with model inputs can effectively controls the generation process. ControlNet [11] aims to add spatial conditioning control to large pre-trained text-to-image DMs, locking model parameters and replicating encoding layers to retain the capability of original model.

Conditional methods [3], [4], [12] include appearance and pose guidance. UPGPT [4] introduces a multimodal LDM that uses 3D information to control poses, while combining reference images of various body parts and text descriptions to guide appearance. Text2Human [3] uses a two-stage architecture converting human pose data into parsing maps to indicate geometry, and controls appearance with predefined labels. It employs a MoE mechanism with hierarchical VQVAEs [13] to retain more information. Multi2Human [12] builds upon the Text2Human parsing-to-human framework and incorporates wavelet embeddings to enhance detail quality.

We express our gratitude to previous work. The design of UPGPT multimodal LDM for the FHIS task and the MoE concept proposed by Text2Human have inspired us. Compared to the MoE in Text2Human, which assigns tasks to experts based on garment textures, our component-driven MoE module assigns tasks to experts based on body regions and integrates the output of each expert. This strategy enables us to synthesize body details beyond garments.

III. Proposed Method

The proposed Human-MoE takes the LDM [5] as the foundational framework and can controllably generate and edit high-fidelity images. As shown in Fig. 1, the generative model uses parsing maps to define geometry, pose maps to distinguish posture, and text descriptions to control appearance. The editing model offers several key functionalities: semantic-region-driven editing for modifying parsing maps and controlling component geometry, text-driven editing for adjusting attributes such as color and style, and component-driven editing for altering components using the MoE mechanism and composition module. Our overview architecture is shown in Fig. 2. We define the image dimensions in latent space as $h = H/8$ and $w = W/8$. To synthesize an image $\hat{x} \in \mathbb{R}^{H\times W\times 3}$, the input to the generative model includes a random noise $z_T \in \mathbb{R}^{h\times w\times 3}$, a parsing map $c_{parsing} \in \mathbb{R}^{H\times W\times 23}$ (where the channels represent the number of categories), a pose map $c_{pose} \in \mathbb{R}^{H\times W\times 3}$, and a text description c_{text}. For the editing model, an original image $x \in \mathbb{R}^{H\times W\times 3}$ and an additional binary mask $c_{mask} \in \mathbb{R}^{H\times W\times 1}$ are required to specify the known and unknown areas.

A. Full-body Human Image Synthesis

Previous work has introduced various innovative methods that have significantly advanced the field of human generation technology. However, due to the richness and diversity of portrait features, previous techniques often struggle to achieve both high fidelity and fine control in FHIS. To balance generation quality and efficiency, we chose the LDM as our framework and combined it with multimodal inputs for fine-grained control. The LDM requires an autoencoder to map data between the pixel space and the latent space. We use the VQGAN as the autoencoder, which consists of an encoder $\mathcal{E}$, a decoder $\mathcal{D}$, a codebook $\mathcal{C}$, and a discriminator $\mathcal{D}_{disc}$. Given an original image x, we obtain the reconstructed image $\hat{x} = \mathcal{D}(\mathcal{Q}(\mathcal{E}(x)))$ and use the trained $\mathcal{E}$ and $\mathcal{C}$ to generate

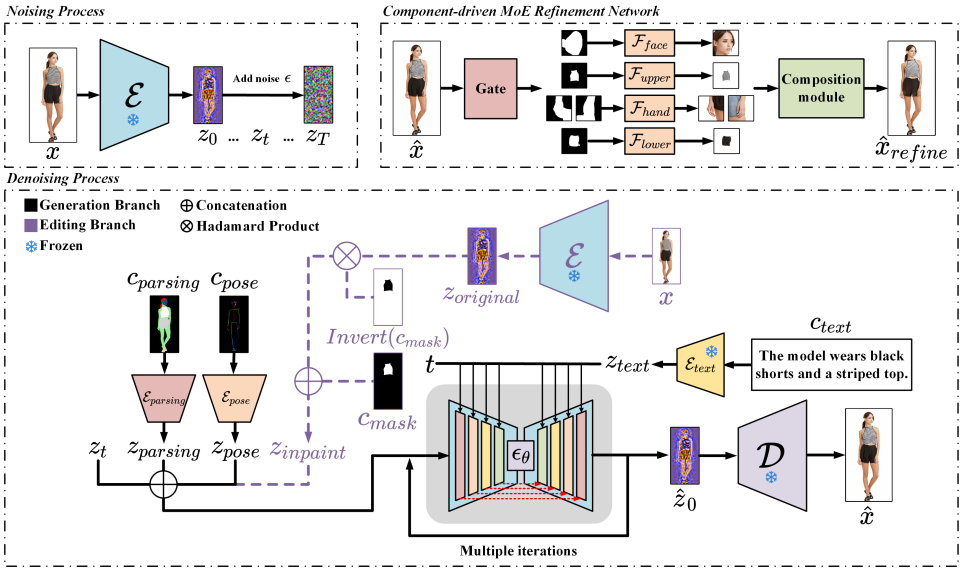

Fig. 2: Overview of Human-MoE architecture. For generative models, during training, parsing maps, pose maps, and text descriptions are used as conditional inputs for the noisy predictor ϵ_θ. The output of ϵ_θ is the predicted noise added to a noisy latent image z_t. During inference, z_t undergoes several denoising steps to obtain a generated latent image $\hat{z}_0$, which is input into the decoder $\mathcal{D}$ to obtain a generated image $\hat{x}$. Editing models introduce binary masks and original images as spatial conditions, while the rest of the operations are similar to generative models. The refinement network combines multiple generative models, and each of them focuses on a specific body part, to collaboratively improve human details.

a latent representation $z = \mathcal{Q}(\mathcal{E}(x))$, where $\mathcal{Q}$ is the vector quantization operation. The LDM consists of a noise predictor, ϵ_θ, which is responsible for estimating noise based on noisy images during the reverse diffusion process. The spatial input is denoted by γ, which is formed by concatenating random noise with other image conditions. The attention conditioning input is denoted by ψ, which aligns with the image dimensions through cross-attention. To explore the role of each condition, we divide the discussion into pose and appearance guidance for generation and regional manipulation for editing.

1) Pose Guidance: For each human image, we adopt a parsing map $c_{parsing}$ and a pose map c_{pose} as pose representations. $c_{parsing}$ tells which category (23 semantic categories) a pixel belongs to. It is used for layout specification and also helps with geometric editing. However, the annotations of $c_{parsing}$ has inherent limitations, which lead to ambiguity. For instance, the model could struggle to determine which limb is in the front and results in distortions at overlapping regions. Therefore, we incorporate c_{pose} to address the aforementioned issue. The pose is predicted by OpenPose [14] model. c_{pose} contains 25 full-body keypoints, 70 facial keypoints, and 21 keypoints for each hand, presented in RGB format. The parsing encoder $\mathcal{E}parsing$ and the pose encoder $\mathcal{E}pose$ consist of pointwise convolutional layers that extract features from the images while preserving spatial information. After inputting $c_{parsing}$ and c_{pose} into $\mathcal{E}_{parsing}$ and $\mathcal{E}_{pose}$, we obtain the parsing embedding $z_{parsing} \in \mathbb{R}^{h\times w\times 3}$ and pose embedding $z_{pose} \in \mathbb{R}^{h\times w\times 1}$. $z_{parsing}$, z_{pose}, and z_t are concatenated as γ to achieve pose guidance.

2) Appearance Guidance: We intuitively control human appearance through a text description c_{text}, which includes details such as hair color, skin tone, and clothing style. Since CLIP [15] has strong text-image alignment capabilities, we use its pre-trained text encoder $\mathcal{E}_{text}$ to convert c_{text} into the text embedding z_{text}. We incorporate z_{text} into ϵ_θ using cross attention, allowing the model to generate images that are highly aligned with the given c_{text} more precisely.

3) Region Manipulation: Our region editing method shares a similar core concept with Palette [10]. As shown in the purple part of Fig. 2, given a original image x and a binary mask c_{mask}, we can obtain a original latent image $z_{original} = \mathcal{Q}(\mathcal{E}(x))$, where $z_{original} \in \mathbb{R}^{h\times w\times 3}$, and then compute the inverse binary mask $Invert(c_{mask})$. In c_{mask}, the white areas represent the unknown area, while the black areas represent the known area. We perform element-wise multiplication between $Invert(c_{mask})$ and $z_{original}$, and then concatenate the product with c_{mask} to obtain $z_{inpainting} \in \mathbb{R}^{h\times w\times 4}$. c_{mask} effectively identifies the target areas that need modification and $z_{original}$ provides crucial information regarding the known area. In this way, our approach not only enhances the flexibility of editing but also improves the consistency of the generated results.

B. Component-driven MoE Refinement Network

Capturing fine details presents a significant challenge for the primary model. Therefore, we incorporate the component-driven MoE module, which consists of a gate and multiple experts. The gate selects the appropriate expert for processing based on the input features. Each expert focuses on a specific synthesis task. Given $\hat{x}$, the gate identifies specific regions

based on the annotations from $c_{parsing}$. The identified regions in $\hat{x}$ correspond to a dilated binary mask m_{roi}. Our experts target the face, hands, upper garment, and lower garment, represented as $\mathcal{F}_{face}$, $\mathcal{F}_{hand}$, $\mathcal{F}_{upper}$, and $\mathcal{F}_{lower}$, and each of them has its own synthesis network, which operates similarly to the holistic one. For example, for a random noisy image, the parsing map and the pose map of $\mathcal{F}_{face}$ are square-cropped versions of the original ones, and the text of $\mathcal{F}_{face}$ is generated by Gemini [16]. The output of $\mathcal{F}_{face}$ is x_{face}. The operations of $\mathcal{F}_{hand}$, $\mathcal{F}_{upper}$, $\mathcal{F}_{lower}$ are similar to $\mathcal{F}_{face}$. We selectively use Poisson blending [17] to smoothly integrate x_{face}, x_{hand}, x_{upper}, x_{lower} with m_{roi} to get the refinement image $\hat{x}_{refine}$, as described in the following formula:

$$\hat{x}_{refine} = Poisson(S, T, M), \tag{1}$$

$$\begin{cases} S = \{x_{face}, x_{hand}, x_{upper}, x_{lower}\} \\ T = \hat{x} \\ M = m_{roi} \end{cases}, \tag{2}$$

where S represents the source image, T represents the target image, and M is used to specify the regions to be extracted from S. We aim to blend S into T based on M. We iteratively compute the edge gradients and dynamically adjust the range of M to prevent color leaks. SSIM [18] is introduced to compare the similarity of the component regions before and after blending. If the similarity score does not exceed the specified threshold, it indicates that the chosen parameters are suboptimal, and thus S will be directly pasted onto T. To improve inconsistencies at the boundaries between S and T, we further finetuned the SD [5] inpainting pre-trained model to handle the edges effectively. This framework successfully alleviates face and finger blurring while enhancing clothing texture preservation. The concept of the component-driven MoE can be extended to handle additional body regions. Compared to common holistic refinement methods, our MoE allows users to achieve more detailed and customized synthesis.

C. Training and Sampling

During the training process, ϵ_θ needs to learn to predict the noise stochastically added to z_0. Our training objective is represented by the loss function below:

$$L = \mathbb{E}_{\mathcal{E}(x), c_{spatial}, c_{cross}, \epsilon \sim N(0,1), t}\left[\|\epsilon - \epsilon_\theta(\gamma, \psi)\|_2^2\right], \tag{3}$$

$$c_{spatial} = \{c_{parsing}, c_{pose}\} \cup \{c_{mask}\}, c_{cross} = \{c_{text}\}, \tag{4}$$

where $c_{spatial}$ and c_{cross} represent the sets of spatial and cross domain conditions. Here, we obtain $\gamma = \{z_t, z_{parsing}, z_{pose}\} \cup \{z_{mask}\}$ and $\psi = \{t, z_{text}\}$.

During inference, we propose a hybrid version of the multi-conditional CFG based on the approach presented by SpaText [9]. We calculate the joint condition $\Delta^t_{joint} = \epsilon_\theta(z_t|\{y_i\}_{i=1}^N) - \epsilon_\theta(z_t|\varnothing)$, and apply the guidance scale ω_{joint} to control the conditioning strength. This approach requires two feedforward executions per denoising step: one for the null condition and one for the joint condition. To control target conditions independently during sampling, we can additionally calculate the direction for target conditions $\Delta^t_j = \epsilon_\theta(z_t|y_j) - \epsilon_\theta(z_t|\varnothing)$ and linearly combine them using M guidance scales ω_{indep}. In the reverse process, the noise predicted at each time step refers to the formula below:

$$\hat{\epsilon}_\theta(z_t|\{y_i\}_{i=1}^N) = \epsilon_\theta(z_t|\varnothing) + \omega_{joint} \cdot \Delta^t_{joint} + \sum_{j=1}^{M} \omega_{indep(j)} \cdot \Delta^t_j. \tag{5}$$

To increase sampling speed while maintaining generation quality, we introduced DDIM [19] sampler, conducting fewer sampling steps. After inference, we obtain the generated latent image $\hat{z}_0$, which is then input into $\mathcal{D}$ to decode and obtain $\hat{x}$.

IV. Experiments

We conducted experiments on the DeepFashion-Multimodal dataset [3]. Following the Text2Human settings, we selected $11,484$ full-body images, which were split into training and testing sets and downsampled to a 512×256 resolution. Parsing maps included the original 23 label categories, excluding the background. Pose maps were extracted by OpenPose [14], covering the body, face, and hands. Text descriptions were generated by BLIP-2 [20] and Gemini [16]. Binary masks were created based on parsing labels or randomly generated strokes.

We evaluated model performance by five metrics. FID [21] quantifies distribution differences between generated and real images to assess quality. CLIP-Score [22] measures alignment between images and text descriptions using cosine similarity in shared embedding space. SSIM [18] calculates pixel-level similarity for reconstruction accuracy. Pose distance (PD) calculates the distance between keypoints to measure posture accuracy. Segmentation intersection (SI) calculates the mIoU between generated and real segmentations output by a fine-tuned SAM [23].

A. Quantitative and Qualitative Comparisons

We compared our method with Pix2Pix-HD [24], SPADE [25], Text2Human [3], and ControlNet [11]. Tab. I summarizes the modalities and functionalities of each method. We trained models for Pix2Pix-HD and SPADE with parsing maps as conditions. For Text2Human, we adopted a pre-trained parsing-to-human model, replacing textual descriptions with provided token labels. ControlNet was used as a subnet to fine-tune the SD [5] model with image-text pairs. During sampling, ControlNet was configured in the same way as in our work. Background regions were removed from each generated image to focus on evaluating the quality of the human subject.

As shown in Tab. II, our model achieves the lowest (best) FID score 19.88, indicating that the distribution of generated images closely align with the distribution of the original dataset. In terms of CLIP-Score, ControlNet achieved the highest score 0.267, and our model reached a score 0.259. It demonstrates that the proposed method is capable of generating high quality images based on specific descriptions without relying on the prior of SD. For SSIM, the proposed model attains the highest score 0.826, indicating that the generated images have high structural and visual consistency. Regarding

TABLE I: Modalities and capabilities checklist. S, T and P denote parsing, text, and pose, respectively. The abbreviations of PC, AC, TE, and SE are pose control, appearance control, text-driven editing, and semantic-driven editing, respectively.

	Modalities			Capabilities			
Methods	S	T	P	PC	AC	TE	SE
Pix2Pix-HD [24]	✓			✓			
SPADE [25]	✓			✓			
Text2Human [3]	✓	✓		✓	✓		
ControlNet [11]	✓	✓	✓	✓	✓	✓	✓
Ours	✓	✓	✓	✓	✓	✓	✓

TABLE II: Quantitative results on DeepFashion-Multimodal dataset. The bold highlight signifies the best-performing score.

Methods	FID↓	CLIP-S↑	SSIM↑	PD↓	SI↑
Pix2Pix-HD [24]	28.26	0.233	0.818	1.39	0.962
SPADE [25]	27.84	0.228	0.817	**1.25**	**0.971**
Text2Human [3]	22.21	0.233	0.783	1.69	0.927
ControlNet [11]	32.13	**0.267**	0.783	1.83	0.938
Ours	**19.89**	0.259	**0.826**	1.47	0.947

PD and SI, our performance is slightly inferior to that of SPADE and Pix2Pix-HD (conditioning with only the parsing map), but surpasses other frameworks. We speculate that the multimodal framework faces greater complexity in learning the relationships among multiple modalities, but our work is still close to the parsing-map-only-based models in terms of pose accuracy.

Next, we conducted a visual comparison between our method and others. In Fig. 3, in scenarios where only parsing maps are used as control conditions, Pix2Pix-HD and SPADE struggle to produce clear images and their results lack diversity. The additional inclusion of text as a generation condition can enhance the control over human appearance. Text2Human can generate images with distinct local details and varied clothing styles using semantic maps and text tokens. However, its reliance on one-hot encoding for text vocabulary limits its flexibility in handling new or synonymous words. It is difficult to specify details such as garment color for their system. ControlNet retains the excellent visual semantic alignment of SD but fails to maintain detailed features during generation. In contrast, our model generates higher realism images using either parsing maps alone or combined with text descriptions. It also shows stronger controllability over appearance when both conditions are used. With the addition of pose conditions, the model achieve higher precision in body posture control.

B. Ablation Study

We conducted an ablation study on the impact of the component-driven MoE refinement network. As shown in Table III, adding MoE improved all metrics except the CLIP-Score compared to the baseline. The visual results after applying MoE are shown in Fig. 4, where clearer details in the face and hands, and more intricate clothing textures are shown. However, color bleeding may still occur at component boundaries, and there could be color discrepancies between the source and target images. The above two situations are the main reason we thought for the drop of CLIP-Score. This issue can be alleviated by manually adjusting the mask boundaries or by re-sampling. Please refer to the supplementary material and codes[1] for more details about experiments and implementation.

TABLE III: Quantitative comparison of the impact of component-based MoE on generation results. The bold highlight indicates the best-performing score.

Methods	FID↓	CLIP-S↑	SSIM↑	PD↓	SI↑
Ours	**19.89**	0.259	**0.826**	**1.47**	**0.947**
w/o MoE	21.64	**0.261**	0.826	1.50	0.943

V. Conclusion

We propose Human-MoE, designed for FHIS tasks. By combining multi-modal conditions: parsing maps, pose maps, and text annotations, our method generates human images consistent with the input conditions that are easily indicated. We innovatively apply the component-based MoE to enhance fidelity in specific areas, including the face, hands, and clothing. From the aspect of editing, our model is highly flexible and allows users to easily adjust the pose and appearance features of humans. We compare the results of our methods and related work on multiple quantitative metrics. The experiments show that our model achieves state-of-the-art performance in FID and SSIM, and is also comparable to related methods in other metrics.

References

[1] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara, "Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on," in Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 8580–8589.

[2] Wenju Xu, Chengjiang Long, Yongwei Nie, and Guanghui Wang, "Disentangled representation learning for controllable person image generation," IEEE Transactions on Multimedia, vol. 26, pp. 6065–6077, 2024.

[3] 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.

[4] Soon Yau Cheong, Armin Mustafa, and Andrew Gilbert, "Upgpt: Universal diffusion model for person image generation, editing and pose transfer," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4173–4182.

[5] 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.

[6] 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.

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," Advances in neural information processing systems, vol. 33, pp. 6840–6851, 2020.

[1]Codes will be available at https://github.com/JoeHuang1999/Human-MoE

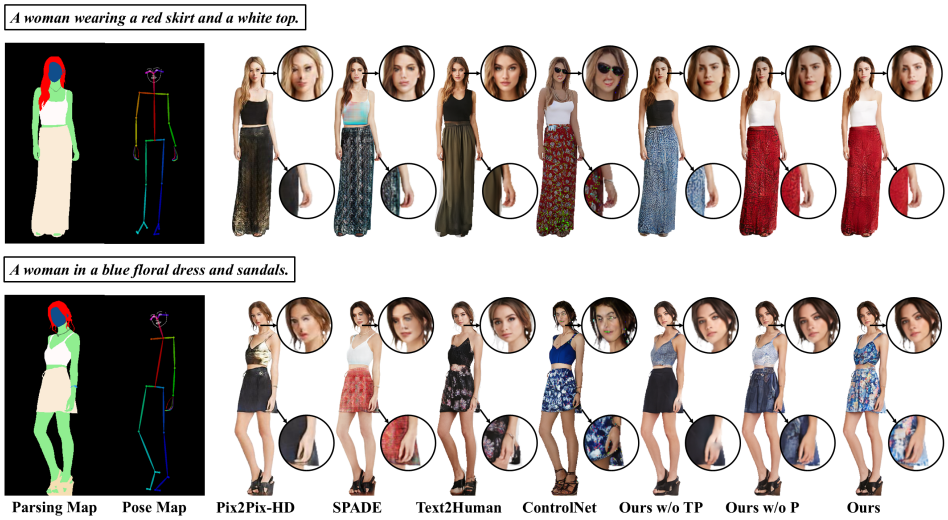

Fig. 3: Qualitative comparison on DeepFashion-Multimodal dataset. P and TP stand for pose and text-and-pose, respectively.

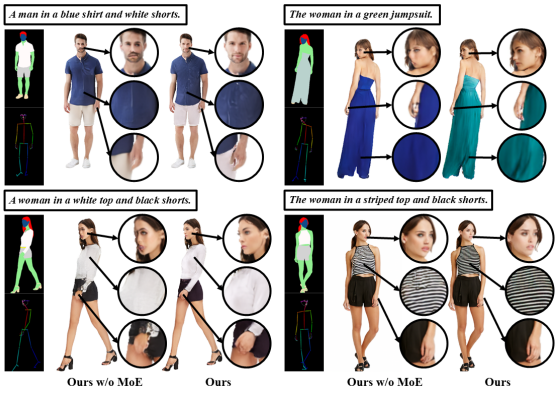

Fig. 4: Comparison of results with and without MoE. It can be observed that visual quality improves with component-based MoE.

[8] Jonathan Ho and Tim Salimans, "Classifier-free diffusion guidance," arXiv preprint arXiv:2207.12598, 2022.

[9] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin, "Spatext: Spatio-textual representation for controllable image generation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 18370–18380.

[10] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi, "Palette: Image-to-image diffusion models," in ACM SIGGRAPH 2022 conference proceedings, 2022, pp. 1–10.

[11] 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.

[12] Xiaoling Gu, Shengwenzhuo Xu, Yongkang Wong, Zizhao Wu, Jun Yu, Jianping Fan, and Mohan S Kankanhalli, "Multi2human: Controllable human image generation with multimodal controls," Neurocomputing, vol. 587, pp. 127682, 2024.

[13] Aaron Van Den Oord, Oriol Vinyals, et al., "Neural discrete representation learning," Advances in neural information processing systems, vol. 30, 2017.

[14] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Real-time 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] 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.

[16] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al., "Gemini: a family of highly capable multimodal models," arXiv preprint arXiv:2312.11805, 2023.

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

[18] 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.

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

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in International conference on machine learning. PMLR, 2023, pp. 19730–19742.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.