
MoLE : Enhancing Human-centric Text-to-image Diffusion via Mixture of Low-rank Experts

Jie Zhu^{1,2}, Yixiong Chen³, Mingyu Ding⁴, Ping Luo⁵, Leye Wang^{1,2*}, Jingdong Wang^{6*}

¹Key Lab of High Confidence Software Technologies (Peking University), Ministry of Education, China

²School of Computer Science, Peking University, Beijing, China, ³Johns Hopkins University

⁴UC Berkeley, ⁵The University of Hong Kong, ⁶Baidu

zhujie@stu.pku.edu.cn, ychen646@jh.edu, myding@berkeley.edu,

pluo@cs.hku.hk, leyewang@pku.edu.cn, wangjingdong@outlook.com



Abstract

Text-to-image diffusion has attracted vast attention due to its impressive image-generation capabilities. However, when it comes to human-centric text-to-image generation, particularly in the context of faces and hands, the results often fall short of naturalness due to insufficient training priors. We alleviate the issue in this work from two perspectives. **1)** From the data aspect, we carefully collect a *human-centric dataset* comprising over one million high-quality human-in-the-scene images and two specific sets of close-up images of faces and hands. These datasets collectively provide a rich prior knowledge base to enhance the human-centric image generation capabilities of the diffusion model. **2)** On the methodological front, we propose a simple yet effective method called **Mixture of Low-rank Experts (MoLE)** by considering low-rank modules trained on close-up hand and face images respectively as experts. This concept draws inspiration from our observation of low-rank refinement, where a low-rank module trained by a customized close-up dataset has the potential to enhance the corresponding image part when applied at an appropriate scale. To validate the superiority of MoLE in the context of human-centric image generation compared to state-of-the-art, we

*Corresponding author



Figure 2: **Inspiration of Mixture of Low-rank Experts.** In the first and second row, we train two low-rank modules on SD v1.5 simply using off-the-shelf Celeb-HQ face dataset [22] and 11k Hands dataset [1], respectively. *With a proper scale weight, low-rank module can refine corresponding part.* We term this phenomenon as low-rank refinement.

construct two benchmarks and perform evaluations with diverse metrics and human studies. Datasets, model, and code are released at [project website](#).

1 Introduction

Human-centric text-to-image generation is an important orientation for realistic applications, *e.g.*, poster design, virtual reality, *etc.* However, current models encounter issues with producing natural-looking results, particularly in the context of faces and hands². To address this issue, we delve into the matter and identify two factors that may contribute to this issue. Firstly, the absence of *high-quality* human-centric data makes diffusion models lack sufficient human-centric prior³; Secondly, in the human-centric context, faces and hands represent the two most complicated parts due to high variability, making them challenging to be generated naturally.

Hence, we alleviate this problem from two perspectives. On the one hand, we collect a human-in-the-scene dataset of high quality and high resolution from the Internet. Basically, the resolution of each image is various and over 1024×2048 . The dataset contains approximately one million images, and covers different races, various gestures, and activities, thereby providing diffusion models with sufficient knowledge to improve the performance of human-centric generation. However, for the second factor, our experiment in Sec 5.3 demonstrates though fine-tuning on above dataset brings an overall enhancement in human quality, human face and hand still exhibit unnatural outcomes, possibly because diffusion models focus more on overall performance during fine-tuning while struggling to accurately capture highly variable parts like face and hand gestures.

To address the second factor, an interesting *low-rank refinement phenomenon* inspires us. As shown in Fig 2, when combined with a customized low-rank module [19] and using a proper scale weight, Stable Diffusion v1.5 (SD v1.5) [40] has the potential to refine the corresponding part of a person, *e.g.*, for hand, $s = 0.4$ subtly refines the appearance of the woman drinking milk’s hand. Thus, inspired by this, to refine the face and hand, we can first gather two customized high-quality datasets (one for face close-ups and one for hand close-ups) to train two low-rank modules, respectively. Then, for the two specialized low-rank modules, we could add a certain assignment to adaptively select which low-rank module to use for a given input and Mixture of Experts (MoE) naturally stands out. Moreover, as face and hand often appear simultaneously in an image for a person, motivated by Soft MoE [35], we could adopt a soft assignment to produce adaptive scale weights, activating multiple experts to handle the input at the same time. We refer to all mentioned three datasets above together as *human-centric dataset* for convenience as shown in Fig 3.

²The HuggingFace website acknowledges that “Faces and people in general may not be generated properly.” Despite some efforts, such as those seen in DALL-E 3 and Midjourney, aimed at addressing this issue, they are generally closed-sourced for business. Our objective is to establish transparent and open-source endeavors for the advancement of the community in generating more realistic human hands/faces.

³We sample 350k human-centric images from LAION2b-en and the average height and width are 455 and 415, among which most are between 320 and 357, limited in providing sufficient human-centric knowledge.



Figure 3: Some showcases of our human-centric dataset.

In the end, we propose a simple yet effective method called **Mixture of Low-rank Experts (MoLE)**. Our key insight is to regard low-rank modules trained on customized datasets as specialized experts of MoE and adaptively activate them in a soft form. By adopting SD v1.5 as a showcase, our method basically contains three stages: We fine-tune SD v1.5 on collected dataset to complement sufficient human-centric prior; Then we use two close-up datasets, *i.e.*, close-ups of face and hand images, to train two low-rank experts separately; Finally, we formulate these two low-rank experts in an MoE form and integrate them with the base model in an adaptive soft assignment manner. To evaluate MoLE, we construct two human-centric benchmarks using prompts from COCO Caption [4] and DiffusionDB [48]. The results based on SD v1.5, v2.1, and XL consistently suggest the superiority and generalization of MoLE. Our contribution can be summarized as:

- We carefully collect a human-centric dataset comprising over one million high-quality images to provide sufficient human-centric prior. Importantly, we include two high-quality close-up of face and hand subsets, especially for hand. To the best of our knowledge, such a high-quality close-up hand dataset is absent in prior related studies. Click [here](#) to have a look.
- We find the low-rank refinement phenomenon and are inspired to propose MoLE, a simple yet effective method. MoLE integrates low-rank modules trained on customized hand and face datasets as experts in an MoE framework with soft assignment⁴ to enable flexible activation.
- We construct two human-centric evaluation benchmarks from DiffusionDB and COCO Caption. MoLE consistently demonstrates improvement over the state-of-the-art and exhibits broad generalization across SD v1.5, v2.1, and XL in human-centric generation.

2 Related Work

Text-to-image generation. Diffusion model [17, 46], especially in text-to-image generation, has been widely used since its proposal, *e.g.*, GLIDE [34] with classifier-free guidance [18], Imagen [42] with a large T5 text encoder [37], Stable Diffusion [40] using VAE encoded latents, and DALL-E 2 [38] using CLIP [36]. Different from them, our work primarily focuses on human-centric text-to-image generation. Though a concurrent work [28] also aims to improve human-centric generation, it does not explicitly consider the issue of face and hand and thereby the related generation is still unnatural. In contrast, MoLE is specially designed for this issue.

Mixture-of-Experts. In MoE [20], different subsets of data or contexts may be better modeled by distinct experts. Theoretically, MoE could scale model capability with little cost by using sparsely-gated MoE layer [44]. Recently, MoE has been adapted in generation tasks [12, 2]. For example, ERNIE-ViLG [12] uniformly divides the denoising process into several distinct stages, with each being associated with a specific model. eDiff-i [2] calculates thresholds to separate the whole process into three stages. Differing from employing experts in divided stages, we consider low-rank modules trained by customized datasets as experts to adaptively refine generation. For more discussion about related work, we put in Appendix A.11.

⁴Considering that the human face and hand are generally the most frequently observed parts in an image and their bad cases are also extensively discussed or complained in image generation communities, thereby in this work, we primarily focus on the two most important and urgent parts. Our work could also easily involve other human parts, *e.g.*, feet, by collecting a close-up of feet dataset, training an extra low-rank feet expert, and accordingly modifying the parameter of the soft assignment.

| | BLIP2 | ClipCap | MiniGPT-4 | LLaVA |
|---|--|---|--|--|
|  | a man in a black jacket leaning against a wall | a man standing in front of a wall with graffiti on it and wearing a black jacket and black shirt | The image shows a young man standing in front of a graffiti-covered brick wall, wearing a black leather jacket and looking off into the distance with a serious expression on his face. | A young man with a black jacket and a black hoodie is standing in front of a graffiti-covered brick wall, looking at the camera with a serious expression. |
|  | a little girl is playing in the water | a little girl is playing in the water by a pond with a rock and grass in the background and a waterfall in the foreground | This image shows a small child sitting on a rock in a pond, looking into the water with a curious expression on their face. The child is wearing a brown dress and has dark hair. The pond is surrounded by trees and rocks, and there is a small waterfall in the background. The water in the pond is clear and reflective, creating a mirror effect. The child's reflection can be seen in the water. The overall mood of the image is peaceful and serene. | A young girl in a brown dress squatting near a small pond, looking into the water and admiring her reflection, while a fountain is also present in the background. |
| | Prompt for MiniGPT-4 and LLaVA: Describe this image in one sentence with details. | | | |

Figure 4: The results of four captioning models. Texts in red are inaccurate descriptions and texts in green are detailed correct descriptions. LLaVA presents a good balance between the level of detail and error rate, and thus is chosen for captioning our dataset.

3 Human-centric Dataset

Overview. Our human-centric dataset involves over one million high-quality images, containing three parts (See Sec 3.1). As shown in Fig 3, these images are diverse w.r.t. occasions, activities, gestures, ages, genders, and racial backgrounds. Specifically, approximately 57.33% individuals identify as White, 14.68% as Asian, 9.98% as Black, 5.11% as Indian, 5.52% as Latino Hispanic, and 7.38% as Middle Eastern⁵. Approximately 58.18% are male and 41.82% are female. For age, approximately 0.93% are babies (0-1 years old), 3.55% are kids (2-11 years old), 4.60% are teenagers (12-18 years old), 84.86% are adults (18-60 years old), and 6.06% are elderly (over 60 years old).

Ethical & legal compliance. Our collection is in compliance with the ethics and law as all images are collected from websites under Public Domain CC0 1.0⁶ license that allows free use, redistribution, and adaptation for non-commercial purposes. To avoid concerns, please see our license and privacy statement in Appendix A.2. Note that this dataset is allowed for academic purposes only. When using it, the users are requested to ensure compliance with ethical and legal regulations.

3.1 Human-centric Dataset Constitution

Human-in-the-scene images. We primarily collect high-resolution human-centric images from Internet and the image resolution is basically over 1024×2048 , providing sufficient priors for diffusion models. To enable training, we use a sliding window (1024×1024) to crop the image to maintain as much information as possible. However, for an image, high resolution does not mean high quality. Therefore, we train a VGG19 [45] to filter out blurred images. Additionally, considering the crop operation could generate images that are full of background or contain little information about people, we train a VGG19 [45] to filter out such bad cases⁷. To ensure the quality, we repeat the two processes multiple times until we do not find any case mentioned above in three times of random sampling. By employing these strategies, we can remove amounts of noise and useless images, thereby guaranteeing the image quality.

Close-up of face images. The face dataset contains two sources: the first is from Celeb-HQ [22] in which we choose images of high quality with size 1024×1024 ; The second is from Flickr-Faces-HQ (FFHQ) [23]. We sample images covering different skin color, age, sex, and race. There are around 6.4k face images. We do not sample more face images as it is sufficient for low-rank expert training.

Close-up of hand images. The hand dataset contains three sources: the first is from 11k Hands [1] where we randomly sample around 1k high-quality images and manually crop them to square; The second is from the Internet where we collect hand images of high quality and resolution with simple backgrounds and use YOLOv5 [9] to detect hands and crop them with details maintained; The third is from human-in-the-scene images (before processing) where we sample 8k images. We check every image and manually crop the hand of the image to square if the image is appropriate and the hand is clear. In this close-up hand dataset, there are abundant hand gestures and scenarios shown in Fig 3, e.g., holding a flower, writing, etc. There are 7k high-quality hand images. To the best of our knowledge, such a high quality close-up hand dataset is absent in prior related studies.

⁵Similar to famous FFHQ dataset [23], our dataset inevitably inherits the potential bias of target websites.

⁶<https://creativecommons.org/publicdomain/zero/1.0/>

⁷See Appendix A.3 for training details of above two filters and illustration of bad samples.

3.2 Image Caption Generation

When collecting the dataset, we primarily consider image quality and resolution, neglecting whether it is text paired so as to increase image amount. Thus, producing a caption for each image is required. We investigate four recently proposed SOTA models including BLIP-2 [26], ClipCap [32], MiniGPT-4 [53], and LLaVA [29]. We show several cases in Fig 4. One can see that BLIP-2 usually produces simple descriptions and ignores details. ClipCap has a better performance but still lacks sufficient details along with the wrong description. MiniGPT4, although gives detailed descriptions, is inclined to spend a long time (17s on average) generating long and inaccurate captions that exceed the input limit (77 tokens) of the Stable Diffusion CLIP text encoder [36]. In contrast, LLaVA produces neat descriptions in one sentence with accurate details in a short period (3-5s). Afterward, we manually streamline long LLaVA caption with a new shorter caption by ourselves while aligning with the content of the image. We also remove unrelated and uninformative text patterns, *e.g.*, “The image features that . . .”, “showcasing . . .”, “creating . . .”, “demonstrating . . .”, *etc.* To further ensure the caption alignment of LLaVA, we use CLIP to filter image-text pairs with lower scores.

4 Method

4.1 Preliminary

Low-rank Adaptation (LoRA). Given a customized dataset, instead of training the entire model, LoRA [19] is designed to fine-tune the “residual” of the model, *i.e.*, ΔW :

$$W' = W + \Delta W \quad (1)$$

where ΔW is decomposed into low-rank matrices: $\Delta W = AB^T$ ($A \in \mathbb{R}^{n \times d}$, $B \in \mathbb{R}^{m \times d}$, $d < n$, and $d < m$). During training, we can simply fine-tune A and B instead of W , making fine-tuning on customized dataset memory-efficient. In the end, we get a small model as A and B are much smaller than W .

Mixture-of-Experts (MoE). MoE [20, 44, 24] is designed to enhance the predictive power of models by combining the expertise of multiple specialized models. Usually, a central “gating” model $G(\cdot)$ selects which specialized model to use for a given input:

$$y = \sum_{i=1} G(x)_i E_i(x). \quad (2)$$

When $G(x)_i = 0$, the corresponding expert E_i will not be activated.

4.2 Mixture of Low-rank Experts

Motivated by the two potential factors discussed in Sec 1, our method contains three stages as shown in Fig 5. We describe each stage below and put the training details in Appendix A.1.

Stage 1: Fine-tuning on Human-centric Dataset. The overall poor performance of human-centric generation could be attributed to the absence of large-scale high-quality datasets. Considering such a pressing need, our work bridges this gap by providing a carefully collected dataset that contains around one million human-centric images of high quality. To learn as much prior as possible, we adopt SD v1.5 as a baseline and leverage the whole human-centric datasets to fine-tune. Concretely, we fine-tuning the UNet modules [41] (and text encoder) while fixing the rest parameters. The well-trained model is then sent to the next stage.

Stage 2: Low-rank Expert Generation. To construct MoE, in this stage, our goal is to prepare two experts that are supposed to contain abundant knowledge about the corresponding part. To achieve this, we train two low-rank modules using two customized datasets. One is the close-up face dataset. The other is the close-up hand dataset that contains abundant hand gestures, full details with simple backgrounds, and interactions with other objects. We then use the two datasets to train two low-rank experts with SD v1.5 trained in stage 1 as the base model. The low-rank experts are expected to focus on the generation of face and hand and learn useful context.

Stage 3: Soft Mixture Assignment. This stage is motivated by the low-rank refinement phenomenon in Fig 2 where a specialized low-rank module using a proper scale weight is able to refine the corresponding part of a person. Hence, the key is to activate different low-rank modules with suitable weights. From this view, MoE naturally stands out and we novelly regard a low-rank module trained

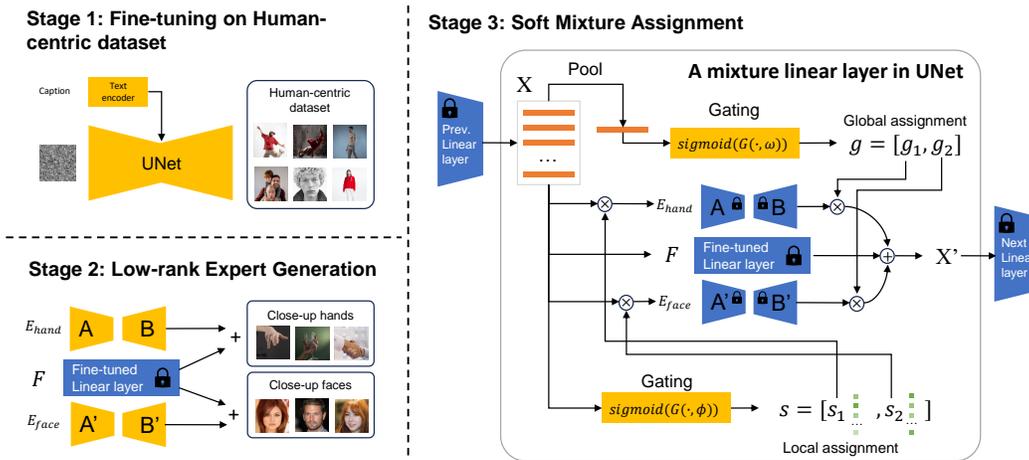


Figure 5: The framework of MoLE. X is the input of any linear layers in UNet. A and B are low-rank matrices.

on a customized dataset, *e.g.*, face dataset, as an expert and formulate in a MoE form. Moreover, for a person, face and hand would appear in an image simultaneously while hard assignment mode in MoE only allows one expert accessible to the given input. Hence, inspired by Soft MoE [35], we adopt a soft assignment, allowing multiple experts to handle input simultaneously. Further, considering that the face and hand would be a part of the whole image (local) or occupy the whole image (global), besides global assignment in original MoE, we novelly introduce local assignment. Specifically, considering a linear layer F from UNet and its input $X \in \mathbb{R}^{n \times d}$ where n is the number of token and d is the feature dimension, we illustrate local and global assignment respectively.

For local assignment, we employ a local gating network that contains a learnable gating layer $G(\cdot, \phi)$ ($\phi \in \mathbb{R}^{d \times e}$, e is the number of experts and here e is 2, below is the same.) and a sigmoid function. The gating network is to produce two normalized score maps $s = [s_1, s_2]$ ($s \in \mathbb{R}^{n \times e}$, here e is 2) for two low-rank experts as formulated:

$$s = \text{sigmoid}(G(X, \phi)) \quad (3)$$

For global assignment, we use a global gating network including an AdaptiveAvePool, a learnable gating layer $G(\cdot, \omega)$ ($\omega \in \mathbb{R}^{d \times e}$, here e is 2), and a sigmoid function. This gating network is to produce two global scalars $g = [g_1, g_2]$ ($g \in \mathbb{R}^e$, e is 2) for two experts as formulated:

$$g = \text{sigmoid}(G(\text{Pool}(X), \omega)). \quad (4)$$

The soft mechanism is built on the fact that each token can adaptively determine how much (weight) should be sent to each expert by the sigmoid function. And intuitively, the weight of every token for two experts is independent as face and hand experts are not competitors during generation. Thus we do not use softmax.

For combination, we first send X to each low-rank expert E_{face} and E_{hand} respectively, use s_1 and s_2 ($\mathbb{R}^{n \times 1}$) to perform element-wise multiplication (local assignment), and also perform global control by scalars g_1 and g_2 (global assignment)⁸:

$$Y_1 = E_{\text{face}}(X \cdot s_1 \cdot g_1) = g_1 \cdot E_{\text{face}}(X \cdot s_1) \quad (5)$$

$$Y_2 = E_{\text{hand}}(X \cdot s_2 \cdot g_2) = g_2 \cdot E_{\text{hand}}(X \cdot s_2) \quad (6)$$

Then we add Y_1 and Y_2 back to the output of a linear layer F from UNet with X as input, producing a new output X' :

$$X' = F(X) + Y_1 + Y_2 \quad (7)$$

In the end, to endow the model with the capability to adaptively activate experts, we use our human-centric dataset to train learnable gating layers while freezing the base model and two low-rank experts.

5 Experiment

5.1 Evaluation Benchmarks and Metrics

Considering that our work primarily focuses on human-centric image generation, before presenting our experiment, we introduce two customized evaluation benchmarks. Additionally, since our

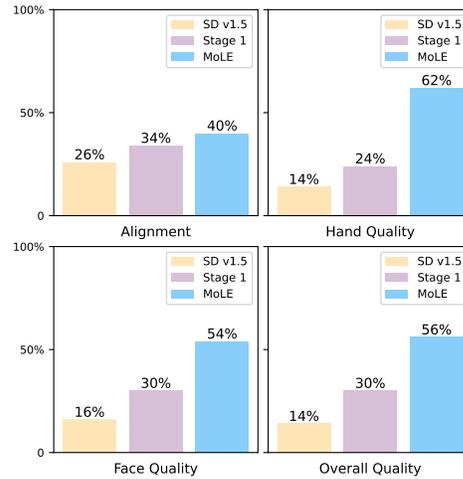
⁸Recalling that each expert is two low-rank matrices, g_1 and g_2 can transition from within E_{face} and E_{hand} to outside of them.

Table 1: The performance of MoLE on COCO Human Prompts and DiffusionDB Human Prompts.

| Model | COCO Human Prompts | |
|---------------------|--------------------|---------------|
| | HPS (%) | IR (%) |
| VQ-Diffusion | 19.21 ± 0.04 | -12.51 ± 2.44 |
| Versatile Diffusion | 19.75 ± 0.09 | -8.81 ± 1.40 |
| SDXL | 20.84 ± 0.06 | 73.34 ± 2.29 |
| SD v1.5 | 19.91 ± 0.09 | 28.34 ± 1.40 |
| MoLE (SD v1.5) | 20.27 ± 0.07 | 33.75 ± 1.49 |
| MoLE (SDXL) | 21.36 ± 0.02 | 98.52 ± 0.61 |

| Model | DiffusionDB Human Prompts | |
|---------------------|---------------------------|---------------|
| | HPS (%) | IR (%) |
| VQ-Diffusion | 19.00 ± 0.02 | -18.42 ± 1.49 |
| Versatile Diffusion | 20.09 ± 0.04 | -29.05 ± 2.72 |
| SDXL | 21.51 ± 0.07 | 87.88 ± 2.53 |
| SD v1.5 | 20.29 ± 0.01 | -2.72 ± 1.66 |
| MoLE (SD v1.5) | 20.62 ± 0.04 | 4.36 ± 1.36 |
| MoLE (SDXL) | 22.35 ± 0.01 | 105.25 ± 1.15 |

Figure 6: User study in four aspects.



generated images are human-centric, they should intuitively meet human preferences. Hence, we primarily adopt two human preference metrics including Human Preference Score (HPS) [49] and ImageReward (IR) [50]. We describe all of them below. Besides the two metrics, we also perform user studies by inviting people to compare the generated images with their own preferences.

Benchmark 1: COCO Human Prompts. We construct this benchmark by leveraging the caption in COCO Caption [4] that has been widely used in previous work [49, 50, 6]. Concretely, we use the captions in the COCO val set, and preserve the caption that contains human-related words, *e.g.*, woman, man, kid, girl, boy, person, teenager, *etc.* In the end, we have around 60k prompts left, dubbed as COCO Human Prompts.

Benchmark 2: DiffusionDB Human Prompts. DiffusionDB [48] is the first real-users-specified large-scale text-to-image prompt dataset containing around 14 million prompts. We construct this benchmark by leveraging its 2M caption set version. Concretely, we first filter out the NSFW prompts by the indicator provided in DiffusionDB. Then we preserve captions containing human-related words. We filter out prompts containing special symbol, *e.g.*, [,], *etc.* In the end, we have around 64k prompts left, dubbed as DiffusionDB Human Prompts.

Metric 1: Human Preference Score (HPS). Human Preference Score (HPS) [49] measures how images present with human preference. It leverages a human preference classifier fine-tuned on CLIP [36].

Metric 2: ImageReward (IR). Different from HPS, ImageReward (IR) [50] is built on BLIP [27]. IR is a zero-shot evaluation metric for understanding human preference in text-to-image synthesis.

5.2 Main Results

Superiority. To evaluate the performance, following previous work [40, 50, 49, 6] we randomly sample 3k prompts from COCO Human Prompts benchmark and 3k prompts from DiffusionDB Human Prompts benchmark to generate images and calculate metrics, HPS and IR, and compare MoLE with open-resource SOTA method with different model structures including VQ-Diffusion [13], Versatile Diffusion [51], our baseline SD v1.5 [40] and its largest variant SDXL. We repeat the process three times and report the averaged results and standard error in Tab 1. MoLE outperforms VQ-Diffusion and Versatile Diffusion and significantly improves our baseline SD v1.5 in both metrics, implying that MoLE could generate images that are more natural to meet human preference. In Appendix A.5, we also evaluate on CLIP-T [25], FID [16], and Aesthetic Score [43], to present a comprehensive comparison. Besides, we illustrate generated images of different models in Fig 1 and Fig 13. Though MoLE is inferior to SDXL in HPS and IR⁹, we qualitatively compare the face and hand of generated images from MoLE and SDXL in Fig 1 and Fig 13, and our results look more realistic with natural hand/face even under high HPS gap (in 2nd row Fig 1 20.39 vs. 22.38). Even though, to provide a more convincing demonstration of the efficacy of our method, we proceed to implement MoLE on SDXL (See Generalization part below). We also compare MoLE with relevant methods like HanDiffuser [33] and HyperHuman [30] in Appendix A.5. Finally, MoLE is resource-friendly and can be trained in a single A100 80G GPU.

⁹MoLE is built on SD v1.5, which has a significant gap in model size (5.1G vs. 26.4G) and output resolution (512 vs. 1024) compared to SDXL.

Generalization. Besides SD v1.5, to verify the generalization of our method, we further construct MoLE on SDXL in Tab 1. We also consider SD v2.1 and transformer-based PixArt- α in Appendix A.5. In Tab 1, we see that based on SDXL, MoLE outperforms SDXL by a large margin. Similar conclusion can be observed in SD v2.1 and PixArt- α . These results demonstrate great generalization of our method. Also, we qualitatively compare images generated by MoLE (SDXL) and SDXL in Appendix A.8 where MoLE generates more natural hands/faces.

5.3 Ablation Study

Stage Enhancement. To figure out how each stage of MoLE enhances the generation performance, we use MoLE built on SD v1.5 to conduct the experiments on COCO Human Prompts by randomly sampling 3k prompts to generate images from each stage with the same seed in Sec 5.2 and calculate the HPS and IR. The whole process repeats three times as well. The results are reported in Tab 2. In Fig 7, we also illustrate examples generated by different stages with the same seeds to intuitively show the role of each stage. We see that fine-tuning on human-centric dataset (Stage 1) is effective in improving overall quality, proved by enhanced HPS and IR, implying the importance of the dataset. However, when adding Stage 2, *i.e.*, both experts are employed, the performance drops. We speculate that, due to training on close-up datasets, the employed experts tend to mimic their training data’s distribution and thereby harm the generation process, *e.g.*, in Fig 7 Stage 2, close-up image of face resemble the distribution of face image in FFHQ dataset [23]. Luckily, as shown in Fig 7 MoLE, adding Stage 3 allows model to adaptively activate two experts with soft assignments and thereby alleviates this issue, verifying its importance. Also, Stage 3 refines human face and hand compared to Stage 1 and thereby outperforms Stage 1 on HPS and IR in Tab 2. Finally, to further verify MoLE’s advantage over SD v1.5 and demonstrate the significance of Stage 3, we conduct user studies by sampling 20 sets of images from SD v1.5, fine-tuned SD (Stage 1), and MoLE, and inviting 50 participants to select the best one from each set according to their preference in four aspects respectively. In Fig 6 we see that Stage 1 improves overall performance but is limited in improving face and hand quality while MoLE obtains highest voting, especially in hand and face quality, indicating that Stage 1 alone is insufficient, and Stage 3 also holds significance.

Table 2: Ablation study on each stage using COCO Human Prompts.

| Stage | HPS (%) | IR (%) |
|----------|------------------|------------------|
| SD v1.5 | 19.91 \pm 0.09 | 28.34 \pm 1.40 |
| +Stage 1 | 20.16 \pm 0.09 | 31.01 \pm 1.75 |
| +Stage 2 | 19.94 \pm 0.07 | 25.66 \pm 2.72 |
| +Stage 3 | 20.27 \pm 0.07 | 33.75 \pm 1.49 |

Table 3: Ablation study on assignment manner using COCO Human Prompts.

| Method | HPS (%) | IR (%) |
|---------|------------------|------------------|
| SD v1.5 | 19.91 \pm 0.09 | 28.34 \pm 1.40 |
| Local | 20.19 \pm 0.04 | 31.98 \pm 2.41 |
| Global | 20.20 \pm 0.02 | 32.20 \pm 0.86 |
| Both | 20.27 \pm 0.07 | 33.75 \pm 1.49 |



Figure 7: Showcases of image generated by different stages. Top prompt: a woman in a purple top pulling food out of a oven. Bottom prompt: smiling woman in red top putting items in a box.

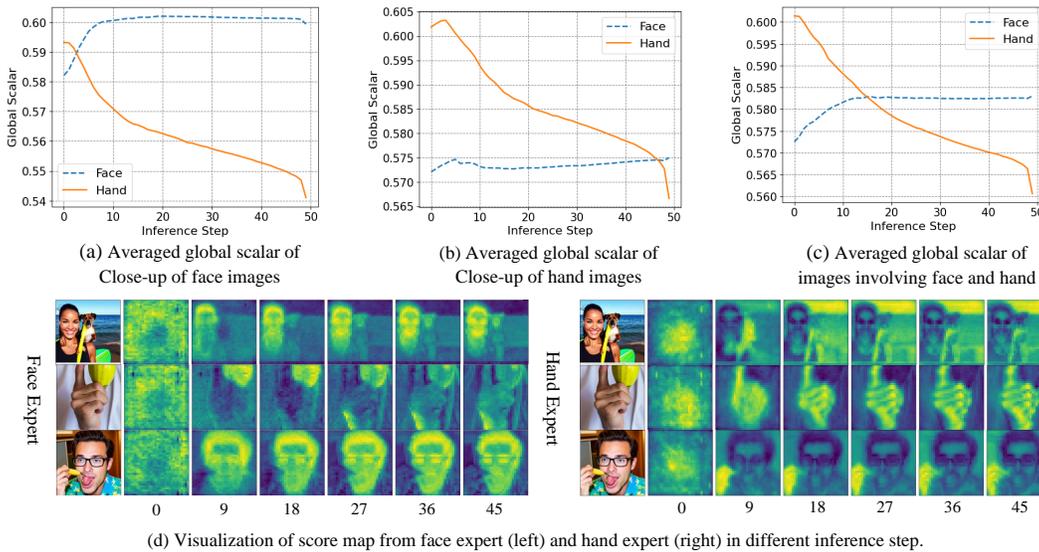


Figure 8: The averaged global and local assignment weights in different inference steps.

Mixture Assignment. In MoLE, we use two kinds of mixture manners including local and global assignment. Hence, we ablate the two assignments and present the results in Tab 3. It can be seen that both local and global assignments can enhance performance. When combining them together, the performance is further improved, indicating the effectiveness of our mixture manners. Moreover, we illustrate how the two assignments work in Fig 8. For global assignment, we average the global scalars of 20 close-up face images, 20 close-up hand images, and 20 normal human images involving hand and face respectively in every inference step in Fig 8 (a), (b), and (c). In (a) and (b), when generating different close-ups, the corresponding expert generally produces higher global value, implying that global assignment is content-aware. In (c), E_{face} and E_{hand} achieve a balance. Besides, as inference progresses the global scalar of E_{hand} always drops while that of E_{face} is relatively flat. We speculate, in light of the diversity of hands (*e.g.*, various gestures), E_{hand} tends to establish general content in the early stage while E_{face} must meticulously fulfill facial details throughout the denoise process due to fidelity requirement. For local assignment, we visualize averaged score maps of sampled images from the two experts respectively in Fig 8 (d). We see that as inference progresses, local assignment of the two experts can highlight and gradually refine corresponding parts, verifying its effectiveness. We also provide the distribution of the local weight sent to each expert in Appendix A.6. Additionally, to understand the importance of using two experts on the model’s performance, we train only one expert using all close-up images and put the comparison results in Appendix A.7.

5.4 More Visualizations and Analysis

We show more human-centric images with natural face and hand in Appendix A.8. Surprisingly, MoLE can also generate non-human-centric images¹⁰ in Appendix A.9, *e.g.*, animals and scenery. In the end, we analyze failure cases in Appendix A.10, which we find are attributed to the large L2 norm of outputs from face and hand experts.

6 Discussion

To further highlight our method contribution, below we present a comprehensive discussion on distinctions between MoLE and conventional MoE methods from three aspects. We also discuss the contribution of our curated human-centric dataset and analysis about ratios of different races in Appendix A.12. Firstly, from the aspect of training, MoLE independently trains two experts to learn completely different knowledge using two customized close-up datasets. In contrast, conventional MoE methods simultaneously train experts and base models using the same dataset. Secondly, from the aspect of expert structure and assignment manner, MoLE simply uses two low-rank matrices while conventional MoE methods use MLP or convolutional layers. Moreover, MoLE combines local

¹⁰We speculate this is because the human-centric dataset also contains these entities that interact with humans in an image. Hence, the model learns these concepts. However, it is worth noting that MoLE may not be better at generic image generation than the generic models as MoLE is trained on a human-centric dataset.

and global assignments together for a finer-grained assignment while conventional MoE methods only use global assignment. Finally, from the aspect of applications in computer vision, MoLE is proposed for text-to-image generation while conventional MoE methods are mainly used in object recognition, scene understanding, *e.g.*, V-MoE [39]. Though MoE recently has been employed in image generation, *e.g.*, ERNIE-ViLG [12] and eDiff-i [2] that employ experts in divided stages, MoLE differs from them – inspired by low-rank refinement in Fig 2, MoLE consider low-rank modules trained by customized datasets as experts to adaptively refine image generation.

7 Conclusion

In this work, we primarily focus on the human-centric text-to-image generation that has important real-world applications but often suffers from producing unnatural results due to insufficient prior, especially the face and hand. To mitigate this issue, we carefully collect and process one million high-quality human-centric images, aiming to provide sufficient prior. Besides, we observe that a low-rank module trained on a customized dataset, *e.g.*, face, has the capability to refine the corresponding part. Inspired by it, we propose a simple yet effective method called Mixture of Low-rank Experts (MoLE) that effectively allows diffusion models to adaptively select experts to enhance the generation quality of corresponding parts. We also construct two customized human-centric benchmarks from COCO Caption and DiffusionDB to verify the superiority of MoLE.

8 Limitation & Future Work

Honestly, although our experiments confirm MoLE’s effectiveness in enhancing human-centric image generation, there is still considerable room for improvement. Our method struggles with scenarios involving multiple individuals, likely due to our dataset being primarily single-person images and uncertainty about the applicability of observations in Fig 2 to such cases. Additionally, in generating images using identical prompts, we observe that only about 25% of MoLE’s results are of high quality, a remarkable improvement over SDXL’s 10%, but still below practical standards. Potential reasons include insufficient close-up data for hands and faces and the need for further tuning of hyperparameters. Future work will focus on model optimization, improving data quality, and enhancing dataset diversity to better represent various demographics and real-world scenarios.

9 Broader Impact

MoLE mainly focuses on enhancing human-centric text-to-image generation in diffusion models. It refrains from introducing any harmful content to the community and society. However, though MoLE may not introduce more biases on race, it also inherits the biases in the training data like pervious methods. Hence it will be more meaningful to enhance the diversity of our collected dataset to represent different demographics and real-world scenarios better. As for other impacts such as fake faces, it also inevitably generates fake faces like other generative models, which requires users to leverage these generated images carefully and legally. We highlight that these issues also warrant further research and consideration. We maintain transparency in our methods with open-source code and dataset composition, allowing for continuous improvement based on community feedback.

Acknowledgments

The authors thank the anonymous reviewers for constructive comments. The authors also thank Qi Zhang, Xin Li, Boqiang Duan, Teng Xi, and Gang Zhang for discussion and help.

References

- [1] M. Afifi. 11k hands: Gender recognition and biometric identification using a large dataset of hand images. *Multimedia Tools and Applications*, 78:20835–20854, 2019.
- [2] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, K. Kreis, M. Aittala, T. Aila, S. Laine, B. Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

- [3] J. Chen, Y. Jincheng, G. Chongjian, L. Yao, E. Xie, Z. Wang, J. Kwok, P. Luo, H. Lu, and Z. Li. Pixart-*alpha*: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations*.
- [4] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.
- [5] X. Chen, C. Liang, D. Huang, E. Real, K. Wang, H. Pham, X. Dong, T. Luong, C.-J. Hsieh, Y. Lu, et al. Symbolic discovery of optimization algorithms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 49205–49233, 2023.
- [6] Y. Chen. X-iqe: explainable image quality evaluation for text-to-image generation with visual large language models. *arXiv preprint arXiv:2305.10843*, 2023.
- [7] Z. Chen, M. Ding, Y. Shen, W. Zhan, M. Tomizuka, E. Learned-Miller, and C. Gan. An efficient general-purpose modular vision model via multi-task heterogeneous training. *arXiv preprint arXiv:2306.17165*, 2023.
- [8] Z. Chen, Y. Shen, M. Ding, Z. Chen, H. Zhao, E. G. Learned-Miller, and C. Gan. Mod-squad: Designing mixtures of experts as modular multi-task learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11828–11837, 2023.
- [9] R. Couturier, H. N. Noura, O. Salman, and A. Sider. A deep learning object detection method for an efficient clusters initialization. *arXiv preprint arXiv:2104.13634*, 2021.
- [10] N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. W. Yu, O. Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR, 2022.
- [11] W. Fedus, B. Zoph, and N. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270, 2022.
- [12] Z. Feng, Z. Zhang, X. Yu, Y. Fang, L. Li, X. Chen, Y. Lu, J. Liu, W. Yin, S. Feng, et al. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10135–10145, 2023.
- [13] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022.
- [14] T. Hang, S. Gu, C. Li, J. Bao, D. Chen, H. Hu, X. Geng, and B. Guo. Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7441–7451, 2023.
- [15] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2022.
- [16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6629–6640, 2017.
- [17] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 6840–6851, 2020.
- [18] J. Ho and T. Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.
- [19] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.
- [20] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.
- [21] X. Ju, A. Zeng, C. Zhao, J. Wang, L. Zhang, and Q. Xu. Humansd: A native skeleton-guided diffusion model for human image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15988–15998, 2023.

- [22] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.
- [23] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.
- [24] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2020.
- [25] D. Li, J. Li, and S. C. Hoi. Blip-diffusion: pre-trained subject representation for controllable text-to-image generation and editing. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 30146–30166, 2023.
- [26] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023.
- [27] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.
- [28] S. Li, J. Fu, K. Liu, W. Wang, K.-Y. Lin, and W. Wu. Cosmicman: A text-to-image foundation model for humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6955–6965, 2024.
- [29] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 34892–34916, 2023.
- [30] X. Liu, J. Ren, A. Siarohin, I. Skorokhodov, Y. Li, D. Lin, X. Liu, Z. Liu, and S. Tulyakov. Hyperhuman: Hyper-realistic human generation with latent structural diffusion. In *The Twelfth International Conference on Learning Representations*.
- [31] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.
- [32] R. Mokady, A. Hertz, and A. H. Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.
- [33] S. Narasimhaswamy, U. Bhattacharya, X. Chen, I. Dasgupta, S. Mitra, and M. Hoai. Handdiffuser: Text-to-image generation with realistic hand appearances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2468–2479, 2024.
- [34] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022.
- [35] J. Puigcerver, C. R. Ruiz, B. Mustafa, and N. Houlsby. From sparse to soft mixtures of experts. In *The Twelfth International Conference on Learning Representations*.
- [36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.
- [37] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.
- [38] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- [39] C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. S. Pinto, D. Keysers, and N. Houlsby. Scaling vision with sparse mixture of experts. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, pages 8583–8595, 2021.
- [40] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

- [41] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.
- [42] C. Saharia, W. Chan, S. Saxena, L. Lit, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. Gontijo-Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 36479–36494, 2022.
- [43] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: an open large-scale dataset for training next generation image-text models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 25278–25294, 2022.
- [44] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2016.
- [45] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations*, 2015.
- [46] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020.
- [47] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023.
- [48] Z. Wang, E. Montoya, D. Munechka, H. Yang, B. Hoover, and P. Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In *Annual Meeting of the Association for Computational Linguistics*, 2023.
- [49] X. Wu, K. Sun, F. Zhu, R. Zhao, and H. Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023.
- [50] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong. Imagereward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 15903–15935, 2023.
- [51] X. Xu, Z. Wang, G. Zhang, K. Wang, and H. Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7754–7765, 2023.
- [52] Y. Zhou, T. Lei, H. Liu, N. Du, Y. Huang, V. Y. Zhao, A. Dai, Z. Chen, Q. Le, and J. Laudon. Mixture-of-experts with expert choice routing. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 7103–7114, 2022.
- [53] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. Minigt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*.
- [54] J. Zhu, L. Wang, and X. Han. Safety and performance, why not both? bi-objective optimized model compression toward ai software deployment. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, pages 1–13, 2022.
- [55] J. Zhu, L. Wang, X. Han, A. Liu, and T. Xie. Safety and performance, why not both? bi-objective optimized model compression against heterogeneous attacks toward ai software deployment. *IEEE Transactions on Software Engineering*, 2024.
- [56] J. Zhu, J. Zha, D. Li, and L. Wang. A unified membership inference method for visual self-supervised encoder via part-aware capability. *arXiv preprint arXiv:2404.02462*, 2024.

A Appendix

A.1 Implementation Details

Stage 1: Fine-tuning on human-centric Dataset. We use Stable Diffusion v1.5 as base model and fine-tune the UNet (and text encoder) with a constant learning rate $2e - 6$. We set batch size to 64 and train with the Min-SNR weighting strategy [14]. The clip skip is 1 and we train the model for 300k steps using Lion optimizer [5]. We use this stage on SD v1.5. For SD v2.1 and SDXL, we do not use this stage to fine-tune the base models as their overall human-centric generation is relatively satisfying but only looks poor on the details of face and hand.

Stage 2: Low-rank Expert Generation. For face expert, we set batch size to 64 and train it 30k steps with a constant learning rate $2e - 5$. The rank is set to 256 and AdamW optimizer is used. For hand expert, we set batch size to 64. Since hand is more sophisticated than face to generate, we train it 60k steps with a smaller learning rate $1e - 5$. The rank is also set to 256 and AdamW optimizer is used. For both experts, we only add low-rank module to UNet. And the two experts are both built on the fine-tuned base model in Stage 1.

Stage 3: Mixture Adaptation. In this stage, we use the batch size 64 and employ AdamW optimizer. We use a constant learning rate $1e - 5$ and train for 50k steps.

A.2 License and Privacy Statement

The human-centric dataset is collected from websites including seeprettyface.com, unsplash.com, gratisography.com, morguefile.com, pexels.com, *etc.* We use web crawler to download images only when it is allowed. Most images in these websites are published by their respective authors under Public Domain CC0 1.0 ¹¹ license that allows free use, redistribution, and adaptation for non-commercial purposes. Seeprettyface.com require giving appropriate credit to the author by adding the sentence (# Thanks to dataset provider:Copyright(c) 2018, seeprettyface.com, BUPT_GWY contributes the dataset.) to the open-source code when using the images. When collecting and filtering the data, we are careful to only include images that, to the best of our knowledge, are intended for free use and redistribution by their respective authors. That said, we are committed to protecting the privacy of individuals who do not wish their images to be included. Besides, for images fetched from other datasets, *e.g.*, Flickr-Faces-HQ (FFHQ) [23], Celeb-HQ [22], and 11k Hands [1], we strictly follow their licenses and privacy. Note that this dataset is allowed for academic purposes only. When using it, the users are requested to ensure compliance with ethical and legal regulations. For the application for the usage of generated images and the dataset, we will carefully review the applicant's qualifications, purpose of use, possible risks [56, 54, 55], *etc.* Finally, we only allow authorized personnel to interact with the data.

A.3 Filter Training and Illustrations of Negative Samples

In both case, to train the VGG19, we manually collect around 300 positive samples and 300 negative samples as training set, and we also collect around 100 positive samples and 100 negative samples as val set. When training the VGG19, we set the batch size to 128, set the learning rate to 0.001, and use random flip as the data augmentation method. We train the model for 200 epochs and use the best-performing model for subsequent classification. In Fig 9, we present the illustrations of negative samples during refining human-in-the scene subset.

A.4 Comparison with Existing Related Datasets

We give a comparison of the differences between existing datasets like CosmicMan [28] and our newly collected dataset, which primarily lie in four aspects:

- From the aspect of image diversity, due to different motivations, CosmicMan only contains human-in-the-scene images while our dataset also involves two close-up datasets for face and hand, respectively. Moreover, to the best of our knowledge, the high-quality close-up hand dataset is absent in prior related studies.

¹¹<https://creativecommons.org/publicdomain/zero/1.0/>



Figure 9: Illustrations of negative samples. Zoom in for a better view.

- From the aspect of image content distribution, there is a relatively severe gender imbalance in CosmicMan where female makes up a large proportion around 75% (See Fig 3 of Appendix in its paper) while our dataset is relatively balanced (58% vs 42%).
- From the aspect of image size, though CosmicMan and our dataset are both of high quality, our collected images (basically over 1024×2048) are relatively larger than CosmicMan whose average size is 1488×1255 .
- From the aspect of data sources, our dataset is legally collected from various websites including unsplash.com, gratisography.com, morguefile.com, and pexels.com, *etc.*, while CosmicMan is sourced from LAION-5B (See <https://huggingface.co/datasets/cosmicman/CosmicManHQ-1.0>). What sets our dataset apart is not just its wide collection, but also the freshness of the data. As a trade-off, the quantity of our dataset (1M) is relatively smaller than that of CosmicMan (5M).

A.5 More Quantitative Comparisons

We perform four kinds of comparisons. We first evaluate MoLE and SD v1.5 on CLIP-T [25], FID [16], and Aesthetic Score [43] to present a comprehensive comparison. Specifically, We follow [25] and [21] using COCO Human Prompts and Human-centric datasets to evaluate the overall quality of generated images on CLIP-T and FID. We also use an aesthetic predictor¹² to generate the Aesthetic Score. All the results are reported in Tab 4. We find MoLE is also superior in CLIP-T and FID. We also find that MoLE is slightly inferior to SD v1.5 in aesthetic score. We deem that it is reasonable as SD v1.5 is especially fine-tuned on laion-aesthetics v2 5+ dataset in which each image's aesthetic score is evaluated with high aesthetics score by exactly the aesthetic predictor we used in this comparison.

Table 4: Comparisons between MoLE and SD v1.5 on CLIP-T, FID, and Aesthetic Score.

| Method | CLIP-T | FID | Aesthetic Score |
|---------|--------|-------|-----------------|
| SD v1.5 | 26.87 | 69.82 | 5.19 |
| MoLE | 27.33 | 64.37 | 5.18 |

Then, we construct MoLE based on SD v2.1 and evaluate MoLE and SD v2.1 using COCO Human prompt to further show the effectiveness and generalization of our method. The experiment shows that MoLE produces 20.45 ± 0.09 for HPS, outperforming SD v2.1 (20.17 ± 0.08). For IR, MoLE produces 62.08 ± 0.86 , outperforming SD v2.1 (58.36 ± 0.51). These results further verifies the effectiveness of our method.

Besides, to verify the generalization of our method on transformer-based models, we also build our MoLE based on PixArt-XL-2-512 \times 512 [3]. To compare the performance, we randomly sample 3k prompts from COCO Human Prompts and calculate HPS for MoLE (PixArt) and PixArt. The evaluation process is repeated three times. Our method achieves 21.79 ± 0.03 HPS (%) and outperforms PixArt (21.33 ± 0.08 HPS). The result demonstrates the generalization of our method.

¹²<https://laion.ai/blog/laion-aesthetics/>



Figure 10: Illustrations of some compared images between MoLE and HyperHuman.



Figure 11: Illustrations of some compared images between MoLE and HanDiffuser.

Finally, we compare MoLE with state-of-the-art methods including HanDiffuser [33] and HyperHuman [30] through user study as their code has not been made available. Specifically, we invite 50 participants to compare the visualization presented in the two methods' papers with our generated images, respectively. In the user study, we prepare 10 MoLE-HyperHuman pairs and ask participants to select the better one from each pair according to their preference in terms of hand quality. Some compared images are presented in Fig 10 and Fig 11 to show the differences between our generated images and theirs. The results show that 58% of participants think our generated images are better than that of HyperHuman. Similarly, for HanDiffuser, we also prepare 10 MoLE-HanDiffuser

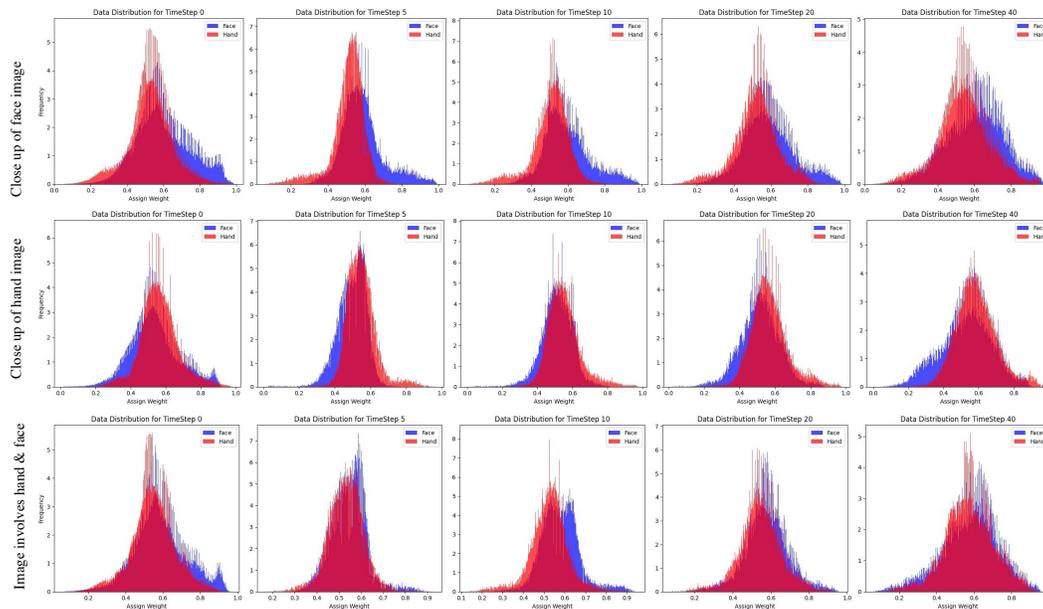


Figure 12: The distribution of the local weight sent to each expert in MoLE.

pairs and ask participants to select the better one. We find that 48% of participants vote for MoLE, slightly inferior to HanDiffuser (52%). All these results demonstrate that our method is effective and competitive with the state-of-the-art methods. More importantly, our method is user-friendly because both HanDiffuser and HyperHuman rely on additional conditions to enhance human and hand generation: HyperHuman takes text and skeleton as input; HanDiffuser needs text, a SMPL-H model, camera parameters, and hand skeleton. In contrast, MoLE only relies on text without the need for any additional conditions, offering greater flexibility and ease of use while maintaining competitive performance.

A.6 Weight Distribution of Local Assignment

We provide the distribution of the local weight sent to each expert in Fig 12. To obtain this, we generate 10 samples for close-up images and normal human images, respectively, and collect local weights for each expert. In Fig 12, one can see that for close-up images, *e.g.*, face, the corresponding expert receives more weights of high value. We think this effectively demonstrates the efficacy of the soft assignment mechanism in MoLE, which adaptively activates the relevant expert to contribute more to the generation of close-up images. When generating normal human images involving face and hand, the two experts contribute equally, and generally, the face expert receives relatively more weights of high value as the area of face is typically larger than that of hand.

A.7 Ablation Study on Experts

To understand the importance of using two experts on the model’s performance, we train only one expert using all close-up images and compare the performance with that of two experts. We find that one expert achieves 20.19 ± 0.03 HPS(%), inferior to that of two experts (20.27 ± 0.07 HPS), which demonstrates the necessity of using one expert for face and hand, respectively.

A.8 More Visualization

We present more generated images and compare with other diffusion models in Fig 13. We provide more full-body images in Fig 14. Additionally, we illustrate more images generated by MoLE in Fig 17, Fig 18, Fig 19, and Fig 20. We also compare MoLE (build on SDXL) with SDXL in Fig 21, Fig 22, and Fig 23 where MoLE further enhances SDXL by generating more natural face/hand.

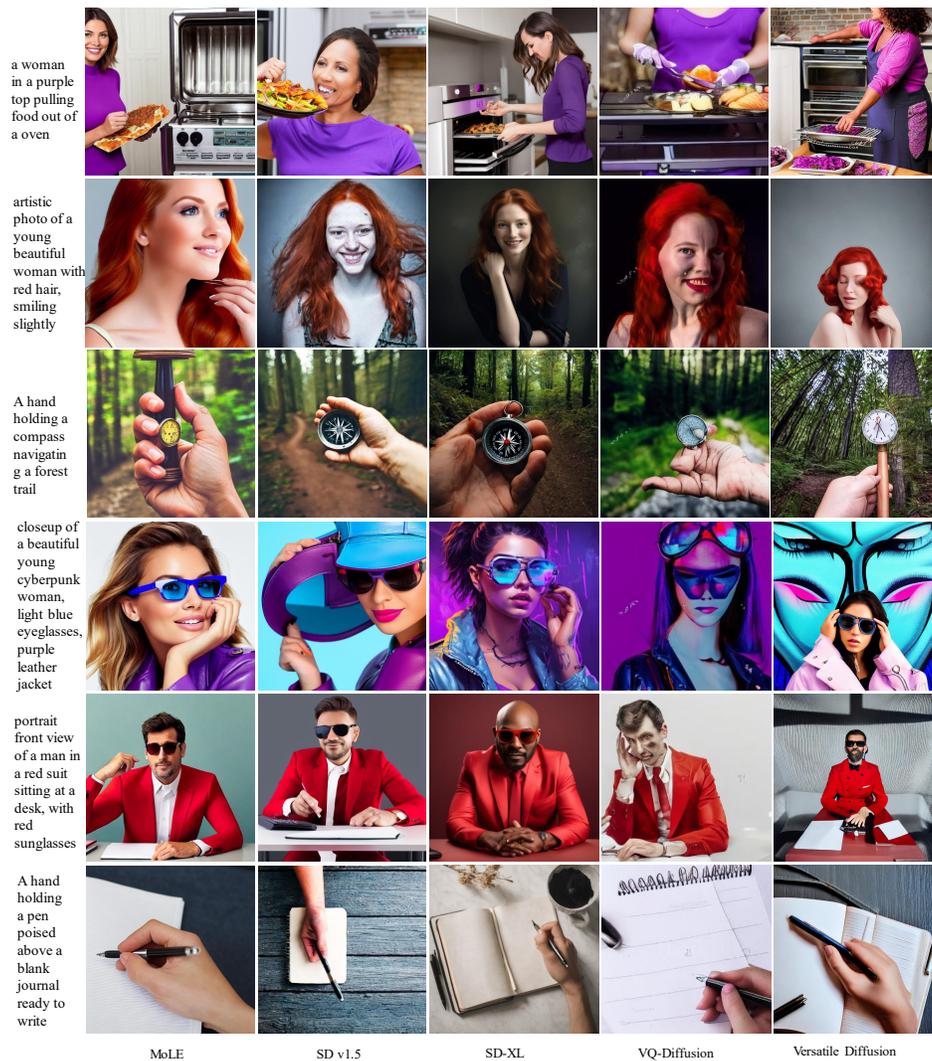


Figure 13: Comparison with other diffusion models. Zoom in for a better view.



Figure 14: Comparisons about full-body images. Zoom in for a better view.

A.9 Generic Image Generation

As shown in Fig 15, MoLE (fine-tuned SD v1.5) can also generate non-human-centric images, *e.g.*, animals, scenery, *etc.* The main reason could be that the human-centric dataset also contains these entities that interact with humans in an image. As a result, the generative model learns these concepts. However, intuitively, MoLE may not be better at generic image generation than the generic generative models as MoLE is trained on a human-centric dataset.

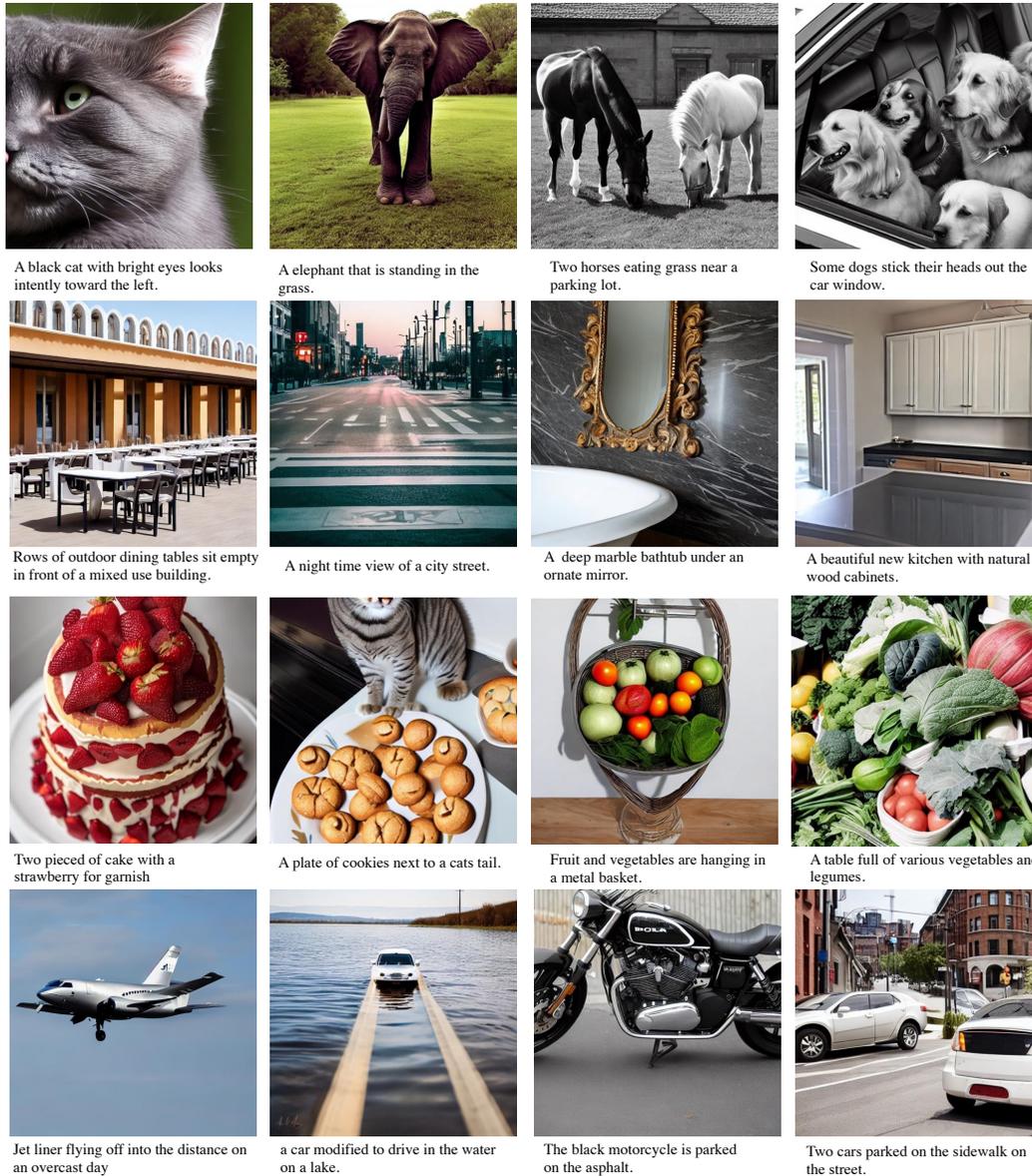


Figure 15: Generic image generation. Zoom in for a better view.

A.10 Failure Case Analysis

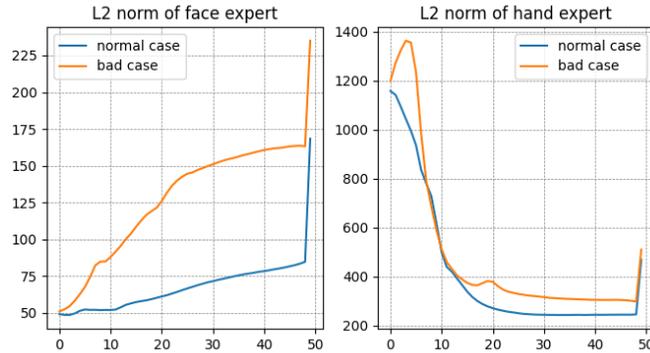
We observe several failure cases involving unrealistic content. For example, we find a generated unrealistic image with a half-man in Fig 16 (a) whose prompt is “a young man skateboarding while listening to music”. To figure out the reason, we sample 10 bad and normal cases, calculate their L2 norm of outputs from face and hand expert respectively, and visualize the averaged L2 norm across timestep as shown in Fig 16 (b). One can see that the bad cases generally have larger L2 norm for both experts, which indicates that the output from linear layer in Eq 7 is strongly influenced by the two experts. As a result, the generated images may be uncoordinated. We leave this as feature work.

A.11 More Related Work

Text-to-image generation. Diffusion model [17, 46] has been widely used in image generation since its proposal. Afterward, a vast effort has been devoted to exploring the applications, especially in text-to-image generation. GLIDE [34] leverages two different kinds of guidance, *i.e.*, classifier-free



(a) An unrealistic case



(b) Averaged L2 norm of outputs

Figure 16: (a) is a showcase of a generated unrealistic image. In (b), the left is averaged L2 norm of outputs from face experts and the right is averaged L2 norm of outputs from hand experts. X is the timestep.

guidance [18] and clip guidance, to match the semantics of the generated image with the given text. Imagen [42] further improves the performance of text-to-image generation via a large T5 text encoder [37]. Stable Diffusion [40] uses a VAE encoder to map image to latent space and perform diffusion on representation. DALL-E 2 [38] transfers text representation encoded by CLIP [36] to image representation via diffusion prior. Besides generation, diffusion models have also been used in text-driven image editing [47]. Inspired by the key observation between text and map in the cross-attention module, Prompt-to-prompt [15] modifies the cross-attention map with prompt while preserving the original structure and content. Further Null-text inversion [31] achieves real image edition via image inversion. Different from them, our work primarily focuses on human-centric text-to-image generation, aiming to alleviate the poor performance of diffusion models in this field.

Mixture-of-Experts. MoE is first proposed in [20]. The underlying principle of MoE is that different subsets of data or contexts may be better modeled by distinct experts. Theoretically, MoE could scale model capability with little cost by using sparsely-gated MoE layer [44]. Researchers extend MoE on transformers by replacing FFN and attention layers with position-wise MoE layers [24, 8, 7]. SwiTh Transformer [11] simplifies the routing algorithm. [52] propose an Expert Choice (EC) routing algorithm to achieve optimal load balancing. GLaM [10] scales transformer model parameter to 1.2T but is inference-efficient. VMoE [39] scale vision model to 15B parameter via MoE. Soft MoE [35] introduces an implicit assignment by passing weighted combinations of all tokens to each expert. MoE has been adapted in generation tasks [12, 2]. For example, ERNIE-ViLG [12] uniformly divides the denoising process into several distinct stages, with each being associated with a specific model. eDiff-i [2] calculates thresholds to separate the whole process into three stages. Differing from employing experts in divided stages, we consider low-rank modules trained by customized datasets as experts to adaptively refine generation.

A.12 More Discussion

Dataset Contribution. To the best of our knowledge, we newly collect and propose a very high-quality and comprehensive human-centric dataset simultaneously meeting legal compliance.

- *High-quality.* Our human-centric dataset consists of images of high resolution (basically over 1024×2048) and large file size, e.g., 1M, 3M, 5M, and 10M, collected from websites (e.g., <https://www.pexels.com/search/people/> and <https://unsplash.com/s/photos/people>). The readers can simply click these links to have a look. Such quality is missing in current widely used image-text datasets, e.g., LAION2b-en. To support it, we sample 350K human-centric images from LAION2b-en. The averaged height and width are approximately 455 and 415 respectively. Most of them are between 320 and 357. Compared to our human-centric dataset, they fail to provide sufficient prior and high-quality details.

- *Comprehensive.* As the natural face and hand are relatively hard to generate compared to other parts as discussed in the community, in addition of the collected human-in-the-scene subset, we also



a smile blond girl lay on chair holding rose



a woman working on her computer while lounging in her bed



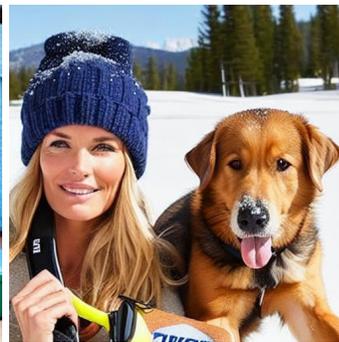
a young man skateboarding while listening to music



a hand is holding a paintbrush and is painting a colorful design



a toddler boy is sitting on a brief case



a beautiful woman with brown hair in a blue hat, standing next to a brown dog in the snow



a man riding on a surfboard on top of a wave



an old man with a beard wearing a blue tie and green suit, smiling



muscular man wearing a leather harness and hat

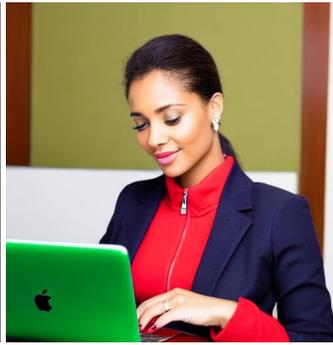
Figure 17: More images generated by MoLE. Zoom in for a better view.

provide two high-quality close-up of face and hand subsets. In particular, close-up images of hand are absent in the human-centric images sampled from LAION2b-en during our observations. To the best of our knowledge, this quantity of the high quality close-up hand images is absent in prior related studies. It will significantly help to address challenges associated with generating natural-looking hands and propel the advancement of human-centric image generation for subsequent researchers.

• *Legal Compliance.* The human-centric dataset is collected from websites including unsplash.com, gratisography.com, seeprettyface.com, morguefile.com, pexels.com, etc. Images on these websites are published by their respective authors under Public Domain CC0 1.0 3 license that allows free use, redistribution, and adaptation for non-commercial purposes. When collecting and filtering the data, we are careful to only include images that, to the best of our knowledge, are intended for free



a hand is leaning over the woman's dress



a woman working on a laptop computer on the table



a boy about to eat a large piece of pizza



lots of men in suits and one boy are sitting together



a man in grey sweater standing at a wooden table with food and coffee



a woman smiling with her surf board on the beach



two small children pose together with skis and poles.



a girl in yellow dress eating a piece of cake

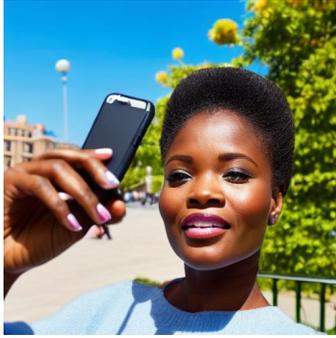


a smiling woman and a dog in the ocean

Figure 18: More images generated by MoLE. Zoom in for a better view.

use. We are committed to protecting the privacy of individuals who do not wish their images to be included. The eventually resulted human-centric dataset can be used for academic purposes only and the users are required to ensure compliance with ethical and legal regulations.

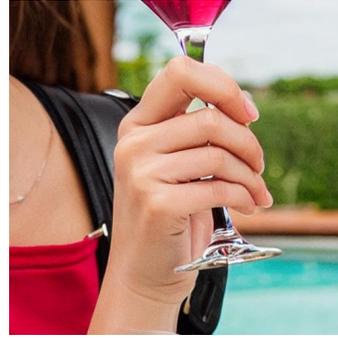
Analysis about Ratios of Different Races. Since our dataset is collected from Internet, it inevitably inherits the race bias of target websites. Hence we provide an analysis about ratios of different races in MoLE with a single prompt “A beautiful woman”. Specifically, we generate 10K images with this prompt. With the help of DeepFace, we find that approximately 51.08% individuals identify as White, 5.29% as Asian, 10.18% as Black, 4.31% as Indian, 24.66% as Latino Hispanic, and 4.48% as Middle Eastern. To verify if the generation of different races can be improved by using a race-balanced dataset, based on our dataset we use DeepFace to reconstruct a new dataset with the same ratio of different races. The newly curated dataset comprises 30k images. We use it to train a MoLE model and generate 10K images using the same prompt “A beautiful woman”. We find that approximately 45.56% individuals identify as White, 7.17% as Asian, 8.32% as Black, 13.41% as Indian, 13.44%



a woman is holding a cell phone in her hand



a man sitting in a kitchen



a close up of a woman holding a wine glass



a little boy sitting in front of a birthday cake



a young girl is smiling and holding a bouquet of roses while laying on a green mat



a man wearing a bow tie is smiling and looking at the camera



smiling woman in red top putting items in a box



a woman with blonde hair wearing glasses



a woman is on the beach, smiling

Figure 19: More images generated by MoLE. Zoom in for a better view.

as Latino Hispanic, and 12.10% as Middle Eastern. This result shows a relatively higher balance of races compared to previous one, demonstrating that MoLE is beneficial to alleviate race biases with the reconstructed dataset. Given the constraints of our race-balanced dataset, which contains only 30k samples, we believe that expanding the size of the race-balanced data could enhance our method's ability to further address the race-related issue.



a girl sitting on a couch with a cat on top of her laptop



a woman with blue hair is drinking from a blue coffee cup



an older woman in a sweater sits at the beach



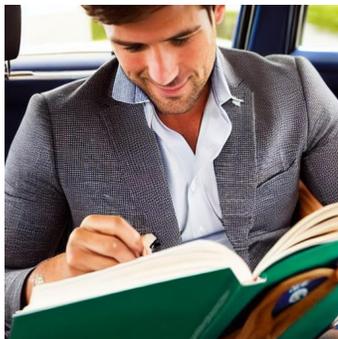
a man sitting in front of a large computer monitor



a woman with long brown hair is holding a pink toothbrush in her hand



a woman in a white dress stands on the edge of a lake



a man is sitting in a car reading a book



a pretty blonde haired girl wrapped up in a sheet and laying down



a young boy wearing a black shirt with a cat on it and cat ears

Figure 20: More images generated by MoLE. Zoom in for a better view.

The woman sits at the table overlooking the pink and white cake with lit candle.



A man and a woman standing behind a fruitstand of bananas at a chinese shop.



A sexy young blonde woman holding a tennis racquet.



A baby girl chews on a stick with a teddy bear in hand.



MoLE

SDXL

Figure 21: Comparing MoLE with SDXL. Zoom in for a better view.

A woman wearing a white shirt holding the face of a white horse.



A woman lays on her side, as she wears beautiful, colorful clothing.



A man wearing a beret while using a laptop computer.



A girl eating a piece of cake from a pink princess plate



MoLE

SDXL

Figure 22: Comparing MoLE with SDXL. Zoom in for a better view.

A woman wearing glasses looking at slices of pizza



An elderly man is using a laptop computer at a desk.



A pregnant woman is in bed reading a large book.



A man dressed very nice posing for a picture.



MoLE

SDXL

Figure 23: Comparing MoLE with SDXL. Zoom in for a better view.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: See our contribution summary and Discussion Section

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: See Discussion Section

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: Our work do not involve this

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide all information needed.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We release the data and code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide all details

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report standard error.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Our method can be trained in only one A100 80G.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: Yes, we verified it.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: See broader impacts part.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: See Section Human-centric Dataset.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: See Section Human-centric Dataset.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: See Section Human-centric Dataset.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: See Section Ablation Study

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This work does not harm participants

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.