HumanRefiner: Benchmarking Abnormal Human Generation and Refining with Coarse-to-fine Pose-Reversible Guidance

Guian Fang[1,*], Wenbiao Yan[4,*], Yuanfan Guo[3,*], Jianhua Han[3],
Zutao Jiang[2], Hang Xu[3], Shengcai Liao[5], and Xiaodan Liang[1,2,**]

[1] Shenzhen campus of Sun Yat-sen University
[2] Mohamed bin Zayed University of Artificial Intelligence
[3] Huawei Noah's Ark Lab
[4] Xi'an Jiaotong University
[5] United Arab Emirates University

Abstract. Text-to-image diffusion models have significantly advanced in conditional image generation. However, these models usually struggle with accurately rendering images featuring humans, resulting in distorted limbs and other anomalies. This issue primarily stems from the insufficient recognition and evaluation of limb qualities in diffusion models. To address this issue, we introduce **AbHuman**, the first large-scale synthesized human benchmark focusing on anatomical anomalies. This benchmark consists of 56K synthesized human images, each annotated with detailed, bounding-box level labels identifying 147K human anomalies in 18 different categories. Based on this, the recognition of human anomalies can be established, which in turn enhances image generation through traditional techniques such as negative prompting and guidance. To further boost the improvement, we propose **HumanRefiner**, a novel plug-and-play approach for the coarse-to-fine refinement of human anomalies in text-to-image generation. Specifically, HumanRefiner utilizes a self-diagnostic procedure to detect and correct issues related to both coarse-grained abnormal human poses and fine-grained anomaly levels, facilitating pose-reversible diffusion generation. Experimental results on the AbHuman benchmark demonstrate that HumanRefiner significantly reduces generative discrepancies, achieving a 2.9x improvement in limb quality compared to the state-of-the-art open-source generator SDXL and a 1.4x improvement over DALL-E 3 in human evaluations. Our AbHuman dataset is available at `https://github.com/Enderfga/HumanRefiner`.

Keywords: Text-to-Image Diffusion Model · Abnormal Human Dataset · Human Anomalies Refinement

1 Introduction

Recent developments in text-to-image diffusion models, as exemplified by SDXL [31], DALL-E 2 [34], and DeepFloyd [31,34,44], have marked significant breakthroughs

---

[*] Equal Contribution.
[**] Corresponding author.

in the generation of high-quality images.These models, benefiting from large-scale pre-training, are capable of generating images for various objects.

However, generating complex and flexible structures, such as the human body, poses significant challenges for current text-to-image models. As illustrated in Figure 1, common issues with these models include the generation of multiple limbs or abnormal limb details (e.g., irregular fingers). These challenges arise because training focuses on overall characteristics, not detailed ones like the number of fingers or hands. Consequently, models have limited control over the fine-grained generation of detailed body parts and lack the knowledge to distinguish between normal and abnormal human body characteristics. Though such knowledge could be introduced by extra pose condition as indicated in HumanSD [18], Pose-ControlNet [47], T2I-Adapter [27], these models are hard to use because of the requirement of providing extra input.

(a) Human images generated by SDXL

(b) HumanRefiner (ours), based on SDXL

Fig. 1: The synthesis of human images by SOTA text-to-image diffusion model SDXL [31] still presents incorrect limb counts, twisted hands, and incorrect limb position problems. We propose *HumanRefiner*, a coarse-to-fine self-diagnosis pose/anomaly-reversible generation pipeline built on the *AbHuman* benchmark to eliminate the anomalies.

Therefore, this paper focuses on enhancing the text-to-image diffusion model's understanding of human anomalies and leveraging this knowledge to drive models for generating accurate and realistic human bodies. To address the above challenges, we introduce the first large-scale synthesized human benchmark for anatomical anomalies, named **AbHuman**. This benchmark comprises 56K synthesized human images, each meticulously annotated with detailed, bounding-box level labels that mark 147K instances of human limbs. Instead of a single "abnormal body" category, these anomalies are categorized into 18 distinct finer

groups, such as "abnormal/normal head", "abnormal/normal hand", "abnormal/normal foot", etc.

With the proposed benchmark, it could be directly employed to provide knowledge about human anomalies by abnormal captions derived from the annotation, which can then be integrated into the diffusion model through the conventional negative prompting technique. Another possible way is to develop an Abnormal Scorer, which provides feedback on generated images to distinguish normal/abnormal images. Such a scorer enables quantitative assessment of anomalies in generated images and could provide guidance on anomalies for the text-to-image diffusion process.

However, these conventional approaches from our AbHuman benchmark do not fully leverage the detailed information available from bounding box-level annotations of human limb anomalies. Therefore, we propose **HumanRefiner**, a novel and adaptable method for addressing human anomalies that can be integrated with any text-to-image generation model. HumanRefiner goes beyond the typical abnormal guidance and negative prompting by implementing a coarse-to-fine, pose-reversible diffusion generation process. This includes a pose-guided generation for global refinement and anomaly detector-guided inpainting for fine-grained refinement. The detailed anomalies are identified using an abnormal detector train on AbHuman. Experimental results on the benchmark illustrate that HumanRefiner markedly reduces generative discrepancies, achieving a $2.9\times$ improvement in limb quality compared to leading open-source generator SDXL and a $1.4\times$ improvement over DALL-E 3 in human evaluations. In summary, our contributions are threefold:

- Introduction of AbHuman, the first large-scale benchmark of synthesized humans with anatomical anomalies, which encompasses 56K images, each annotated with detailed, bounding-box level human anomaly labels.
- We propose HumanRefiner, a novel plug-and-play approach for coarse-to-fine refinement of human anomalies via pose-reversible guidance.
- Comprehensive experimental results show that HumanRefiner significantly reduces generative discrepancies compared to SOTA text-to-image models. Human evaluation results indicate a $2.9\times$ preference for our HumanRefiner in generating high-quality human images compared to SDXL and a $1.4\times$ preference over DALL-E 3.

2 Related Work

2.1 Text-to-Image/Human Generation

Text-To-Image Generation. The goal of text-to-image synthesis is to generate realistic images based on given descriptions. Fueled by large-scale image-text datasets [4, 42], various training and inference techniques [12, 13, 45], and advancements in scalability [34], text-to-image models have made significant progress. Existing text-to-image models can be broadly categorized into three

main approaches: those utilizing diffusion models [2, 11, 29, 34, 37–39], autoregressive models [7, 35, 46], and generative adversarial models [19, 40, 49]. The latest models, such as DALL-E 3 [30], and SDXL [31], have shown substantial improvements in image generation quality. However, the challenge of generating human-like images from text remains difficult due to the diversity of human body poses and insufficient alignment between image and text. Our work focuses on improving the quality of generated human images.

Text-To-Human Generation. Numerous models have made considerable efforts to enhance the generation quality of realistic human images. They have discovered that injecting additional conditions into diffusion models provides better guidance for human image generation. Models such as [15, 27, 47] have incorporated additional training modules, introducing human pose conditions to finely control the generation of human images. Additionally, [21] has the capability not only to introduce key points but also to incorporate bounding box detection data. Building upon the introduction of additional poses, [18] has implemented a strategy involving heatmap alignment to achieve higher-quality human images.

However, existing works dedicated to human generation face two main challenges: (1) the introduction of poses significantly restricts the diversity of generated human poses, and (2) pose conditions still exhibit limitations in addressing fine-grained details of human body parts, such as hands, feet, and facial features.

Differing from this line of research, HumanRefiner is designed to address the issue of human generation without the requirement of human-provided pose conditions.

2.2   Human Generation Datasets.

Currently, human datasets primarily consist of two types. The first focuses on authentic depictions of individuals in natural settings, such as human photos and images from human video data [1, 6, 8, 20, 22, 28, 42, 50]. These widely available datasets offer abundant human-pose pairs, with annotations limited to key points or bounding boxes. Given that these data are derived from real photos, the humans depicted are inherently natural, without limb anomalies. The second category involves artificial scenes featuring human-like figures in settings such as paintings, cartoons, and similar environments. The Human-Art dataset [17] provides 50,000 images centered around humans, encompassing five natural scenes and fifteen artificial scenes, accompanied by precise pose and text annotations. None of these datasets provide large-scale synthesized images or potential limb abnormal annotations.

To address the challenge of limb anomalies in the text-driven human image generation task, we have developed a novel AbHuman dataset as a component of this research effort. This dataset serves as a valuable resource for effectively handling limb abnormalities within the context of generating human images from textual descriptions.

We provide a large-scale dataset of abnormal human images called AbHuman. Firstly, this is the first dataset to focus on limb abnormalities, as existing text-to-human algorithms are primarily based on the MS-COCO [22] and

LAION-5B [42] datasets. Notably, models trained on real images often face challenges in accurately generating limb abnormalities in humans. To address this issue in general human image generation scenarios, we constructed a new benchmark. Secondly, this dataset collects a significant number of failure cases based on existing generative models. We observed that existing AI-generated human images often feature limb abnormalities, which we filtered, annotated, and analyzed. With the guidance of this dataset, the quality of generative models has been greatly improved.

2.3   Data Collection

We designed a standardized pipeline for abnormal human limb image collection, including prompt collection, image generation, and abnormal annotation, as shown in Fig. 2.
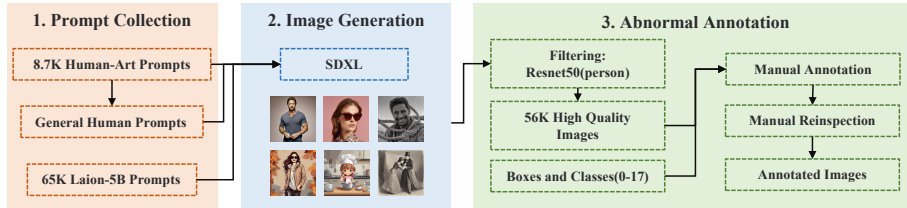

Fig. 2: Data generation pipeline of our AbHuman dataset.

3   The AbHuman Dataset

– **Prompt Collection:** Our prompts are derived from three sources: textual descriptions related to humans in the Laion-5B [42], textual descriptions from the Human-Art dataset [17], and prompts generated using ChatGPT 3.5 [3]. To extract text-generating human-related content from the Laion dataset, We used keyword filtering to extract human-related content from the Laion dataset, resulting in 100,000 high-quality textual prompts. Subsequently, we utilized the collected textual content as the foundation of our dataset.
– **Image Generation:** With the rapid advancements in text-to-image models, the quality of human image generation has significantly improved. However, even state-of-the-art generation models like SDXL [32], Midjourney, and DALL-E 3 [30] encounter issues with limb abnormalities when generating human images. To enhance the quality of human generation using the best existing generation models as a baseline, we employed the SDXL model for generating our images.
– **Abnormal Annotation:** After collecting a large number of images, a comparative analysis revealed the presence of non-human images within the dataset. Initially, we applied a ResNet50-based classifier to filter all images labeled as "person".After removing low-quality images, trained human

annotators annotated the remaining 56,000 images. Two human evaluators reviewed the dataset, resulting in a curated AbHuman dataset with 56,000 images and 147,000 annotations. Details of the annotated categories are provided in the appendix.

– **Data Split:** The dataset is partitioned into training and testing splits in 4:1, with 52k and 13k image-text pairs, respectively. We further develop a testing split for human evaluation, where prompts are from HumanArt [17], each describing humans performing difficult actions such as acrobatics, dancing, and drama.

As illustrated in Figure 3, the AbHuman dataset provides bounding boxes and classification information for abnormal humans. Unlike existing keypoint-based annotation methods, the bounding box annotation enables the identification of both the location and type of anomalies in limbs. By detecting and inpainting anomalies in human limbs, our dataset is more conducive to improving generation algorithms.


Fig. 3: The visualization of limb detection in the test set using the fine-tuned AbHuman detection model. Red boxes and labels are used to annotate the abnormal limbs detected, and white boxes and labels are used to annotate the normal limbs detected.

3.1   Dataset Statistics and Analysis

After reviewing a large number of generated samples, we conducted an analysis of human abnormalities and identified 18 distinct detection classes (*e.g., Normal human, Abnormal Head, Abnormal Neck, Abnormal Body, Abnormal Arm, Abnormal Hand...*). To illustrate the class distribution of anomalous human samples generated by the SDXL model, we performed a statistical analysis of both anomalous humans and their corresponding annotations. As depicted in Figure 4, the model predominantly generates images with anomalous hands, aligning with the highest count of annotations for hand anomalies. Thus, it is evident that hand anomalies remain a challenge for text-to-image models. Additionally,

the number of samples with anomalies in the head and foot regions is relatively higher compared to other parts of the human body.
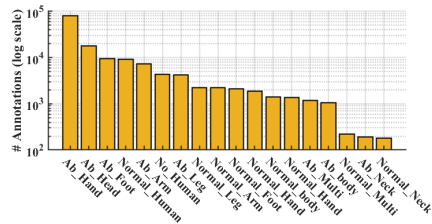
Fig. 4: Ab-human annotations statistics.
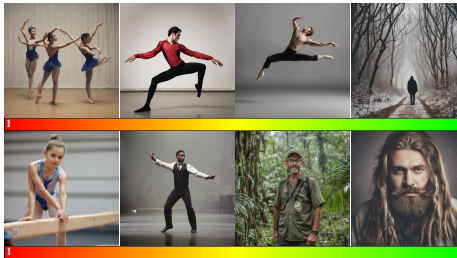
Fig. 5: Visualization of abnormal scores. red indicates images with high abnormal scores while green indicates images with low abnormal scores.

Table 1 illustrates a comparison of datasets concerning human images. Compared to other human datasets [1, 16, 17, 20, 25, 28] that solely annotate images of authentic and normal individuals, existing datasets have limited annotations for genuinely abnormal human limbs. On the contrary, the AbHuman dataset introduces anomalies in human-generated images and provides ample annotations for anomalous body postures. In the context of human-focused tasks such as generation and detection, AbHuman proves to be more suitable. Additionally, compared to datasets with relatively monotonous scenes, AbHuman encompasses a more diverse range of scenarios, enhancing its applicability in various contexts.

Table 1: Comparison of human-centric datasets, including human generation and detection tasks. **AbHuman provides more granular annotation boxes and information to address the abnormality issues in the human generation.**

Dataset	Images	Bbox	Normal	Abnormal-Limb	General
BodyHands [28]	20K	✓	✓	✗	✗
CrowdPose [20]	20K	✓	✓	✗	✗
Multi-Person PoseTrack [16]	23K	✓	✓	✗	✗
PoseTrack [1]	23K	✓	✓	✗	✗
ClassArch [25]	1.5K	✓	✓	✗	✗
Human-Art [17]	50K	✓	✓	✗	✓
AbHuman (Ours)	56K	✓	✓	✓	✓

3.2 Abnormal Scorer

As indicated, the lack of a metric measuring the generated limb quality makes it difficult to (1) quantitatively compare different models; and (2) develop generators for the improvement of limb quality. Based on the proposed AbHuman

Dataset, we present a quantitative metric of limb anomalies in a generated or real human image. To this end, inspired by the aesthetic scorer based on the Laion dataset [41], we train an Abnormal Scorer $S_\theta(\cdot)$. For the model structure of Abnormal Scorer $S_\theta(\cdot)$, we adhere to the architecture utilized by Aesthetic Scorer [41]. Building upon a pretrained CLIP ViT-Large with patch size of 14 [33] with a hidden dimension of 768, a Multi-Layer Perceptron (MLP) is constructed, as depicted in Table 2.

Table 2: Model architecture of MLP part of our abnormal scorer. Linear(A, B) means a linear layer with an input feature dimension of A and an output feature dimension of B. Dropout (α) means a dropout layer with the probability of α.

Layer index	0	1	2	3
Layer	Linear (768, 1024)	Dropout (0.2)	Linear (1024, 128)	Dropout (0.2)
Layer index	4	5	6	7
Layer	Linear (128, 64)	Dropout (0.1)	Linear (64, 16)	Linear (16, 1)

Given an image $x \in \mathcal{R}^{3 \times H \times W}$, the Abnormal Scorer assigns a score $S_\theta(x) \in [0, 1]$ reflecting the severity of limb aberrations or distortions. Specifically, this model is architecturally based on a Multi-Layer Perceptron (MLP) and is built upon the foundation of the pre-trained CLIP Vision Transformer Large (ViT-L) model [33]. As illustrated in Figure 5, high-scoring images are characterized by pronounced limb distortions or the absence of limbs, whereas images with lower scores correspond to either normal human depictions or non-human subjects.

3.3   Abnormal Detector

To locate and detect abnormal limbs in human images for subsequent fine-grained refinement, we developed an abnormal limb detector. Specifically, we fine-tuned YOLOv8 [36] and RT-DETR [24] on the training set of AbHuman for 100 epochs. Subsequently, we conduct corresponding detections on the test set, obtaining consistently stable results in the detection of abnormal limbs. Visualization results are provided in Figure 3, showing the capability of identifying abnormal/normal regions and corresponding categories.

4   HumanRefiner

Based on the trained detector as illustrated in sec 3.3, we can develop a flexible human image synthesis and refinement mechanism, **HumanRefiner**, which can be integrated into any pipeline for human image generation. As shown in Figure 6, in addition to the basic improvement on generation gained from negative prompting learned from Abnormal Guidance, HumanRefiner performs a coarse-to-fine pose-reversible refinement based on the initially generated images. An off-the-shelf pose detector [6] is employed to inject pose guidance into the

text-to-image diffusion model, which serves as conditions for coarse refinement of the initial generated image. Then, the Abnormal Detector identifies abnormals that are further rectified through an inpainting diffusion model, refining the remained abnormals at a fine-grained level.
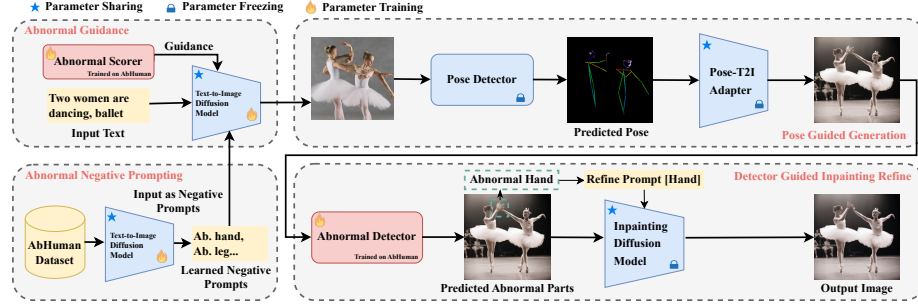
Fig. 6: Overview of our HumanRefiner pipeline empowered by the proposed AbHuman benchmark. It consists of a Text-to-Image Diffusion Model trained with negative prompts on the AbHuman benchmark, which generates an initial human image with guidance from Abnormal Scorer, a **Coarse-to-Fine Refinement Process** where the generated image is globally refined by pose-guided generation and locally refined by detector-guided inpainting. For both the Pose-T2I Adapter [27] and the Inpainting Diffusion models, the Stable Diffusion parameters are adopted from the Text-to-Image model, which was refined through training in the Abnormal Guidance Stage.

The following sections demonstrate the basic control of human generation from negative prompting and abnormal guidance, followed by the coarse refinement module of pose-guided generation and the fine-grained refinement module of detector-guided inpainting.

4.1   Negative Prompts

The diffusion model is designed to generate data through a sequence of denoising autoencoders. The Latent Diffusion Model (LDM) [37] conducts the diffusion process with a VAE encoder $\mathcal{E}$ and decoder $\mathcal{D}$. Given an image x, the forward diffusion process incrementally adds Gaussian noise ϵ to the latent $z_0 = \mathcal{E}(x)$ following a T-step noise schedule $\{\alpha_t\}_{t=1}^{T}$ to produce a noisy latent z_t where the noise level increases over timesteps t:

$$z_t = \sqrt{1 - \overline{\alpha}_t} z_0 + \sqrt{\overline{\alpha}_t}\epsilon, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, I)$ and $\overline{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. A text-to-image diffusion model learns a noise predictor $\epsilon_\theta(z_t, t, y)$ conditioned on text y with the objective:

$$L = \mathbb{E}_{\mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1), t}[||\epsilon - \epsilon_\theta(z_t, t, y)||_2^2]. \tag{2}$$

The text condition y ensures that the model's denoising is conditioned and controlled. Consequently, based on our AbHuman Benchmark, we can train the

Latent Diffusion Model with the abnormal text conditions y_a(e.g., Ab. Hand, Ab. Feet...), making the model aware of the abnormal limb information. During the generation process, Classifier-Free Guidance [14] could be applied by using the abnormal text as a negative prompt to eliminate the generation of abnormal limbs:

$$\hat{\epsilon}_\theta(z_t, t, y) = \epsilon_\theta(z_t, t, y_a) + s \cdot (\epsilon_\theta(z_t, t, y) - \epsilon_\theta(z_t, t, y_a)), \tag{3}$$

where s is a scaling factor that modulates the influence of the condition.

4.2   Abnormal Guidance

In addition to adding information on human abnormals through negative prompting, we can introduce a general form of guidance to improve control over the denoising process. As shown in section 3.2, the Abnormal Scorer $S_\phi(x)$ gives a score to an image that measures the severity of limb abnormals. This scorer can provide further guidance using a general Classifier Guidance [10]. Specifically, since $\epsilon_\theta(z_t, t, y)$ predicts the added noise, we can estimate the clean latent variable z_0 by:

$$\hat{z_0} = \frac{z_t - \sqrt{\overline{\alpha}_t}\epsilon_\theta(z_t, t, y)}{\sqrt{1 - \overline{\alpha}_t}}. \tag{4}$$

After this, we can calculate an estimated abnormal score with $S_\phi(\mathcal{D}(\hat{z_0}))$. Consequently, the noise prediction guided by this method is:

$$\ddot{\epsilon}_\theta(z_t, t, y) = \epsilon(z_t, t, y) + \mu \cdot \nabla S_\phi(\mathcal{D}(\hat{z_0})). \tag{5}$$

We can combine classifier-free guidance through negative prompting with abnormal guidance easily by:

$$\begin{aligned}\hat{\epsilon}_t =&\epsilon_\theta(z_t, t, y_a) + s \cdot (\epsilon_\theta(z_t, t, y) - \epsilon_\theta(z_t, t, y_a)) \\ &+ \mu \cdot \nabla S_\phi(\mathcal{D}(\hat{z_0})).\end{aligned} \tag{6}$$

4.3   Coarse-to-fine Pose-reversible Guidance

In contrast to unguided approaches, our HumanRefiner now yields human images with more accurate poses initially. To further enhance the generation quality, we refine the generated image with pose-reversible guidance at both coarse and fine-grained levels.

Pose-Guided Coarse Refine The abnormal guided text-to-image diffusion model could be combined with off-the-shelf pose detector [6] and T2I-Adapter [27] to further inject global human pose information for pose-guided refinement. For an initially generated image $\hat{x}$ from a guided text-to-image diffusion model, a pose keypoint map M could be detected by a pose detector $M = P_\psi(\hat{x})$. The pose-guided generation could be expanded from the text-to-image diffusion model with T2I-Adapter [27] by injecting the pose condition information into intermediate encoder features:

$$F_c = \mathcal{F}_{AD}(M), \tag{7}$$

$$F_{enc} = F_{enc} + F_c, \tag{8}$$

where $\mathcal{F}_{AD}$ is T2I-Adapter and F_{enc} is the intermidate feature of the encoder part of $\epsilon_\theta(\cdot)$. Forming the pose conditioned denoising as $\epsilon_\theta(z_t, t, y, \mathcal{F}_{AD}(M))$, the advantage of abnormal guidance and negative prompts could also be integrated as:

$$\begin{aligned} \hat{\epsilon}_t =& \epsilon_\theta(z_t, t, y_a) + s \cdot (\epsilon_\theta(z_t, t, y, \mathcal{F}_{AD}(M)) - \epsilon_\theta(z_t, t, y_a)) \\ & + \mu \cdot \nabla S_\phi(\mathcal{D}(\hat{z}_0)). \end{aligned} \tag{9}$$

Detector-Guided Inpainting Refine The pose-guided generation refines the initial image in a global view, performing a coarse refinement. Besides, a fine-grained refinement could be established with our abnormal detector providing region-level abnormal information.

Specifically, for a refined image by applying diffusion denoise process in Eq. (9), $x_r = \mathcal{D}(\hat{z}_0)$, abnormal detector $D_{Ab}(\cdot)$ outputs a set of K bounding boxes and corresponding categories $\{B_i, C_i\}_{i=1}^{K} = D_{Ab}(x_r)$, where $B_i = \{x_i, y_i, w_i, h_i\}$ denotes the detected regions, $C_i \in \{0, ..17\}$ denoting the detected categories defined in sec 3.1. For detected abnormal classes, the local region B_i is inpainted with an input text conditioned on abnormal category $\hat{y}(C_i)$. The local region in the original image is then pasted by the inpainted output.

5 Experiments

5.1 Implementation Details

Negative Prompt Fine-tuning: we finetune SDXL [31] released checkpoint at a resolution of 1024 x 1024 for 10 epochs with the learning rate of 1.0e-5 on the training split, with a batch size of 32 in total. The images are paired with the original text description followed by abnormal captions (e.g. Ab.hand, Ab.leg). More details can be found in the appendix.

5.2 Quantitative Experiments

Evaluation Setup We present quantitative comparison with baseline methods including SOTA text-to-image diffusion models SDXL [31], DeepFlyod-IF [44], LCM [23]. For methods require extra pose condition inputs such as HumanSD [18], Pose-ControlNet [47] and Pose-T2I-Adapter [27], we first generate an initial image from SDXL [31], then generate pose condition from it with pose detector [5], which operates similarly to our pose-guided generation. Test prompts are provided from the AbHuman full test split (prompts from LAION) and hard test split (prompts from HumanArt). We quantitatively evaluate the generated images with Abnormal Score, CLIP score and FID, respectively. Abnormal Score measures the severity of anomalies.

Results As indicated in Table 3, HumanRefiner consistently achieves the best performance in terms of Abnormal score, FID, and CLIP score. Furthermore, when compared to Pose-ControlNet [47] and Pose-T2I-Adapter [27] applied on SDXL, HumanRefiner demonstrates a significant improvement, indicating the effectiveness of HumanRefiner on integrating off-the-shelf methods such as pose detectors and T2I-adapter [27].

Table 3: Abnormal score, CLIP Score and FID on AbHuman full test split (prompts from LAION) and hard test split (prompts from HumanArt).

Dataset	LAION			HumanArt		
Model	Abnormal Score ↓	CLIP Score↑	FID Score↓	Abnormal Score↓	CLIP Score↑	FID Score↓
SSD-1B [43]	0.612	34.02	16.465	0.832	33.37	13.304
PixArt-XL-2-1024-MS [9]	0.701	34.32	19.604	0.807	32.68	15.452
DeepFloyd-IF [44]	0.595	32.72	25.926	0.849	32.67	21.512
LCM [23]	0.644	32.99	25.175	0.831	32.40	21.179
SDXL [32]	0.659	34.13	30.024	0.857	32.40	20.796
Pose-ControlNet [48]	0.663	33.51	15.368	0.838	31.88	13.047
Pose-T2I-Adapter [26]	0.624	34.79	15.658	0.807	32.72	12.582
HumanSD [18]	0.661	33.78	50.546	0.801	33.03	38.032
HumanRefiner (ours)	**0.590**	**34.85**	**13.634**	**0.778**	**33.90**	**9.145**

Human Evaluation We conduct a human preference study on generated images with prompts from HumanArt [17], asking 38 users for a preference comparison between DALL-E 3 [30], SDXL [31], and the proposed HumanRefiner.Users compared the images based on fidelity and limb quality. Figure 7 shows that HumanRefiner surpasses DALL-E 3 and SDXL in both fidelity and limb quality for the human generation. For limb quality, HumanRefiner gets $2.9\times$ preference over SDXL (32.3% vs. 11.20%), and $1.4\times$ over DALL-E 3 (51.42% vs. 36.03%), indicating that HumanRefiner can generate high-quality human pose.
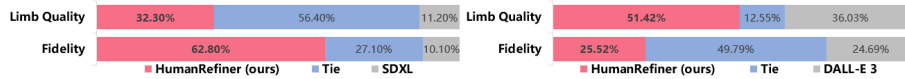
Fig. 7: Human Evaluation on prompts from HumanArt. The percentage indicates the frequency of human preference. "Tie" means having a similar level of preference.

Ablation Study We perform a component analysis to assess the contribution of each element utilizing the Abnormal Score as a metric for generation quality, as shown in Table 4. For instance, the incorporation of the negative prompt learned from the AbHuman dataset significantly reduces the Abnormal Score from 0.659 to 0.614.

Figure 8 shows several ablation experiment findings. The **Baseline** represents the unaltered output of the SDXL model, while **AbnormalGuided** and **PoseGuided** illustrate results achieved through the application of distinct guiding modules. As depicted in the first row of figures, discernible enhancements are observed in the characters' outline. Moreover, the second and fourth rows

| Baseline | AbnormalGuided | PoseGuided | HumanRefiner (ours) |

A couple is paddleboarding on the calm lake.

A woman is doing kickboxing in the martial arts studio.

A man is doing slackline tricks in the park.

**Fig. 8:** Visualization outcomes from ablation experiments.

Table 4: Ablation Study on the proposed components, evaluated on the AbHuman test split (prompts from Laion) with Abnormal score.

Negative Prompt	Abnormal Guidance	Detector Inpaint	Pose Guide	Abnormal Score ↓
✗	✗	✗	✗	0.659
✓	✗	✗	✗	0.614
✓	✓	✗	✗	0.605
✓	✓	✓	✗	0.601
✓	✓	✓	✓	**0.590**

showcase varying degrees of improvement in hand details. Notably, our AbnormalGuided module demonstrates corrective efficacy even in scenes where hand details typically appear blurry. Importantly, the ablation of our AbnormalGuided module mirrors an ablation of the dataset, underscoring the dataset's pivotal role in our approach.

5.3   Qualitative Results

We also present showcases in the challenging test split, featuring humans engaged in activities such as dancing or performing acrobatics. As illustrated in Figure 9, HumanRefiner exhibits superior limb quality compared to the baseline methods. It accurately captures the human pose, represents natural limb attributes, and presents visually pleasing arrangements.
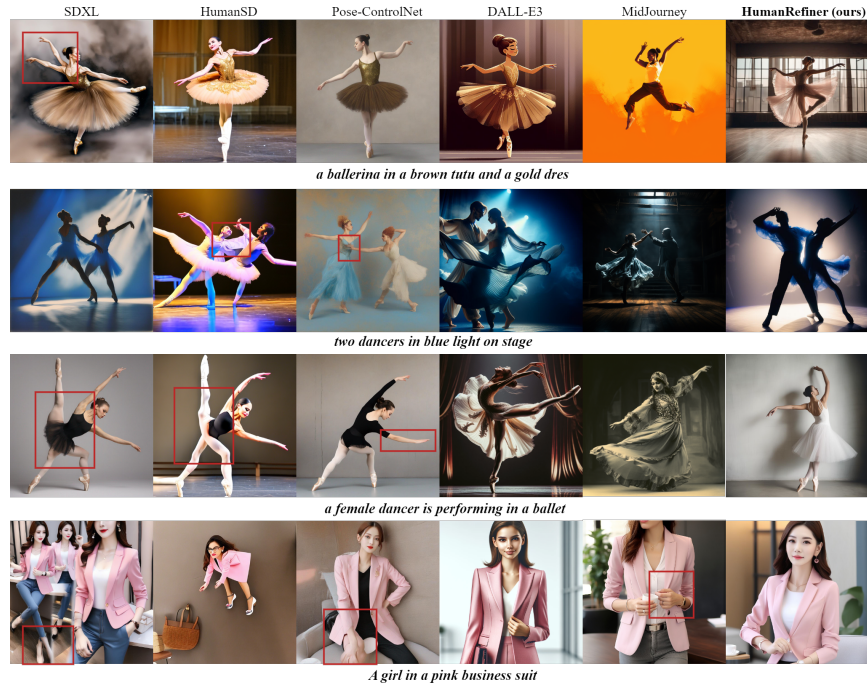
Fig. 9: Qualitative comparison with baselines conducted on humans engaged in activity prompts.

6   Conclusion and Future work

In this paper, we introduce AbHuman, a benchmark including fine-grained annotations of abnormal limbs. AbHuman comprises 56K generated human images along with 147K annotations for abnormal limbs at the bounding box level. Leveraging this benchmark, we construct HumanRefiner, a plug-and-play human refinement module that could be integrated into any text-to-image diffusion model.

Integraed with abnormal guidance, pose-reversible guidance, and detector-guided inpainting, HumanRefiner identifies abnormal limbs and removes them during the generation process, resulting in high-quality human image synthesis. Quantitative experimental results on the AbHuman test split, along with human evaluation results, demonstrate the superiority of HumanRefiner. We envision that the AbHuman Benchmark will facilitate further exploration of methods in the field of human image generation.

Limitation and Future Work While HumanRefiner is capable of generating high-quality human images, the process involves dual generation steps and inpainting, resulting in a limited inference speed. Furthermore, the AbHuman Dataset lacks broader scene annotations for normal human images. We would continue to improve the diversity of the AbHuman benchmark. We believe that richer annotations will further enhance the performance of HumanRefiner and foster the development of more effective models.

Acknowledgments

References

1. Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., Schiele, B.: Posetrack: A benchmark for human pose estimation and tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5167–5176 (2018)
2. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems pp. 1877–1901 (2020)
4. Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., Kim, S.: Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset (2022)
5. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
6. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7291–7299 (2017)
7. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. arXiv preprint arXiv:2301.00704 (2023)
8. Charles, J., Pfister, T., Magee, D., Hogg, D., Zisserman, A.: Personalizing human video pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3063–3072 (2016)
9. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis (2023)
10. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis (2021)
11. Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., Taigman, Y.: Make-a-scene: Scene-based text-to-image generation with human priors. In: European Conference on Computer Vision. pp. 89–106 (2022)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851 (2020)
13. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. The Journal of Machine Learning Research pp. 2249–2281 (2022)

14. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
15. Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., Zhou, J.: Composer: Creative and controllable image synthesis with composable conditions. arXiv preprint arXiv:2302.09778 (2023)
16. Iqbal, U., Milan, A., Gall, J.: Posetrack: Joint multi-person pose estimation and tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2011–2020 (2017)
17. Ju, X., Zeng, A., Wang, J., Xu, Q., Zhang, L.: Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
18. Ju, X., Zeng, A., Zhao, C., Wang, J., Zhang, L., Xu, Q.: Humansd: A native skeleton-guided diffusion model for human image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15988–15998 (2023)
19. Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10124–10134 (2023)
20. Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.S., Lu, C.: Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10863–10872 (2019)
21. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22511–22521 (2023)
22. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
23. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference (2023)
24. Lv, W., Xu, S., Zhao, Y., Wang, G., Wei, J., Cui, C., Du, Y., Dang, Q., Liu, Y.: Detrs beat yolos on real-time object detection (2023)
25. Madhu, P., Villar-Corrales, A., Kosti, R., Bendschus, T., Reinhardt, C., Bell, P., Maier, A., Christlein, V.: Enhancing human pose estimation in ancient vase paintings via perceptually-grounded style transfer learning. ACM Journal on Computing and Cultural Heritage **16**(1), 1–17 (2022)
26. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
27. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
28. Narasimhaswamy, S., Nguyen, T., Huang, M., Hoai, M.: Whose hands are these? hand detection and hand-body association in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4889–4899 (2022)
29. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)

30. OpenAI: Improving image generation with better captions (2023), `https://cdn.openai.com/papers/dall-e-3.pdf`, online; accessed 16-November-2023

31. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis (2023)

32. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis (2023)

33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

34. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)

35. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)

36. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)

37. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)

38. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)

39. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)

40. Sauer, A., Karras, T., Laine, S., Geiger, A., Aila, T.: Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. arXiv preprint arXiv:2301.09515 (2023)

41. Schuhmann, C.: Laion-aesthetics predictor v2 (2023), `https://laion.ai/blog/laion-aesthetics/`, online; accessed 16-November-2023

42. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)

43. Segmind: Announcing ssd-1b: A leap in efficient t2i generation (2023), `https://blog.segmind.com/introducing-segmind-ssd-1b/`, online; accessed 16-November-2023

44. Shonenkov, A., Konstantinov, M., Bakshandaeva, D., Schuhmann, C., Ivanova, K., Klokova, N.: Deepfloyd if: A powerful text-to-image model that can smartly integrate text into images (2023), `https://www.deepfloyd.ai/deepfloyd-if`, online; accessed 16-November-2023

45. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)

46. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789 **2**(3), 5 (2022)
47. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
48. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)
49. Zhang, P., Yang, L., Lai, J.H., Xie, X.: Exploring dual-task correlation for pose guided person image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7713–7722 (2022)
50. Zhang, S.H., Li, R., Dong, X., Rosin, P., Cai, Z., Han, X., Yang, D., Huang, H., Hu, S.M.: Pose2seg: Detection free human instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 889–898 (2019)