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

1 More Details on Dataset Collection

In this section, we provide more details on dataset collection including the image generation process of our AbHuman dataset, the annotation process, and a detailed illustration of class definition with visualized examples.

Image Generation Our dataset prompts are derived from three sources: Laion dataset, utilizing prompts extracted through keyword analysis (e.g., "paper airplane in children's hands on a yellow background and blue sky on a cloudy day"); the second portion comprises textual descriptions from Human-Art (e.g., "acrobatics, a woman performing aerial acrobatics on stage"); and the third part consists of language text generalized by GPT-3.5 based on scene descriptions from Human-Art (e.g., "movie, a woman smoking a cigarette while sitting on a couch").

After obtaining a substantial amount of text related to humans, we employed SDXL for image generation with a resolution of 1024*1024. Subsequently, we conducted a preliminary screening of the generated images. Using the Resnet50 classification model with the category 'person,' we identified images exhibiting a higher degree of relevance to humans.

Class Definition After observing a large number of generated images, we analyzed anomalies in existing human images. Eventually, we categorized the anomalies into 18 classes, annotated with indexes ranging from (0-17). Concerning the generation of images, we categorized them into ten classes: Normal People (0), Non-human (9), and Abnormal Annotations (1-8). To enhance the discriminative capabilities of the detector for both normal and abnormal instances, we introduced a set of authentic images. Notably, we annotated the normal parts as (10-17). The specific annotation results are illustrated in the accompanying Table 3 and 4.

⋆ Equal Contribution.
⋆⋆ Corresponding author.

Annotation Process An annotator employs labeling tools labelme [4] to annotate all objects within an image on a bounding box level, based on the category definitions and examples provided in Table 3 and 4. The categories labeled in the annotated image are then preserved as annotation text, as depicted in the Class & explain column. This categorical text would be concatenated with the original descriptive text of the image, and such aggregated text is utilized for Negative prompt training.

2 More Details on Abnormal Classifier and Abnormal Detector

2.1 Abnormal Detector

We split the dataset into a training set and a test set with a ratio of 4:1. Subsequently, we fine-tuned YOLOv8 [3] and RT-DETR [2] to serve as our detectors. The model parameters used for fine-tuning YOLOv8 are detailed in the table 1. The results obtained on the test set are summarized in the table 2. Among them, YOLOv8n and YOLOv8x are two models in the YOLOv8 series, and YOLOv8x contains more layers and parameters. In our pursuit of enhanced detection performance, we incorporated YOLOv8x as the primary detector in our pipeline.

Table 1: The parameter Settings of the YOLOv8 model.

Parameter	**nc**	**batch**	**patience**	**pre-trained**	**lr0**	**weight_decay**	**NMS**	**iou**
Value	18	32	50	true	0.01	0.0005	false	0.7
Parameter	**epochs**	**images**	**workers**	**optimizer**	**momentum**	**augment**	**conf**	**max_det**
Value	120	640	8	auto	0.937	false	null	300

Table 2: The detection results of abnormal limbs on the test set are presented for the AbHuman dataset.

Model	Yolov8n	Yolov8x	RT-DETR
mAP50 ↑	0.333	**0.426**	0.331
mAP50-95 ↑	**0.282**	0.235	0.188

2.2 Abnormal Scorer

Training Details We finetune the Abnormal Scorer with a learning rate of 1.0e-4 using AdamW [1] optimizer with a batch size of 256 for 30epochs on the AbHuman training split. Images assigned with abnormal labels are positive samples while normal ones are negative samples. A sigmoid layer is followed by the last linear layer and binary cross-entropy is utilized as the objective function.

3 More Visualization of HumanRefiner

We present the intermediate results of our HumanRefiner pipeline including the output of abnormal guidance, detected pose, pose-guided outputs, and the output of inpainting in Figure 1.

Table 3: Annotations examples, class definition and corresponding textual explanation of AbHuman (Part 1).

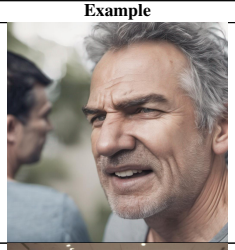
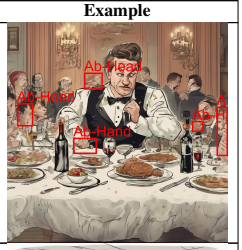
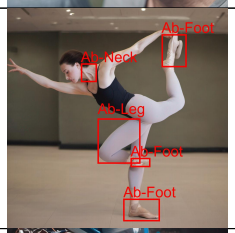
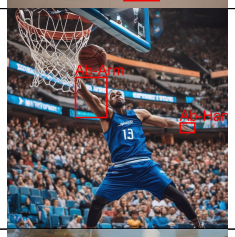
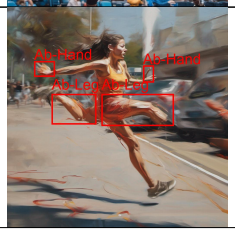
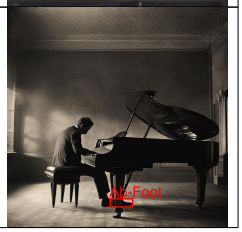
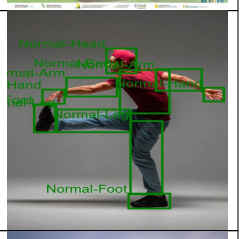
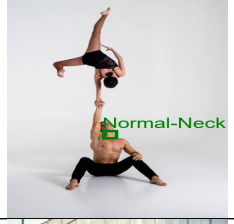
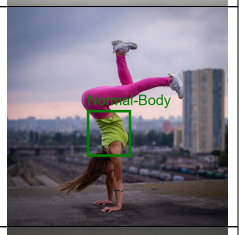
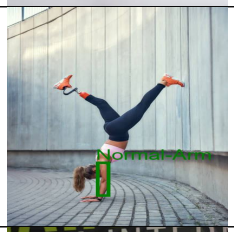
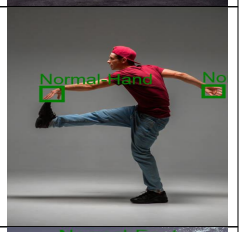
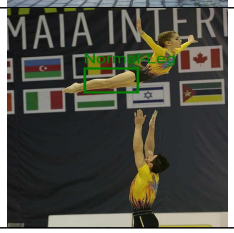
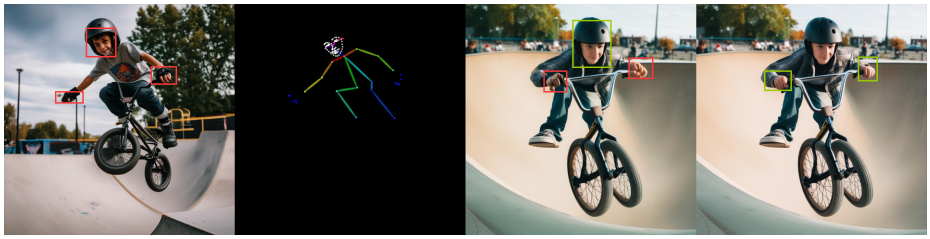
Class & explain	Index	Example	Class & explain	Index	Example
Normal Human: The content of this picture is a normal image with no abnormal human limbs.	"0"		Abnormal Head: The content of this picture is a human image with an abnormal head.	"1"	
Abnormal Neck: The content of this picture is a human image with an abnormal neck.	"2"		Abnormal Body: The content of this picture is a human image with an abnormal body.	"3"	
Abnormal Arm: The content of this picture is a human image with an abnormal arm.	"4"		Abnormal Hand: The content of this picture is a human image with an abnormal hand.	"5"	
Abnormal Leg: The content of this picture is a human image with an abnormal leg.	"6"		Abnormal Foot: The content of this picture is a human image with an abnormal foot.	"7"	

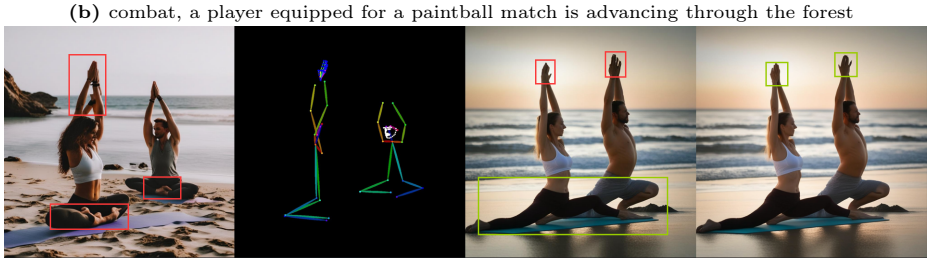
Table 4: Continuation of annotations examples, class definition and corresponding textual explanation of AbHuman (Part 2).

Abnormal Multi: This painting is about the abnormal physical contact of multiple people.	"8"		Non Human: There are no human beings in the contents of this painting.	"9"	
Normal Multi: This painting is about the normal physical contact of multiple people.	"10"		Normal Head: The content of this picture is a human image with a normal head.	"11"	
Normal Neck: The content of this picture is a human image with a normal neck.	"12"		Normal Body: The content of this picture is a human image with a normal body.	"13"	
Normal Arm: The content of this picture is a human image with a normal arm.	"14"		Normal Hand: The content of this picture is a human image with a normal hand.	"15"	
Normal Leg: The content of this picture is a human image with a normal leg.	"16"		Normal Foot: The content of this picture is a human image with a normal foot.	"17"	

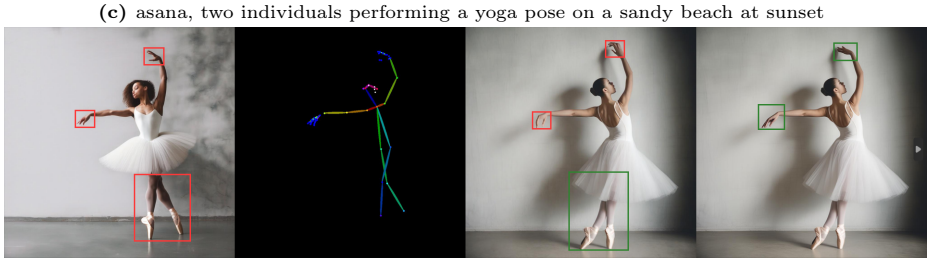


(a) airtime, a young BMX rider in casual attire is catching air at a skatepark

(b) combat, a player equipped for a paintball match is advancing through the forest



(c) asana, two individuals performing a yoga pose on a sandy beach at sunset



(d) dance, a ballet dancer in a white dress is standing in front of a wall

Fig. 1: Intermediate outputs of the HumanRefiner Pipeline from left to right: (1) Abnormal Guidance, (2) Pose Detection, (3) Pose-Guided Coarse Refinement, and (4) Inpainting Output. Highlight all the abnormal parts with a red box and mark the corresponding fixed normal parts with a green box.

References

1. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
2. 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)
3. 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)
4. Wada, K.: labelme: Image polygonal annotation with python. `https://github.com/wkentaro/labelme` (2018)