

# Is this Generated Person Existed in Real-world? Fine-grained Detecting and Calibrating Abnormal Human-body

## Supplementary Material

We highly recommend watching the supplementary video, as it comprehensively demonstrates our proposed task and the results of our proposed HumanCalibrator.

**Disclaimer:** The supplementary material includes images that may be unsettling or discomforting to some readers. We have removed all personal information from the cases and applied mosaic to some images that may cause discomfort.

### A. Details in HumanCalibrator

#### A.1. Model Usage

In addition to using LLaVAv1.5-7B as the base model for the Absent Human-body Detector, the other models employed in the HumanCalibrator are as follows: (1) The inpainting model  $R$ , which utilizes StableDiffusion2-Inpainting<sup>1</sup>, (2) the grounding model  $G$ , which adopts GroundingDINO<sup>2</sup>, and (3) the video interpolation model, which employs CogVideoX-Interpolation<sup>3</sup> based on CogVideoX [68].

#### A.2. Other Implementation Details

Additional details in our HumanCalibrator are as follows: (1) To improve the repair quality of the overall human photo, we expand the bounding box of the abnormal region before applying inpainting. This ensures better visual quality between the inpainted and surrounding regions. (2) Since the inpainting model inevitably leads to a decline in overall image quality, we apply  $2\times$  super-resolution processing to the inpainted images. It is worth noting that, for a fair comparison, no super-resolution processing is applied in any of the comparisons in Table 2. (3) To better adapt the Absent Human-body Detector, trained on real-world COCO datasets, for application in AIGC, we perform semantic detection on each absent region identified by the AHD using the Grounding Model  $G$ . If the same semantic content is detected, the result from this iteration of the Absent Human-body Detector is discarded.

<sup>1</sup><https://huggingface.co/stabilityai/stable-diffusion-2-inpainting>, Rombach, R. et al. (2022). High-Resolution Image Synthesis With Latent Diffusion Models. In Proc. CVPR2022.

<sup>2</sup><https://github.com/IDEA-Research/GroundingDINO>, Liu, S. et al. (2023). Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv

<sup>3</sup><https://huggingface.co/feizhengcong/CogvideoX-Interpolation>

### B. Details of Baselines and Analysis

#### B.1. Baseline for COCO Human-Aware Val

The COCO Human-Aware Val dataset only contains absent abnormalities resulting from masking out body parts. Since it is derived from real-world images and includes only the “absent” category of abnormalities, our comparisons on this dataset primarily focus on two objectives: (1) demonstrating the deficiency of existing VLMs in abnormality perception and (2) evaluating the performance of our trained Absent Human-body Detector (AHD).

Evaluating the baseline of CLIP on COCO Human-Aware Val: Similar to other methods, we transform the different types of abnormalities into a classification problem. The CLIP model selects the text with the highest matching score to the image as its predicted answer. The specific text categories are as follows:

- “The person in the picture has absent head.”
- “The person in the picture has absent ear.”
- “The person in the picture has absent arm.”
- “The person in the picture has absent hand.”
- “The person in the picture has absent foot.”
- “The person in the picture has absent leg.”
- “The person in the image has no abnormalities.”

Evaluating the baseline of Generative VLMs, we use the following prompt to like VQA tasks [8, 34] to prompt the VLMs:

- “Are there any absent body parts in the person shown in the image? If yes, please answer from ‘head’, ‘arm’, ‘leg’, ‘foot’, ‘hand’, or ‘ear’; otherwise, please answer ‘no’. Answer the question using a single word:”

#### B.2. Baseline for AIGC Human-Aware 1K

Unlike the COCO Human-Aware Val dataset, the AIGC Human-Aware 1K dataset includes all categories of abnormalities. For CLIP, we directly add additional abnormal categories and use a similar classification approach to evaluate its performance. For Generative VLMs, we adopt a simpler method tailored to VLMs. Specifically, we separately ask whether there were abnormalities in the “redundant” category and the “absent” category. Additionally, since the abnormalities in AIGC Human-Aware 1K are diverse in number, we do not constrain the model’s responses to a single word, i.e., we do not use “answer the question using a single word”. After receiving the responses, we use an LLM for post-processing to produce formatted data suitable for accuracy calculation. Since these baseline VLMs perform

weakly on FHAD, we try various prompts per model to optimize performance in our experiments. The prompts that yielded the best performance are displayed below:

- For **LLaVA-34B**:
  - In “Absent Abnormality Detection”: *“Are there any missing body parts in the person shown in the image? If so, please answer the precise part:”*
  - In “Redundant Abnormality Detection”: *“Are there any extra body parts in the person shown in the image? If so, please answer the precise part:”*
- For **Intern VL2-26B**:
  - In “Absent Abnormality Detection”: *“According to the human anatomical structure, are there any missing body parts in the person shown in the image? If so, please answer the precise part:”*
  - In “Redundant Abnormality Detection”: *“According to the human anatomical structure, are there any extra body parts in the person shown in the image? If so, please answer the precise part:”*
- For **GPT-4o**:
  - In “Absent Abnormality Detection”: *“It is a common sense that all human being has one head, two ears, two hands, two arms, two legs and two foots, are there any missing body parts which I discussed in the person shown in the image? If so, please answer the precise part:”*
  - In “Redundant Abnormality Detection”: *“It is a common sense that all human being has one head, two ears, two hands, two arms, two legs and two foots, are there any extra body parts which I discussed in the person shown in the image? If so, please answer the precise part:”*

For the post-process for the response of VLMs (Note that, we use the GPT4o-mini to post-process the response) as shown in Figure S1.

### B.3. Baseline Analysis

Our work is based on a key assumption: that existing powerful VLMs fail to perform abnormality detection, a task that is exceptionally simple for humans. We provide a detailed case analysis of their poor performance. Specifically, there are two primary reasons for this under-performance: (1) a lack of understanding of human body structure, and (2) a misinterpretation of abnormalities. We present examples from **real test** in Figure S2.

### B.4. Pose Condition

Since the code for HumanRefiner [14] is unavailable and our objective differs fundamentally, we only reproduce its step of using pose as an additional constraint to ensure no abnormalities in the number of body parts. Specifically,

Please analyze the model's response about extra or missing body parts and output only a list of the specifically mentioned body parts that are confirmed as extra or missing. Return the result as a simple list of individual words wrapped in <output> tags (e.g. <output>['arm']</output>, <output>['leg', 'hand']</output>). If the response indicates uncertainty, normal body parts, or no abnormalities, return <output>[]</output>.

Input: "The image depicts a person with one visible arm. The other arm appears to be missing or obscured."  
Output: <output>['arm']</output>

Input: "The image shows the upper half of a person. All visible body parts like the head, ears, arms, and hands seem present. Legs and feet are not visible in this image, so a determination about them cannot be made."  
Output: <output>[]</output>

Input: "The person in the image appears to have an extra hand"  
Output: <output>['hand']</output>

Input: "The person in the image appears to have an extra arm and an extra leg"  
Output: <output>['arm', 'leg']</output>

Input:

Figure S1. Prompt for post-processing the VLM output.

for the input human photo, we use MMPose<sup>4</sup> to extract the human pose and then employ Stable-Diffusion-v1.5<sup>5</sup> with t2iadapter\_keypose<sup>6</sup> as a pose-conditioned method to regenerate the entire image.

## C. Why do current VLMs lack the ability to perceive abnormality?

Our extensive experiments demonstrate that existing VLMs are unable to perceive human abnormalities (some cases are shown in Figure S2), even though this task is very simple for humans, and both we humans and the models are trained on a large amount of normal data. We believe that the drawbacks arise from the simplistic image-text alignment approach of existing VLMs, which lacks perception of content and, consequently, an understanding of human body structure. Additionally, the existing VLMs underutilize the data and are undertrained, and the proportion of human subjects in the training data may not be substantial. In our work, we utilize the correlation among human body structures to train our absent human-body detector.

## D. AIGC Human-Aware 1K Annotation

The target of our proposed task, “Fine-grained Human Abnormality Detection”, is to detect whether the human photos in AIGC exhibit abnormalities that render them impossible to exist in the real world. This imposes two requirements on

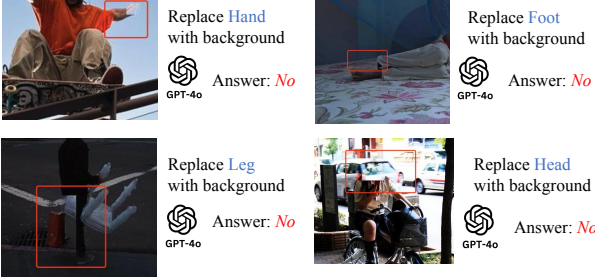
<sup>4</sup>MMPose Contributors. (2020). OpenMMLab Pose Estimation Toolbox and Benchmark. Retrieved from <https://github.com/open-mmlab/mmpose>

<sup>5</sup><https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5>, Rombach, R. et al. (2022). High-Resolution Image Synthesis With Latent Diffusion Models. In Proc. CVPR2022.

<sup>6</sup><https://github.com/TencentARC/T2i-Adapter>, Mou, C. et al. (2023). T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv

### Failure Cases In COCO Human-Aware Val

Question: Are there any absent body parts in the person shown in the image? If yes, please answer from 'head', 'arm', 'leg', 'foot', 'hand', or 'ear'; otherwise, please answer 'no'. Answer the question using a single word:



### Failure Cases In AIGC Human-Aware 1K

Question: It is a common sense that all human being has one head, two ears, two hands, two arms, two legs and two feet, are there any missing body parts which I discussed in the person shown in the image? If so, please answer the precise part:

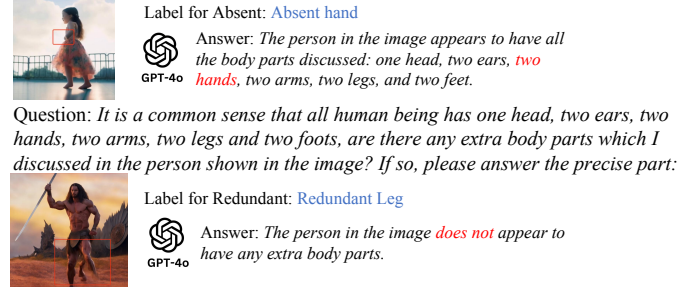


Figure S2. Failure cases of the powerful VLM (GPT-4o) on COCO Human-Aware Val and AIGC Human-Aware 1K. For COCO Human-Aware, it is observed that despite generating distinctly anomalous images, GPT-4o still responds with a definitive “No”. In the case of AIGC Human-Aware 1K, even though GPT-4o is aware of the components that constitute a normal human body, it fails to recognize or respond to abnormalities. Note that our prompt includes the category of abnormalities, which simplifies the task; however, GPT-4o still struggles to perform effectively, resulting in poor baseline performance.

our evaluation data: (1) The annotated abnormalities must be objective, avoiding controversial cases caused by ambiguity or other factors. (2) The human photos in the annotations must appear in real-world environments, which necessitates selecting realistic styles for annotation and excluding sci-fi or cartoon-style images. In Figure S3, we demonstrate examples of cases that are manually filtered out during the annotation process. After the initial annotation, to ensure data objectivity, we conduct multi-round and multi-reviewer checks on the data labels, removing any remaining controversial annotations. This process ensures the quality of our proposed AIGC Human-Aware 1K dataset. We provide statistics on the number of different annotation types in AIGC Human-Aware 1K, as shown in Table S3.

## E. Metric Details

It is essential to emphasize that a comprehensive evaluation of our proposed task requires the integration of multiple metrics. Specifically, we employ Accuracy (ACC) and False Discovery Rate (FDR) as detection metrics to ascertain the correct identification of existing abnormalities. Furthermore, we utilize the CLIP score, the Human CLIP Score, and the Human Concept Score to evaluate the reasonableness of the identified abnormal locations and to assess the quality of the repairs to these abnormalities. Additionally, we use the Fréchet Inception Distance (FID) and Latent Consistency to examine the similarity between our repaired

human photos and the original human photos, demonstrating the granularity of our repairs; i.e., we only repair the abnormal areas while preserving the other content.

### E.1. Human CLIP Scores

Since our task focuses on repairing the human body in a given human photo, directly using the original prompt to calculate the CLIP score is not ideal, as it includes substantial background and camera-related information. Instead, we utilize GPT4o-mini to extract prompts specifically related to the human body to evaluate whether our repair improves the correlation with human-related prompts, a metric we refer to as the Human CLIP Score. An example case is shown below:

- Original Prompt: “A girl with long hair is walking on the avenue in the forest, with a gentle breeze blowing her hair and falling leaves fluttering in the wind. The girl looks melancholy in the distance.”
- Processed Human Prompt: “A girl with long hair is walking on the avenue in the forest, looking melancholy into the distance.”

### E.2. Human Concept Scores

Compared to the Human CLIP Score, which focuses more on the quality of the human body in the repaired human photo, the Human Concept Score emphasizes evaluating whether the repaired human conceptually aligns more closely with the distribution of “human” as understood by CLIP, trained on extensive real-world data. To verify this, we use a straightforward method: calculating the similarity between the human photo before and after repair and the prompt “an image contains human” to examine whether the repaired human better matches CLIP’s concept of a human existing in the real-world which learned from diverse real-world training data.

Type	Absent	Redundant	No Abnormality
Number	649	158	343

Table S3. Statistics on the Number of Annotation Types in AIGC Human-Aware 1K

### ⊗ Filtered samples



Figure S3. Categories and examples filtered out during the annotation process for AIGC Human-Aware 1K. The goal of our proposed task, “Fine-grained Human-body Abnormality Detection”, is to determine whether the body structure in a given human photo could exist in the real world. Thus our annotated data are grounded in real-world contexts, leading us to exclude images of genres such as science fiction and cartoons. Additionally, to enhance the dataset’s quality, we filter out samples where the specific abnormality cannot be ascertained or where the abnormality is controversial, labeled as “Too Low-quality” or “Abnormality is not objective”. All NSFW images have also been excluded, and *the displayed samples have been processed with mosaic*. These rigorous criteria not only ensure the quality of our AIGC Human-Aware 1K dataset but also explain why annotating a large number of data for the training process directly from AIGC is costly.

### E.3. Visual Consistency

For the FID, we treat the repaired images as generated images and calculate the distributional discrepancy between them and the original images. For Latent Consistency, we encode the images into the latent space via the CLIP Visual Encoder and compute the cosine similarity between the original and repaired images.

### F. Cases in COCO Human-Aware Val

We also provide examples of the Absent Human-body Detector’s performance on the COCO Human-Aware Val, as shown in Figure S4. It shows that the trained Absent Human-body Detector accurately identifies the locations and the type of artificially created abnormalities.

### G. More Cases

In Figure S6, we provide additional examples, including results for test cases with no abnormalities and some cases in complex scenarios (more than one abnormality). Additionally, we present several failure cases, which primarily fall into four categories, as shown in Figure S7: (1) **missing detection** (2) **error identifying similar body parts (such as hand and arm, foot and leg)** (3) **false detection in the normal human figure** (4) **correct detection but the wrong area** (5) **all good but low-quality repairing**. It is worth to note that, not only does the failure happen in the detection phase, but the inpainting model may also fail to repair the figure even with

Inpainting Model	Absent		Redundant	
	Avg Acc%↑	Avg FDR%↓	Avg Acc%↑	Avg FDR%↓
Kandinsky2	59.5	8.9	47.7	7.6
SDXL	64.9	11.5	48.4	3.7
SD2	80.7	8.5	58.6	2.5
LAMA	—	—	65.1	15.6

Table S4. Performance with different inpainting models.

a good detection result.

### H. Performance with other Inpainting Models

We obtain quantitative and qualitative results with different kinds/ability inpainting models in Table S4 and Figure S5, including the inpainting model, such as Kandinsky2<sup>7</sup> and SDXL<sup>8</sup>. Furthermore, we also test an inpainting model that focuses on removing objects, LAMA<sup>9</sup>. For LAMA, although it increases the acc (6.5%) in detecting the redundant body parts, it tends to generate a rich background and further remove the normal body parts, leading to a significant increase in FDR (13.1%).

<sup>7</sup><https://github.com/ai-forever/Kandinsky-2>

<sup>8</sup><https://huggingface.co/papers/2307.01952>

<sup>9</sup><https://github.com/advimman/lama>

### Cases in COCO Human-Aware Val

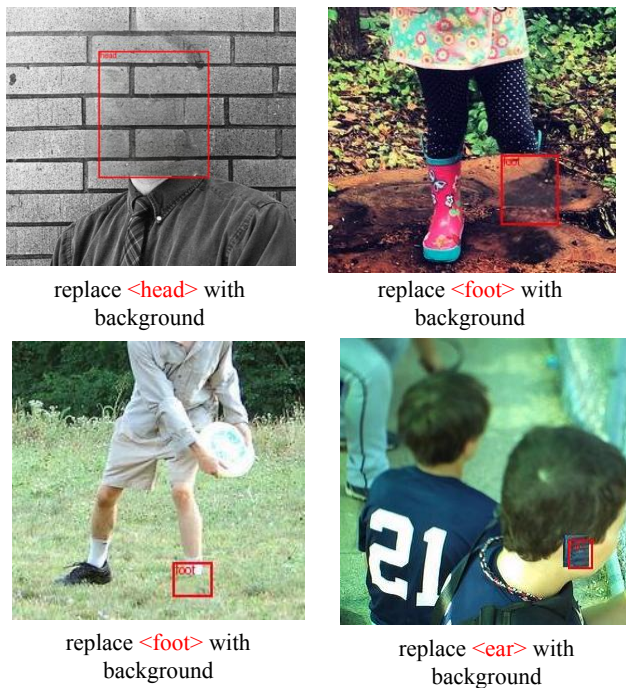


Figure S4. Examples of the AHD on COCO Human-Aware Val. The red boxes indicate the predictions made by AHD. It is observable that AHD, trained utilizing the correlation within human body structures, can accurately identify the location and type of artificially created abnormalities. Note that all personal information has been removed from the cases displayed. The training set created from the COCO Train Split is in a similar format.

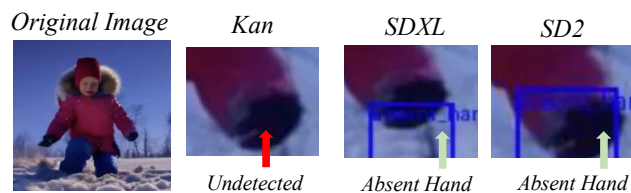


Figure S5. Qualitative results with different inpainting models. Kan denotes Kandinsky2.





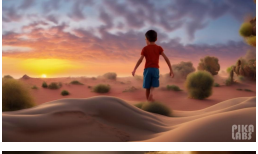
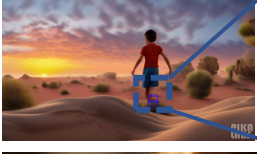

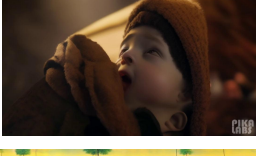

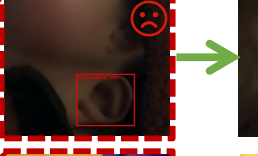



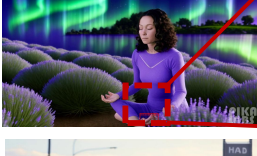


Origin Image	Human Calibrator		
	Detect		Repair
			
			
			
			
			
			
			

Figure S6. More Cases in HumanCalibrator. The anomalies in human figures exhibit significant diversity, appearing in various locations, differing in categories, and varying in quantity. Moreover, under occlusions and other challenging conditions, determining whether an anomaly is present can be difficult. To better capture these characteristics, we aim to enrich our dataset by increasing the proportion of diverse samples.

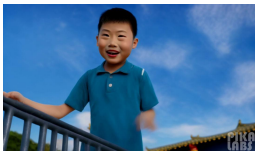
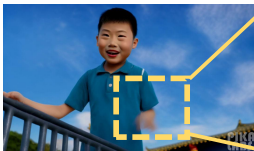

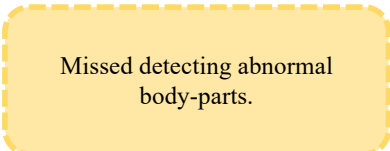



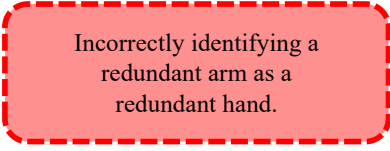

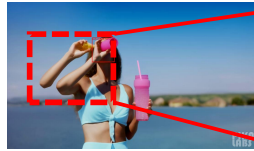

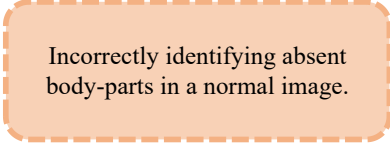

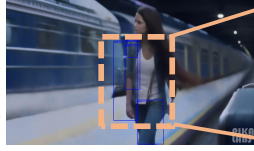

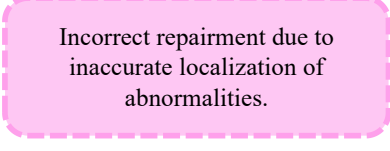

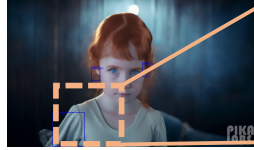
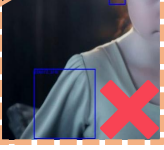
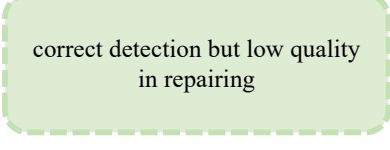

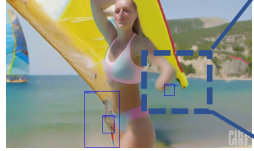

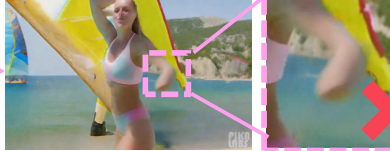

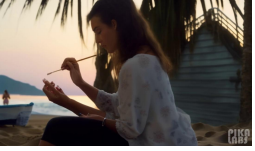




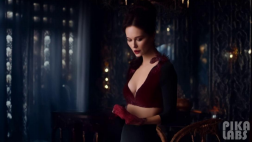

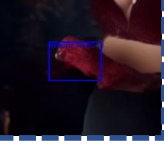
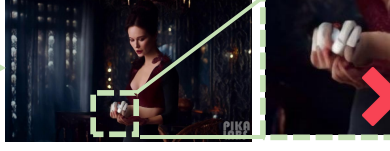

Origin Image	Detect	Human Calibrator	Repair
<p>Missed Detection</p> 	<p>'Absent arm'</p>  	 <p>Missed detecting abnormal body-parts.</p>	
<p>Missed Detection</p> 	<p>'Redundant Arm'</p>  	 <p>Incorrectly identifying a redundant arm as a redundant hand.</p>	
<p>Incorrect abnormality identification</p> 	<p>'Redundant Arm'</p>  	 <p>Incorrectly identifying absent body-parts in a normal image.</p>	
<p>Fake positive detection</p> 	<p>'No Abnormality'</p>  	 <p>Incorrect repairment due to inaccurate localization of abnormalities.</p>	
<p>Fake positive detection</p> 	<p>'Absent ear'</p>  	 <p>correct detection but low quality in repairing</p>	
<p>Inaccurate localization of abnormality</p> 	<p>'Absent hand Absent ear'</p>  	 	
<p>Correct class but inaccurate localization of abnormality</p> 	<p>'Redundant Hand'</p>  	 	
<p>Correct detection but low quality in repairing</p> 	<p>'Absent hand'</p>  	 	

Figure S7. Failure cases in HumanCalinratpr. For such a difficult and challenging task, it is necessary to further investigate its failure cases under different scenes. In the process of HumanCalibrator, failure mainly comes from five aspects (1) missing detection (2) error identifying similar body parts (such as hand and arm, foot and leg) (3) false detection in the normal human figure (4) correct detection but the wrong area (5) all good but low-quality repairing.