We Still See Broken Limbs: Towards Anatomical Realism in GenAI via Human Preference Learning

Supplementary Material

A. Additional Experiments

This supplementary document provides additional experiments, ablations, and diagnostic analyses that further contextualize the findings presented in the main paper. It includes: (i) Correlation analysis between human preference rating and ViTPose-H Score, (ii) ablation studies on Best-of-N sampling, (iii) a pilot study assessing whether GPT-5 can function as an anatomical judge.

A.1. ViTPose-H Score for Human Preference

We show in Sec. 5.2.1, that we can approximate anatomical realism using the ViTPose-Huge score model. We assess its correllation with preference win rates collected in BACON. Fig. 1 shows the preference win rate of different image generation pipelines over SDXL in relation to the ViTPose-H Score obtained for these images.

Result. We observe a linear correlation between the learned ViTPose-H Score and preference winrate with a Person correlation coefficient of $r = 0.94$. This implies that this score is a solid proxy for predicting human preference.

A.2. Ablation Studies on Best-of-N Sampling

In Best-of-N sampling, we evaluate pose-based scoring functions derived from keypoint confidences. To suppress noise from occluded or invisible limbs, we apply a minimum confidence threshold θ and compare scores for thresholds between 0.0 and 0.5. Additionally, the aggregation of per-person confidence scores into a single anatomical plausibility score was tested using minimum, mean, and maximum operators.

Results. As shown in Tab. 1, thresholds in the range 0.2–0.4 achieve the highest agreement with human preference: lower values preserve noisy low-confidence joints, while higher values remove too much pose information. This trend is consistent across OpenPose, Sapiens, and the ViTPose family. We further see that OpenPose benefits most from minimum aggregation, while ViTPose-Large and ViTPose-Huge show slightly better performance with maximum aggregation. Sapiens and ViTPose-Base exhibit no strong dependence on aggregation method. We state that minimum aggregation is the most feasible, as an image will be rated as unrealistic even if a only single part (or person) depicted shows anatomical errors. Despite these differences, all confidence-based approaches remain below 60% agreement with human preference, indicating that confidence-only scoring is fundamentally limited for anatomical realism.

A.3. Pilot Study: ChatGPT as Anatomical Judge

In a small pilot study, we examined whether a vision-enabled large language model (LLM), specifically ChatGPT [4], can function as an automatic judge of anatomical correctness.

Motivation. Recent work demonstrates that multimodal large language models (MLLMs), like ChatGPT, can serve as effective evaluators of text-to-image generation quality. VisionPrefer [6] uses an MLLM to construct large-scale

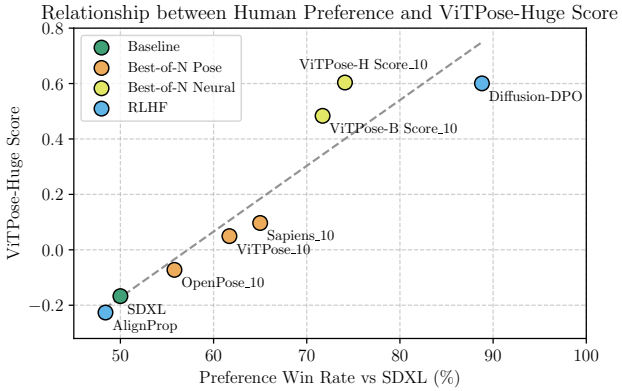

Figure 1. Model preference vs SDXL compared to the average ViTPose-H score of the model. The strong linear trend (Pearson correlation coefficient $r = 0.94$) confirms that the ViTPose-H score model reliably predicts relative human preference across models.

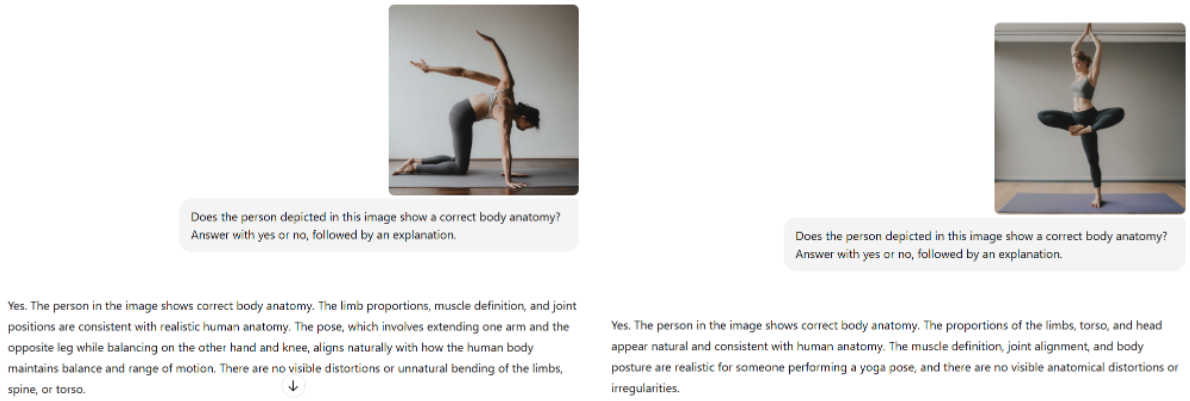

Figure 2. Example conversations with GPT-5 [4]. The model was asked whether the depicted person shows anatomically correct body proportions and structures. Despite visible deformations, GPT-5 often assessed the images as anatomically correct.

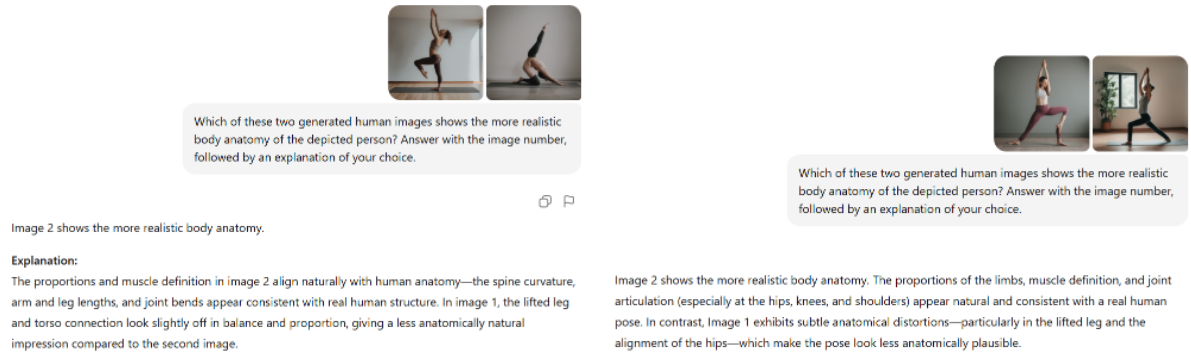

Figure 3. Example conversations with GPT-5 [4]. The model was asked to identify which of two generated human images exhibits more realistic anatomy. Even when one image showed clear structural errors, GPT-5 frequently failed to select the correct option. Even when choosing the correct option, the provided reasoning missed the major anatomical errors.

preference annotations covering prompt adherence, aesthetics, fidelity, and harmlessness. EvalAlign [7] further shows that fine-tuning multimodal LLMs with human-aligned supervision yields reliable metrics for text–image alignment and image faithfulness. Complementarily, TIFA [8] formulates faithfulness evaluation as visual question answering, enabling interpretable and fine-grained assessment of prompt compliance. Motivated by this evidence, we investigate whether MLLMs can be employed to assess the anatomical realism of AI-generated human images. If successful, such models could function both as automatic evaluation metrics and as scalable annotators.

Experiment & Results. For each generated image from a small subset of BACON, the model was prompted to identify anatomical mistakes or confirm their absence. As seen in Fig. 2 and Fig. 3, ChatGPT failed to detect the major anatomical anomalies present in the images. Severe distortions—such as extra or missing limbs, incorrect joint geometry, disconnected body parts, or globally implausible body poses—were frequently overlooked. Even images labeled by human annotators as *completely deformed* were often described by ChatGPT as showing *no anatomical issues*. When the model did mention problems, its descriptions were typically inaccurate or unrelated to the deformation.

Conclusion These findings align with recent work highlighting limitations of vision–language models for fine-grained visual error detection. In our analysis, we consider ChatGPT as an illustrative example of a contemporary LLM-based vision system, without claiming that the observed behavior necessarily generalizes to other models. Prior studies have reported that such systems can recognize high-level semantic content but often struggle with subtle geometric inconsistencies, structural defects, or human-body–specific anomalies. While LLM-based vision models appear effective at global scene understanding, they may lack the specialized inductive biases required for reliably assessing pose correctness or human anatomical coherence.

Table 1. Accuracy of pose-estimator (OpenPose [1], ViTPose [2], Sapiens [3]) confidence scoring in predicting human preference. Values indicate how often the image with the higher score matches the image preferred by annotators. Best-performing configuration per model backbone is highlighted in bold. All values are in [%].

Aggr.	θ	**OpenPose**	**Sapiens**	**ViTPose-B**	**ViTPose-L**	**ViTPose-H**
Min	0.0	55.1	58.3	55.6	53.9	54.8
	0.1	52.7	58.5	52.4	49.9	50.8
	0.2	52.5	**59.0**	55.0	52.6	53.3
	0.3	52.4	59.0	**55.5**	53.1	53.4
	0.5	51.6	58.5	55.4	53.9	54.8
Mean	0.0	**52.9**	54.2	52.9	52.4	53.9
	0.1	52.7	55.0	52.6	52.1	53.1
	0.2	52.9	55.3	53.1	52.3	53.1
	0.3	52.8	55.5	53.2	52.4	53.3
	0.5	51.8	55.3	52.7	52.2	52.8
Maximum	0.0	52.4	55.9	53.3	54.1	55.3
	0.1	52.7	56.3	53.5	54.0	55.0
	0.2	52.8	56.8	53.8	54.3	55.4
	0.3	52.8	57.2	54.0	**54.8**	**55.8**
	0.5	52.5	56.8	53.9	54.5	55.6

References

[1] Z. Cao *et al.*, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," in *CVPR*, 2017. 3

[2] H. Xia *et al.*, "VitPose: Multi-view 3D Human Pose Estimation with Vision Transformer," in *ICCC*, 2022. 3

[3] R. Khirodkar *et al.*, "Sapiens: Foundation for Human Vision Models," in *ECCV*, 2024. 3

[4] OpenAI, "ChatGPT-5: Large-Scale Multimodal Language Model," Tech. Rep., 2025. 1, 2

[5] Z. Wang *et al.*, "Is This Generated Person Existed in the Real World? Fine-Grained Detecting and Calibrating Abnormal Human Body," in *CVPR*, 2025.

[6] X. Wu *et al.*, "Multimodal Large Language Models Make Text-to-Image Generative Models Align Better," in *NeurIPS Poster*, 2025. 1

[7] Z. Tan *et al.*, "EvalAlign: Supervised Fine-Tuning Multimodal LLMs with Human-Aligned Data for Evaluating Text-to-Image Models," arXiv preprint, 2024. 2

[8] Y. Hu *et al.*, "TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering," arXiv preprint, 2023. 2