ChatPose: Chatting about 3D Human Pose - Supplemental Material

Yao Feng[1,2,3] Jing Lin[3,4] Sai Kumar Dwivedi[1] Yu Sun[3] Priyanka Patel[1] Michael J. Black[1]
[1]Max Planck Institute for Intelligent Systems - Tübingen [2]ETH Zürich
[3]Meshcapade [4]Tsinghua University

1. Training Data Details

As described in the Method, we construct question and answer pairs to finetune a multi-modal LLM; specifically we use text-to-SMPL pose and image-to-SMPL pose pairs. Details of the question list are illustrated in Table 3 and Table 1, while example answers are shown in Table 2.

- "<image> Can you predict the SMPL pose of the person in this image?"
- "<image> There is a person in the middle of the image, please output this person's SMPL pose."
- "<image> What is the human pose in this image? Please respond with SMPL pose."
- "<image> What is the person doing in this image? Please output SMPL pose."
- "<image> There is a person in the middle of the image, use SMPL to describe the pose."

Table 1. The list of questions for training ChatPose with image-to-SMPL pose pairs.

- "The SMPL pose is <POSE>."
- "It is <POSE>."
- "The SMPL format of this person's pose is <POSE>."
- "Sure, it is <POSE>."
- "Sure, the SMPL pose is <POSE>."
- "<POSE>."
- "The SMPL pose of the person is <POSE>."
- "Sure, <POSE>."

Table 2. The list of answers for training ChatPose with SMPL pose as the output.

2. Benchmark Details

We introduce two benchmarks, speculative pose generation (SPG) and reasoning-based pose estimation (RPE), to evaluate the performance on reasoning about human poses.

SPG Benchmark. Unlike traditional text-to-pose generation tasks, speculative pose generation requires the model to reason about, and interpret, indirect pose descriptions and to generate appropriate 3D poses. Consequently, a novel benchmark for evaluation is necessary. We utilize the PoseScript dataset [2], which provides direct pose descriptions, as a starting point. Subsequently, we visualize the pose from four viewpoints and feed the visual result along with the direct pose description into GPT-4V [6], prompting it to generate implicit descriptions of associated activities, as shown in Figure 1. To improve the generation quality, we design a chain-of-thought mechanism, in which we ask GPT-4V to answer four questions before generating the speculative pose descriptions. The details of the query input are presented in Table 4. We then manually check these labels and construct instruction data containing 780 text-pose pairs formatted as follows: "USER: {descriptions_implicit}, can you give the SMPL pose of this person? ASSISTANT: Sure, it is <POSE>." Here, {description_implicit} represents the speculative queries generated by GPT4.

RPE Benchmark. To establish the reasoning-based pose estimation benchmark, we begin by selecting 50 multiple-person images from the 3DPW [7] test set. Subsequently, we employ GPT4V to generate descriptions of the individuals depicted in these images, covering attributes like behavior, outfits, pose, shape, summary, with summary summarizing all the other attributes. Notably, during our experiments, we observe that GPT4V [6] consistently confuses left and right body parts. Inspired by [9], we incorporate a visual prompt to assist the model in distinguishing between left and right body parts. Specifically, we utilize ViTPose [8] for body keypoint detection, and then visually differentiate left and right body parts with distinct colors on the image and explicitly specify them in the text prompt provided to GPT4V, as shown in Figure 2. The details of the query input are represented in Table 5. After generating these descriptions,

- "I have a word description of a person's pose, can you give the SMPL pose of this person? {`description`}"
- "There is a person: {`description`} Please output this person's SMPL pose."
- "{`description`} Give the SMPL pose."
- "What's the SMPL pose of this person? {`description`}"
- "Use SMPL pose to describe this person's behavior. {`description`}"
- "There is a person doing this: {`description`} Can you use SMPL pose to describe the pose?"
- "A person is described as: {`description`} Use the SMPL pose to reflect this."
- "Human pose is described as words: {`description`} The SMPL pose is?"
- "Human pose can be described as words: {`description`} And it can also be described in SMPL pose format, can you output this?"

Table 3. The list of questions for training ChatPose with text-to-SMPL pose pairs. Where {`description`} is the text description from the dataset.

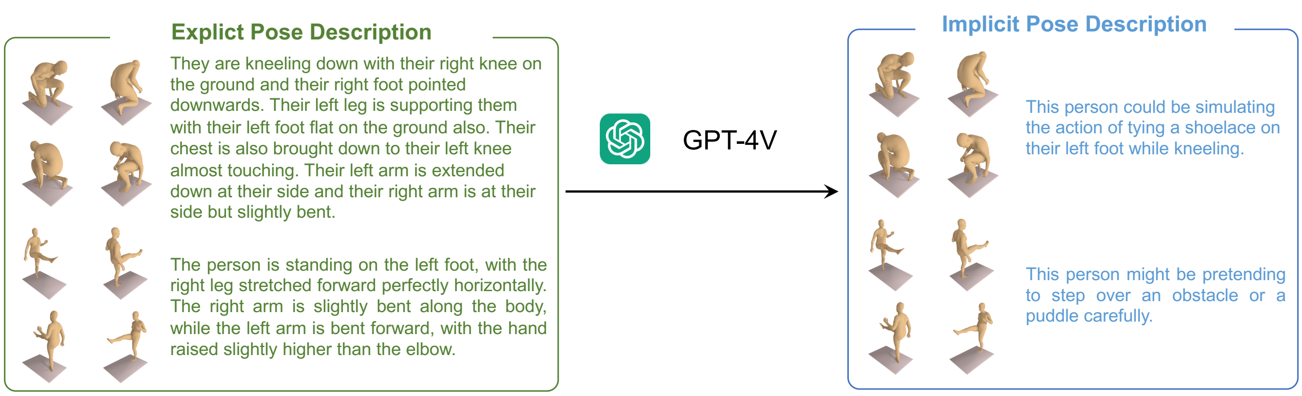

Figure 1. Illustration of the annotation pipeline that generates implicit pose description for our SPG benchmark. We take the fine-grained explicit pose descriptions from PoseScript [2] and visualize the described pose from four viewpoints, and then query GPT4 to reformulate them into indirect pose descriptions.

we manually refine them and create 250 question-answer pairs in the following format: "`USER:<IMAGE>` {`descriptions_person`}, `can you give the SMPL pose of this person? ASSISTANT: Sure, it is <POSE>`." Here, {`descriptions_person`} represents the person description from a specific aspect.

3. Ablation Study Details

Representations of Human Pose. Instead of utilizing the pose token `<POSE>`, an alternative approach to representing human poses involves using natural language, specifically textual descriptions specifying keypoint locations. To facilitate a comparison between these two pose representations, we use the same dataset pairs as in ChatPose and formulate Visual Question Answering (VQA) pairs for training. The question-answer template is structured as follows: "`USER: <Image> There is a person in the image, please estimate the visible keypoints coordinates. The output format should be Nose:(x1,y1),Neck:(x2,y2),... ASSISTANT: The detected visible keypoints are {KEYPOINT_NAME1}:{X1, Y1}, {KEYPOINT_NAME2}:{X2, Y2}, ...`". In this template, `<IMAGE>` represents the image patch token placeholder, {`KEYPOINT_NAME`} denotes the name of the visible keypoint, and {`X, Y`} indicates the discretized keypoint coordinates. Figure 3 provides some examples of these training pairs. We then fine-tune the base model, LLaVA [5], referred to as LLaVA *, to estimate keypoints and then use SMPLify to transform the keypoints into a SMPL pose for comparison with our pose token `<POSE>` representation. Visual results of LLaVA * are displayed in Figure 4. As shown, using textual descriptions as pose representation causes the network to often struggle to accurately estimate human poses and to often predict symmetrical poses, which may stem from the discretized nature of language signals.

Effects of Various Datasets. For training, we utilize three data types: text-to-SMPL pose (Text2Pose), image-

As an AI visual assistant specializing in human pose analysis, you will receive a visual depiction of a person, captured from multiple views, and a detailed, fine-grained textual description of his/her pose.
Your task is to infer the possible daily activities, ball games, and other behaviors that the person is mimicking. Offer a high-level interpretation without delving into the minutiae of joint positions. Concentrate on high-level descriptions of daily activities, ball games, and behaviors evident from the visual and textual information provided. If the pose resembles a specific yoga pose, be sure to mention the name of the yoga pose.

Ensure that your answers are clear enough to allow users to accurately mimic and replicate the pose based on your description. Avoid overly vague and ambiguous descriptions such as ”This person is doing a balancing behavior” or ”The person is warming up”. Your answer should be as diverse as possible and minimize the use of terms like ”balance”, ”stretch”, ”warm-up”, and ”flexibility”.

Prior to formulating the pose description, think and answer the following questions:

1. Which yoga pose the person might be doing? What are the differences between the visualized pose and standard yoga pose?
2. What everyday activity might the individual be engaging in?
3. Which sporting activity appears to be mimicked by the individual?
4. Could there be other actions the person is undertaking?

Based on your responses to the above questions, craft 5 responses describing the pose, each starting with ”{number}. This person,” accompanied by a succinct one or two sentences. Example answers and pose descriptions:

Answer to the questions:
1. The individual seems to be adopting a yoga pose, resembling the ”Natarajasana” or ”Lord of the Dance Pose.”
2. The individual could be reaching for an item on a high shelf.
3. It appears the individual is imitating a basketball player.
4. Additionally, the person might be engaging in an activity such as watching a movie with a friend.

Pose descriptions:
1. This person is executing the ”Downward-Facing Dog” yoga pose.
2. This person is making a marriage proposal.
3. This person is kneeling on one knee, potentially in a protest.
4. This person is participating in basketball, performing a jump shot.
5. This person seems to be looking for something on the ground.

Table 4. Example to query GPT4 for implicit pose descriptions.

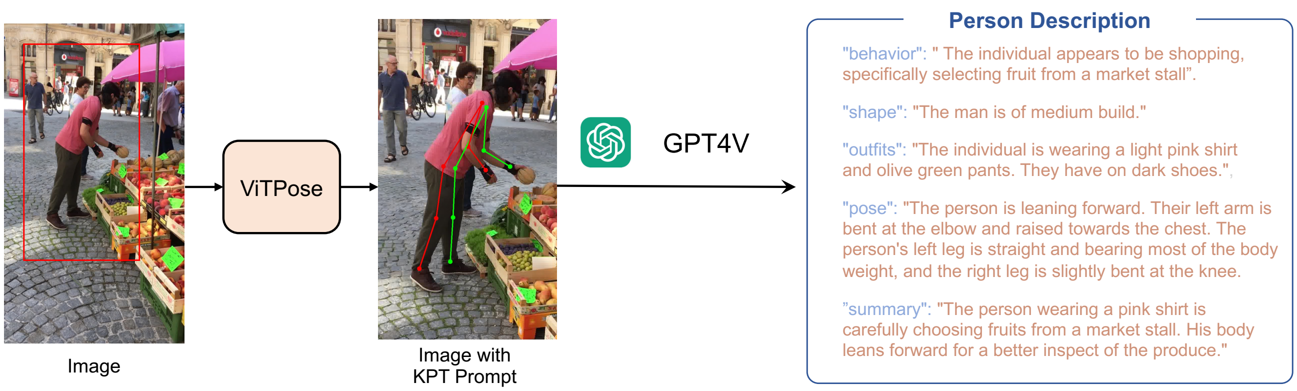

Figure 2. Illustration of our method to generate person descriptions for the RPE benchmark. We use ViTPose [8] to detect the body keypoints and mark the left-body and right-body joints with different colors as visual prompts, and then query GPT4V for person descriptions.

(a) You serve as an AI visual analyst for image examination. Your input will be an image containing humans. Your task is to provide descriptions of this individual. Your analysis should focus on four attributes: the individual's overall behavior, shape, outfits, and detailed pose. For the overall behavior, if this person is doing specific activities like yoga or sports, provide a detailed name. For the outfits, specify the color of the clothes. For the detailed pose, describe as detail as possible, looking into the torso, left, right arms, hands, and legs. To help you distinguish the left arms/legs from the right arms/legs, we have drawn the left body joints with green color, while the right body joints with red color. Don't mention the lines/marks/joints color in your answer! Please output the attributes (behavior, shape, outfits, and pose) as keys in a JSON file format, each value should be one or two sentences.

(b) You serve as an AI assistant. Your input will be a description of a person from four attributes: overall behavior, shape, outfits, and detailed pose. Your task is to understand the provided descriptions and then use your reasoning ability to generate one comprehensive short description in a manner that requires an advancing logical reasoning ability to understand and distinguish the correct individual. Remember, the comprehensive description should be shorter than 30 words and do not need to cover all the details, and require a strong reasoning ability to understand.

Table 5. Example to query GPT4 for person description. Prompt (a) is used to request GPT4V for detailed behavior, shape, outfits, and pose descriptions. Prompt (b) then instruct GPT4 to integrate and summarize these elements into a comprehensive description.

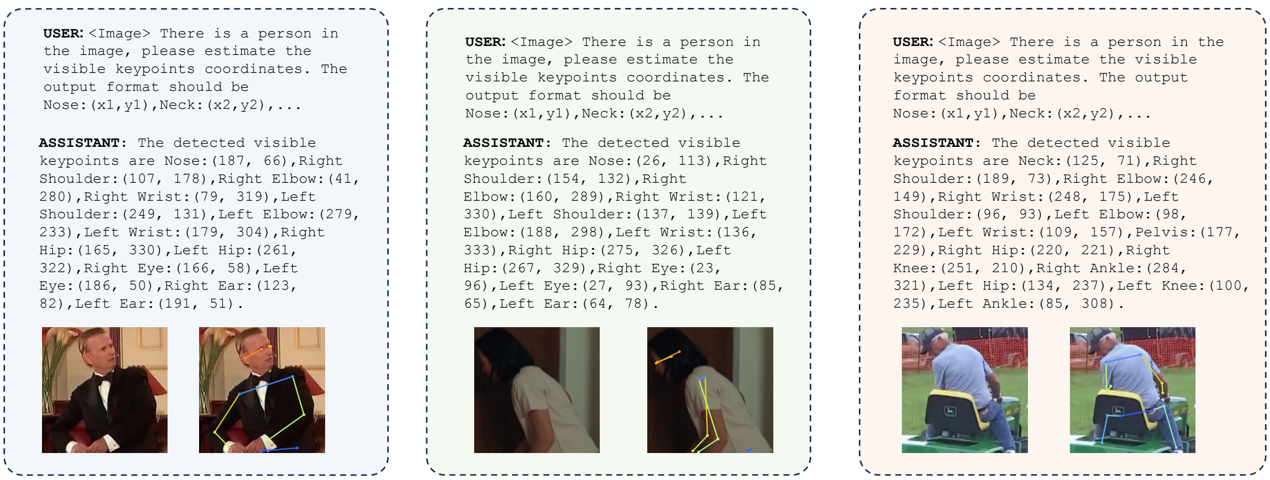

Figure 3. Examples of VQA data used to fine-tune the LLaVA model for pose estimation with textual descriptions of 2D keypoints.

Method	VQA [5]	Image2Pose	Text2Pose	Pose Estimation		Reasoning-based Pose Estimation
				3DPW [7]	H36M [3]	
LLaVA-P	✓			172.3	172.5	186.8
ChatPose w/o Image2Pose	✓		✓	115.1	121.6	123.7
ChatPose w/o Text2Pose	✓	✓		87.8	89.2	109.8
ChatPose full data	✓	✓	✓	**81.9**	**82.4**	**101.8**

Table 6. **Ablation study: effect of different training data.** PA-MPJPE (in mm) is reported. Lower is better.

Pretrained Model	Pose Estimation		Reasoning-based Pose Estimation
	3DPW [7]	H36M [3]	
LLaVA-V1.5-7B [4]	84.5	82.9	102.5
LLaVA-V1.5-13B [4]	81.9	82.4	101.8

Table 7. **Ablation study: effect of multimodal LLM backbones.** PA-MPJPE (in mm) is reported. Lower is better.

to-SMPL pose (Image2Pose), and general instruction-following data for visual question answer (VQA). To maintain the model's reasoning capabilities comparable to other LLMs, the VQA dataset is consistently used. For evaluating the effects of Text2Pose and Image2Pose, we fine-tune the model separately with each dataset. Table 6 presents the quantitative results. In contrast to the original LLaVA, which solely trains on VQA data, incorporating either Image2Pose or Text2Pose data into our model enhances pose estimation accuracy. Utilizing all data types, our model achieves optimal performance.

Multimodal LLM backbones. To evaluate how the LLM affects the performance of ChatPose, we employ both the LLaVA-V1.5-7b[1] and LLaVA-V1.5-13B[2] models, which

[1] liuhaotian/llava-v1.5-7b
[2] liuhaotian/llava-v1.5-13b

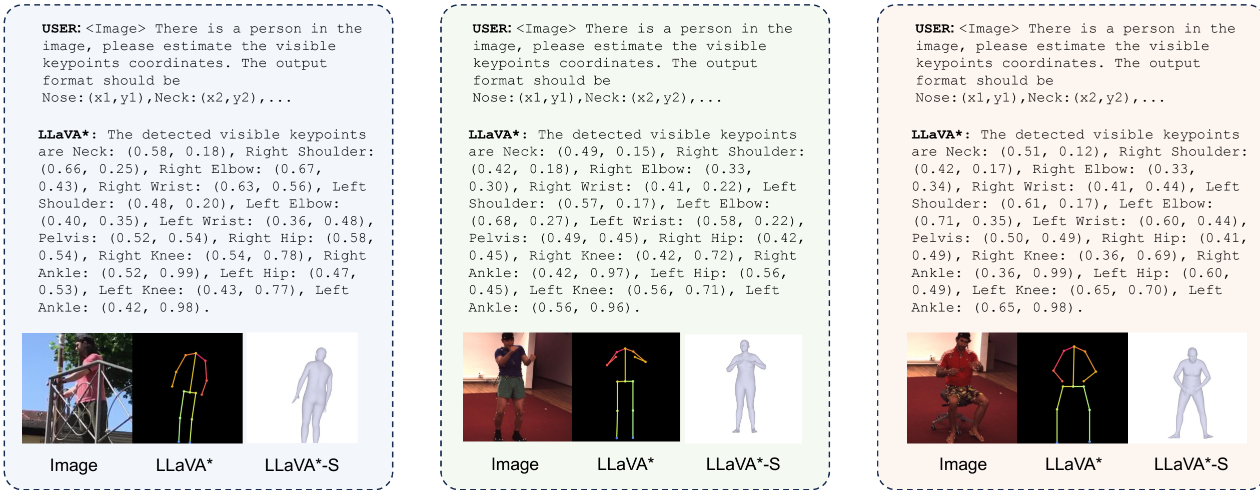

Figure 4. Visual results of LLaVA *. Given an RGB image, LLaVA * generates textual descriptions about keypoint locations. We then extract the keypoints from the textual descriptions and adopt SMPLify [1] to fit the SMPL pose.

are based on the LLaMA-7b and LLaMA-13b backbones, respectively. Table 7 shows the comparisons between 7b and 13b models. The 13b model, despite needing more training time, delivers superior accuracy over the 7b model. This suggests that our method's effectiveness is contingent on the capabilities of the LLM models and also benefits from their rapid advancements.

3.1. More Results

Generalization to Strong Occlusions. Even without any data augmentation during training, our model surprisingly still performs well on images with severe occlusions. Figure 5 shows pose estimation results for such cases. Even when half of the images are missing, ChatPose can still produce reasonable human poses. This suggests that it is able to leverage its general visual knowledge about occlusion in solving the human pose estimation problem.

Comparisons Details. For pose estimation, when comparing with other multi-modal LLMs that do not directly output 3D human poses, we adopt two approaches: firstly, generating keypoint coordinates followed by SMPLify [1] optimization of the 3D pose, and secondly, producing textual descriptions of the pose that are then processed by PoseScript [2] to create SMPL pose parameters. The workflow for the first method is illustrated in Figure 4, and for the second method in Figure 6.

FID for pose generation. We evaluated FID on real poses from the PoseScript and 3DPW test sets, generating text descriptions for the latter using PoseScript Rules; see Tab. 8. FID reflects distribution similarity more than generation quality. Since PoseScript trains only on its data and our model uses data from PoseScript and HMR (w/o text); the scores reflect this.

Method	FID (PoseScript) ↓	FID (3DPW) ↓
PoseScript	**0.50**	1.21
PoseGPT	1.51	**0.75**

Table 8. FID Scores on PoseScript and 3DPW dataset.

More analysis of T2P results As shown Table 1 in main paper, ChatPose lags behind for classical pose-to-text (P2T) retrieval while being on par with PoseScript [2] for classical text-to-pose (T2P) retrieval. We delve deeper into this analysis here. We start by visualizing instances where ChatPose underperforms while PoseScript succeeds, with one such example illustrated in Figure 7. Further analysis of failures did not reveal a distinct pattern. The contributing factors include: 1) Training strategy differences – PoseScript employs a VAE model with KL loss to ensure relative symmetry for T2P and P2T, whereas we employ LLMs with inherent strong priors about languages. 2) Varied training data – Unlike PoseScript's consistent use of AMASS, our multi-modal training employs a mix of AMASS, HMR, and general VQA data, leading to a varied training-test distribution. 3) Bias in the retrieval models with P2T being less accurate than T2P (as noted in the PoseScript paper Tab. 1). We reevaluated P2T and T2P using a higher-accuracy retrieval model from the PoseScript journal version. Top 5/10/50/100 P2T and T2P results are detailed in Tab. 9.

Method	R^{P2T} ↑	R^{T2P} ↑
PoseScript	**22.6/31.0/57.9**/70.8	22.4/32.1/58.7/71.5
ChatPose	17.6/25.3/57.6/**71.2**	**28.0/39.0/70.4/83.5**

Table 9. TOP 5/10/50/100 T2P and P2T results with retrieval model from PoseScript journal version.

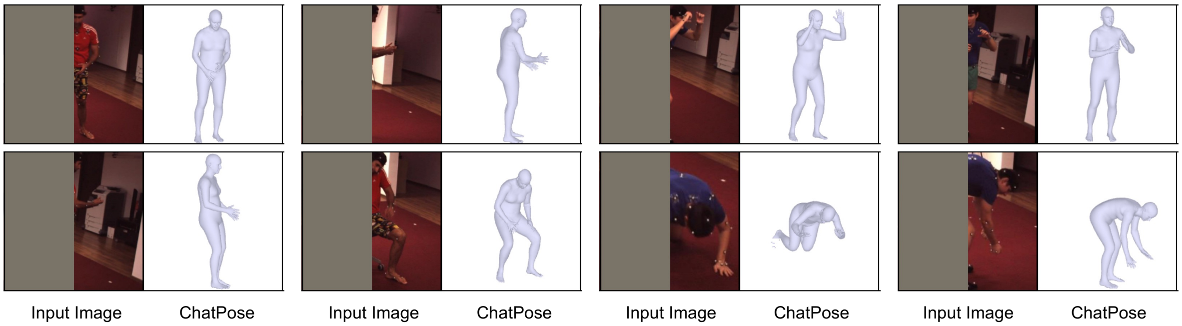

Figure 5. Pose estimation on images with significant occlusion. Without training for occlusion cases, ChatPose is surprisingly robust.

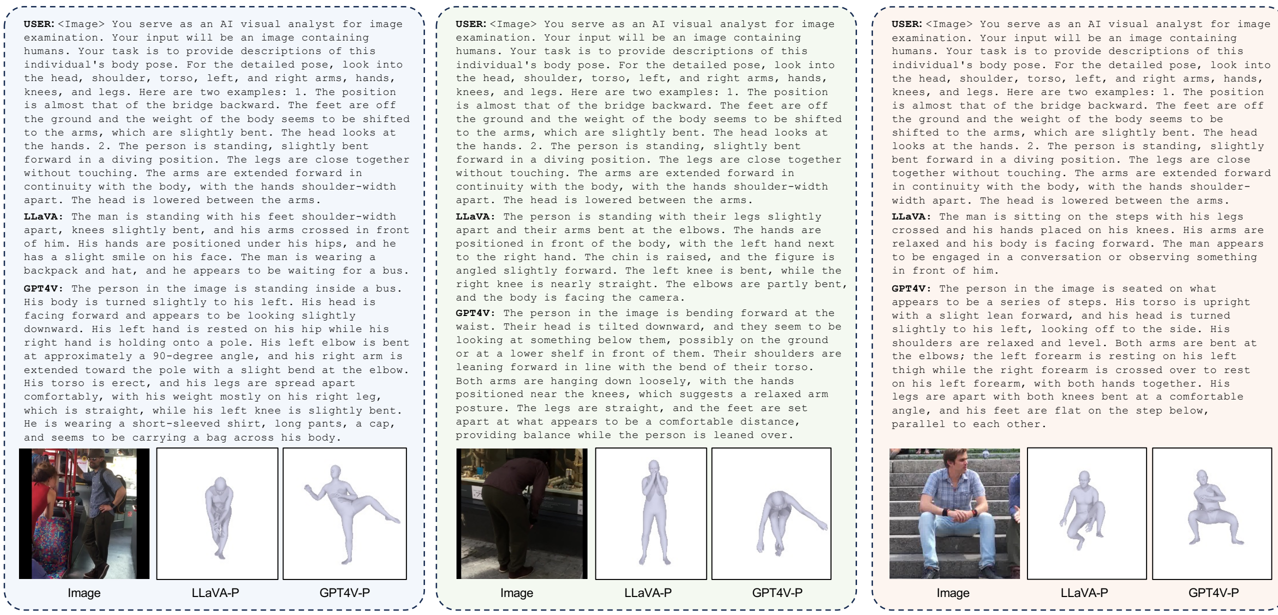

Figure 6. Visual results of LLaVA and GPT4. Given an RGB image, LLaVA and GPT4 generate textual descriptions about human poses. We then use PoseScript [2] to generate SMPL poses based on the text descriptions.

Figure 7. From left to right: GT, PoseScript, ChatPose. This illustrates a comparison in pose generation between PoseScript and our approach. In instances where T2P retrieval is correct, PoseScript's P2T is also correct, whereas ChatPose's P2T is incorrect.

Method	SPG R^{P2T} ↑	SPG R^{T2P} ↑
LLaVA-P	5.0/8.6/13.8	5.8/9.7/14.7
LLAVA-P (w/ ICL)	2.6/5.3/9.2	3.5/6.3/10.5
GPT4-P	3.5/6.9/11.3	4.1/7.3/11.9
GPT4-P (w/ ICL)	3.7/7.6/13.1	5.1/8.1/13.5
PoseScript finetuned with SPG	6.0/9.6/15.4	7.4/12.1/18.5
ChatPose (ours)	8.6/14.2/20.8	10.9/16.9/25.3

Table 10. Results of suggested baselines. ICL means "in context learning", where we teach LLaVA/GPT4 with a few examples of converting our SPG text to more detailed PoseScript descriptions.

Other baselines for RPE and SPG. We show more baselines in Table 10. Using LLaVA/GPT4 to convert SPG texts into PoseScript texts (LLaVA/GPT4+PoseScript) preforms poorly. To improve results we add in-context learning (w/ ICL) but this remains less accurate than ChatPose. We finetuned PoseScript with SPG data; the results in are also less accurate than ChatPose.

Failure Cases. We also show some limitations of the current model in Figure 8. It is important to note that the global orientation can be significantly off, even when the body pose is approximately correct. This global orientation issue might be improved by using a superior vision backbone, particularly one that excels at localization.

Figure 8. Failures cases of ChatPose on the human pose estimation task. Note that a common failure mode is to estimate the articulated pose correctly but to output the incorrect global orientation.

References

[1] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, 2016. 5

[2] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. Posescript: 3D human poses from natural language. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 5, 6

[3] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2014. 4

[4] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 4

[5] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 2, 4

[6] OpenAI. GPT-4 technical report. 2023. 1

[7] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 4

[8] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 1, 3

[9] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. *arXiv preprint arXiv:2310.11441*, 2023. 1