

# ChatPose: Chatting about 3D Human Pose

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

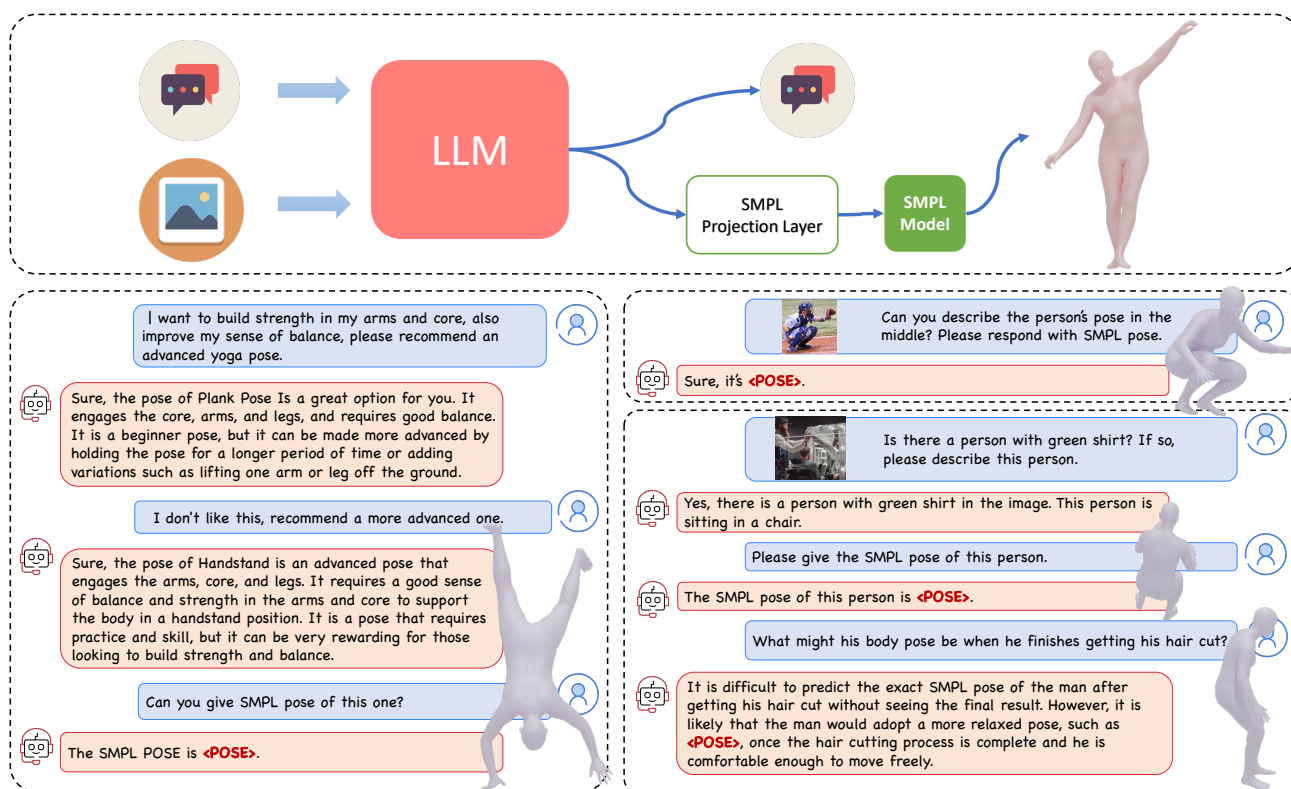


Figure 1. We introduce ChatPose, a multimodal LLM designed for chatting about human pose that produces 3D human poses (SMPL pose parameters) upon user request. ChatPose features a specialized SMPL projection layer trained to convert language embeddings into 3D human pose parameters. Our demonstration includes conversations both without (left) and with (right) an image input. Upon detection of a pose token, the token is used to estimate the SMPL pose parameters and subsequently generate the corresponding 3D body mesh.

## Abstract

We introduce ChatPose, a framework employing Large Language Models (LLMs) to understand and reason about 3D human poses from images or textual descriptions. Our work is motivated by the human ability to intuitively understand postures from a single image or a brief description, a process that intertwines image interpretation, world knowledge, and an understanding of body language. Traditional human pose estimation and generation methods often operate in isolation, lacking semantic understanding and reasoning abilities. ChatPose addresses these limitations by embedding SMPL poses as distinct signal tokens within a

multimodal LLM, enabling the direct generation of 3D body poses from both textual and visual inputs. Leveraging the powerful capabilities of multimodal LLMs, ChatPose unifies classical 3D human pose and generation tasks while offering user interactions. Additionally, ChatPose empowers LLMs to apply their extensive world knowledge in reasoning about human poses, leading to two advanced tasks: speculative pose generation and reasoning about pose estimation. These tasks involve reasoning about humans to generate 3D poses from subtle text queries, possibly accompanied by images. We establish benchmarks for these tasks, moving beyond traditional 3D pose generation and estimation methods. Our results show that ChatPose out-

*performs existing multimodal LLMs and task-specific methods on these newly proposed tasks. Furthermore, ChatPose’s ability to understand and generate 3D human poses based on complex reasoning opens new directions in human pose analysis. Code and data are available for research at <https://yfeng95.github.io/ChatPose>.*

## 1. Introduction

We address the problem of understanding and reasoning about 3D human pose from an image or a text description via large language models. For humans, a quick glance at a picture or a brief description of a person allows us to form an impression of their articulated body posture. For instance, one might wonder, “What is the girl in the dress doing?” or “How might she behave if she feels tired?”. This involves interpreting the image, employing general knowledge about the world, and understanding human body language. Current methods that estimate 3D human poses from images [9, 12, 20, 21, 25, 61], usually detect individuals, segment them from the image, then use a neural network to predict 3D pose and shape in terms of the parameters of a body model like SMPL [32]. Other approaches [40, 47, 48] regress poses of all individuals by analyzing the full image. However, these processes lack a comprehensive understanding of the scene, failing to fully consider the interactions between humans and their environment, as well as their intentions. Methods for text-driven pose generation have also progressed rapidly [7, 18] but the text instructions are typically “explicit,” precisely describing the pose with words.

Thus, existing specialized systems for 3D pose estimation and generation are constrained to narrow tasks. This is in contrast to the general-purpose reasoning exhibited by large language models (LLMs). Existing multimodal LLMs [23, 30, 36, 54] demonstrate proficiency in perceiving and interpreting information from images and reasoning based on a wealth of world knowledge. They are particularly adept at describing scenes, including the appearance of people, their activities, and high-level behaviors. If the LLM could relate this generic world knowledge to 3D human pose and motion, it would have powerful reasoning capabilities beyond existing solutions. That is, the LLM could bring to bear all that it has learned from both images and language for a richer and more nuanced understanding of human pose. Existing LLMs, however, have not yet demonstrated the ability to interpret 3D human pose.

Our hypothesis is that, long term, general purpose multimodal LLMs will subsume special-purpose methods. Estimating 3D pose from a 2D image is fundamentally ambiguous and must use prior information or contextual cues. Generating pose or motion from language, likewise, is ambiguous and open to interpretation. By formulating these problems in the context of LLMs, the solutions can theo-

retically benefit from the LLM’s broad general knowledge. The solutions can also benefit from interaction with a user through a language interface. For our hypothesis to be true, LLMs must be able to understand and interpret 3D human pose. What do they already understand about 3D pose and how can we teach them about 3D human pose?

To investigate these questions, we introduce ChatPose, an approach that finetunes multimodal Large Language Models for predicting human pose, represented as SMPL [32] pose parameters. Our method embeds SMPL poses as a unique <POSE> token, prompting the LLM to output these when queried about SMPL pose-related questions. We extract the language embedding from this token, and use an MLP (multi-layer perceptron) to directly predict the SMPL pose parameters. This enables the model to take either text or images as input and subsequently output 3D body poses, as shown in Fig. 1. We maintain the vision components in a frozen state while training the SMPL projection layers and fine-tuning the LLM models with LoRA [15]. Our training strategy involves constructing question and answer pairs derived from image-to-SMPL and text-to-SMPL pose pairings, originating from pose estimation and text-driven pose generation tasks. Additionally, we integrate general multi-modal instruction-following data throughout the end-to-end training process of our model.

We evaluate ChatPose on a variety of diverse tasks, including the traditional task of 3D human pose estimation from a single image and pose generation from text descriptions. While the metric accuracy on these classical tasks does not yet match that of specialized methods, we see this as a first proof of concept. More importantly, once the LLMs are able to understand SMPL poses, they can utilize their inherent world knowledge to relate to, and reason about, human poses without the need for extensive additional data or training. For example, as demonstrated by the right example in Fig. 1, ChatPose is capable of inferring the body pose following the action depicted in the image. This capability gives rise to two innovative tasks concerning human poses: (1) **Speculative Pose Generation (SPG)**: In contrast to methods that generate poses like “sitting” based on text like “the person is sitting,” in SPG we ask the LLM to speculate, for example, about “how would the person’s pose change if they were tired?” Such data is not in classic pose training datasets and requires an understanding of (i) what being tired does to a body and (ii) how this translates into 3D pose. This is a significantly harder task than is considered by prior work. (2) **Reasoning-based Pose Estimation (RPE)**: Contrary to conventional approaches in pose regression, our methodology does not involve providing the multimodal LLM with a cropped bounding box surrounding the individual. Instead, the model is exposed to the entire scene, enabling us to formulate queries regarding the individuals and their respective poses within that

context. For example, “what are the poses of all the people wearing glasses?” This requires an integration of scene understanding with 3D human pose that does not exist in current human pose regression systems. To successfully address these tasks, the model needs two primary capabilities: 1) the ability to reason through complex and implicit text queries, integrating them with image data when available; 2) the ability to generate SMPL pose parameters based on its understanding of high-level concepts.

In summary, for the first time, we demonstrate the ability of a large vision-language model to reason about 3D human pose from images or text and to connect this with 3D SMPL parameters. Our key contributions are as follows: (1) We present *ChatPose*, a multimodal Large Language Model (LLM) that can directly generate SMPL poses. This enables the generation and estimation of human poses through reasoning from text or images. (2) We introduce two innovative tasks: speculative pose generation and reasoning-based pose estimation. These tasks necessitate an accurate understanding of human poses and the ability to reason using world knowledge. We have also established new benchmarks that can drive research on this topic. (3) Our model, *ChatPose*, demonstrates superior performance compared with other multimodal LLM baselines on the tasks of pose generation and estimation.

## 2. Related Work

Our work spans multiple research areas. Consequently, we briefly review 3D human pose estimation from images, language and pose, and large language models.

**Human Pose Estimation.** Human pose estimation in 2D, 3D, or over time, has a long history, which we do not review here. Instead, we focus on work that estimates the pose of a 3D parametric body model from a single image. Here we use the SMPL model [32], which produces a 3D triangulated mesh given relative body part rotations and body shape (though we ignore shape here). SMPL is widely used, in part because it is compatible with graphics engines and because there is a large amount of training data available in SMPL format. SMPL parameters are typically estimated from an image using one of two techniques. Optimization-based approaches solve for the parameters such that, when the model’s 3D joints are projected into the image, they match detected 2D keypoints, subject to various priors [3, 10, 19, 37]. Regression-based approaches [12, 20–22, 25, 61] directly infer the pose parameters from a cropped image. When provided with a full image, these methods typically first detect each person in the image and then apply the regression network to tight crops. The best regression methods are now quite accurate and robust except when there is significant occlusion, poor image quality, or unusual poses. Additionally,

there are methods designed for multi-person pose estimation [40, 47, 48], which are capable of directly generating body meshes for multiple people within a single image. The above methods, however, do not “understand” the semantics of human pose or relate pose to language.

**Language and Human Pose.** Given a textual description of a person’s attributes, advanced image generation methods like Stable Diffusion [43] and DALL-E 2 [42] generate realistic 2D images of people. These can further be conditioned on information like 2D human pose [64]. Such methods clearly understand properties of the human body and human pose but they output pixels and not 3D representations. Recent language-to-3D generation methods [5, 14, 26, 39, 63] create 3D human shapes from textual descriptions. Yet, these methods struggle to represent complex body poses. Other approaches exist that can take text input and directly produce parameters of a parametric body model like SMPL. For example, BodyTalk [45] takes human shape attributes (such as “broad shoulders” or “skinny”) and outputs SMPL shape parameters. Similarly, [4] employs text annotations to describe a person’s general action and the surrounding scene, which it uses to generate SMPL pose parameters. PoseScript [7] creates SMPL pose parameters from fine-grained textual descriptions of 3D human poses. While these methods are effective when test descriptions closely match the word distributions of their training data, they often lack the capability to understand or reason based on complex textual inputs. For example, PoseScript’s training data lacks descriptions that relate human poses with scenes. Since our method leverages LLMs, it can deal with more complex text queries even when trained only with the same text-to-SMPL pose pairs as PoseScript.

Unlike all the task-specific approaches to pose estimation, action recognition, and pose generation, we develop a single, unified, model capable of reasoning about 3D humans from images, text, or both by leveraging its general knowledge of the visual world. Additionally, it can interact with users through conversations, discussing human poses and providing relevant responses.

**Multimodal Large Language Models.** Large Language Models (LLMs) are rapidly changing multiple fields. While the most powerful models like OpenAI’s ChatGPT [35] and GPT-4 [36] are private, a range of open-source LLMs such as Vicuna [6], LLaMA [50], and Alpaca [49] enable research like ours. In particular, we exploit the ability to fine-tune LLMs on multimodal tasks. There are two primary ways to do this. The first leverages LLMs for decision-making guidance. Research such as [16, 31, 38, 44, 53, 57, 58] typically employs prompt engineering or instruction tuning. In this approach, LLMs connect separate modules via API calls. The LLM generates API calls to solve tasks and retrieve results. Such an approach falls short of achieving a comprehensive understanding of new modalities.

An alternative approach maps modality-specific information into the language embedding space of the LLM. The visual modality has been a major focus in this area. Recent initiatives like LLaVA [29, 30] and MiniGPT-4 [67] incorporate vision encoders to interpret images and use projection layers that align image features with language embeddings. Work like LISA [23] generates visual information in the output, processing both images and questions to yield text and masks. In addition to images, MM-LLMs (Multi-Modal Large Language Models) are rapidly being developed for video [24, 62] and audio [60]. Notably, models such as PandaGPT [46], ImageBind [11], and NeXT-GPT [54] demonstrate the capability to handle a wide array of modalities, including text, image, audio, and video. Specifically, NeXT-GPT aligns embeddings from these four modalities with language, both as input and output.

In this work, we investigate 3D body pose as a new modality for LLMs to process. We explore (1) the ability of LLMs to generate 3D pose from text or image input, and (2) whether LLMs can comprehend 3D body poses and integrate this understanding into their overall functionality. To our knowledge, this has not previously been explored.

### 3. Method

Our goal is to enable Large Language Models (LLMs) to comprehend human poses, represented as SMPL [32] pose parameters in our case. Drawing inspiration from recent advancements in multi-modal LLMs [13, 23, 54, 59], we approach human pose as a distinct modality. In this framework, the LLM generates a unique token representing this modality, which is subsequently mapped to SMPL pose parameters via an MLP projection layer<sup>1</sup>. Leveraging the SMPL parametric model [32], we can then decode this information into a three-dimensional body mesh. Here we describe the architecture and training strategy that integrates SMPL pose as a modality within LLMs. Once the LLM grasps the concept of 3D body pose, it gains the dual ability to generate human poses and to comprehend the world, enabling it to reason through complex verbal and visual inputs and subsequently generate human poses. This leads us to introduce novel tasks that are made possible by this capability, along with benchmarks to assess performance.

#### 3.1. Architecture

The architecture of ChatPose is illustrated in Fig. 2. Our approach takes text or images (if provided) as input and produces textual output. Also, when users request human pose information, it also returns the corresponding SMPL pose. Our model consists of a multi-modal LLM model,  $f_\phi$ , an embedding projection layer,  $g_\Theta$ , and a parametric

<sup>1</sup>Readers familiar with the geometric “projection” of SMPL into images should not confuse that with the use of projection in this context, which effectively means “aligning” one representation with another.

human body model, SMPL [32], represented by pose and shape parameters  $\theta$  and  $\beta$ , respectively. Here, we assume the  $\beta$  values are all zero, corresponding to the average body shape. Given a text string  $X_q$  and an image  $X_v$  as input, the model produces a textual response  $Y_t = f_\phi(X_q, X_v)$  or  $Y_t = f_\phi(X_q)$  in the absence of an image. The language embedding corresponding to  $Y_t$  is represented as  $H_t$ . If  $\langle \text{POSE} \rangle$  is present in the textual output  $Y_t$ , its corresponding embedding  $H_{pose}$  is retrieved from  $H_t$ . The pose embedding, processed by the SMPL projection layer  $g_\Theta$ , yields the SMPL pose parameters  $\theta = g_\Theta(H_{pose})$ . The 3D vertices and triangles of the body mesh are then determined using the standard SMPL function  $M(\theta, \beta)$  (see [32]).

#### 3.2. Training

We keep both the vision encoder and vision projection frozen and train the SMPL pose projection layer  $g_\Theta$ . Additionally, we employ LoRA [15] to finetune the LLM, with its parameters denoted as  $\phi_{lora}$ . The final set of optimizable parameters is  $\{\phi_{lora}, \Theta\}$ . With the provided ground truth textual output  $\hat{Y}_t$  and SMPL pose parameters  $\hat{\theta}$ , we optimize the model using the following objective function:

$$\mathcal{L} = \lambda_t \text{CE}(\hat{Y}_t, Y_t) + \lambda_\theta |\hat{\theta} - \theta|. \quad (1)$$

The first term is the cross-entropy loss, while the second, pose loss, is the L1 difference between the ground truth and estimated pose parameters.  $\lambda_t$  and  $\lambda_\theta$  serve as the weights for their respective loss terms. To train our multi-modal LLM model, we construct data by leveraging existing task-specific datasets below.

**Text to Pose Generation.** A 3D human pose can be generated from a detailed textual description of the pose. The data pairs in this case are SMPL pose parameters and detailed text description labels  $\{X_q, \hat{\theta}\}$ . To fit this data into a question-answer format, we employ templates such as “USER: {description}, can you give the SMPL pose of this person. ASSISTANT: Sure, it is  $\langle \text{POSE} \rangle$ .”, where {description} contains the pose descriptions  $X_q$  from the dataset.

**Human Pose Estimation.** Conventional methods of 3D human pose estimation [12, 22] typically involve using cropped images to regress SMPL body shape and pose parameters. Similarly, we use pairs of cropped images and SMPL pose parameters  $\{X_v, \hat{\theta}\}$ . To format the data suitably for visual question answering, similar to text to pose generation, we use a question-answer template like “USER:  $\langle \text{IMAGE} \rangle$  Can you provide the SMPL pose of the person in the center of this image? ASSISTANT: Sure, the SMPL pose of this person is  $\langle \text{POSE} \rangle$ .”, where  $\langle \text{IMAGE} \rangle$  is a placeholder for the input image tokens. The corresponding ground truth SMPL pose parameters  $\hat{\theta}$  are used to calculate the pose loss as in Equation 1. During training, we also use



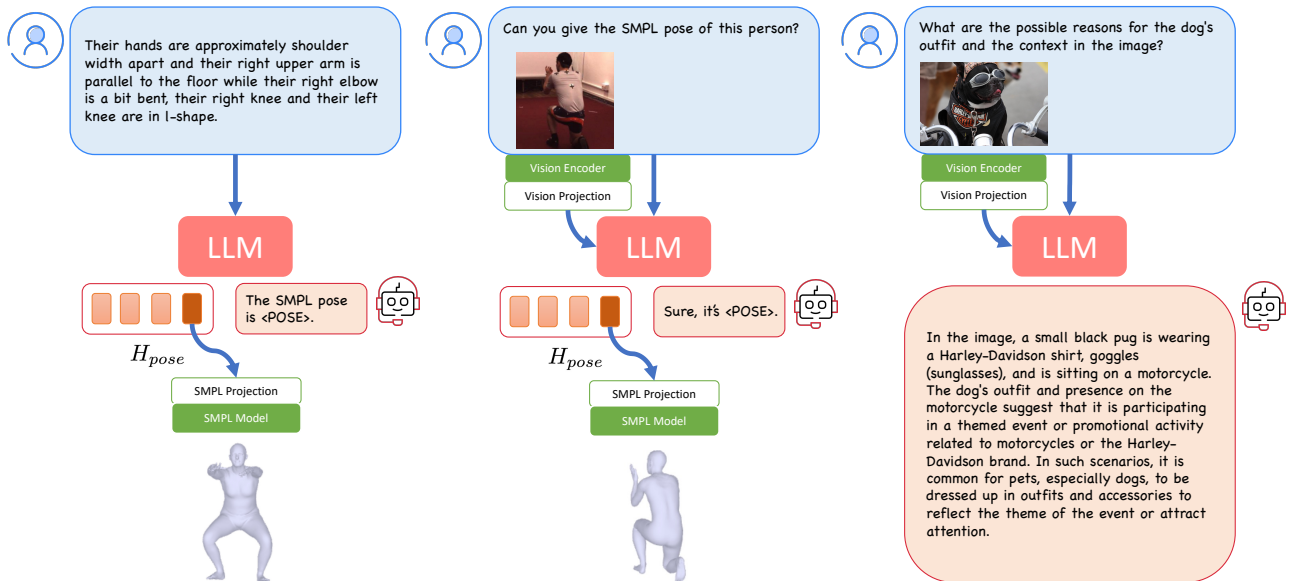


Figure 2. Method and Training Overview. Our model is composed of a multi-modal LLM (with vision encoder, vision projection layer and LLM), a SMPL projection layer, and the parametric human body model, i.e. SMPL [32]. The multi-modal LLM processes text and image inputs (if provided) to generate textual responses. In the training phase, we focus on training the SMPL projection layer and fine-tuning the LLM, while keeping the other components frozen. The three data types used for the end-to-end training are: text-to-3D pose generation, image-to-pose estimation, and multi-modal instruction-following data. When an image is available, its information is used by the LLM to deduce an answer. If the user inquires about a SMPL pose, the LLM responds with a `<pose>` token. The embedding related to this token is then used to predict the SMPL pose parameters, leading to the generation of a body mesh, as visualized.

other templates to generate question-answer data to ensure diversity; please see *Sup. Mat.* for details.

**Multi-Modal Instruction-following.** In order to maintain the multi-modal LLM’s inherent capability for multi-turn conversations, we use a multi-modal instruction-following dataset during training. Following LLaVA-V1.5 [29], we utilize the LLaVA-V1.5-MIX665K<sup>2</sup> dataset, which is created through queries made to GPT-4.

### 3.3. Reasoning about Human Pose

After training, our model is capable of estimating SMPL poses from single images, generating poses based on detailed descriptions, and facilitating question-and-answer conversations. Remarkably, even without integrating SMPL pose into multi-turn conversations or linking complex phrases with SMPL pose, our model demonstrates a *zero-shot* capability for reasoning about human poses within multi-turn dialogues. This suggests that the model is able to interweave reasoning and world knowledge with the SMPL pose representation. Therefore, in addition to conventional evaluation approaches for human pose and generation tasks, we introduce two new tasks that require reasoning skills: Speculative Pose Generation and Reasoning-based Pose Estimation. These new tasks leverage the model’s ability to apply reasoning in the context of human pose analysis.

**Speculative Pose Generation (SPG).** In this task, rather than using explicit pose descriptions from the text-to-pose generation dataset, users pose indirect questions about a person’s state, requiring the LLM to deduce and generate the appropriate pose. For instance, a user might ask, “USER: {descriptions\_implicit}, can you give the SMPL pose of this person? ASSISTANT: Sure, it is <POSE>.” Here, {description\_implicit} represents speculative queries such as “This man is proposing marriage, what pose might he be in?”. This kind of inquiry requires an understanding of global concepts such as “marriage” and the capacity to logically deduce the individual’s pose, followed by the generation of SMPL pose parameters. To create an evaluation dataset, we use pose descriptions from the PoseScript [7] dataset as a source. We then query GPT4 to reformulate these descriptions into questions about the activities associated with each pose, generating a total of 20k responses, of which 780 examples are used for evaluation. These responses are then manually reviewed and corrected as needed.

**Reasoning-based Pose Estimation (RPE).** Standard human pose estimation methods typically first run a person detector and then only process a cropped image around the person. This ignores scene context, which can be useful in reasoning about human pose. In contrast, RPE lets users make inquiries about

<sup>2</sup>liuhaotian/LLaVA-Instruct-150K

Method	PoseScript [7]		SPG Benchmark	
	$R^{P2T} \uparrow$	$R^{T2P} \uparrow$	$R^{P2T} \uparrow$	$R^{T2P} \uparrow$
PoseScript [7]	22.6 / 31.0 / 42.3	22.4 / 32.1 / 43.6	1.9 / 3.8 / 6.5	2.8 / 4.3 / 7.2
ChatPose	17.6 / 25.3 / 35.8	28.0 / 39.0 / 54.4	8.6 / 14.2 / 20.8	10.9 / 16.9 / 25.3

Table 1. **Comparison of classical and speculative pose generation.** Arrows show whether higher or lower values are better. Top 5 / 10 / 20 retrieval recall rates are reported for pose generation on the PoseScript test set and our new SPG Benchmark.

an image before requesting details about a person’s pose. Specifically, we define RPE as: “USER:<IMAGE> {description\_person}, can you give the SMPL pose of this person? ASSISTANT: Sure, it is <POSE>.” In this case, {description\_person} could be queries about a particular individual, such as “The man with black hair”, or “the woman near the stairs”. The model is required to interpret the scene context and generate the SMPL pose parameters for the individual fitting the description. To evaluate this task, we start with image-to-SMPL pose pairs from standard pose estimation evaluation datasets. We then use GPT4V to generate descriptions of the individuals in these images. The generated descriptions are subsequently refined manually. Specifically, we sample 50 multiple-person images from the 3DPW [51] test set. For each individual, we collect descriptions that cover behavior, outfits, pose, shape, summary, where summary summarizes all the other attributes. This process leads to a total compilation of 250 question and answer pairs for evaluation. For more details of the collection pipeline, please see the *Sup. Mat.*

## 4. Experiments

We employ LLaVA-1.5V-13B [30] as the multimodal LLM backbone, with CLIP [41] for vision encoding and Vicuna-13B [65], finetuned from Llama 2 [50] on conversational data, for the LLM backbone. We maintain the CLIP encoder and vision projection layer, while training the SMPL projection layer from scratch and fine-tuning the LLM using LoRA. The SMPL projection layer is an MLP with layer dimensions of [5120, 5120, 144]. Following previous work [12, 22], our network predicts 6D rotations [66] for the SMPL pose, which are converted into rotation matrices for loss computation. For further implementation details, training details, our ablation study, and details about LLM backbones, please see *Sup. Mat.*

### 4.1. Datasets

**Text to Pose Generation.** We use the text-to-SMPL pose pairs from PoseScript [7], which features textual descriptions of 20k diverse human poses derived from the AMASS [33] dataset. Within this dataset, 6.5k texts are human-annotated and there are six types of automated labels for the entire set of 20k poses. Our training employs their designated training set of approximately 14k pairs. Addition-

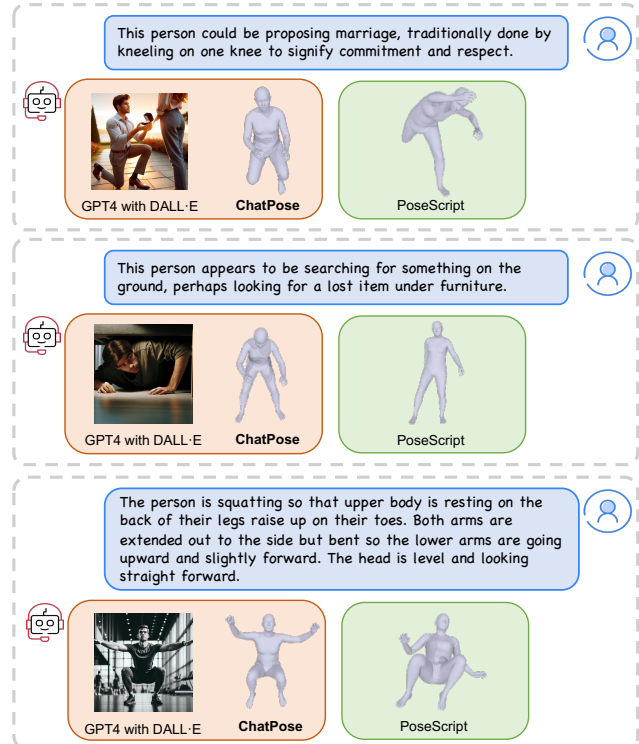


Figure 3. **Pose Generation.** GPT-4 (DALL-E) [36] generates images that depict the correct pose but does not explicitly generate 3D poses. In contrast, PoseScript [7] is a task-specific method for 3D pose from language but it is not able to relate high-level concepts like “searching under furniture” with 3D pose. In contrast, ChatPose, understands high-level concepts and how to relate them to 3D pose. The methods in orange address SPG, while the green region indicates the “classical” approach. The first two query examples are sourced from our SPG benchmark, which offers implicit text queries regarding human poses. The third example is derived from the PoseScript test set, which has detailed descriptions of human poses.

ally, we observe that the automatically generated labels in the dataset exhibit significant noise. Thus, we prioritize human labels when available; in their absence, we randomly select one of the automated labels for each pose.

**Human Pose Estimation.** In line with prior research on “classical” 3D human pose and shape regression, we employ datasets from Human3.6M [17], MPI-INF-3DHP [34], COCO [28], and the MPII dataset [1] for training. These datasets include training pairs of images with ground-truth or pseudo-ground-truth SMPL pose parameters. Note that we ignore the SMPL shape parameters here. Unlike previous methods, which typically use significant data augmentation (e.g. [21, 27]), our approach solely uses tightly cropped images without any additional augmentation such as blur or occlusion. Despite this, our model still demonstrates good generalization to these scenarios, suggesting that the network is able to leverage its general visual capabilities.

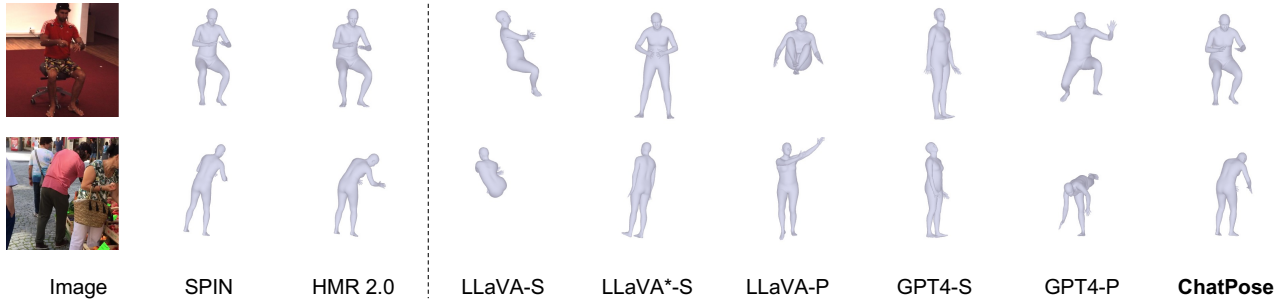


Figure 4. We compare multi-modal LLMs (LLaVA [30], GPT-4 [36]) and traditional HMR-style methods (HMR2.0 [12], SPIN [22]) for **classical human pose estimation**. LLaVA\* is LLaVA fine-tuned with keypoint data.

Stage 1	Stage 2	Text Description					Averaged
		Behavior	Shape	Outfit	Pose	Summary	
SPIN [22]	-	244.9 / 107.3 / 12.4	244.9 / 107.3 / 12.4	244.9 / 107.3 / 12.4	244.9 / 107.3 / 12.4	244.9 / 107.3 / 12.4	244.9 / 107.3 / 12.4
HMR 2.0 [12]	-	<b>225.2</b> / 105.7 / 12.1	<b>225.2</b> / 105.7 / 12.1	<b>225.2</b> / 105.7 / 12.1	<b>225.2</b> / 105.7 / 12.1	<b>225.2</b> / 105.7 / 12.1	<b>225.2</b> / 105.7 / 12.1
LLaVA [30]	SMPLify [3]	490.7 / 200.6 / 20.9	462.3 / 204.3 / 20.2	481.1 / 198.7 / 20.0	480.9 / 207.4 / 21.1	490.7 / 207.4 / 21.1	481.1 / 203.7 / 20.7
LLaVA [30]	PoseScript [7]	370.8 / 182.3 / 17.5	407.8 / 191.3 / 18.0	440.7 / 190.4 / 17.6	363.2 / 177.9 / 17.4	391.5 / 191.9 / 17.8	394.8 / 186.8 / 17.7
ChatPose (Ours)	-	307.9 / <b>102.9</b> / <b>12.1</b>	269.9 / <b>103.7</b> / <b>12.0</b>	265.6 / <b>102.6</b> / <b>11.8</b>	277.9 / <b>96.0</b> / <b>11.7</b>	253.6 / <b>103.8</b> / <b>11.7</b>	275.0 / <b>101.8</b> / <b>11.9</b>

Table 2. Comparison of **reasoning-based pose estimation** with different text descriptions. MPJPE / PA-MPJPE/ MPJRE ( $\times 100$ ) on the RPE benchmark are reported. Examples of each description type are in the *Sup. Mat.* Bold shows the best model for each metric.

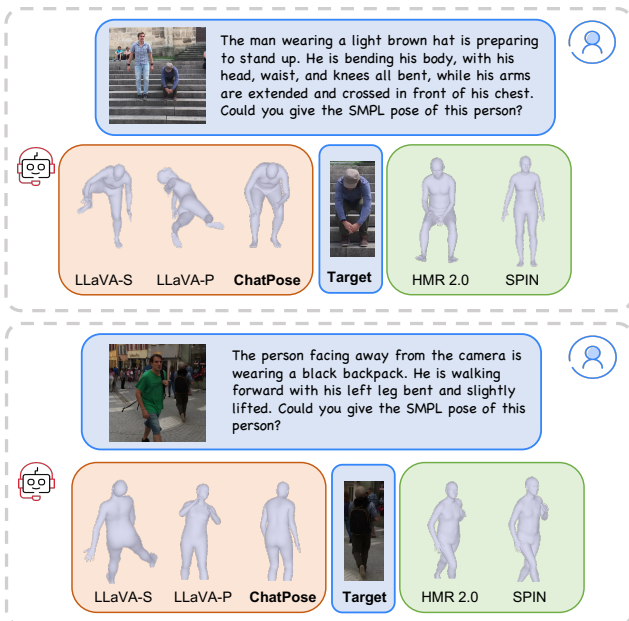


Figure 5. Comparison with LLaVA [30] and classical HMR-style methods (HMR2.0 [12] and SPIN [22]) on **reasoning-based human pose estimation**. For each method, we utilize the entire image provided by the user as input, without applying cropping. Methods involving LLMs are highlighted in orange, while those that are purely task-specific methods, are marked in green.

## 4.2. Evaluation Metrics and Baselines

**Generation.** For both the standard text-to-pose generation task and our new speculative pose generation (SPG) task, we use the evaluation metrics established in PoseScript [7].

Method	3DPW [52]			H3.6M [17]	
	MPJPE ↓	PA-MPJPE ↓	MPJRE ↓	MPJPE ↓	PA-MPJPE ↓
SPIN [22]	102.9	62.9	10.1	61.9	42.6
HMR 2.0 [12]	91.0	58.4	9.2	50.0	33.6
LLaVA-S [30]	440.8	205.4	21.8	461.3	195.4
LLaVA*-S [30]	232.1	101.1	12.8	246.0	118.2
GPT4-S [36]	322.0	136.7	16.0	336.9	144.0
LLaVA-P [30]	335.2	172.3	16.5	334.1	172.5
GPT4-P [36]	396.5	203.4	18.6	354.1	203.5
ChatPose (Ours)	163.6	81.9	10.4	126.0	82.4

Table 3. **Comparison on Human Pose Estimation.** MPJPE (mm), PA-MPJPE (mm), and MPJRE ( $\times 100$ ) are reported.

We report the text-to-pose recall rate  $R^{T2P}$  and the pose-to-text recall rate  $R^{P2T}$  of the retrieval models trained on real poses and evaluated on generated poses. Following previous work [7], for the SPG task, the retrieval model is re-trained for evaluation using SPG training data.

**Estimation.** To evaluate traditional and reasoning-based 3D pose estimation, we use the traditional metrics: Mean Per-Joint Position Error (MPJPE) and this error after rigidly aligning the posed body with the ground truth (PA-MPJPE). Additionally, we introduce the Mean Per-Joint Rotation Error (MPJRE) to more directly evaluate body pose accuracy. To evaluate human pose estimation, we select 200 samples from the 3DPW [51] and Human3.6M [17] test sets. To assess ChatPose’s performance on SPG and RPE tasks, we introduce several baseline methods:

- **LLaVA\***. Instead of utilizing the pose token  $\langle \text{POSE} \rangle$ , human poses can be represented through language, such as textual descriptions of keypoint locations. Using the same dataset pairs as in ChatPose, we formulate VQA pairs as described in *Sup. Mat.* for training. We then fine-

tune the base model LLaVA, referred to as LLaVA \*, with results shown in Table 3 and Fig. 4.

- **LLaVA-S, LLaVA\*-S, and GPT4-S.** For the RPE task, we initially request LLMs, such as LLaVA [30], LLaVA\*, and GPT4 [36], to provide textual descriptions of the keypoint locations for the target individual, and then apply SMPLify [3] to optimize the human poses based on these keypoint locations.
- **LLaVA-P and GPT4-P.** Similarly, for RPE and SPG tasks, we use LLMs like LLaVA [30] and GPT4 [36] to describe human poses in response to questions, and then generate SMPL poses with PoseScript [7] from these descriptions. We show the RPE results in Figure 5 and SPG comparison in *Sup. Mat.*

### 4.3. Pose Generation

We evaluate ChatPose’s pose generation capabilities on both the classical task and the new SPG task. Figure 3 shows how ChatPose handles detailed and speculative queries, outperforming PoseScript in complex scenarios involving reasoning. While ChatPose and DALL-E produce different output modalities (3D poses vs images), they both “understand” the concepts. Quantitatively, as Table 1 shows, ChatPose performs comparably to PoseScript on classical tasks (with detailed pose descriptions) and outperforms it on speculative pose generation.

### 4.4. Pose Estimation

Figure 4 and Table 3 show qualitative and quantitative results on classical human pose estimation. ChatPose outperforms other Multi-modal LLMs, yet it does not match the performance of methods designed and trained specifically to estimate 3D human pose. This is not surprising and we see these results as a first proof-of-concept. For failure cases, please see the *Sup. Mat.* In reasoning-based human pose estimation, ChatPose outperforms both task-specific and multi-modal LLM methods. This is illustrated in Fig. 5 and Table 2. Notably, the MPJPE is heavily affected by the global orientation, while PA-MPJPE lessens this impact, offering a truer reflection of body pose accuracy. ChatPose has trouble estimating global orientation of the person; this could likely be addressed by additional training.

We also found that ChatPose generalizes well to strong occlusions. Even without any data augmentation during training. This suggests that it is able to leverage its general visual knowledge about occlusion in solving the human pose estimation problem. See *Sup. Mat.* for examples.

### 4.5. GPT-Assisted Evaluation

When training ChatPose to understand 3D pose, it is critical that it does not forget its general knowledge. To evaluate this, we follow LLaVA’s [30], using GPT4-Assisted evaluation. Table 4 shows ChatPose slightly lags LLaVA, indicat-

Method	Conv	Detail	Complex	All
LLaVA-V1-13B [30]	83.1	75.3	96.5	85.1
LLaVA-V1.5-13B [29]	84.4	81.0	93.9	86.5
ChatPose (Ours)	78.8	76.2	96.7	84.0

Table 4. **GPT4-Assisted Evaluation.** “Conv,” “Details,” and “Complex” signify three categories of questions produced by the LLaVA data generation pipeline, covering conversation, detailed description, and complex reasoning.

ing ChatPose successfully combines 3D pose abilities with its vision and language understanding.

### 4.6. Ablation study

We evaluate the impact of different aspects of ChatPose, including human pose representations, multi-modal LLM backbones, and various datasets. Please refer to *Sup. Mat.*

## 5. Conclusions

ChatPose makes a first step towards integrating 3D human pose estimation with the general reasoning capabilities of LLMs. This study teaches us several things. First, multi-modal LLMs can be fine-tuned to infer 3D human pose from images. In particular, they are able to infer the real-valued rotations of human body parts. To our knowledge, this is the first demonstration that such models can directly solve this task. Second, the model can connect 3D human pose with language. This is important because it opens up many possibilities both for applications and for training. Third, we have demonstrated new use cases in which a user can chat with the language model about 3D human pose using text and images. We think this is the beginning of a rich space that will open up new ways of training and using LLMs to reason about 3D human pose.

**Limitations.** The accuracy of our 3D pose estimation from images is below recent specialized regressors. Better quality data relating language to pose is needed. A key lesson of recent LLM research is that the scale and quality of the data is key. Additionally, freezing the vision encoder is a limitation which could be overcome with a more powerful backbone or by fine-tuning the whole model on more data.

**Future work.** Future work should also improve the ability of ChatPose to have multi-turn conversations about 3D pose. It should also be possible to enable pose editing, cf. [8]. It should be straightforward to extend our work to infer and reason about 3D body shape and human movement. The extension to video input is particularly promising given recent progress on video models, which have broad knowledge about the 3D world and human behavior, e.g. [2].

**Acknowledgements.** We thank Weiyang Liu, Haiwen Feng and Longhui Yu for discussions and proofreading. We also thank Nareen Mahmood and Nicolas Keller for support with data. This work was partially supported by the Max Planck ETH Center for Learning Systems. CoI disclosure: [https://files.is.tue.mpg.de/black/CoI\\_CVPR\\_2024.txt](https://files.is.tue.mpg.de/black/CoI_CVPR_2024.txt).



# ChatPose: Chatting about 3D Human Pose

## Supplementary Material

### 6. 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 7 and Table 5, while example answers are shown in Table 6.

- “<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 5. 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 6. The list of answers for training ChatPose with SMPL pose as the output.

### 7. 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 [7], 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 [36], prompting it to generate implicit descriptions of associated activities, as shown in Figure 6. 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 8. 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 [52] 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 [36] consistently confuses left and right body parts. Inspired by [56], we incorporate a visual prompt to assist the model in distinguishing between left and right body parts. Specifically, we utilize ViTPose [55] 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 7. The details of the query input are represented in Table 9. After generating these 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.

### 8. 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

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

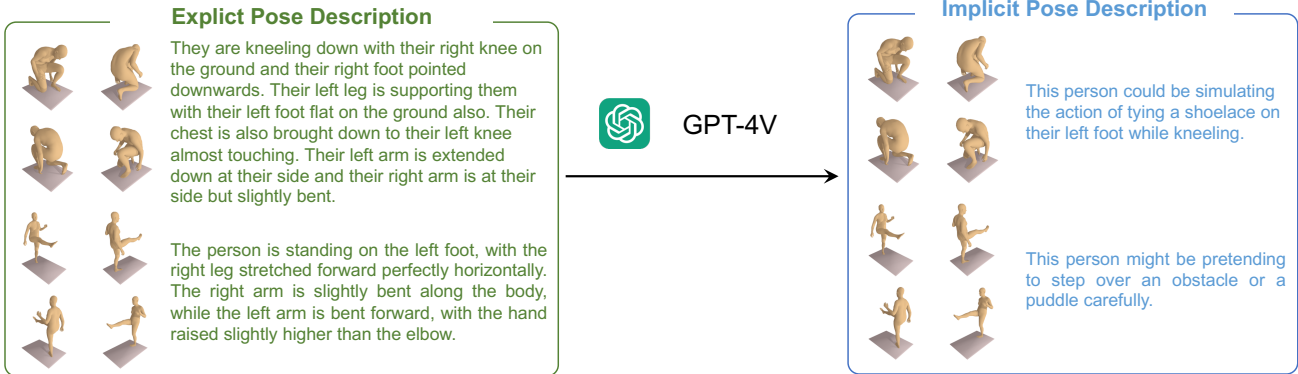


Figure 6. Illustration of the annotation pipeline that generates implicit pose description for our SPG benchmark. We take the fine-grained explicit pose descriptions from PoseScript [7] and visualize the described pose from four viewpoints, and then query GPT4 to reformulate them into indirect pose descriptions.

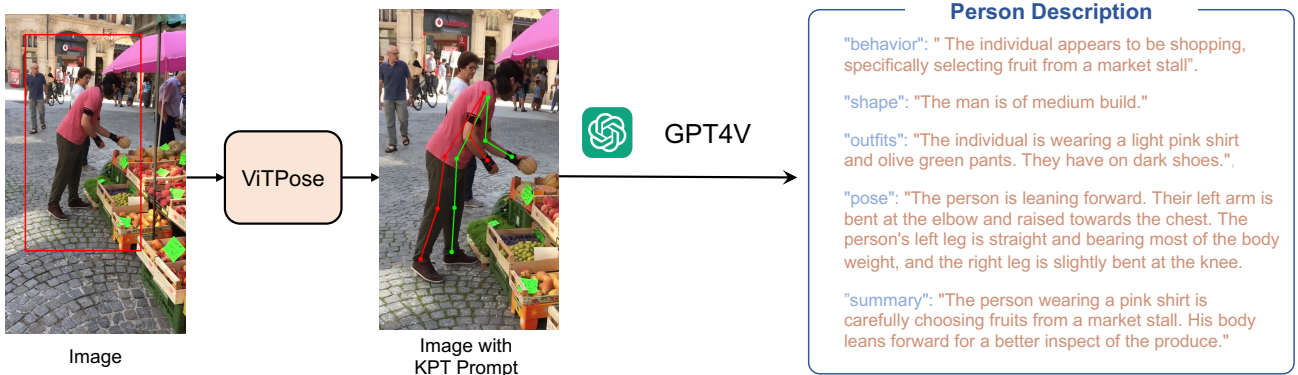


Figure 7. Illustration of our method to generate person descriptions for the RPE benchmark. We use ViTPose [55] to detect the body keypoints and mark the left-body and right-body joints with different colors as visual prompts, and then query GPT4V for descriptions.

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},

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 8. Example to query GPT4 for implicit pose descriptions.

{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 8 provides some examples of these training pairs. We then fine-tune the base model, LLaVA [30], 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 9. 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-

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 10 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<sup>3</sup> and LLaVA-V1.5-13B<sup>4</sup> models, which

<sup>3</sup>liuhaotian/llava-v1.5-7b

<sup>4</sup>liuhaotian/llava-v1.5-13b

(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 9. 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.

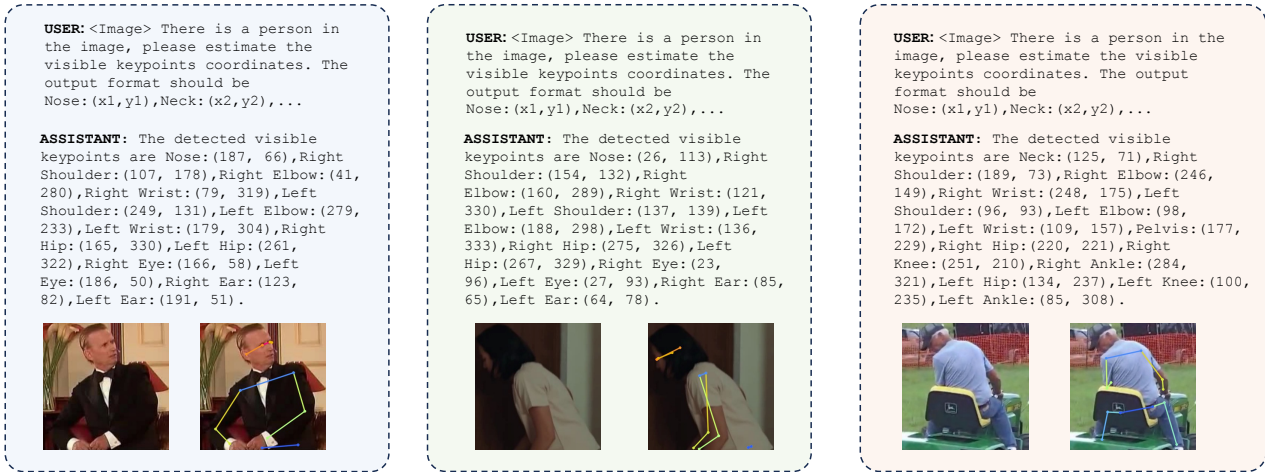


Figure 8. Examples of VQA data used to fine-tune the LLaVA model for pose estimation with textual descriptions of 2D keypoints.

Method	VQA [30]	Image2Pose	Text2Pose	Pose Estimation		Reasoning-based Pose Estimation
				3DPW [52]	H36M [17]	
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 10. 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 [52]	H36M [17]	
LLaVA-V1.5-7B [29]	84.5	82.9	102.5
LLaVA-V1.5-13B [29]	81.9	82.4	101.8

Table 11. Ablation study: effect of multimodal LLM backbones. PA-MPJPE (in mm) is reported. Lower is better.

are based on the LLaMA-7b and LLaMA-13b backbones, respectively. Table 11 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.

## 8.1. 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 10 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



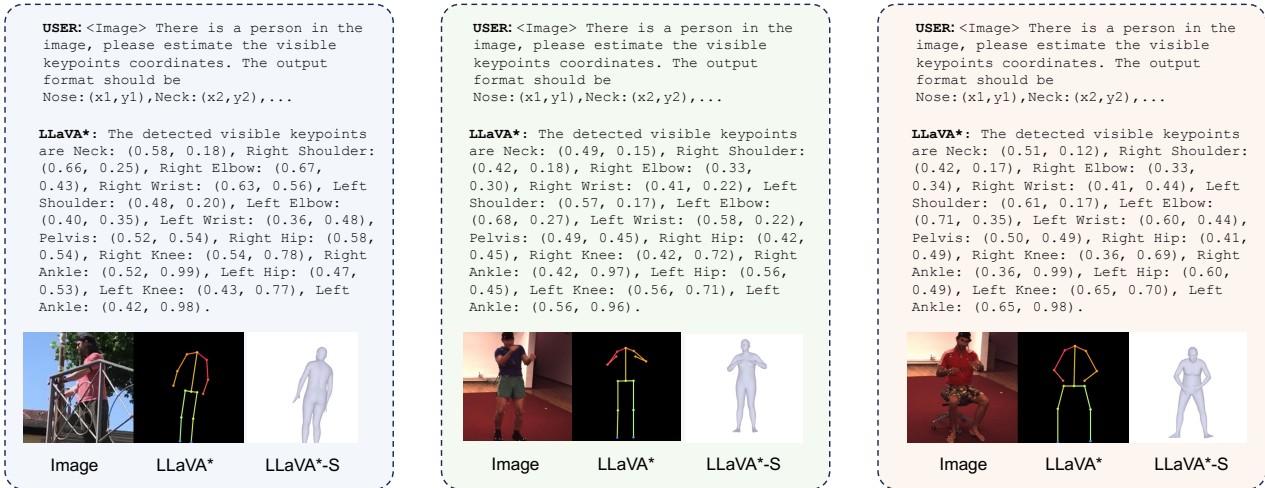


Figure 9. Visual results of LLaVA \*. Given an RGB image, LLaVA \* generates textual descriptions about keypoint locations. We then extract the keypoints from the textual descriptions and adopt SMPLify [3] to fit the SMPL pose.

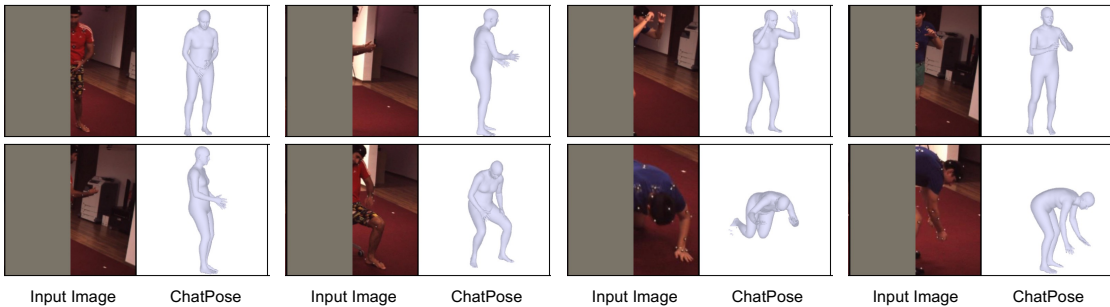


Figure 10. Pose estimation on images with significant occlusion. Without training for occlusion cases, ChatPose is surprisingly robust.

output 3D human poses, we adopt two approaches: firstly, generating keypoint coordinates followed by SMPLify [3] optimization of the 3D pose, and secondly, producing textual descriptions of the pose that are then processed by PoseScript [7] to create SMPL pose parameters. The workflow for the first method is illustrated in Figure 9, and for the second method in Figure 11.

**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. 12. 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	<b>0.50</b>	1.21
ChatPose	1.51	<b>0.75</b>

Table 12. 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 [7] 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 13. 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. 13.

**Other baselines for RPE and SPG.** We show more baselines in Table 14. 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 fine-

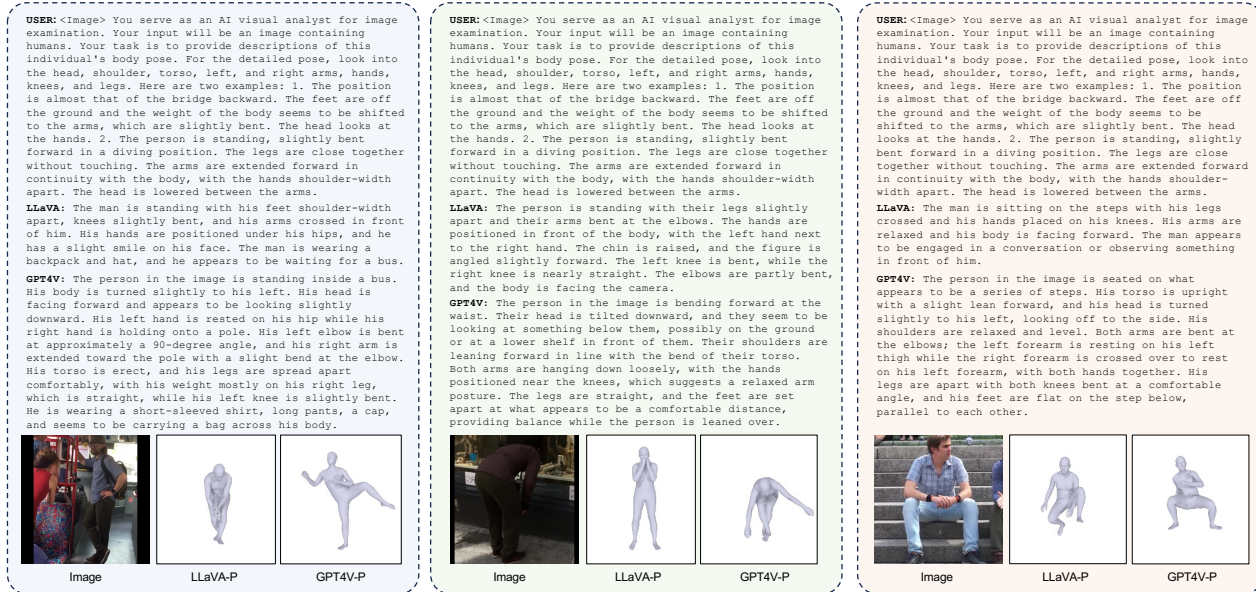


Figure 11. Visual results of LLaVA and GPT4. Given an RGB image, LLaVA and GPT4 generate textual descriptions about human poses. We then use PoseScript [7] to generate SMPL poses based on the text descriptions.

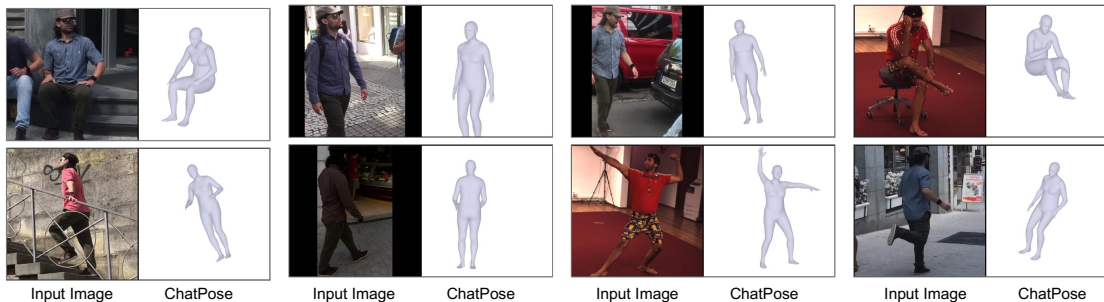


Figure 12. Failures cases of ChatPose on the human pose estimation task. Note that a common failure mode is to estimate the articulated pose correctly but to output the incorrect global orientation.

Method	$R^{P2T} \uparrow$	$R^{T2P} \uparrow$
PoseScript	<b>22.6/31.0/57.9/70.8</b>	22.4/32.1/58.7/71.5
ChatPose	17.6/25.3/57.6/71.2	<b>28.0/39.0/70.4/83.5</b>

Table 13. TOP 5/10/50/100 T2P and P2T results with retrieval model from PoseScript journal version.

Method	SPG $R^{P2T} \uparrow$	SPG $R^{T2P} \uparrow$
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 14. 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. accurate than ChatPose.



Figure 13. 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.

finetuned PoseScript with SPG data; the results in are also less

**Failure Cases.** We also show some limitations of the current model in Figure 12. 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.

## References

- [1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D Human Pose Estimation: New benchmark and state of the art analysis. In *Computer Vision and Pattern Recognition (CVPR)*, 2014. 6
- [2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. 8
- [3] 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. 3, 7, 8, 13
- [4] Rania Briq, Pratika Kochar, and Juergen Gall. Towards better adversarial synthesis of human images from text. *arXiv preprint arXiv:2107.01869*, 2021. 3
- [5] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. Dreamavatar: Text-and-shape guided 3D human avatar generation via diffusion models. *arXiv preprint arXiv:2304.00916*, 2023. 3
- [6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality, 2023. 3
- [7] 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. 2, 3, 5, 6, 7, 8, 9, 10, 13, 14
- [8] Delmas, Ginger and Weinzaepfel, Philippe and Moreno-Noguer, Francesc and Rogez, Grégory. PoseFix: Correcting 3D Human Poses with Natural Language. In *ICCV*, 2023. 8
- [9] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, 2021. 2
- [10] Yao Feng, Weiyang Liu, Timo Bolkart, Jinlong Yang, Marc Pollefeys, and Michael J. Black. Learning disentangled avatars with hybrid 3d representations. *arXiv*, 2023. 3
- [11] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 4
- [12] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *International Conference on Computer Vision (ICCV)*, 2023. 2, 3, 4, 6, 7
- [13] Siavash Golkar, Mariel Pettee, Michael Eickenberg, Alberto Bietti, Miles Cranmer, Geraud Krawezik, Francois Lanasse, Michael McCabe, Ruben Ohana, Liam Parker, et al. xVal: A continuous number encoding for large language models. *arXiv preprint arXiv:2310.02989*, 2023. 4
- [14] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. AvatarCLIP: Zero-shot text-driven generation and animation of 3D avatars. In *Transactions on Graphics (TOG)*, 2022. 3
- [15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 2, 4
- [16] Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Zhou Zhao, and Shinji Watanabe. AudioGPT: Understanding and generating speech, music, sound, and talking head. *arXiv preprint arXiv:2304.12995*, 2023. 3
- [17] 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. 6, 7, 12
- [18] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. MotionGPT: Human motion as a foreign language. *arXiv preprint arXiv:2306.14795*, 2023. 2
- [19] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. In *International Conference on 3D Vision (3DV)*, 2020. 3
- [20] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3
- [21] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *International Conference on Computer Vision (ICCV)*, 2021. 2, 6
- [22] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision (ICCV)*, 2019. 3, 4, 6, 7
- [23] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 2, 4
- [24] Kunchang Li, Yanan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. VideoChat: Chat-centric video understanding. *arXiv preprint arXiv:2205.06355*, 2023. 4
- [25] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 3
- [26] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiayang Tang, Yangyi Huang, Justus Thies, and Michael J Black. TADA! text to animatable digital avatars. In *International Conference on 3D Vision (3DV)*, 2024. 3

- [27] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. *CVPR*, 2023. 6
- [28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. 2014. 6
- [29] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 4, 5, 8, 12
- [30] 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, 7, 8, 11, 12
- [31] Zhaoyang Liu, Yanan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Zeqiang Lai, Yang Yang, Qingyun Li, Jiashuo Yu, et al. InternGPT: Solving vision-centric tasks by interacting with chatbots beyond language. *arXiv preprint arXiv:2305.05662*, 2023. 3
- [32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. In *Transactions on Graphics (TOG)*, 2015. 2, 3, 4, 5
- [33] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, 2019. 6
- [34] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved cnn supervision. In *International Conference on 3D Vision (3DV)*, 2017. 6
- [35] OpenAI. Introducing chatgpt. 2022. 3
- [36] OpenAI. GPT-4 technical report. 2023. 2, 3, 6, 7, 8, 9
- [37] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 3
- [38] Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, Lingpeng Kong, and Tong Zhang. DetGPT: Detect what you need via reasoning. *arXiv:2305.14167*, 2023. 3
- [39] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3D using 2D diffusion. 2022. 3
- [40] Zhongwei Qiu, Qiansheng Yang, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Chang Xu, Dongmei Fu, and Jingdong Wang. Psvt: End-to-end multi-person 3d pose and shape estimation with progressive video transformers. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3
- [41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 6
- [42] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 3
- [43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 3
- [44] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. HuggingGPT: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023. 3
- [45] Stephan Streuber, M. Alejandra Quiros-Ramirez, Matthew Q. Hill, Carina A. Hahn, Silvia Zuffi, Alice O’Toole, and Michael J. Black. Body Talk: Crowdshaping realistic 3D avatars with words. *Transactions on Graphics (TOG)*, 2016. 3
- [46] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. PandaGPT: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023. 4
- [47] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *International Conference on Computer Vision (ICCV)*, 2021. 2, 3
- [48] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J. Black. Putting people in their place: Monocular regression of 3D people in depth. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3
- [49] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. 2023. 3
- [50] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 3, 6
- [51] Timo von Marcard, Roberto Henschel, Michael 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. 6, 7
- [52] 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. 7, 9, 12
- [53] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. VisionLLM: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv:2305.11175*, 2023. 3
- [54] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NEX-T-GPT: Any-to-any multimodal LLM. *arXiv preprint arXiv:2309.05519*, 2023. 2, 4
- [55] Yufeì 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. 9, 10



- [56] 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. 9
- [57] Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. GPT4Tools: teaching large language model to use tools via self-instruction. *arXiv preprint arXiv:2305.18752*, 2023. 3
- [58] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. MM-ReAct: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 3
- [59] Ruosong Ye, Caiqi Zhang, Runhui Wang, Shuyuan Xu, and Yongfeng Zhang. Natural language is all a graph needs. *arXiv:2308.07134*, 2023. 4
- [60] Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*, 2023. 4
- [61] Hongwen Zhang, Yating Tian, Xinchu Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In *International Conference on Computer Vision (ICCV)*, 2021. 2, 3
- [62] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 4
- [63] Hao Zhang, Yao Feng, Peter Kulits, Yandong Wen, Justus Thies, and Michael J. Black. TECA: Text-guided generation and editing of compositional 3d avatars. In *International Conference on 3D Vision (3DV)*, 2024. 3
- [64] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *International Conference on Computer Vision (ICCV)*, 2023. 3
- [65] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023. 6
- [66] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 6
- [67] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*, 2023. 4