PoseLLaVA: Pose Centric Multimodal LLM for Fine-Grained 3D Pose Manipulation

Dong Feng[1*], Ping Guo[2*†], Encheng Peng[3], Mingmin Zhu[1], Wenhao Yu[4], Peng Wang[2]

[1] inchitech [2] Intel Labs China
[3] Nanjing University of Posts and Telecommunications [4] Beijing Jiaotong University
ustcfd2018@gmail.com , ping.guo@intel.com , 1322059105@njupt.edu.cn

Abstract

Manipulating human poses based on natural language is an emerging research field that has traditionally focused on coarse commands such as "walking" or "dancing." However, fine-grained pose manipulation, like instructing "put both hands in front of the stomach," remains underexplored. In this paper, we introduce PoseLLaVA, a pioneering model that integrates SMPL-based pose representations into the multimodal LLaVA framework. Through a novel pose encoder-decoder mechanism, PoseLLaVA achieves precise alignment between pose, textual, and visual modalities, enabling detailed control over pose manipulation tasks. PoseLLaVA excels in three key tasks: pose estimation, generation, and adjustment, all driven by detailed language instructions. We further introduce a fine-grained pose adjustment dataset PosePart, where each sample contains an initial pose and a target pose, along with specific instructions for adjustments, mimicking the guidance a human instructor might provide. Extensive evaluations across these tasks demonstrate significant improvements over existing methods, including metrics such as MPJPE and PA-MPJPE, which measure SMPL reconstruction errors, and Recall rates, which assess feature alignment across modalities. Specifically, PoseLLaVA reduces MPJPE errors by more than 20% compared to state-of-the-art methods in pose adjustment and generation tasks. Additionally, we demonstrate the feasibility of combining PoseLLaVA with generative models, such as diffusion, for pose image editing, highlighting its potential applications in language-controlled pose manipulation.

Code — https://github.com/ustcfd/PoseLLaVA

Introduction

Accurately capturing and generating 3D human poses is crucial for lifelike animations in fields such as movie production, gaming, virtual reality, and robotics. Despite numerous advancements, the process often requires high-end motion capture systems and skilled professionals to design intricate human poses and movements. To democratize this technology, it is essential to develop a pose manipulation model that communicates seamlessly with users, particularly through natural language interfaces.

Manipulating human poses based on natural language is an emerging research area that remains largely unsolved. Existing approaches predominantly focus on motion generation from coarse semantic commands like "walking" or "running" (Zhang et al. 2024). However, fine-grained pose manipulation—such as instructing "put both hands in front of the stomach" or adjusting a posture from "vertical arms" to "horizontal arms"—presents significant challenges and has been underexplored.

These challenges can be viewed from two perspectives. First is Model Structure. A straightforward approach to natural language-driven pose manipulation might involve using existing multimodal large language models like MiniGPT-4 (Zhu et al. 2023) or LLaVA (Liu et al. 2024c), which connect image and text modalities. However, images alone often lack the precision needed for detailed adjustments. Subtle changes in joint angles or limb orientations are difficult to extract and interpret. To address this, we propose incorporating SMPL-based pose representations into the LLaVA model. By introducing a pose encoder-decoder mechanism and a three-stage training strategy, we achieve well-aligned pose, image, and text modalities. Second is Data Availability. Fine-grained pose-language datasets are scarce. Recently, PoseFix (Delmas et al. 2023) provided the first language descriptions for pose adjustment. However, these descriptions are often redundant or overly complex due to rigid rules. Additionally, PoseFix's selection of pose pairs involves significant changes across multiple body parts, making it challenging for models to learn subtle movements. Drawing inspiration from how human instructors adjust poses—focusing on one body part at a time and breaking down complex movements—we introduce a body-part-level pose adjustment dataset, PosePart. This dataset complements PoseFix by offering small changes in single body part, which is critical for fine-grained pose manipulation.

We present PoseLLaVA, a model designed to handle three pose manipulation tasks guided by language instructions: pose estimation, pose generation, and pose adjustment, as illustrated in Figure 1. For pose estimation, the model predicts a 3D human pose from a monocular RGB image. In pose generation, it creates a 3D pose from a detailed language description. In pose adjustment, an initial pose is modi-

[*]Co-first authors contributed equally
[†]Corresponding author.

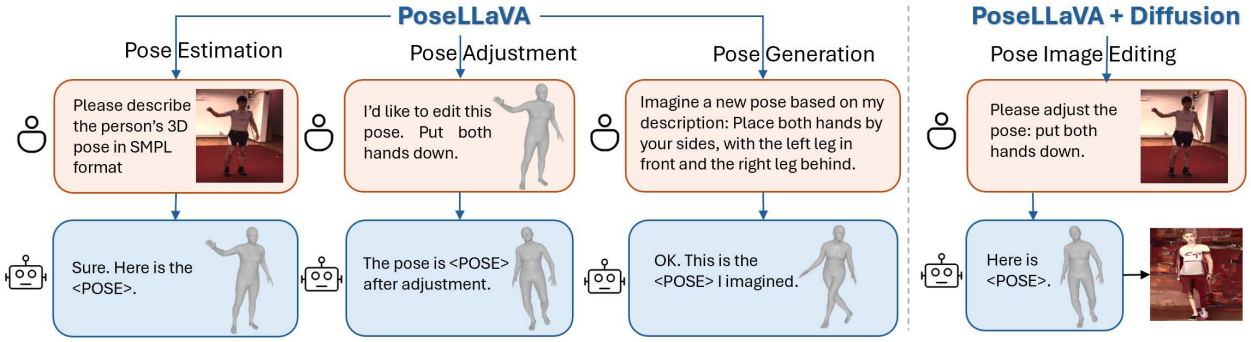

Figure 1: We introduce PoseLLaVA, a pose centric multimodal LLM designed for pose estimation, pose adjustment and pose generation. PoseLLaVA accepts three types of input modalities: images, text and SMPL pose parameters. Left: Example of using PoseLLaVA for 3D pose manipulation. Right: Example of combining PoseLLaVA with Diffusion models such as PIDM(Bhunia et al. 2023) for pose image editing.

fied based on a language instruction, and generate the target pose. Unlike existing approaches that address each task separately, PoseLLaVA employs a unified framework and results in more coherent outcomes. The model's advanced NLP capabilities further enhance its ability to interpret complex language instructions. In summary, our contributions are threefold:

- Pioneering Pose Centric Multimodal LLM (PoseLLaVA) : It integrates a pose encoder and decoder into the LLaVA framework. It aligns pose, image, and text modalities through a three-stage training strategy. Unlike existing works that specifically designed for each one task, PoseLLaVA offers a unified framework capable of handling multiple pose manipulation tasks guided by language.
- Fine-Grained Pose Adjustment Dataset (PosePart): In PosePart, each sample consists of a pose pair with a single body part change and a fine-grained language instruction for adjustment. This dataset complements existing ones by enhancing data diversity and instruction quality, thereby improving the accuracy of pose manipulation.
- Leading Performance Across Metrics and Tasks: Our model outperforms multimodal LLM baselines in the pose estimation task, and significantly surpasses existing methods in pose generation and adjustment. Additionally, when combined with diffusion models, our method enables language guided pose image generation, outperforming standalone diffusion models.

Related Work

Human Pose Estimation Human pose estimation has a long history in computer vision. This paper focuses on estimating 3D human poses from a single image using the SMPL model. State-of-the-art methods are typically optimization-based, refining poses iteratively to minimize the difference between projected and detected points (Joo, Neverova, and Vedaldi 2021), or regression-based, directly inferring pose parameters from images using deep learning (Choutas et al. 2022). However, most works focus solely on single pose estimation without incorporating language interaction, limiting practical usability. Recently, ChatPose (Feng et al. 2024) leveraged a vision-language model to enhance user interaction by reasoning about 3D human poses from images and text, though it still lags in accuracy compared to traditional methods.

Human Pose Generation Existing works in pose generation mainly focused on generating pose sequences conditioned on coarse text descriptions, like "running" or "dancing." Methods like MotionDiffuse (Zhang et al. 2024) use diffusion models for motion prediction, while T2M-GPT (Zhang et al. 2023b) employs transformers to generate motion based on text context. Although these models produce coherent sequences, they struggle with fine-grained poses, such as "a person walking with hands extended horizontally to the sides." PoseScript (Delmas et al. 2022) and ChatPose (Feng et al. 2024) have recently shown potential in generating specific 3D poses from detailed descriptions. Despite progress, fine-grained pose generation remains challenging, with accuracy much lower than pose estimation(Feng et al. 2024; Delmas et al. 2024).

Human Pose Adjustment Adjusting 3D human poses via natural language has applications in fitness coaching and animation editing, but research in this area is limited. Some methods generate new images with pose controls, like ControlNet (Zhang, Rao, and Agrawala 2023), while others allow manual pose editing via tools like dragging (Yenphraphai et al. 2024). However, research on language-driven pose editing is scarce. PoseFix (Delmas et al. 2023), which predicts the modified pose, introduced a triplet dataset comprising initial pose, target pose, and the textual description of the change. However, PoseFix emphasizes significant pose differences, which makes learning subtle transformations more difficult and less applicable to real-world scenarios.

Multimodal Large Language Models Multimodal Large Language Models have enhanced the interpretation of various modalities, particularly in vision and language tasks like visual question answering and image captioning. Models like MiniGPT-4 (Zhu et al. 2023) and LLaVA (Liu et al.

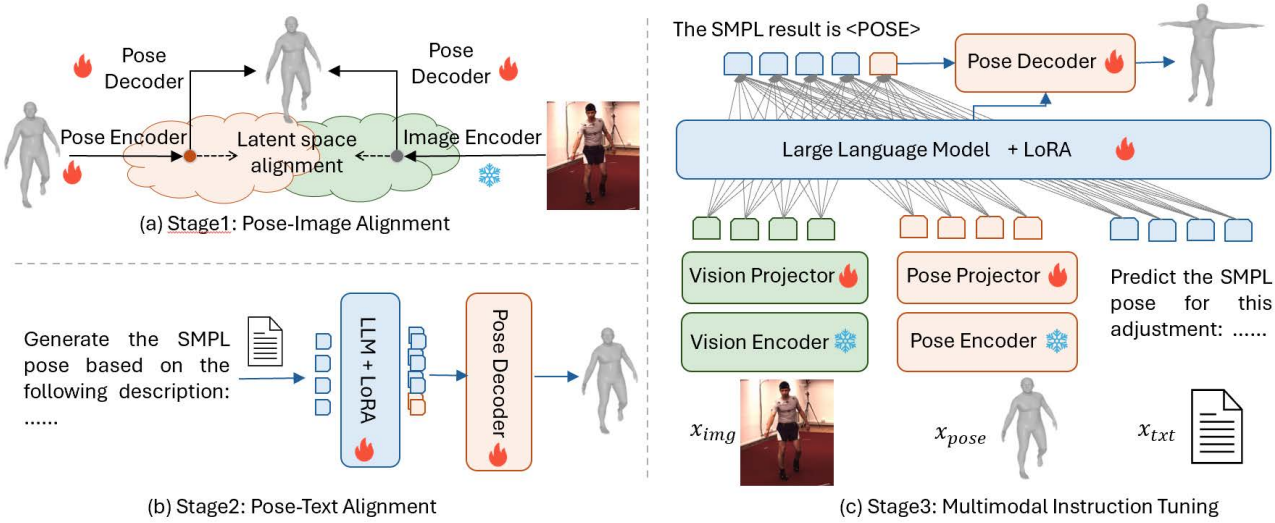

Figure 2: We developed a three-stage training pipeline for PoseLLaVA. First, align the pose and vision encoder in latent space. Second, the LLM is pre-trained for pose generation task. Finally, we perform instruction tuning across all three task.

2024c) integrate vision encoders with LLMs for vision language interaction. Despite recent advances in exploring multiple modalities such as video and audio (Zhang et al. 2023a; Zhang, Li, and Bing 2023; Wu et al. 2024), the integration of the pose modality remains underexplored. ChatPose (Feng et al. 2024) recently enabled interaction with 3D poses through language, and MotionLLM (Chen et al. 2024) integrated video and motion inputs for motion understanding.

In this paper, we explore pose estimation, adjustment, and generation within a unified framework. We integrate a pose modality with an encoder and decoder into the LLaVA model and develop a three-stage training strategy for aligning pose, image, and text modalities. We also introduce the PosePart dataset for fine-grained pose adjustments.

Methodology

Architecture

The architecture and training pipeline of PoseLLaVA is illustrated in Figure 2. We integrate the pose modality into the vision-language model LLaVA (Liu et al. 2024a) to create a pose centric multimodal LLM. Our model accepts text, images, and poses as inputs and produces text and SMPL pose parameters as outputs. We use CLIP-ViT (Radford et al. 2021) as the vision encoder, followed by a vision projector same with that in LLaVA. The pose modality is represented by pose orientations in 6D representation (Zhou et al. 2019) for both pose encoder input and pose decoder output. In our implementation, we fix the shape parameters and only estimate rotation parameters. For the pose encoder and decoder, we propose a transformer architecture, along with a pose projector with the same design as the vision projector.

Let the image input be denoted as x_{img}, the pose input as x_{pose}, and text prompts as x_{txt}. Let the multimodal LLM be denoted as M. The LLM generates an output sequence of tokens $Y_{txt} = [t_1, ..., t_n]$ associated with corresponding hidden states $[h_1, ..., h_n]$ such that:

$$Y_{txt} = \mathcal{M}(f(x_{img}, x_{pose}), x_{txt}) \quad (1)$$

where $f(\cdot)$ is a modality selector that select image or pose embedding according to the modality placeholder $\langle image \rangle$ or $\langle pose \rangle$. We follow ChatPose (Feng et al. 2024) and Lisa (Lai et al. 2024) by expanding the text vocabulary with a $\langle POSE \rangle$ token. If one of the LLM output tokens $t_n \in Y_{txt}$ is the $\langle POSE \rangle$ token, we extract the hidden state of the specical token as $h_{pose} \in R^d$, where d is the dimension of the LLM hidden states. The hidden state h_{pose} is then projected by the PoseDecoder into the pose rotation in 6D representation that $Y_{pose} \in R^{144}$.

Training Strategy

Pose-Image Alignment The PoseEncoder is a critical component of PoseLLaVA, addressing the limitations in understanding fine-grained pose semantics. Unlike image and text modalities, which benefit from extensive public datasets, fine-grained pose datasets are scarce. To overcome this, we employ a CLIP-like contrastive learning approach combined with a pose encoder-decoder structure to align the pose modality with the well-pretrained image modality.

As illustrated in Figure 3, the model consists of three main components: *i)* a pre-trained CLIP vision encoder, *ii)* a pose encoder-decoder, and *iii)* a lightweight cross-attention Resampler designed to learn compact visual representations.

The vision and pose encoders are aligned at both global and local scales. At the global scale, both encoders use an additional learnable global token as input, producing global feature embeddings V_{avg} and Z_{avg}. At the local scale, since tokens in the visual and pose modalities are not directly consistent, we apply a resampler to compress visual features V_{local} to match that of Z_{local}. The compressed local visual features V_{local} is denoted as $V_{local} = [V_1, V_2, \ldots, V_n]$, and the local pose features are denoted

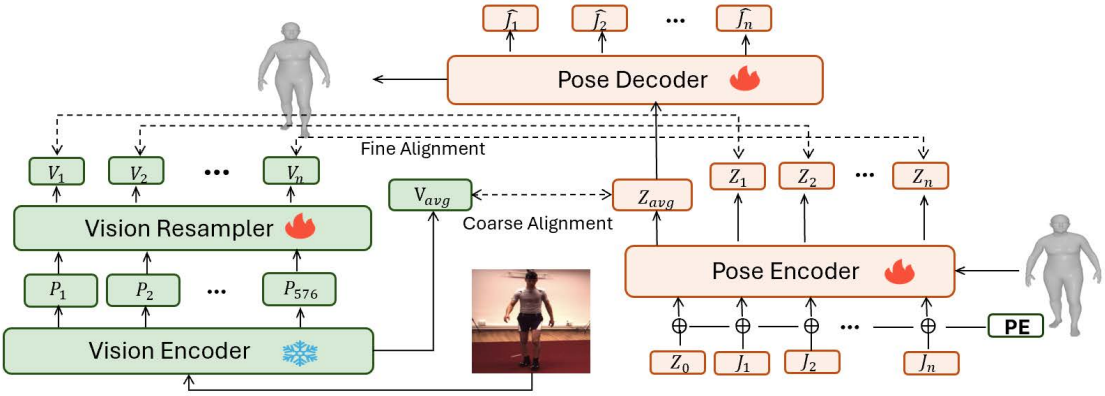

Figure 3: Multi-scale Alignment between Pose and Image.

as $Z_{local} = [Z_1, Z_2, \ldots, Z_n]$. We use contrastive learning method to align the global features of the vision and pose encoders. Additionally, we employ regression loss $\mathcal{L}_{reg}$ in the pose encoder-decoder to achieve accurate pose reconstruction.

$$\begin{aligned}\mathcal{L}_{stage1} =& \mathcal{L}_{cl}(Z_{avg}, V_{avg}) + \mathcal{L}_{cosin}(Z_{local}, V_{local}) \\ &+ \mathcal{L}_{reg}(x_{pose}, \hat{x}_{pose})\end{aligned} \tag{2}$$

where $x_{pose} = [J_1, J_2, \ldots, J_n]$ and $\hat{x}_{pose} = \left[\hat{J}_1, \hat{J}_2, \ldots, \hat{J}_n\right]$ are the input and groundtruth of pose representations that $n = 24$ indicating a global body orientation and 23 SMPL joints, and $J_n \in R^{1\times 6}$ is the 6D pose rotations. For the contrastive learning loss $\mathcal{L}_{cl}$, we employ SigLip (Zhai et al. 2023) for global feature alignment.

Pose-Text Alignment Since existing multimodal models like LLaVA (Liu et al. 2024a) and language models like Mistral (Jiang et al. 2023) are trained on natural language data, there is a significant distribution mismatch with the feature space of our pose manipulation tasks. To address this, we pre-train the LLM component to better adapt to pose tasks. We achieve this by training the LLM and pose decoder on the pose generation task, where input is text data and output is SMPL pose parameters. To expand the dataset scale and variety, we combine text data from both PoseFix (Delmas et al. 2023) and PoseScript (Delmas et al. 2022). For the PoseFix data which was original created for pose adjustment, we exclude the initial pose data, using only text data as input and target pose as output.

Multi-Task Instruction Tuning In this step, we use a multi-modal instruction-following strategy using all data from the three tasks. The model is trained end-to-end using an auto-regressive cross-entropy loss $\mathcal{L}_{txt}$ for text generation and a pose loss $\mathcal{L}_{pose}$ for generating SMPL pose parameters. The pose loss is the L2 loss between the predicted and target pose parameters. The overall objective $\mathcal{L}$ is the weighted sum of these losses, with weights λ_{txt} and λ_{pose} set to 1 and 10, respectively.

$$\begin{aligned}\mathcal{L}_{stage3} =& \lambda_{txt} \times \mathcal{L}_{txt}(Y_{txt}, \hat{Y}_{txt}) \\ &+ \lambda_{pose} \times \mathcal{L}_{pose}(Y_{pose}, \hat{Y}_{pose})\end{aligned} \tag{3}$$

Where Y_{txt} and Y_{pose} are the predicted text and pose, $\hat{Y}_{txt}$ and $\hat{Y}_{pose}$ are the ground truth text and pose, respectively.

PosePart Dataset Construction

We introduce a new dataset for fine-grained pose adjustment, called PosePart. This dataset focuses on changes in a single body part per sample to simulate fine-grained pose manipulation. The construction pipeline is illustrated in Figure 4.

We divide the SMPL parameters into six body parts: left upper limb, right upper limb, left lower limb, right lower limb, trunk, and head. Pose pairs are generated by adjusting only one body part at a time, enabling the model to learn subtle changes. We begin by selecting pose pairs PoseA and PoseB from AMASS (Mahmood et al. 2019), where PoseA serves as the initial pose and PoseB as the reference pose. Our process for creating triplet data samples $\langle PoseA, PoseC, Text_{modifier}\rangle$ consists of three steps: First, we compare PoseA and PoseB to identify the body part with the most significant pose change. Second, the selected change from PoseB is applied to PoseA to simulate the pose adjustment a human tutor might request. Subtle noise is added to the non-selected body parts to introduce data variation. Third, pose difference descriptions between PoseA and PoseC are generated using PoseFix. Given the limitations of rigid rules in PoseFix, the resulting descriptions may be redundant or complex, so we refine them using GLM-4 (GLM et al. 2024). The prompt to GLM-4 includes two components: specifying the highlighted body part and selecting language descriptions for this part from the PoseFix results. Ultimately, we construct a dataset of 135,000 samples, matching the size of PoseFix, to ensure balanced data variation.

Experiments

In this section, we present a series of experiments designed to evaluate the performance of PoseLLaVA. We compare our model against SOTA baselines and demonstrate its effectiveness in both qualitative and quantitative assessments.

Implementation Details The training dataset is constructed by converting each task-specific datasets into

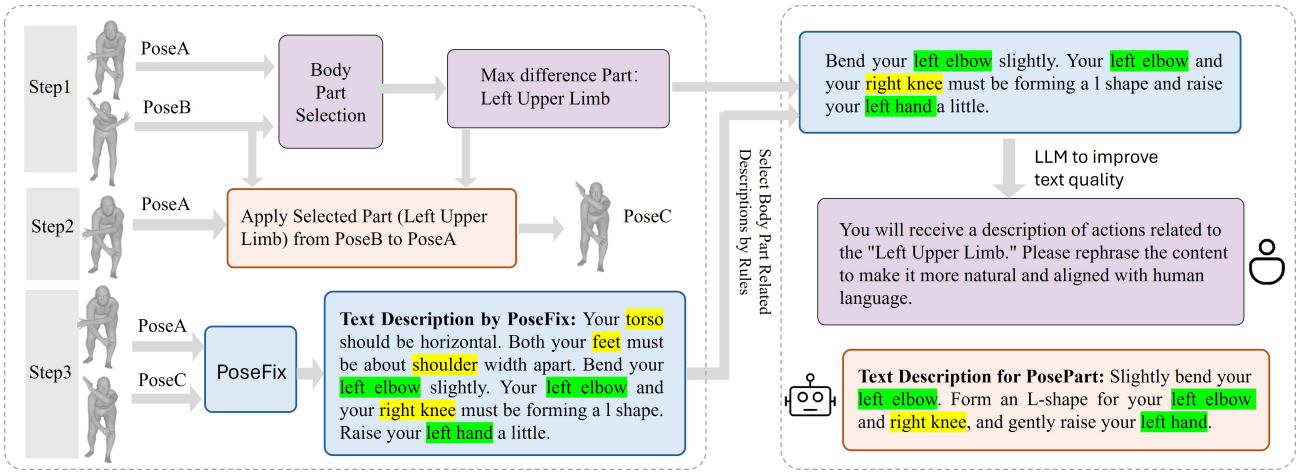

Figure 4: The pipeline for PosePart dataset construction. We highlight redundant text (unrelated to pose changes) in yellow and relevant text in green. Compared to PoseFix, PosePart produces language that is more concise and natural.

Task	Dataset	Annotations
Pose Estimation	Human3.6M	300,000
Pose Generation	PoseScript	100,000
Pose Adjustment	PoseFix	135,000
	PosePart	135,000

Table 1: Datasets for PoseLLaVA training.

Method	Human3.6M		3DPW
	MPJPE ↓	PA-MPJPE ↓	PA-MPJPE ↓
SPIN	60.2	41.7	**52.3**
HRM2.0	**52.1**	**41.0**	58.4
Chatpose	177.8	55.7	79.0
PoseLLava	71.7	51.2	81.0

Table 2: Comparison on Human Pose Estimation.

instruction-following datasets, including Human3.6M (Ionescu et al. 2013), PoseScript (Delmas et al. 2022), PoseFix (Delmas et al. 2023) and the new introduced PosePart. We follow ChatPose (Feng et al. 2024) to use a question-answer template such as: "Question: $\langle image \rangle$ / $\langle pose \rangle$ Kindly review the provided description. $\{description\}$. Given the initial pose of a human, can you predict and adjust the SMPL pose according to the textual description? Sure, the SMPL pose is $\langle POSE \rangle$. All the data used to train our model is listed in Table 1.

We use pre-trained weights from Llava-v1.6-mistral-7b (Liu et al. 2024b), which adopt CLIP-VIT-L/14-336 model for vision encoding and mistral for the LLM. The vision projector and the pose projector of representation alignment are two MLP layers. We employ 8 NVIDIA 40G Tesla A100 GPUs for training. The LLM is tuned by LoRA(Hu et al. 2022) with a rank of 128 and an alpha of 256. The batch size per device is set to 16 with gradient accumulation step of 4 and the training process include 2 epochs in total.

Results on Pose Estimation

For the pose estimation task, we trained our model using the Human3.6M (Ionescu et al. 2013) dataset. Unlike traditional methods such as (Lin et al. 2023), which often rely on extensive data augmentation, our approach does not employ any data augmentation. Instead, we sampled only 300,000 instances from the original data to demonstrate the effectiveness of PoseLLaVA. Following the methodology used in ChatPose (Feng et al. 2024), we randomly selected 200 samples from the Human3.6M (Ionescu et al. 2013) and 3DPW (Von Marcard et al. 2018) test set for evaluation. We utilized two widely recognized metrics: Mean Per-Joint Position Error (MPJPE) and Procrustes-Aligned MPJPE (PA-MPJPE). The former measures the average Euclidean distance between the predicted 3D joints and the ground truth after root alignment, while the latter calculates MPJPE after rigid alignment. For both metrics, lower values indicate higher accuracy.

As shown in Table 2, our method significantly outperforms the recent multimodal LLM ChatPose (Feng et al. 2024) on the Human3.6M dataset and achieves comparable performance on 3DPW, even though ChatPose was trained on much larger datasets. While our MPJPE and PA-MPJPE errors are higher than those of traditional task-specific methods like SPIN (Kolotouros et al. 2019) and HRM (Goel et al. 2023), it's important to note that we use significantly less training data than all baseline methods. For example, PoseChat and HRM employ four different datasets, including Human3.6M (Ionescu et al. 2013), MPI-INF-3DHP (Mehta et al. 2017), COCO (Lin et al. 2014), and MPII (Andriluka et al. 2014), while SPIN is trained on three datasets including Human3.6M (Ionescu et al. 2013), MPI-INF-3DHP (Mehta et al. 2017), and LSP (Johnson and Everingham 2010). Moreover, traditional methods like SPIN and

Methods	PoseFix Dataset					PosePart Dataset				
	MPJPE	PA-MPJPE	R@1	R@5	R@10	MPJPE	PA-MPJPE	R@1	R@5	R@10
PoseFix	130.5	94.09	25%	38%	50%	-	-	-	-	-
ChatPose	180.5	100.7	3%	21%	31%	181.66	107.33	3%	15%	28%
ControlNet	98.7	71.5	-	-	-	70.3	52.7	-	-	-
PoseLLaVA-Image	88.8	63.2	14%	43%	52%	189.43	112.06	12%	41%	60%
PoseLLaVA-Pose	**71.7**	**52.1**	**37%**	**56%**	**63%**	**61.6**	**42.3**	**31%**	**56%**	**64%**

Table 3: Comparison Results on the Pose Adjustment Task. For the MPJPE and PA-MPJPE metrics, lower values indicate better performance. For the R@K recall rates, higher values are better.

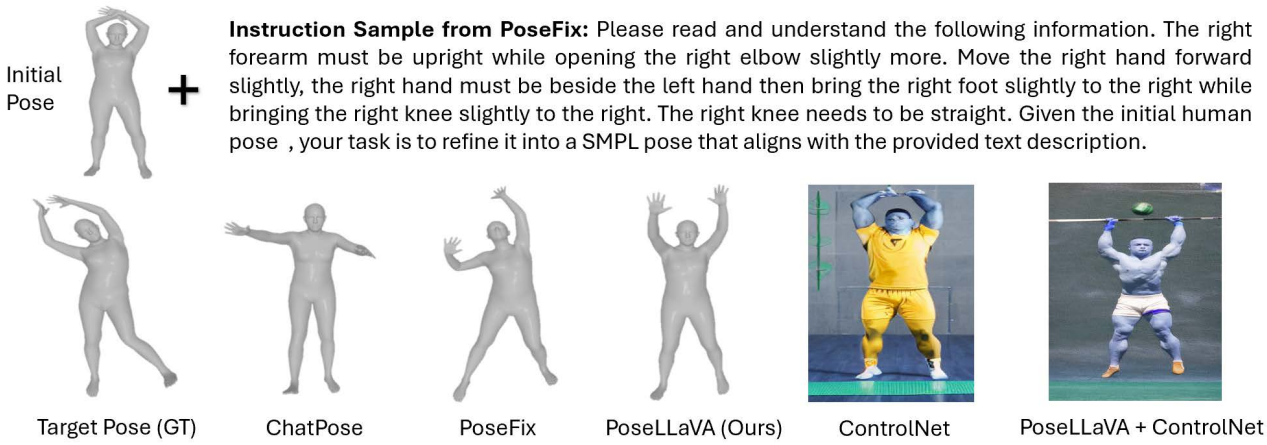

Figure 5: We compared different methods for the pose adjustment task and demonstrated the feasibility of combining our approach with image generation techniques like ControlNet for pose image editing. In this example, our method accurately captures fine-grained instructions such as 'right forearm must be upright while opening the right elbow slightly more.' In contrast, the pose generated directly by ControlNet remains almost identical to the initial pose, failing to adjust the human pose according to the language guidance. By combining PoseLLaVA with ControlNet, we first generate the adjusted pose using PoseLLaVA, then use this pose as a control for image generation, resulting in more precise pose control for image editing.

HRM2.0 focus exclusively on pose estimation, whereas our method handles multiple pose manipulation tasks within a unified framework, leveraging language instructions for improved human-machine interaction.

Results on Pose Adjustment

For the pose adjustment task, we used a training dataset composed of two parts: the PoseFix (Delmas et al. 2023) dataset and the newly introduced PosePart dataset. The PoseFix dataset is generated by sampling data from AMASS (Mahmood et al. 2019) and contains only pairs of SMPL poses and corresponding language descriptions, without the original images. To accommodate both pose and image modalities, we used mmhuman3d (Contributors 2021) to render SMPL pose images for both the PoseFix and PosePart datasets. The final dataset for pose adjustment consists of 270,000 samples.

We used multiple metrics, including MPJPE, PA-MPJPE, and retrieval recall rates, to compare our method with SOTA methods. Table 3 presents the comparison results. Our approach surpasses all existing SOTA methods, including the GAN-based PoseFix (Delmas et al. 2023), which is specifically designed for this task, and the recent multimodal LLM method PoseChat (Feng et al. 2024). Notably, we achieved a 45% reduction in MPJPE error compared to PoseFix and a 60% reduction compared to ChatPose, demonstrating significant improvements over the current SOTA methods.

We also compared our method with the diffusion-based approach ControlNet (Zhang, Rao, and Agrawala 2023), a SOTA method for image generation conditioned on language and other inputs. For the pose adjustment task, where the input is <PoseA, PoseC, $Text_{modifer}$ >, we provided PoseA and the language description to ControlNet to generate an image with the adjusted pose. We then extracted the SMPL pose from the generated image using SPIN (Kolotouros et al. 2019) and SMPLify (Bogo et al. 2016) for quantitative comparison. The results, as illustrated in Figure 5, show that using ControlNet alone does not yield satisfactory outcomes, highlighting the challenges of language-guided pose image manipulation. However, our PoseLLaVA method can be combined with image generation methods to achieve better results. Specifically, we first use PoseLLaVA to generate the adjusted pose and then employ ControlNet again for pose-guided image generation, achieving

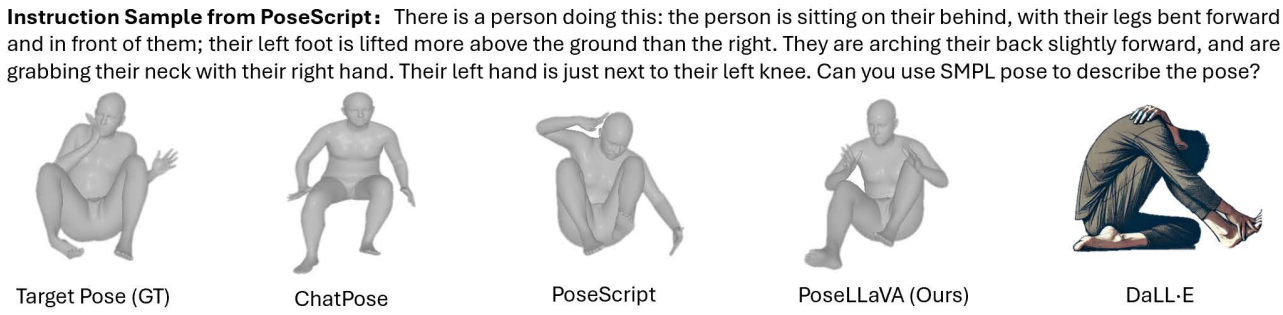

Figure 6: We compared various methods for language-guided pose generation, including both pose generation approaches and one image generation method. In this example, the language instruction describes a fine-grained pose involving multiple body parts, specifying details such as the positioning of the legs, arms, and hands. Our method accurately captures multiple fine-grained details, such as 'legs bend,' 'left foot lifted higher above the ground than the right,' and 'left hand next to the left knee'.

language-guided fine-grained pose adjustment. These results underscore the potential of our method for both language-guided pose editing and pose image editing tasks.

Additionally, we evaluated the feature representation capability by reporting the pose adjustment recall rate in a retrieval task. Specifically, after generating the adjusted pose, we used it to retrieve the corresponding true adjusted pose, leveraging the embedding models of each method. We sampled 100 instances for this retrieval task and reported the R@K recall rates, representing top-K retrieval performance. Higher values of R@K signify better performance. Across all metrics, our method consistently achieved the best results, demonstrating significantly superior performance.

Results on Pose Generation

Method	MPJPE	PA-MPJPE	R@1	R@5	R@10
PoseScript	214.3	145.1	9%	20%	24%
Chatpose	237.1	139.9	8%	14%	22%
PoseLLava	**169.72**	**111.03**	**11%**	**23%**	**30%**

Table 4: Comparison on Human Pose Generation Task.

For the pose generation task, we used the PoseScript (Delmas et al. 2024) dataset, which includes textual descriptions for 100,000 diverse human poses sourced from the AMASS dataset. We followed the dataset splits used in PoseScript and PoseChat to create training and evaluation sets. We compared our method against state-of-the-art approaches using the same metrics as in the pose adjustment task. For the MPJPE and PA-MPJPE metrics, we compare the predicted SMPL pose with the target SMPL pose. For the retrieval task, we use the generated pose to retrieve the target pose by comparing the embedding similarity between the two pose sets. We sampled 100 instances for this retrieval task.

As shown in Table 4, our method outperforms all existing SOTA methods across all metrics, including the VAE-based PoseScript (Delmas et al. 2024) and the MLLM method PoseChat (Feng et al. 2024). Visual examples are provided in Figure 6, along with results from Dall·e3 (OpenAI 2023) for image generation.

Effectiveness of the PosePart Dataset

	PoseFix		PosePart	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
PoseFix Only	74.6	**51.9**	65.5	45.9
PosePart Only	83.5	57.2	62.9	44.8
Both DB	**71.7**	52.1	**61.6**	**42.3**

Table 5: Training on Different Dataset.

We evaluated the impact of the PosePart dataset, specifically designed to facilitate fine-grained pose adjustments by focusing on single body part modifications. We trained PoseLLaVA using PoseFix, PosePart, and a combination of both, then compared their performance on the pose adjustment task. The results, summarized in Table 5, show that the model trained solely on PoseFix performs better on PoseFix, while the model trained solely on PosePart excels on PosePart. When combining the two datasets, the model achieves strong performance on both.

Conclusion

In this paper, we introduced PoseLLaVA, a multimodal large language model that integrates human pose as a distinct modality alongside images and text. PoseLLaVA is capable of performing pose estimation, adjustment, and generation guided by textual instructions. With a pose centric model structure, a fine-grained pose-language dataset, and a sophisticated pose encoder-decoder architecture, PoseLLaVA sets a new benchmark for multimodal LLMs in pose manipulation tasks. Our integration with diffusion models also highlights its potential for both pose manipulation and pose image editing.

Limitations: Our dataset mainly focuses on single-person poses with simple backgrounds due to limited public datasets and the high cost of creating more complex ones. Future work could involve collecting data in more complex scenarios, such as multi-person interactions in cluttered scenes, to further improve the model's capabilities.

References

Andriluka, M.; Pishchulin, L.; Gehler, P.; and Schiele, B. 2014. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, 3686–3693.

Bhunia, A. K.; Khan, S.; Cholakkal, H.; Anwer, R. M.; Laaksonen, J.; Shah, M.; and Khan, F. S. 2023. Person image synthesis via denoising diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5968–5976.

Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.; Romero, J.; and Black, M. J. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, 561–578. Springer.

Chen, L.-H.; Lu, S.; Zeng, A.; Zhang, H.; Wang, B.; Zhang, R.; and Zhang, L. 2024. MotionLLM: Understanding Human Behaviors from Human Motions and Videos. arXiv:2405.20340.

Choutas, V.; Müller, L.; Huang, C.-H. P.; Tang, S.; Tzionas, D.; and Black, M. J. 2022. Accurate 3D body shape regression using metric and semantic attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2718–2728.

Contributors, M. 2021. OpenMMLab 3D Human Parametric Model Toolbox and Benchmark. https://github.com/open-mmlab/mmhuman3d. Accessed: 2025-01-06.

Delmas, G.; Weinzaepfel, P.; Lucas, T.; Moreno-Noguer, F.; and Rogez, G. 2022. Posescript: 3d human poses from natural language. In *European Conference on Computer Vision*, 346–362. Springer.

Delmas, G.; Weinzaepfel, P.; Lucas, T.; Moreno-Noguer, F.; and Rogez, G. 2024. PoseScript: Linking 3D Human Poses and Natural Language. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Delmas, G.; Weinzaepfel, P.; Moreno-Noguer, F.; and Rogez, G. 2023. Posefix: correcting 3D human poses with natural language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15018–15028.

Feng, Y.; Lin, J.; Dwivedi, S. K.; Sun, Y.; Patel, P.; and Black, M. J. 2024. Chatpose: Chatting about 3d human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2093–2103.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; et al. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793.

Goel, S.; Pavlakos, G.; Rajasegaran, J.; Kanazawa, A.; and Malik, J. 2023. Humans in 4D: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14783–14794.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7): 1325–1339.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. arXiv:2310.06825.

Johnson, S.; and Everingham, M. 2010. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In *British Machine Vision Conference*.

Joo, H.; Neverova, N.; and Vedaldi, A. 2021. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *2021 International Conference on 3D Vision (3DV)*, 42–52. IEEE.

Kolotouros, N.; Pavlakos, G.; Black, M. J.; and Daniilidis, K. 2019. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2252–2261.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9579–9589.

Lin, J.; Zeng, A.; Wang, H.; Zhang, L.; and Li, Y. 2023. One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21159–21168.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024b. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024c. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5442–5451.

Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; and Theobalt, C. 2017. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, 506–516. IEEE.

OpenAI. 2023. Dall·e 3 system card. https://cdn.openai.com/papers/DALL_E_3_System_Card.pdf. Accessed: 2025-01-06.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Von Marcard, T.; Henschel, R.; Black, M. J.; Rosenhahn, B.; and Pons-Moll, G. 2018. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, 601–617.

Wu, S.; Fei, H.; Qu, L.; Ji, W.; and Chua, T.-S. 2024. NExT-GPT: Any-to-Any Multimodal LLM. arXiv:2309.05519.

Yenphraphai, J.; Pan, X.; Liu, S.; Panozzo, D.; and Xie, S. 2024. Image sculpting: Precise object editing with 3d geometry control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4241–4251.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11975–11986.

Zhang, D.; Li, S.; Zhang, X.; Zhan, J.; Wang, P.; Zhou, Y.; and Qiu, X. 2023a. SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. arXiv:2305.11000.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. arXiv:2306.02858.

Zhang, J.; Zhang, Y.; Cun, X.; Zhang, Y.; Zhao, H.; Lu, H.; Shen, X.; and Shan, Y. 2023b. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14730–14740.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2024. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, Y.; Barnes, C.; Lu, J.; Yang, J.; and Li, H. 2019. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5745–5753.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv:2304.10592.