ChatHuman: Chatting about 3D Humans with Tools

Jing Lin[3,4,⋆] Yao Feng[2,3,⋆] Weiyang Liu[1,5] Michael J. Black[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen [2]Stanford University
[3]Meshcapade [4]Tsinghua University [5]University of Cambridge
chathuman.github.io

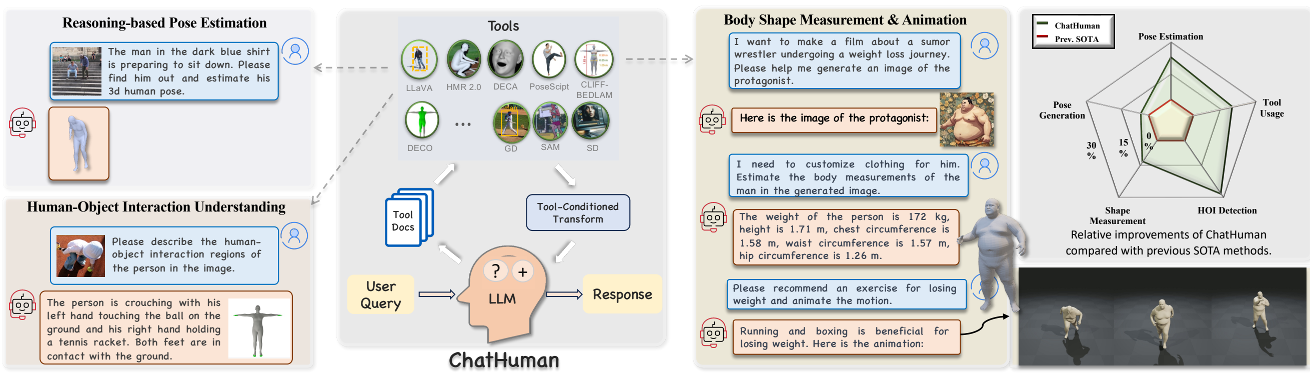

Figure 1. ChatHuman is a LLM-based agent that uses a multimodal LLM to exploit and combine tools, discriminate their results, and integrate the results to solve tasks related to 3D Humans.

Abstract

Numerous methods have been proposed to detect, estimate, and analyze properties of people in images, including 3D pose, shape, contact, human-object interaction, and emotion. While widely applicable in vision and other areas, such methods require expert knowledge to select, use, and interpret the results. To address this, we introduce ChatHuman, a language-driven system that integrates the capabilities of specialized methods into a unified framework. ChatHuman functions as an assistant proficient in utilizing, analyzing, and interacting with tools specific to 3D human tasks, adeptly discussing and resolving related challenges. Built on a Large Language Model (LLM) framework, ChatHuman is trained to autonomously select, apply, and interpret a diverse set of tools in response to user inputs. Our approach overcomes significant hurdles in adapting LLMs to 3D human tasks, including the need for domain-specific knowledge and the ability to interpret complex 3D outputs. The innovations of ChatHuman include leveraging academic publications to instruct the LLM on tool usage, employing a retrieval-augmented generation model to create in-context learning examples for managing new tools, and effectively discriminating between and integrating tool results by transforming specialized 3D outputs into comprehensible formats. Experiments demonstrate that ChatHuman surpasses existing models in both tool selection accuracy and overall performance across various 3D human tasks, and it supports interactive chatting with users. ChatHuman represents a significant step toward consolidating diverse analytical methods into a unified, robust system for 3D human tasks. Code and data are available at chathuman.github.io.

1. Introduction

Research on 3D humans has progressed rapidly, resulting in the creation of many tools that can perform tasks like estimating a human's 3D pose from a single image [18, 25, 27, 36, 37], predicting face/body shapes [8, 15], capturing emotions [9, 15], and identifying regions of touch/contact [46, 68], generating human poses from text descriptions [10], and animating human images [84]. Each of these tools, however, focuses on a specific problem, functioning as isolated "specialists". Moreover, these separate tools cannot benefit from the expertise of others, and combining them to solve more complex tasks requires significant domain expertise. Ideally, we would have a single model that can adaptively leverage different tools to solve complex 3D human-related problems while offering intuitive user interaction through natural language input. Recent work such as ChatPose [16] has taken initial steps in this direction, unify-

⋆Equal contribution. This work was done while YF and JL were at Meshcapade.

ing pose generation, estimation, and general understanding within an LLM framework. Unfortunately, ChatPose lacks the accuracy of the best specialist methods.

To address these issues, we build a multi-modal LLM, called ChatHuman, that specializes in using digital human modeling tools, enabling it to autonomously interpret instructions and complete diverse tasks related to 3D humans; see Fig. 1. Specifically, we teach an LLM to use a wide range of specialized human-related models for tasks like 3D pose estimation, emotion recognition, contact reasoning, and more, effectively extending the LLM's capabilities to the domain of 3D humans. This goes beyond providing a natural-language interface to these tools, as the LLM can use its broad understanding of humans to augment tool results or to analyze and integrate their outputs, providing better responses than any single tool alone.

With ChatHuman, we introduce a novel approach by fine-tuning an LLM to act as an agent that autonomously calls appropriate tools based on user inputs, completing tasks and enhancing responses with tool-generated results. Similar in spirit, recent works have employed off-the-shelf or fine-tuned LLMs for tasks like basic vision (e.g., Visual ChatGPT [53]), mobile applications (e.g., AppAgent [77]), biology (e.g., AmadeusGPT [78]) and system automation (e.g., GPT4Tools [74]). Our work, however, differs by focusing specifically on the unique challenges of 3D human understanding. This domain requires precise, specialized terminology and a nuanced understanding of 3D-specific tools, which conventional LLMs lack. To teach the network this specialized terminology, we do what we would do as humans – we have the LLM read the papers describing the methods. Even with that knowledge, the LLM needs to understand the task goals, select an appropriate tool or tools, interpret results, and resolve differing results. These skills are all beyond the abilities of general LLMs.

To address these challenges, we design the following training pipeline: 1) We utilize relevant literature about the tools to familiarize the LLM with domain knowledge, helping it know when and how to use these tools; 2) After using a tool, the LLM evaluates the reliability of the outcome using its "judgment" and compares different methods to identify the most reliable results; 3) It combines these results with its general knowledge to create response. This pipeline represents several key innovations, laying a foundation for LLMs to effectively handle complex, tool-driven 3D human tasks.

Retrieval-Augmented Tool Use: Details about tools are typically present in corresponding academic paper. We give the LLM access to these papers and demonstrate that "reading the paper" improves tool use performance. We further analyze which paper sections are most effective for instructing tool use. When encountering a new tool, users often turn to the user guide for assistance. We compile the complete documentation for these tools and utilize a paper-based Retrieval-Augmented Generation (RAG) mechanism to improve the LLM's understanding and management of new tools. This means that, although the LLM has not encountered such tools during fine-tuning, it can still effectively use the tools with the help of the paper-based RAG mechanism. In some cases, tasks require combining multiple tools. To address a broader range of tool usage scenarios, we employ a graph-based invocation system, which includes a node for single-tool use, a chain for sequential tool execution, and a DAG for multi-tool combinations as shown in Fig. 3.

3D Human-Related Tool Result Integration: Analyzing outputs from tools is crucial, as these outputs, such as body meshes, model parameters (e.g., SMPL pose), or motion sequences, are highly varied and complex. To make these results compatible with our LLM analysis system, we convert them into visual formats that the LLM can easily interpret. Guided by Cognitive Load Theory [63], we present these outputs as multiple-choice options, streamlining the selection process and enhancing the LLM's effectiveness in handling 3D human-related tasks. Combined with the LLM's extensive general knowledge, these integrated results enable it to generate sophisticated responses about 3D humans.

Specifically, ChatHuman consists of a multimodal LLM LLaVA [41], and 26 tools involving 3D Humans and general vision tasks. The LLM is finetuned to use these tools and incorporate their results. User requests can be in the form of text descriptions, images or other 3D information (if applicable), and the model produces text descriptions, images, or other 3D outputs after tool reasoning. Extensive evaluations demonstrate that ChatHuman not only surpasses previous models in tool-use accuracy but also improves performance on various human-related tasks. It achieves this by reasoning about multiple outputs, evaluating their veracity, and combining them with its own knowledge.

Summarizing, our key contributions include: (1) a framework that leverages LLMs to interact with users and address human-centric tasks using specialist tools; (2) a scientific-paper-based RAG mechanism that ensures precise tool use by comprehending tool descriptions from scholarly articles, enhancing tool applications and interactions; and (3) the integration of tool outcomes with LLMs, enabling the LLM to effectively explain tool results and interact with users. Additionally, the LLM is fine-tuned to distinguish between optimal and suboptimal tool results, improving overall accuracy. ChatHuman achieves superior performance in tool use and human-related tasks compared with other LLM-based methods or task-specific methods. The code, trained models, and datasets are available for research purposes.

2. Related work

3D Humans: Many 3D human analysis tasks leverage parametric models like SMPL [43], SMPL-X [48], or GHUM [73] for the body, BFM [49] or FLAME [35] for faces, and MANO [55] for hands. These models enable the representation of the human body, face, and hands as low-dimensional vectors, facilitating subsequent applications in estimation and generation. Estimation of human pose and shape either relies on optimization methods [4, 24] or regression methods [8, 14, 18, 25, 27, 32, 36, 37, 56, 81], which estimate body shape and pose parameters from a single image. Similarly, face reconstruction methods [12, 15, 66] estimate shape and expression parameters of the face model from single images. The analysis of contact, vital for understanding human-environment interaction and social touch, has seen recent attention [19, 46, 68]. Generative modeling techniques such as PoseScript [10] and PoseFix [11] provide methods for synthesizing and correcting 3D human poses based on textual descriptions, while language-to-3D generation methods [5, 20, 82] facilitate 3D avatar creation. Additionally, numerous methods generate human motions [1, 23, 33, 34, 45, 50, 64, 65], and recent language-to-video models are even able to generate humans moving [3, 84].

These basic methods excel in their respective scenarios, but are typically treated in isolation. When mature, such tools are often incorporated into software systems for animators that require significant domain knowledge. In contrast, recent generative models provide language interfaces to image, video, and 3D generation tools, making them accessible to novices. Until recently, such language-based control has not been possible for 3D humans. ChatPose [16] makes a step in this direction, unifying pose generation, estimation, and an LLM's general understanding into one model, but remains limited in its task capabilities. In contrast, our model integrates the performance of 26 3D human-related tasks into a single, LLM-based model. ChatHuman enables non-experts to solve real-world tasks by invoking appropriate tools and adding an extra layer of language-driven understanding that effectively leverages the tool outputs.

Large Language Models and Tool Use: To expand LLM capabilities without expensive retraining, recent work has focused on enabling them to use specialized tools. In this approach, a tool library is constructed and LLMs act as planners to coordinate tool usage. Various tools have been adopted, *e.g.*, vision modules [53, 62, 76], mobile applications [77], community tools [58], special tools [79], and system tools [74]. However, general-purpose LLMs often lack a deep understanding of specific tools, especially those requiring domain knowledge. To address this, recent work [13, 28, 71] proposes to fintune general-purpose LLMs (*e.g.*, LLaMA [67], LLaVA [40, 42]) with domain-specific tool-use data. Some methods use additional tool documentation to improve accuracy [21, 80], while others compose different tools to accomplish complex tasks [29, 57, 72] . Distinct from previous work, ChatHuman focuses on 3D human tasks through language interaction by leveraging off-the-shelf human-related tools.

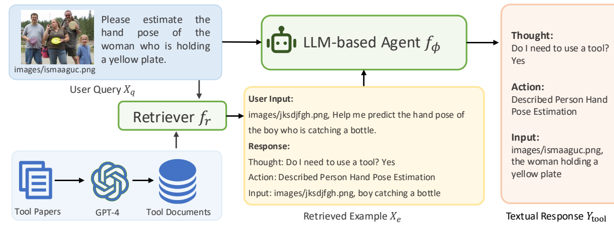

Figure 2. Paper-based Retrieval-Augmented Tool Use. We feed academic papers describing each tool to GPT-4 to build a document for each tool. During inference, given a user query, a relevant sample is retrieved from the documents and provided to the LLM-based agent as an in-context example to improve tool use accuracy.

Retrieval Augmented Generation: RAG [17, 31, 83] is a technique to enhance generative tasks by retrieving relevant information from external databases, allowing for continual knowledge update. Here, we design a RAG mechanism to facilitate the use of new unseen tools.

3. Method

3.1. Overall Pipeline

ChatHuman consists of a multimodal LLM $f_\phi(\cdot)$, along with a set of 3D human-related functions $f_{\theta_1}(\cdot), f_{\theta_2}(\cdot), \ldots$. These functions serve as tools for various tasks, such as 3D human pose estimation, pose generation, and 3D face reconstruction. Our model takes input text queries X_q, images X_v, and optionally X_m representing other 3D human-related modalities (e.g., SMPL parameters for 3D human poses). Then it invokes tools and integrates their results to generate outputs as text Y_t, images Y_v, or 3D human-related modalities Y_m.

3.2. Retrieval-Augmented Tool Usage

Teaching LLMs to decide when and how to use tools effectively is a significant challenge. A basic approach [53, 74] might involve including tool usage scenarios and input arguments within the LLM prompt, represented as $Y_{\text{tool}} = f_\phi(X_q, X_t)$, where X_t denotes tool definitions. However, this approach often falls short for specialized tools, especially given the variety of advanced tools for 3D human tasks. Many tools require background knowledge for correct use and have multiple application scenarios. For instance, the HMR tool [18] may be queried with requests like, "Can you estimate this person's pose?", "What are the SMPL parameters?", or "Provide the 3D mesh of this person." Capturing all possible usage scenarios succinctly in a prompt is difficult, and as tools proliferate, prompt descriptions become unwieldy. To address these challenges, we introduce paper-based Retrieval-Augmented Generation (RAG) [30] and build a tool graph for tool combination. As shown in Fig. 2, we feed academic papers associated with each tool into GPT-4, prompting it to summarize the tool's functions and generate possible user queries for tool activation. These

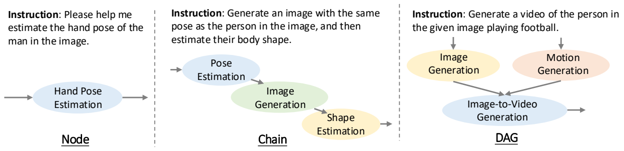

Figure 3. Examples of tool use graphs. Tool use patterns include: a Node, which uses a single tool; a Chain, which requires sequential tool execution; and a DAG, which combines multiple tools.

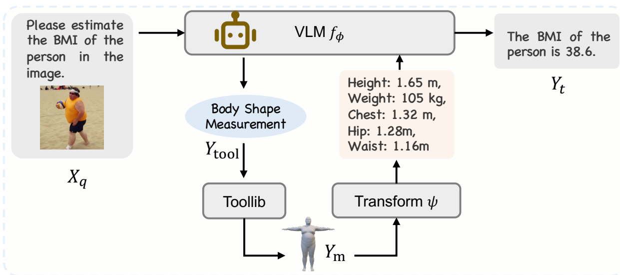

Figure 4. Illustration of tool result integration. After tool invocation, results are transformed into VLM-compatible representations. The VLM then incorporates these transformed outputs to generate accurate, informative responses to user queries.

papers, with their rich background and detailed instructions, enable the generation of user queries that cover diverse use cases. By combining these queries with each tool's structured arguments, we compile a document of question-answer pairs for each tool's operation. Figure 2 provides an example from one of these documents. These documents serve as an auxiliary knowledge base during inference, from which we retrieve a relevant example X_e in response to a user query X_q. The retrieval process matches the text embedding of the query with embeddings in the tool documents using a text embedding model [61]. The retrieved sample is then presented to the agent f_ϕ as an in-context learning example:

$$X_e = f_r(X_q), \;\; Y_{\text{tool}} = f_\phi(X_q, X_e, X_t), \tag{1}$$

where f_r is the retrieval function, and Y_{tool} is a textual description of the tool invocation, specifying tool selection, names, and input arguments for tool calls.

Graph-based Tool Invocation. Note that the tool use description Y_{tool} varies depending on task settings, as shown for a single tool case in Fig. 2. However, some complex tasks require combining multiple tools. To handle this, we introduce a graph-based mechanism for tool invocation. We construct a tool graph with three structure types: nodes (single tool calls), chains (tool sequences for dependent tasks), and directed acyclic graphs (DAGs) [59] for complex multi-branch operations. For each user query, the model predicts an appropriate tool graph and invokes the tools accordingly. Examples of tool graphs are shown in Fig. 3.

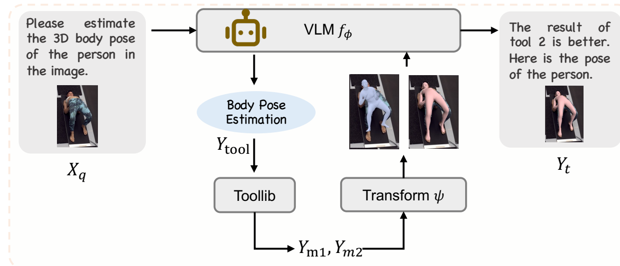

Figure 5. Illustration of tool results discrimination. When multiple plausible tools exist for a task, ChatHuman discriminates and chooses the best result as the final response.

3.3. Tool Result Integration

After using tools, integrating their results is essential to effectively engage with users and solve problems. However, outputs from different tools vary widely, appearing as language, images, or vectors (like SMPL poses), which can challenge current multimodal LLMs, such as LLaVA [41], that process only text and images. To utilize these varied results and enhance the LLM's understanding of 3D humans, thereby improving its ability to apply world knowledge to problem-solving, we introduce a tool-conditioned transformation, $\Psi(\cdot)$. As shown in Fig. 4, this transformation converts tool outputs Y_m into textual or visual formats the LLM can process. For example, we transform the vertex-wise contact label from DECO [68] into body part-level descriptions using SMPL's [43] vertex-to-part mapping dictionary, and render the mesh generated by PoseScipt [10] into an RGB image using rendering techniques. See Sup. Mat. for more details. The transformed results are then merged with the user query as context for response generation:

$$Y_t = f_\phi(X_q, \Psi(Y_m)). \tag{2}$$

In scenarios where multiple tools can address a request (Fig. 5), we present outcomes as multiple-choice questions, prompting the model to select the most relevant answer:

$$Y_t = f_\phi(X_q, \Psi(Y_{m1}), \Psi(Y_{m2}), ...), \tag{3}$$

where Y_{mi} denotes the i-th tool result. Since different tools have different failure modes, this process enables ChatHuman to identify the best method case by case, producing more accurate output than any individual method alone.

3.4. Training Data Construction

Tool Usage Instruction-following Data. To teach the LLM-based agent to correctly use tools, we construct 90K instruction-response pairs about tool usage. Following GPT4Tools [74], we provide GPT-4 [47] with a textual description of COCO training images [39] and a tool-related prompt containing a tool description. To improve efficiency,

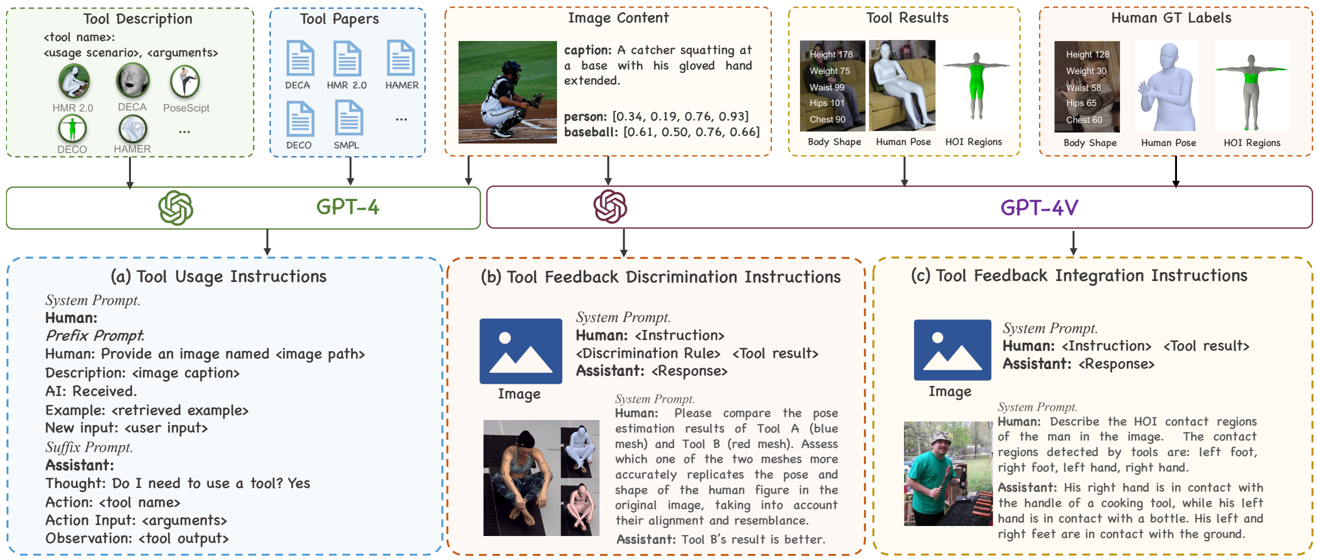

Figure 6. Illustration of our instruction-following data construction pipeline. We construct tool usage and feedback data by providing GPT-4 with various tool-related information, image content, and ground truth labels. Gray text shows some example instructions.

we first prompt GPT-4 to summarize paper content, re-articulate tool functions, and enumerate 50 potential user queries for tool activation (see Fig. 6(a)).

Tool Feedback Instruction-following Data. To help the multimodal LLM model discriminate and integrate the tool results, we construct 88K pairs of instruction-following data based on existing datasets 3DPW [70], MOYO [69], PoseScript [10] and SHAPY [8] (see Fig. 6(b)(c)). Please see Sup. Mat. for more details about data construction.

3.5. Model Training

Once we have data, we use LoRA [22] to finetune the LLM $f_\phi(\cdot)$ with the cross entropy loss. More specifically, with the ground truth tool invocation labels $\hat{Y}_{tool}$ and response label $\hat{Y}_t$, we optimize the model using the following objective function: $\mathcal{L} = \textbf{CE}(\hat{Y}_{\text{tool}}, Y_{\text{tool}}) + \textbf{CE}(\hat{Y}_t, Y_t)$, where **CE** denotes the cross-entropy loss. See Sup. Mat. for details.

4. Experiments and Results

4.1. Implementation Details

We use LLaVA-1.5 [41] as the VLM backbone, with CLIP [51] for vision encoding and Vicuna [7] for the LLM backbone. For retrieval, we adopt INSTRUCTOR [61] for text embedding and utilize Chroma's vector similarity searching algorithm to identify relevant examples. To preserve the generalization of the pretrained multi-modal LLM, we use LoRA [22] to perform efficient finetuning, with rank 128 and alpha 256. We implement tool utilization with LangChain [6], which enables automatic parsing of tool names and input parameters, followed by tool invocation Optimization uses AdamW [44], with a learning rate of 2e-4 and weight decay of 0. All models are finetuned over 2 epochs with a mixture of tool usage, tool feedback, and LLaVA multimodal instruction-tuning data, using 8 Nvidia A100-80G GPUs with the DeepSpeed [52] engine. Unless otherwise specified, we use LLaVA-1.5-7B as the base model for the ablation study. We support 26 tools, as listed in Tab. 1.

Perception	**Reasoning**	**Generation**
Body Pose Estimation [18]	Selective Person Pose Detection [18, 41]	Text-to-Pose Generation [10]
Body Shape Measurement [2]	Specific Person Shape Measurement [2, 41]	Speculative Pose Generation [10, 41]
Hand Pose Estimation [38]	Targeted Hand Pose Estimation [38, 41]	Text-to-Image Generation [54]
Face Reconstruction [15]	Described Person Face Reconstruction [15, 41]	Text-based Pose Editing [11]
Human Segmentation [26]	Described Person Segmentation [26, 41]	Remove Someone From The Photo [26, 41, 54]
HOI Detection [68]	Selective Person Contact Estimation [41, 68]	Replace Someone From The Photo [26, 41, 54]
Pose Description [10]	Visual Question Answering [41]	Instruct Image Using Text [54]
Image Caption [41]		Text-to-Motion Generation [50]
Motion Capture [60]		Text-to-Video Generation [50, 54, 84]
		Image-to-Video Generation [50, 84]

Table 1. ChatHuman supports 26 human-related tools, including 9 perception tools, 10 generation tools, and 7 reasoning tools. Tools in grey are unseen tools that are not included in the training data.

	Seen Tools					Unseen Tools				
Method	SR_t	SR_{act}	SR_{args}	SR	IoU	SR_t	SR_{act}	SR_{args}	SR	IoU
GPT4Tools [75]	0.609	0.547	0.525	0.520	0.566	0.612	0.546	0.542	0.525	0.573
GPT4Tools-FT [75]	0.825	0.710	0.687	0.690	0.741	0.904	0.807	0.690	0.747	0.800
Visual ChatGPT-3.5 [53]	0.498	0.319	0.237	0.251	0.791	0.507	0.314	0.226	0.293	0.803
Visual ChatGPT-4 [53]	0.892	0.802	0.715	0.753	0.797	0.998	0.913	0.801	0.872	0.907
ChatHuman	**1.000**	**0.974**	**0.950**	**0.970**	**0.975**	**0.999**	**0.967**	**0.893**	**0.954**	**0.953**

Table 2. Tool use accuracy comparison. Successful rate of thought (SR_t), action (SR_{act}), arguments (SR_{args}), execution (SR), and IoU are reported.

4.2. Evaluation on Tool Usage

Tool Usage Benchmark. To evaluate tool usage accuracy, we construct a validation and test set. The validation set has 1000 samples with the same tools as the training set, while the test set includes 689 samples related to 3 tools unseen during training. Split of seen and unseen tools are detailed in Table 1. Similar to our training data construction, we feed a textual description of COCO validation set image, a

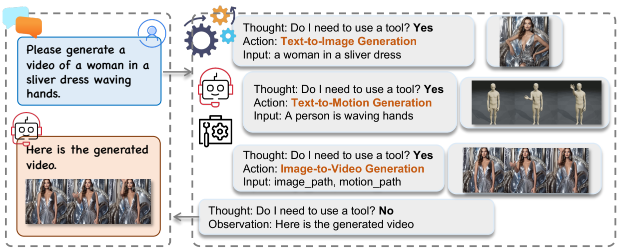

Figure 7. Visualization of Animation Processing. Left: Conversation between the user and ChatHuman. Right: ChatHuman automatically calls tools to solve the task. ChatHuman can handle tasks beyond the capabilities of individual tools.

tool description, and some examples summarized from the tool paper into GPT-4 and prompt it to generate instruction-following data. We use the image description captioned by LLaVA [41] instead of the original image captions to ensure differences between training and test sets. Finally, we manually check all question-answering pairs for accuracy.

Baselines. We compare our method with Visual ChatGPT [53] and GPT4Tools [75] on the proposed evaluation set and report 5 metrics proposed in GPT4Tools [75]. See Sup. Mat. for details of the metrics. For Visual ChatGPT, we experiment with two versions of GPT: "gpt-3.5-turbo-1106" and "gpt-4-turbo-preview". For GPT4Tools, we adopt the official pretrained 13B model. For a fair comparison, we also finetune GPT4Tools with our training data using the official training code and obtain a variant, GPT4Tools-FT.

Table 2 shows that the original GPT4Tools struggles on our benchmark due to mismatches between its tools and our human-centric ones. Visual ChatGPT-4 exhibits impressive tool use accuracy, showing its powerful zero-shot ability to follow a standardized format and use tools accurately. ChatHuman consistently outperforms other methods, particularly with tools not seen in training, thanks to our paper-based RAG mechanism.

4.3. Evaluation on 3D Human Related Tasks

Character Animation. ChatHuman employs tools for text-to-motion and image-to-video generation. We demonstrate how these tools are utilized to interact with users and reason about motions based on conversations in Fig. 7 and Fig. 1. ChatHuman can also tackle tasks that cannot be resolved with a single tool. For instance, text-to-human video generation poses significant challenges due to the complexity of motion. Therefore, another option is to first generate a motion sequence via text-to-motion generation, then apply a video generation model conditioned on this sequence. The internal processing within ChatHuman, detailing how it analyzes and solves tasks, is visualized in Fig. 7. We also compare our text-to-video generation results with those of Pika[1]. The qualitative comparisons are shown in Fig. 8.

[1]We use the demo available at https://pika.art/ (as of May 2025) to obtain the results.

Method	3DPW [70]			RPE Benchmark [16]		
	MPJPE ↓	PA-MPJPE ↓	MPJRE ↓	MPJPE ↓	PA-MPJPE ↓	MPJRE ↓
SPIN [27]	102.9	62.9	10.1	244.9	107.3	12.4
HMR 2.0 [18]	**91.0**	**58.4**	**9.2**	225.2	105.1	12.1
LLaVA-S [41]	440.8	205.4	21.8	490.7	207.4	21.1
LLaVA*-S [41]	232.1	101.1	12.8	-	-	-
GPT4-S [47]	322.0	136.7	16.0	-	-	-
LLaVA-P [41]	335.2	172.3	16.5	391.5	191.9	17.8
GPT4-P [47]	396.5	203.4	18.6	-	-	-
ChatPose [16]	163.6	81.9	10.4	253.6	103.8	11.7
ChatHuman	91.3	58.7	**9.2**	**147.2**	**79.1**	**10.3**

Table 3. Comparison of vanilla human pose estimation and reasoning-based pose estimation on 3DPW [70] and RPE [16]. LLaVA* is fine-tuned with human keypoint data. "S" uses multimodal LLMs for keypoint detection and SMPLify [4] for pose optimization. "P" utilizes multimodal LLMs for textual pose descriptions, processed by PoseScript [10] to generate poses. MPJPE (in mm), MP-MPJPE (in mm), and MPJRE ($\times 100$) are reported.

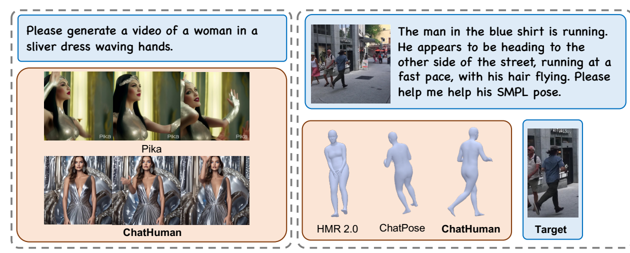

Figure 8. Left: Comparison to Pika[1] on text to video generation. Right: Qualitative comparison with ChatPose [16] and HMR 2.0 [18] on reasoning-based human pose estimation task.

Pose Estimation. Following ChatPose [16], we evaluate the performance of our method on both classical and reasoning-based pose estimation (RPE) tasks. MPJPE, PA-MPJPE, and MPJRE on the 3DPW [70] and RPE [16] benchmarks are reported. For the reasoning-based pose estimation task, ChatHuman first grounds a human based on a textual description and feeds it into the pose estimation tool to get the result. As shown in Table 3, ChatHuman achieves comparable performance to the task-specific models on the classical pose estimation task. For reasoning-based human pose estimation, which involves both reasoning ability and advanced human pose estimation ability, ChatHuman outperforms both task-specific and multi-modal LLM methods by a large margin ($34.6\% \downarrow$ in MPVPE). As shown in Fig. 8, only our method achieves a satisfactory result. The multimodal LLM competitor ChatPose finds the correct person but fails to obtain an accurate pose due to its limited perception ability, while the task-specific tool does not match the correct person due to the lack of reasoning ability. This demonstrates the advantage of ChatHuman, which combines task-specific tool use expertise with the general reasoning ability of an LLM.

Pose Generation. Here we evaluate the pose generation capability of ChatHuman on the classical text-to-pose generation task and the speculative pose generation task (SPG) [16]. Following previous work [10, 16], we report the text-to-pose recall rate R^{T2P} and pose-to-text recall rate R^{P2T}

Method	PoseScript [10]						SPG Benchmark [16]					
	$R^{P2T}\uparrow$			$R^{T2P}\uparrow$			$R^{P2T}\uparrow$			$R^{T2P}\uparrow$		
PoseScript [10]	40.4	52.3	65.0	41.4	**54.1**	65.9	1.5	3.5	6.2	1.4	2.3	5.1
ChatPose [16]	17.6	25.3	35.8	28.0	39.0	54.4	**3.3**	**5.5**	8.2	**3.5**	5.8	**11.0**
LLaVA-P [41]	-	-	-	-	-	-	2.1	4.0	7.1	2.1	3.3	6.1
GPT4-P [47]	-	-	-	-	-	-	2.7	4.7	9.2	2.7	5.3	8.2
ChatHuman	**41.8**	**52.6**	**65.1**	**42.1**	52.3	**66.5**	3.2	5.0	**9.9**	**3.5**	**6.5**	10.6

Table 4. Comparison of classical and speculative pose generation on PoseScript [10] and SPG [16]. "P" denotes using LLMs to rephrase textual pose descriptions, which are then processed by PoseScript [10] to generate poses. Top 5, 10, 20 recall rates are reported.

Method	Height ↓	Chest ↓	Waist ↓	Hip ↓
LLaVA [41]	**6.7**	16.5	22.9	17.6
CLIFF-BEDLAM [2]	7.8	8.6	13.5	7.0
ChatHuman	**6.7**	**6.1**	**13.0**	**6.4**

(a) Comparison of body shape measurement. Measurement errors (in cm) on HBW [8] are reported.

Method	Precision ↑	Recall ↑	F1 Score ↑
LLaVA [41]	0.26	**0.81**	0.39
GPT-4 [47]	0.61	0.48	0.49
ChatHuman	**0.67**	0.67	**0.63**

(b) Comparison of HOI estimation. Precision, Recall, and F1 score on DECO [68] are reported.

Table 5. Comparison of body shape and HOI estimation.

of the retrieval models trained on real poses and evaluated on generated poses. For the SPG task, ChatHuman first rephrases the indirect pose descriptions into explicit ones and adopts PoseScript (journal version) [10] to generate a pose. As shown in Table 4, our method archives comparable performance to the SOTA methods on both benchmarks. In contrast, the previous LLM-based method, ChatPose, performs poorly on the classical pose generation benchmark, while the task-specific model, PoseScript, lags in the SPG benchmark due to limited reasoning ability.

Body Shape Measurement. We evaluate the body shape measurement accuracy of ChatHuman. We randomly sample 100 images from the HBW validation set [8] and compare our method with a multimodal LLM, LLaVA [41], and a SOTA body shape estimation method, CLIFF-BEDLAM [2]. For LLaVA and ChatHuman, we ask them the same question to inquire about the height, weight, chest, waist, and hip circumferences of a person and then prompt GPT-3.5 to extract the value from the model output. The details of the question and prompt are available in Sup. Mat. CLIFF-BEDLAM predicts the body shape parameters, which are then converted to measurements based on the shape-to-measurement function from SHAPY [8]. Anthropometric measurement errors are reported in Table 5(a). ChatHuman achieves superior performance in most measurements, outperforming the multimodal LLM competitor LLaVA by 42% and CLIFF-BEDLAM by 15.7% in average metric accuracy.

Human-Object Interaction (HoI). We evaluate the human-object interaction understanding ability of ChatHuman on the DECO [68] test set. The ground truth (GT) labels are obtained by converting the vertex-level contact labels into body part-level contact labels with SMPL's vertex-to-part mapping dictionary. Given a human image, we ask the multimodal LLM to detect the body parts contacting objects and prompt GPT-3.5 to extract the body part labels from the answer. Subsequently, we compare the predicted body parts with the GT label and compute the average detection precision, recall rate, and F1 Score. Table 5(b) shows that ChatHuman achieves SOTA precision and F1 score, demonstrating superior human-object interaction understanding ability. Notably, although LLaVA has a high recall rate, its precision and F1 score are rather poor, which means that it tends to predict all the body parts to be in contact with objects.

4.4. Ablation Study

Paper-based RAG Mechanism. To improve tool use accuracy, we design a paper-based RAG mechanism. We perform a breakdown ablation to investigate the effect of each component and their interactions. The baseline model, created by removing the RAG operation and trained with instruction-following data without paper content, is compared in Table 6(a). The baseline model's success rate (SR) is 0.96 for seen tools and 0.82 for unseen tools. Adding RAG increases the SR for unseen tools to 0.89, demonstrating its effectiveness in zero-shot settings. Further incorporating articles into training data boosts the performance: the successful rate of arguments (SR_{args}) rises from 0.93 to 0.95 for the seen tools and 0.84 to 0.94 for the unseen tools. This suggests that scholarly articles can help create high-quality instruction-following data and tool documents due to their detailed use instructions and diverse application scenarios. For a detailed analysis of each component's effect on instructing tool usage, please see Sup. Mat.

Multiple Tools Invocation. One of the advantages of using a VLM as an agent is its powerful generalization capacity. To test the robustness and generalization ability of ChatHuman, we conduct the following ablation study. During training, we only include the tool graphs with no more than three tools, while during evaluation, the user queries might need up to five tools to solve. Table 6(b) depicts the results. As shown, ChatHuman exhibits an excellent robustness in this out-of-domain setting (more than three tools combination) with an action accuracy higher than 90%.

Tool Result Integration. We first conduct an ablation to study how the tools can enhance the human understanding capacity of multimodal LLM. The model without tools is our multimodal LLM backbone, LLaVA-1.5-7B [41], and the model with tools is our ChatHuman. The quantitative results are listed in Table 7. When equipped with tools, the HOI contact detection F1 score increases from 0.39 to 0.63 and the average body shape measurement error declines by 38%. These results demonstrate the effectiveness of tools in enriching the LLM's comprehension of human models and behaviors. Additionally, we study whether ChatHuman can utilize its world knowledge to discriminate and improve the tool performance. We design two discrimination schemes, i.e., selection and modification, and conduct an ablation study on two human-related tasks by comparing ChatHuman with the SOTA task-specific tools. For the selection scheme, we ex-

Paper	RAG	Seen Tools SR_t	SR_{act}	SR_{args}	SR	IoU	Unseen Tools SR_t	SR_{act}	SR_{args}	SR	IoU
×	×	0.998	0.967	0.928	0.960	0.964	0.946	0.894	0.775	0.822	0.872
×	✓	**1.000**	0.967	0.928	0.961	0.965	0.996	0.945	0.842	0.891	0.927
✓	✓	**1.000**	**0.974**	**0.950**	**0.970**	**0.975**	**0.999**	**0.967**	**0.893**	**0.954**	**0.953**

(a) Ablation study of paper-based RAG mechanism.

Tool Numer	SR_t	SR_{act}	SR_{args}	SR	IoU
2	0.997	0.960	0.943	0.928	0.973
3	0.998	0.959	0.931	0.932	0.974
4	0.998	0.943	0.928	0.875	0.968
5	0.996	0.929	0.899	0.847	0.950

(b) Ablation study of tool number.

Table 6. Ablations related to tool usage. Successful rate of thought, action, arguments, execution, and IoU are reported.

Method	Precision ↑	Recall ↑	F1 Score ↑
w/o Tool	0.26	**0.81**	0.39
w/ Tool	**0.67**	0.67	**0.63**

(a) HOI Contact Detection.

Method	Height ↓	Chest ↓	Waist ↓	Hip ↓
w/o Tool	**6.7**	16.5	22.9	17.6
w/ Tool	**6.7**	**6.1**	**13.0**	**6.4**

(b) Body Shape Measurement.

Table 7. Ablation study on the impact of tool usage for human-object contact detection and body shape estimation.

Method	MPJPE ↓	PA-MPJPE ↓	PA-MPVPE ↓
Tool A	126.2	81.4	101.9
Tool B	124.0	84.6	104.7
ChatHuman	**119.6**	**78.2**	**98.3**

(a) Mesh Error (in mm) on MixPose.

Method	Height ↓	Chest ↓	Waist ↓	Hip ↓
Tool [2]	7.8	8.6	13.5	7.0
ChatHuman	**6.7**	**6.1**	**13.0**	**6.4**

(b) Body Shape Measurement Error (in cm) on HBW [8].

Table 8. Study revealing how tool use improves human understanding on pose estimation and body shape measurement tasks.

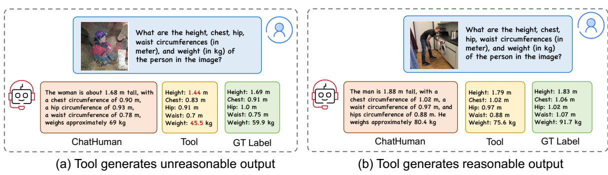

Figure 9. Illustration of how ChatHuman discriminates and integrates tool results. The Agent will fix the unreasonable tool result and integrate the reasonable tool result to generate a final response.

periment with the pose estimation task and select two SOTA methods, HMR 2.0 [18] and CLIFF-SMPLify [4, 36], as our tools to generate two poses of each person. We then prompt the LLM-based agent to discriminate the results and choose the better one as the final response. Different tools excel in different scenarios and, to cover more diverse human poses and camera views, we build a new benchmark MixPose by selecting 100 images with extreme camera views from the MoYo [69] test set, 100 full-body samples and 100 severely-truncated samples from 3DPW [70] test set. Details of the prompt and MixPose benchmark are in Sup. Mat. As shown in Table 8 (a), ChatHuman archives a lower reconstruction error on the MixPose benchmark, validating the agent's effectiveness as a discriminator to improve tool performance. For the modification scheme, we validate on the body shape measurement task. We use CLIFF-BEDLAM [2] as tool and prompt the agent to discriminate and modify the tool result. The result is reported in Table 8 (b), and Fig. 9. The LLM-based agent enhances tool performance by using its general world knowledge to identify and correct unreasonable tool results, such as height and weight in Fig. 9 (a).

5. Discussion and Concluding Remarks

We introduce ChatHuman, an LLM-based model designed to learn the use of tools related to 3D humans and assist users

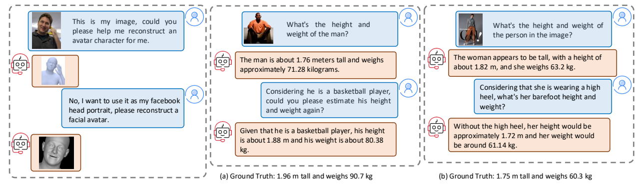

Figure 10. Human interaction can improve the performance and tool usage accuracy of ChatHuman.

in solving tasks associated with 3D humans. ChatHuman processes requests from users, analyzes the needs, and utilizes the necessary tools. It then evaluates the tools' outputs to respond to the user's queries.

Limitations. ChatHuman may fail in certain calling scenarios, particularly when the user request is vague, and subsequent LLM internal analysis cannot rectify an incorrect initial function call. However, further interaction with users can remedy this if they provide additional information. Fig. 10 illustrates an instance of using body estimation and face reconstruction tools for avatar creation. Even with the application and analysis of the tool, outcomes like height estimation may not be entirely precise. One contributing factor is the accuracy of the training data; for instance, most height labels in datasets use the official height of models or celebrities, which may not account for variations like shoe height, such as a 7-inch heel. Incorporating more cues from users, combined with the LLM's knowledge of the world and reasoning capabilities, can enhance result accuracy, as shown in Fig. 10. Our system is currently limited by the academic methods used. Incorporating better academic methods will enhance model performance. Notably, adding new tools requires no additional training, allowing our method to evolve and improve as new techniques are developed.

Future Work. ChatHuman offers several avenues for future development. In particular, user interaction/dialog offers opportunities to learn from user feedback. This could exploit reinforcement learning to refine the model's understanding.

Acknowledgements and Disclosures

We thank Naureen Mahmood, Nicolas Keller and Nicolas Heron for the support of the data collection. While MJB is a co-founder and Chief Scientist at Meshcapade, his research in this project was performed solely at, and funded solely by, the Max Planck Society.

References

[1] Kfir Aberman, Rundi Wu, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Learning character-agnostic motion for motion retargeting in 2d. *arXiv preprint arXiv:1905.01680*, 2019. 3

[2] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023. 5, 7, 8

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3

[4] 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 *ECCV*, 2016. 3, 6, 8

[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] Harrison Chase and LangChain Contributors. Langchain, 2022. 5

[7] 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. 5

[8] Vasileios Choutas, Lea Müller, Chun-Hao P Huang, Siyu Tang, Dimitrios Tzionas, and Michael J Black. Accurate 3d body shape regression using metric and semantic attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2718–2728, 2022. 1, 3, 5, 7, 8

[9] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022. 1

[10] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. Posescript: 3d human poses from natural language. In *ECCV*, 2022. 1, 3, 4, 5, 6, 7

[11] Ginger Delmas, Philippe Weinzaepfel, Francesc Moreno-Noguer, and Grégory Rogez. Posefix: Correcting 3d human poses with natural language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15018–15028, 2023. 3, 5

[12] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Computer Vision and Pattern Recognition Workshops*, 2019. 3

[13] Yu Du, Fangyun Wei, and Hongyang Zhang. Anytool: Self-reflective, hierarchical agents for large-scale api calls. *arXiv preprint arXiv:2402.04253*, 2024. 3

[14] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *3DV*, 2021. 3

[15] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics*, 40(4):1–13, 2021. 1, 3, 5

[16] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J. Black. ChatPose: Chatting about 3d human pose. In *CVPR*, 2024. 1, 3, 6, 7

[17] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023. 3

[18] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 1, 3, 5, 6, 8

[19] Sookwan Han and Hanbyul Joo. Chorus: Learning canonicalized 3d human-object spatial relations from unbounded synthesized images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15835–15846, 2023. 3

[20] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*, 2022. 3

[21] Cheng-Yu Hsieh, Si-An Chen, Chun-Liang Li, Yasuhisa Fujii, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. Tool documentation enables zero-shot tool-usage with large language models. *arXiv preprint arXiv:2308.00675*, 2023. 3

[22] 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. *arXiv:2106.09685*, 2021. 5

[23] Yifeng Jiang, Jungdam Won, Yuting Ye, and C Karen Liu. Drop: Dynamics responses from human motion prior and projective dynamics. *SIGGRAPH Asia*, 2023. 3

[24] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. In *3DV*, 2020. 3

[25] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 3

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 5

[27] 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 *ICCV*, pages 2252–2261, 2019. 1, 3, 6

[28] Yilun Kong, Jingqing Ruan, Yihong Chen, Bin Zhang, Tianpeng Bao, Shiwei Shi, Guoqing Du, Xiaoru Hu, Hangyu Mao, Ziyue Li, et al. Tptu-v2: Boosting task planning and tool

usage of large language model-based agents in real-world systems. *arXiv preprint arXiv:2311.11315*, 2023. 3

[29] Yilun Kong, Jingqing Ruan, Yihong Chen, Bin Zhang, Tianpeng Bao, Shiwei Shi, Guoqing Du, Xiaoru Hu, Hangyu Mao, Ziyue Li, et al. Tptu-v2: Boosting task planning and tool usage of large language model-based agents in real-world systems. *arXiv preprint arXiv:2311.11315*, 2023. 3

[30] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020. 3

[31] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. 3

[32] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation. In *CVPR*, 2021. 3

[33] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023. 3

[34] Peizhuo Li, Kfir Aberman, Rana Hanocka, Libin Liu, Olga Sorkine-Hornung, and Baoquan Chen. Learning skeletal articulations with neural blend shapes. *ACM Transactions on Graphics (TOG)*, 40(4):1, 2021. 3

[35] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 2

[36] 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 *ECCV*, 2022. 1, 3, 8

[37] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. *CVPR*, 2023. 1, 3

[38] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1954–1963, 2021. 5

[39] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 4

[40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 3

[41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2, 4, 5, 6, 7

[42] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang, Jianfeng Gao, and Chunyuan Li. Llava-plus: Learning to use tools for creating multimodal agents, 2023. 3

[43] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. In *ACM TOG*, 2015. 2, 4

[44] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[45] Thomas Lucas*, Fabien Baradel*, Philippe Weinzaepfel, and Grégory Rogez. Posegpt: Quantization-based 3d human motion generation and forecasting. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[46] Lea Muller, Ahmed AA Osman, Siyu Tang, Chun-Hao P Huang, and Michael J Black. On self-contact and human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9990–9999, 2021. 1, 3

[47] OpenAI. GPT-4 technical report. 2023. 4, 6, 7

[48] 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 *CVPR*, 2019. 2

[49] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D face model for pose and illumination invariant face recognition. Genova, Italy, 2009. IEEE. 2

[50] Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *International Conference on Computer Vision (ICCV)*, 2023. 3, 5

[51] 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. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[52] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. *arXiv preprint arXiv:2002.11681*, 2020. 5

[53] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 2, 3, 5, 6

[54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 5

[55] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 3

[56] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *IEEE International Conference on Computer Vision Workshops*, 2021. 3

[57] Jingqing Ruan, Yihong Chen, Bin Zhang, Zhiwei Xu, Tianpeng Bao, Hangyu Mao, Ziyue Li, Xingyu Zeng, Rui Zhao, et al. Tptu: Task planning and tool usage of large language model-based ai agents. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023. 3

[58] 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

[59] Yongliang Shen, Kaitao Song, Xu Tan, Wenqi Zhang, Kan Ren, Siyu Yuan, Weiming Lu, Dongsheng Li, and Yueting Zhuang. Taskbench: Benchmarking large language models for task automation. *arXiv preprint arXiv:2311.18760*, 2023. 4

[60] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *CVPR*, 2024. 5

[61] Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, and Tao Yu. One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint arXiv:2212.09741*, 2022. 4, 5

[62] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2023. 3

[63] John Sweller, Paul Ayres, and Slava Kalyuga. *Cognitive Load Theory*. Springer, New York, NY, 2011. 2

[64] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 358–374. Springer, 2022. 3

[65] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[66] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 1274–1283, 2017. 3

[67] 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

[68] Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J Black. Deco: Dense estimation of 3d human-scene contact in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8001–8013, 2023. 1, 3, 4, 5, 7

[69] Shashank Tripathi, Lea Müller, Chun-Hao P Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. 3d human pose estimation via intuitive physics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4713–4725, 2023. 5, 8

[70] 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 *ECCV*, 2018. 5, 6, 8

[71] Chenyu Wang, Weixin Luo, Qianyu Chen, Haonan Mai, Jindi Guo, Sixun Dong, Xiaohua (Michael) Xuan, Zhengxin Li, Lin Ma, and Shenghua Gao. Mllm-tool: A multimodal large language model for tool agent learning. *arXiv preprint arXiv:2401.10727*, 2024. 3

[72] Qinzhuo Wu, Wei Liu, Jian Luan, and Bin Wang. Toolplanner: A tool augmented llm for multi granularity instructions with path planning and feedback. *arXiv preprint arXiv:2409.14826*, 2024. 3

[73] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. In *CVPR*, 2020. 2

[74] Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. GPT4Tools: Teaching llm to use tools via self-instruction. *arXiv preprint arXiv:2305.18752*, 2023. 2, 3, 4

[75] 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, 2023. 5, 6

[76] 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

[77] Zhao Yang, Jiaxuan Liu, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. Appagent: Multimodal agents as smartphone users. *arXiv preprint arXiv:2312.13771*, 2023. 2, 3

[78] Shaokai Ye, Jessy Lauer, Mu Zhou, Alexander Mathis, and Mackenzie Mathis. Amadeusgpt: a natural language interface for interactive animal behavioral analysis. *Advances in neural information processing systems*, 36:6297–6329, 2023. 2

[79] Shaokai Ye, Jessy Lauer, Mu Zhou, Alexander Mathis, and Mackenzie Mathis. Amadeusgpt: a natural language interface for interactive animal behavioral analysis. *Advances in neural information processing systems*, 36, 2024. 3

[80] Siyu Yuan, Kaitao Song, Jiangjie Chen, Xu Tan, Yongliang Shen, Ren Kan, Dongsheng Li, and Deqing Yang. Easytool: Enhancing llm-based agents with concise tool instruction. *arXiv preprint arXiv:2401.06201*, 2024. 3

[81] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11446–11456, 2021. 3

[82] Longwen Zhang, Qiwei Qiu, Hongyang Lin, Qixuan Zhang, Cheng Shi, Wei Yang, Ye Shi, Sibei Yang, Lan Xu, and Jingyi Yu. Dreamface: Progressive generation of animatable 3d faces under text guidance, 2023. 3

[83] Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, et al. Retrieving multimodal information for augmented generation: A survey. *arXiv preprint arXiv:2303.10868*, 2023. 3

[84] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ:

Controllable and consistent human image animation with 3d parametric guidance, 2024. 1, 3, 5