ChatHuman: Language-driven 3D Human Understanding with Retrieval-Augmented Tool Reasoning

Jing Lin[3,4,⋆], Yao Feng[1,2,3,⋆,†], Weiyang Liu[1,5], and Michael J. Black[1]

[1]Max Planck Institute for Intelligent Systems-Tübingen [2]ETH Zürich
[3]Meshcapade [4]Tsinghua University [5]University of Cambridge

chathuman.github.io

Abstract. Numerous methods have been proposed to detect, estimate, and analyze properties of people in images, including the estimation of 3D pose, shape, contact, human-object interaction, emotion, and more. Each of these methods works in isolation instead of synergistically. Here we address this problem and build a language-driven human understanding system – ChatHuman, which combines and integrates the skills of many different methods. To do so, we finetune a Large Language Model (LLM) to select and use a wide variety of existing tools in response to user inputs. In doing so, ChatHuman is able to combine information from multiple tools to solve problems more accurately than the individual tools themselves and to leverage tool output to improve its ability to reason about humans. The novel features of ChatHuman include leveraging academic publications to guide the application of 3D human-related tools, employing a retrieval-augmented generation model to generate in-context-learning examples for handling new tools, and discriminating and integrating tool results to enhance 3D human understanding. Our experiments show that ChatHuman outperforms existing models in both tool selection accuracy and performance across multiple 3D human-related tasks. ChatHuman is a step towards consolidating diverse methods for human analysis into a single, powerful, system for 3D human reasoning.

1 Introduction

Humans are distinguished from other animals by their extensive creation and use of tools. We exploit tools to solve specialized tasks and integrate these solutions into our general understanding of the world. Recent work on Large Language Models (LLMs) has shown that such models can also learn to use tools to extend their skills. Inspired by this, we develop ChatHuman, a multi-modal LLM that is specialized for understanding humans and their 3D behavior. ChatHuman learns to exploit a wide range of specialized human-related models for tasks such as 3D pose estimation, emotion recognition, reasoning about contact, and more. We introduce novel training methods and show that ChatHuman is able to navigate and exploit multiple tools to achieve state-of-the-art results on various tasks regarding 3D human understanding. The power of ChatHuman makes a

⋆Equal contribution. †Corresponding author.

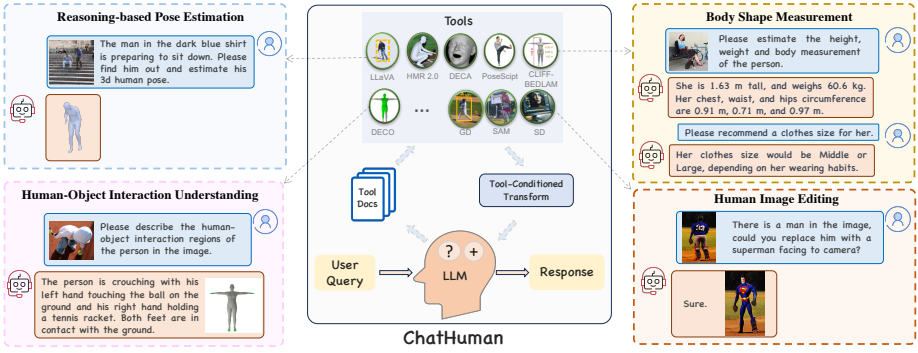


Fig. 1: ChatHuman is a novel system that uses a multimodal LLM to exploit tools, discriminate their results, and integrate the results to solve tasks focused on 3D humans.

step closer towards a foundation model for holistic human understanding with applications in domains like AR/VR [25,64], games [25,57], and fashion [13,65].

Research on 3D humans has progressed rapidly in recent years, resulting in the creation of many tools that can perform tasks like estimating a human's 3D pose from a single image [15, 21, 23, 29], predicting face/body shapes [5, 11], capturing emotions [6, 11], and identifying regions of touch/contact [40, 60]. Each of these tools, however, focuses on a specific problem; that is, they are all "specialists". Moreover, each of these separate tools cannot benefit from the expertise of the others. Combining these tools to solve more complex tasks requires significant domain expertise. To address complex, real-world, tasks, we would like a "generalist" model that can solve a wide range of problems with equal or better accuracy than the specialists. Such a system should be able to use a wide range of specialist tools, know how to apply them to appropriate tasks, and be able to synthesize the results of several tools to solve new problems. For example, reasoning-based pose estimation (Figure 1) can be addressed through the combination of text-guided detection (LLaVA [32]), cropping, and human pose estimation tools (HMR2 [15]), rather than relying on a single tool alone.

ChatHuman provides one of the earliest solutions to this problem by finetuning a LLM to serve as an agent to call the appropriate tools in response to user inputs, thereby completing real-world tasks and enhancing its responses to users with the results from the tools. Sharing a similar spirit, some recent approaches have utilized off-the-shelf or finetuned LLMs to solve basic vision problems (e.g., Visual ChatGPT [49]), to combine mobile applications (e.g., AppAgent [69]), and to understand complex visual inputs (e.g., GPT4Tools [68]). In contrast to these works, we focus on the challenging task of 3D human understanding. This is challenging because there are many disparate tools for analyzing different aspects of humans in images. Our model has to select the appropriate tool, for a given user query, use the tool correctly, and combine the output with its broader knowledge to respond to the user. The tools themselves produce a variety of output formats including images, text, and 3D parametric meshes.

We draw inspiration from how humans use tools: 1) we read related papers to understand the tools, which helps us know when and how to use them; 2)

after using the tool, we evaluate whether the outcome is reliable, or compare the results of several methods to determine which is most trustworthy; and 3) we then incorporate these results to form our response. We apply the same steps to our LLM-based model, resulting in a novel approach with several contributions:

Paper-based Retrieval-Augmented Tool Usage: Details of a model are described in a research paper. We give the LLM access to various parts of these papers and show that "reading the paper" improves tool use performance. Furthermore, we analyze which sections of the paper are most valuable for instructing tool use. Additionally, when encountering a tool for the first time, people often turn to the user guide for assistance. We compile documentation for these tools and utilize a paper-based Retrieval-Augmented Generation (RAG) mechanism to improve 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 aid of the paper-based RAG model.

Tool Result Discrimination and Integration: After using the tools, analyzing the outcomes is crucial. In scenarios with various outcomes, the LLM should discern the most reliable result. Inspired by Cognitive Load Theory [56], which posits that choosing from options is faster and demands less cognitive effort than creating detailed explanations, we've designed a method to streamline the LLM's decision-making process. We present the results from various tools as multiple-choice questions, making it easier for the LLM to evaluate and select the best outcome, thus enhancing the tool's output. These results, combined with the LLM's general knowledge, are used to generate responses about 3D humans. Specifically, ChatHuman consists of a multi-modal LLM, and 22 human-related tools. The LLM is finetuned for tasks that involve using these tools and incorporating 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 usage accuracy but also enhances performance on a variety of human-related tasks.

To summarize, our key contributions include: (1) a new framework that leverages LLMs to address tasks related to 3D human understanding with human-centric tools; (2) a scientific-paper-based RAG mechanism, which ensures precise tool usage by comprehending tool descriptions from scholarly articles, enhancing tool application and contextual understanding; and (3) the integration of tool outcomes with LLMs, where the LLM is finetuned to discriminate between good and bad tool results to provide more accurate results than any single tool. ChatHuman achieves superior performance in tool usage and human-related tasks compared to other LLM-based methods or task-specific work. We will make available the code, trained models, and datasets for research purposes.

2 Related work

3D Humans: There is an extensive literature on the analysis of 3D humans, which we only sample here. Reasoning about 3D humans typically leverages parametric models like SMPL [36], SMPLX [43], or GHUM [66] for the body,

BFM [19] or FLAME [28] for faces, and MANO [51] 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-based methods [2, 20] or regression-based methods [15, 21, 23, 27, 29, 71], which estimate SMPL shape and pose parameters from a single image. Similarly, face reconstruction methods [9, 11, 58] estimate shape and expression parameters of the face model from single images. The analysis of contact, vital for understanding human-environment interaction and social properties, has seen innovations in methods [16, 40, 60]. In the generation domain, PoseScript [7] and PoseFix [8] offer methods for synthesizing and correcting 3D human poses from text descriptions. Recent language-to-3D generation methods [3, 17] create 3D human shapes. For understanding, there are studies focusing on classifying action labels in video sequences [42, 47] or recognizing human emotions [6, 62], enhancing our comprehension of human behavior. These basic methods excel in their respective scenarios. Recently, ChatPose [12] unifies pose generation, estimation, and LLM's general understanding into one model, but is still limited in full-body pose tasks. In contrast, our model leverages and integrates the performance of 22 3D human-related tasks into a single, LLM-based model. We emphasize the significance of ChatHuman, as it 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 output of different tools.

Large Language Models and Tool Use: To expand the capabilities of LLMs without expensive retraining, recent work has focused on enabling them to use specialized off-the-shelf tools. In this line of work, a library of tools is constructed and LLMs act as the planner to navigate the use of the tools. Diffrent types of tools have been adopted, e.g., vision modules [49,55,70], mobile applications [69], community tools [52] and system tools [68]. However, general-purpose LLMs often lack a deep understanding of specific tools, especially those containing domain knowledge. To address this, a few recent works [10, 24, 63] propose to fintune general-purpose LLMs (e.g., LLaMA [59], LLaVA [33, 34]) with domain-specific tool use data. Distinct from previous works, ChatHuman focuses on understanding 3D humans through language interaction by leveraging the power of off-the-shelf human-related tools.

Retrieval Augmented Generation: RAG [14,26,72] is a technique to enhance generative tasks by retrieving relevant information from external databases, allowing for continuous knowledge updates. Based on this feature, we design a RAG mechanism to facilitate the usage of newly introduced, unseen tools.

3 Language-driven 3D Human Understanding

The goal of ChatHuman is to leverage the capabilities of LLMs for tasks regarding 3D human understanding. The model is designed to interact with users and deal with various user requests through input like text, image, or any information about 3D humans. This includes details on how a person is posed

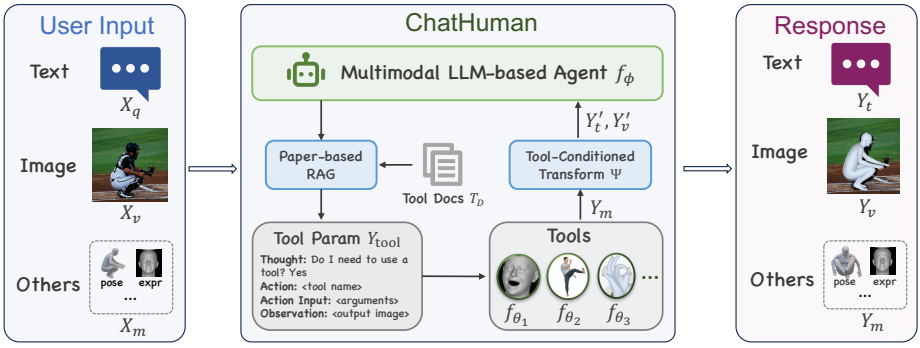


Fig. 2: Method Overview. Given a user query, the multimodal LLM-based agent adopts a paper-based RAG mechanism to determine whether to employ tools and identify the optimal way to utilize them. After applying the tools, the tool results are transformed into a text or visual format and fed back to the agent to formulate responses.

(using SMPL [36] pose parameters), their emotions (using FLAME [28] expressions),and how they touch or interact with the environment (contact), etc. Figure 2 gives an overview of the proposed ChatHuman model. We start by elaborating the architecture design and the detailed training process (Section 3.1). This includes the introduction of the paper-based retrieval-augmented generation for accurate tool usage (Section 3.2) and a feedback mechanism aimed at enhancing the LLM's responses to human-related questions (Section 3.3). Finally, we provide the details of our training data construction by employing GPT-4V [41] with a range of datasets associated with different tools.

3.1 Overall Pipeline

We refer to our ChatHuman model as Φ, which consists of a multi-modal LLM $f_\phi(\cdot)$, along with a set of 3D human-related functions $f_{\theta_1}(\cdot), f_{\theta_2}(\cdot), ...$ 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, or X_m representing other 3D human-related modalities. For instance, X_m could represent 3D human poses encoded as SMPL parameters. Then it produces outputs in the form of text Y_t, images Y_v, or modalities Y_m related to 3D humans, as described by $Y_t, Y_v, Y_m = \Phi(X_q, X_v, X_m)$. For training, we keep all the tool functions fixed, and only finetune the LLM $f_\phi(\cdot)$. Specifically, we use LoRA [18] to finetune the LLM, where the trainable parameters are denoted as ϕ_{lora}. With the ground truth textual output $\hat{Y}_t$, we optimize the model using the following objective function: $\mathcal{L} = \mathbf{CE}(\hat{Y}_t, Y_t)$, where $\mathbf{CE}$ denotes the cross-entropy loss. In the following sections, we introduce the details of the model function Φ and the construction of our training data.

3.2 Paper-based Retrieval-Augmented Tool Reasoning

Teaching LLMs to accurately discern when and how to utilize tools presents a significant challenge. A straightforward approach [49, 68] might involve sum-

marizing the usage scenarios and input arguments of each tool within the input prompt of the LLM $f_\phi(\cdot)$, formulated as $Y_{\text{tool}} = f_\phi(X_q, X_t)$, where X_t represents the tool definitions. But this often falls short for specialized tools, particularly when there exist numerous professional tools related to 3D human tasks. Firstly, some tools have various usage scenarios and require background knowledge for proper utilization. For instance, for HMR tool [15], relevant inquiries might include "Can you estimate the pose of this person?", "What are the SMPL parameters?", or "I want to get the 3D mesh of this person." It is difficult to concisely summarize all potential application scenarios and requisite background knowledge in a few sentences within the prompt. Furthermore, as the number of tools grows, the descriptions of these tools in the prompt get longer and more complex. This makes it harder for the LLM to use different tools, especially new ones for which it was not trained.

To solve these problems, we introduce a paper-based Retrieval-augmented Generation (RAG) mechanism to facilitate tool usage. As illustrated in Figure 3, we first input the academic papers associated with each tool to GPT-4 and prompt it to rearticulate the tool functions and enumerate potential user queries for tool activation. Academic papers, with their extensive background, detailed instructions, and varied applications, enable the generation of user queries covering a wide range of application scenarios. By combining these queries with the structured arguments of each tool, we compile a document consisting of question-answering pairs about tool operation for each tool. Figure 3 presents one example from the tool document. These documents X_d 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 is implemented by matching the text embedding of the query and those within the tool documents using a pretrained text embedding model [54]. The retrieved sample is then merged with the user query and provided to the agent f_ϕ as an in-context learning example,

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

where f_r represents the retrieval function and Y_{tool} is a textual description specifying whether to use a tool or not, tool names, and input arguments for tool function calling. This textual description, alongside the text query, vision, or inputs from other modalities, enables the tool function f_θ to compute the results, which is denoted by $Y_m = f_\theta(Y_{\text{tool}}, X_q, X_v, X_m)$.

3.3 Tool Result Discrimination and Integration

After using the tools, it's important to integrate the outcomes to solve problems. However, outputs from various tools come in different forms (e.g., language, images, vectors like SMPL poses), some of which might be difficult for current multimodal LLMs to understand, e.g. LLaVA [32] only accepts text and image as input. To leverage these results and enhance the LLMs' understanding of 3D humans, we introduce a tool-conditioned transformation $\Psi(\cdot)$ that converts tool outcomes Y_m into textual or visual formats. For example, we transform the vertex-wise contact label predicted by DECO [60] into a body part-level

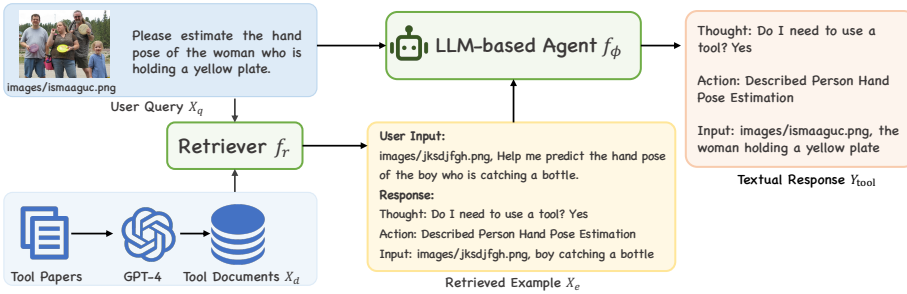


Fig. 3: Paper-based Retrieval-Augmented Tool Usage. We first feed the academic papers of each tool to GPT-4 and build a document for each tool. During inference, given a user query, a relevant sample is retrieved from the documents and provided to the LLM-based agent as an in-context example to improve the tool usage accuracy.

description based on the vertex-to-part mapping dictionary of SMPL [36], and we render the mesh generated by PoseScipt [7] into an RGB image using rendering techniques. After that, the transformed results will be merged with the user query as a clue to help the agent generate a response:

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

In scenarios where multiple tools can address a user's request, leading to various potential outcomes, we aim for the LLM to identify the most accurate result. To achieve this, we present the outcomes as multiple-choice questions, prompting the model to select the most relevant answer to the user query:

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

where Y_{mi} denotes the i-th tool result, X_d is a prompt specifying the discrimination rule. Details of the prompt are available in Appendix.

3.4 Training Data Construction

Tool Usage Instruction-following Data. Given a user query, the LLM-based agent aims to select appropriate tools and formulate the arguments for their operation. Following GPT4Tools [68], we provide GPT-4 [41] with an image content X_I and a tool-related prompt P_T to compile the instruction-following data (see Figure 4(a)). The image content X_I consists of a textual description of an image, detailing captions and object locations. The tool-related prompt includes a system message, and a description of tools delineated as `"<tool name>: <usage scenario>, <arguments>"`. As proposed in Section 3.2, we further incorporate the paper content X_P into GPT-4 to generate the tool usage instruction-following data I_T, formulated as $I_T \sim f_g(P_T|X_I, X_P)$, where $f_g(\cdot)$ denotes the GPT-4 model. I_T is a question-answering data pair consisting of a user query X_q and a response $\hat{Y}_t$ specifying whether to use a tool $\hat{Y}_{\text{thought}}$, the tool name $\hat{Y}_{\text{act}}$, and the tool input arguments $\hat{Y}_{\text{args}}$.

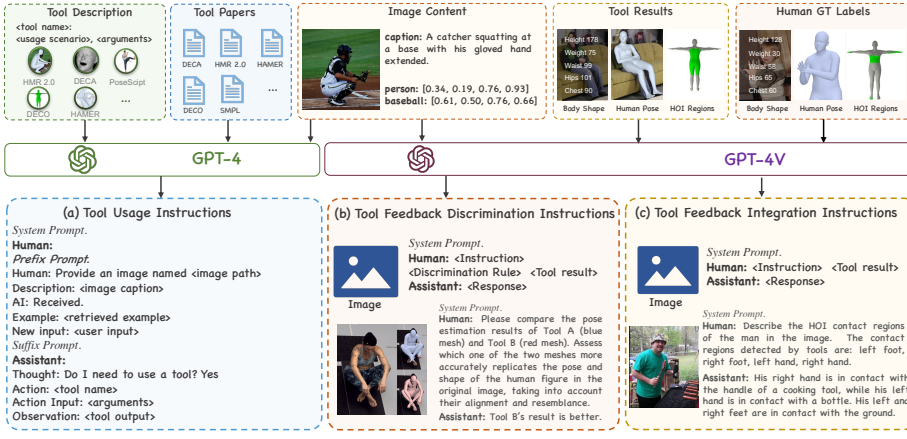


Fig. 4: Illustration of data construction. We construct instruct-following data about tool usage and feedback by providing GPT-4 with multiple tool-related information, image content, and ground truth labels. Gray text shows example of the instruction.

Tool Feedback Instruction-following Data. After calling the tools, the LLM-based agent should be able to discriminate and integrate the tools' outcomes. To achieve this, we construct instruction-following data based on the tool results and corresponding ground truth labels. Specifically, we transform the tool results and ground truth human labels into textual or visual formats using specialized techniques or rendering tools. Subsequently, we feed GPT-4V with the image content X_I, the transformed tool feedback X_T, and the ground truth label X_L to generate the instruction-following data about tool feedback, denoted as $I_F \sim f_g(P_F|X_I, X_T, X_L)$, where P_F is a system prompt, I_F consists of an instruction X_q and a corresponding answer $\hat{Y}_t'$. We instruct GPT-4V to curate two types of data. The first as illustrated in Figure 4(b), is about identifying the most suitable tool response. The second (Figure 4(c)) focuses on the integration of tool results, featuring a query posed by a user, the tool's output serving as a hint, and the assistant's reply.

Multi-Modal Instruction-following Data. To preserve the LLM's inherent capability for multi-turn conversations, we also incorporate the multi-modal instruction-following data I_M from LLaVA [32] during training.

Finally, We consolidate the tool usage instruction-following data I_T, tool feedback instruction-following data I_F, and multi-modal instruction-following data I_M into a unified format consistent with LLaVA [32]:

$$\texttt{Human} : X_{\text{instruct}} \texttt{<STOP> Assistant} : Y_t \texttt{<STOP>}, \tag{4}$$

where <STOP> is sequence termination token and X_{instruct} is input instruction. We set X_{instruct} as X_q for a language-only instruction, and as $X_q < \backslash n > X_v$ for a multimodal instruction that incorporates both text and visual elements.

Method	Seen Tools					Unseen Tools				
	SR_t	SR_{act}	SR_{args}	SR	IoU	SR_t	SR_{act}	SR_{args}	SR	IoU
GPT4Tools [67]	0.609	0.547	0.525	0.520	0.566	0.612	0.546	0.542	0.525	0.573
GPT4Tools-FT [67]	0.825	0.710	0.687	0.690	0.741	0.904	0.807	0.690	0.747	0.800
Visual ChatGPT-3.5 [49]	0.498	0.319	0.237	0.251	0.791	0.507	0.314	0.226	0.293	0.803
Visual ChatGPT-4 [49]	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 1: Comparison of tool usage accuracy. Successful rate of thought (SR_t), action (SR_{act}), arguments (SR_{args}), execution (SR), and IoU are reported.

4 Experiments and Results

4.1 Implementation Details

We employ LLaVA-1.5V [32] as the multimodal LLM backbone, with CLIP [46] for vision encoding and Vicuna [4] for the LLM backbone. For retrieval, we adopt INSTRUCTOR [54] for text embedding and utilize the vector similarity searching algorithm from Chroma to identify a relevant example. To preserve the generalization ability of the pre-trained multi-modal LLM, we leverage LoRA [18] to perform efficient fine-tuning, with a rank of 128 and an alpha of 256. Alternatively, orthogonal finetuning [35,45] can be used to improve performance. For optimization, we use AdamW [37], with a learning rate and weight decay set to 2e-4 and 0, respectively. All models are fine-tuned over 2 epochs with a mixture of tool usage, tool feedback, and LLaVA multimodal instruction-tuning data. During the training phase, we use 8 Nvidia 80G A100 GPUs and use the DeepSpeed [48] engine to enhance the training efficiency. Unless otherwise specified, we used LLaVA-1.5V-7B as the base model for the ablation study.

4.2 Datasets

Tool Usage Instruction-Tuning Data. To teach the agent to correctly use tools, we construct 90K instruction-response pairs about tool usage. Our tool library consists of 22 human-related tools: 8 for perception, 7 for reasoning, and 7 for generation. We further construct a validation and test set for evaluation. 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 not presented during training. To ensure the difference between the training and test sets, we use varied image caption sources for input prompts. More details are provided in the Appendix.

Tool Feedback Instruction-Tuning Data. To help the multimodal LLM model discriminate and integrate the tool results, we construct 88K instruction-following data based on existing 3D human datasets, including 61K tool result discrimination instructions built with MoYo [61], 3DPW [38], and PoseScript [7], and 27K tool result integration instructions from SHAPY [5] and DECO [60].

LangChain/Chroma

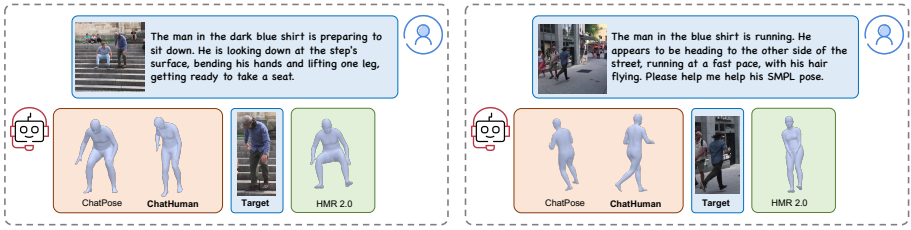


Fig. 5: Qualitative comparison with ChatPose [12], HMR 2.0 [15] for reasoning-based human pose estimation on SPG [12] benchmark.

4.3 Evaluation on Tools Usage

We compare our method with Visual ChatGPT [49] and GPT4Tools [67] on the proposed evaluation set and report 5 metrics proposed in GPT4Tools [67]. Please refer to Appendix 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". Regarding 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 1 shows that the original GPT4Tools does not perform well in our benchmark due to differences between the common visual tools in GPT4Tools and our human-centric tools. Visual ChatGPT-4 exhibits impressive tool usage accuracy, showing its powerful zero-shot ability to follow a standardized format and use tools accurately. Our ChatHuman consistently outperforms the competitors, particularly with tools not seen in training, thanks to our paper-based RAG mechanism.

4.4 Evaluation on 3D Human Understanding

In the following, we evaluate the performance of ChatHuman on four representative human-related tasks and report the performance on six benchmarks.

Pose Estimation. Following ChatPose [12], we evaluate the performance of our method in both classical and reasoning-based pose estimation tasks. MPJPE, PA-MPJPE, and MPJRE on the 3DPW [38] and RPE [12] 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 2, 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 Figure 5, only our method achieves a satisfactory result. The multimodal LLM competitor ChatPose could find the correct person but fails to obtain an accurate pose due to the limited perception ability, while the task-specific tool could not match the correct person due to the lack of reasoning ability. This demonstrates the advantages of ChatHuman, which combines task-specific tool use expertise with the general reasoning ability of an LLM.

Method	3DPW [38] MPJPE ↓	PA-MPJPE ↓	MPJRE ↓	RPE Benchmark [12] MPJPE ↓	PA-MPJPE ↓	MPJRE ↓
SPIN [23]	102.9	62.9	10.1	244.9	107.3	12.4
HMR 2.0 [15]	**91.0**	**58.4**	**9.2**	225.2	105.1	12.1
LLaVA-S [32]	440.8	205.4	21.8	490.7	207.4	21.1
LLaVA*-S [32]	232.1	101.1	12.8	-	-	-
GPT4-S [41]	322.0	136.7	16.0	-	-	-
LLaVA-P [32]	335.2	172.3	16.5	391.5	191.9	17.8
GPT4-P [41]	396.5	203.4	18.6	-	-	-
ChatPose [12]	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 2: Comparison of vanilla human pose estimation and reasoning-based pose estimation on 3DPW [38] and RPE [12] benchmarks. LLaVA* is fine-tuned with human keypoints data. "S" uses multimodal LLMs for keypoint detection and SMPLify [2] for pose optimization. "P" utilizes multimodal LLMs for textual pose descriptions, processed by PoseScript [7] to generate poses. MPJPE (in mm), MP-MPJPE (in mm), and MPJRE (×100) are reported.

Method	PoseScript [7] R^{P2T} ↑			R^{T2P} ↑			SPG Benchmark [12] R^{P2T} ↑			R^{T2P} ↑		
PoseScript [7]	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 [12]	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 [32]	-	-	-	-	-	-	2.1	4.0	7.1	2.1	3.3	6.1
GPT4-P [41]	-	-	-	-	-	-	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 3: Comparison of classical and speculative pose generation on PoseScript [7] and SPG [12] benchmarks. "P" utilizes LLMs for textual pose descriptions rephrase, processed by PoseScript [7] to generate poses. Top 5,10,20 recall rates are reported.

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) [12]. Following previous work [7,12], we report the text-to-pose recall rate R^{T2P} and pose-to-text recall rate R^{P2T} 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) [7] to generate a pose. As shown in Table 3, our method archives comparable performance to the SOTA methods on both benchmarks. In contrast, the previous LLM-based method, ChatPose, performances 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. Here we evaluate the body shape measurement accuracy of ChatHuman. We randomly sample 100 images from the HBW validation set [5] and compare our method with a multimodal LLM, LLaVA [32], and a SOTA body shape estimation method, CLIFF-BEDLAM [1]. For LLaVA and ChatHuman, we ask them the same question to inquire about the height, weight, chest, waist, and hip circumferences of a person in the image and then prompt GPT-3.5 to extract the value from the model output. The details of the question and prompt are available in Appendix. CLIFF-BEDLAM predicts the

Method	Height ↓	Weight ↓	Chest ↓	Waist ↓	Hip ↓
LLaVA [32]	**6.7**	**10.1**	16.5	22.9	17.6
CLIFF-BEDLAM [1]	7.8	13.9	8.6	13.5	7.0
ChatHuman	**6.7**	10.4	**6.1**	**13.0**	**6.4**

Table 4: Comparison of body shape measurement. Measurement errors (in cm and kg) on HBW validation set [5] are reported.

Method	Precision ↑	Recall ↑	F1 Score ↑
LLaVA [32]	0.26	**0.81**	0.39
GPT-4 [41]	0.61	0.48	0.49
ChatHuman	**0.67**	0.67	**0.63**

Table 5: Comparison of HOI understanding. Precision, Recall Rate, and F1 Score on DECO [60] are reported.

		Seen Tools					Unseen Tools				
Paper	RAG	SR_t	SR_{act}	SR_{args}	SR	IoU	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**

Table 6: Ablation study of paper-based RAG mechanism. Successful rate of thought (SR_t), action (SR_{act}), arguments (SR_{args}), execution (SR), and IoU are reported.

body shape parameter, which is then converted to measurements based on the shape-to-measurement function from SHAPY [5]. Anthropometric measurement errors are reported in Table 4. As shown, ChatHuman achieves superior performance in most measurements, outperforming the multimodal LLM competitor LLaVA by 42% and CLIFF-BEDLAM by 15.7% in average metrology accuracy.

Human-Object Interaction. Here we evaluate the human-object interaction understanding ability of ChatHuman on the DECO [60] test set. The textual ground truth (GT) labels are obtained by converting the vertex-level contact labels into body part-level contact labels with the vertex-to-part mapping dictionary of SMPL. Given a human image, we ask the multimodal LLM to detect the body parts that contact with 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 body part label and compute the average detection precision, recall rate, and F1 Score. From Table 5, ChatHuman achieves SOTA performance in the precision and F1 score metrics, demonstrating the superior human-object interaction understanding ability of our method. 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.

4.5 Ablation Study

Paper-based RAG Mechanism. To improve tool use accuracy, we design a paper-based RAG mechanism. Here we perform a break-down ablation to investigate the effect of each component and their interactions. The baseline model is derived by removing the RAG operation and is trained with the instruction-following data constructed without referring to paper content. The results are listed in Table 6. 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 scholarly 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 the use of scholarly articles can help construct

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 ↓	Weight ↓	Chest ↓	Waist ↓	Hip ↓
w/o Tool	**6.7**	**10.1**	16.5	22.9	17.6
w/ Tool	**6.7**	10.4	**6.1**	**13.0**	**6.4**

(b) Body Shape Measurement.

Table 7: Ablation study about how can tools improve human understanding on human-object contact detection and body shape measurement tasks.

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 ↓	Weight ↓	Chest ↓	Waist ↓	Hip ↓
Tool [1]	7.8	13.9	8.6	13.5	7.0
ChatHuman	**6.7**	**10.4**	**6.1**	**13.0**	**6.4**

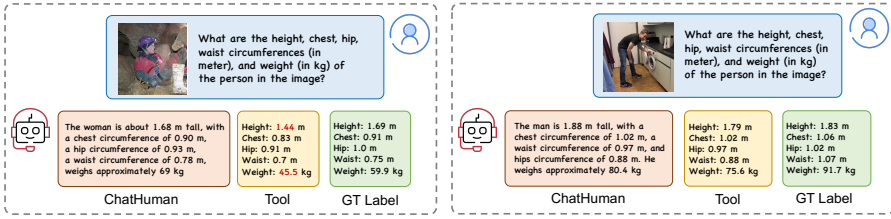
(b) Body Shape Measurement Error (in cm or kg) on HBW [5] Validation Set.

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

high-quality instruction-following data and tool documents, stemming from the detailed usage instructions and diverse application scenarios within the paper. We further analyze the effects of each component of the paper for instructing tool usage, please see Appendix for details.

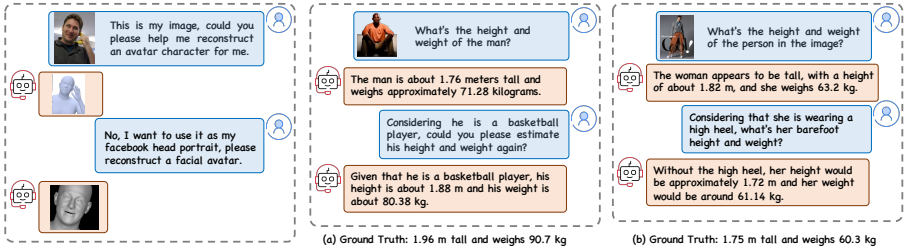
Tool Result Integration and Discrimination. We first conduct an ablation to study how can the tools enhance the human understanding capacity of multimodal LLM. The model without tools is our multimodal LLM backbone, LLaVA-1.5-7B [32], and the model with tools is our ChatHuman. The quantitative results are listed in Table 7. When equipped with tools, the HOI contact detection F1 score increases from 0.39 to 0.63 and the averaged 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 the multimodal LLM-based agent 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 experiment with the pose estimation task and select two SOTA methods, HMR 2.0 [15] and CLIFF-SMPLify [2, 29], 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 [61] test set, 100 full-body samples and 100 severely-truncated samples from 3DPW [38] test set. Details of the prompt and MixPose benchmark are in Appendix. As shown in Table 8(a), ChatHuman archives a lower reconstruction error on the MixPose benchmark, validating the effectiveness of the agent as a discriminator to improve the tool performance. For the modification scheme, we validate on the body shape measurement task. We use CLIFF-BEDLAM [1] as the tool and prompt the agent to discriminate and modify the tool result. The result is re-



(a) Tool generates unreasonable output (b) Tool generates reasonable output

Fig. 6: Illustration about how the multimodal LLM-based agent discriminates and integrate tool results. The Agent will fix the unreasonable tool result and integrate the reasonable tool result to generate a final response.



(a) Ground Truth: 1.96 m tall and weighs 90.7 kg (b) Ground Truth: 1.75 m tall and weighs 60.3 kg

Fig. 7: Human interaction can improve the performance and tool usage accuracy.

ported in Table 8(b) and Figure 6. As shown, the LLM-based agent can improve the tool performance by utilizing its general world knowledge to discriminate and fix the unreasonable tool results, e.g., the height and weight in Figure 6(a).

5 Discussion and Concluding Remarks

In conclusion, we introduce ChatHuman, an LLM-based model designed to learn the use of tools related to 3D humans and assist users in solving tasks associated with 3D humans. The model processes requests from users, analyzes the needs, and applies the necessary tools. It then evaluates and synthesizes the tools' outputs to respond to the user's queries and address the problems effectively. Having computers better understand 3D humans has many potential benefits in healthcare, fitness, human-robot interaction, entertainment, etc. Possible negative impacts include unwanted surveillance, body shaming, and deep fakes. The risk-benefit ratio should be monitored and, like other LLMs, guardrails may be needed to prevent malicous use cases.

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. For example, Figure 7 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 mores cues from users, combined with the LLM's knowledge of the world and reasoning capabilities, can enhance the accuracy of results, as shown in Figure 7.

Future Work. ChatHuman offers several exciting avenues for future development: 1) Integrated Learning and Self-Improving: Merging tool use learning with user feedback to continuously refine the model's understanding and approach to 3D human tasks. 2) User Feedback for Enhanced Training: As shown in Figure 7, user interaction has a tangible impact on improving outcomes. Future work could explore how ongoing dialogues with users might provide valuable feedback for refining and expanding the system's capabilities in 3D human-related research.

Acknowledgement

We thank Naureen Mahmood, Nicolas Keller and Nicolas Heron for support with data. MJB has received research gift funds from Adobe, Intel, Nvidia, Meta/Facebook, and Amazon. MJB has financial interests in Amazon and Meshcapade GmbH. 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. Black, M.J., Patel, P., Tesch, J., Yang, J.: Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8726–8737 (2023) 11, 12, 13, 20, 23
2. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: ECCV (2016) 4, 11, 13, 22
3. Cao, Y., Cao, Y.P., Han, K., Shan, Y., Wong, K.Y.K.: Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. arXiv preprint arXiv:2304.00916 (2023) 4
4. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 909
5. Choutas, V., Müller, L., Huang, C.H.P., Tang, S., Tzionas, D., Black, M.J.: Accurate 3d body shape regression using metric and semantic attributes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2718–2728 (2022) 2, 9, 11, 12, 13, 23
6. Daněček, R., Black, M.J., Bolkart, T.: Emoca: Emotion driven monocular face capture and animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20311–20322 (2022) 2, 4
7. Delmas, G., Weinzaepfel, P., Lucas, T., Moreno-Noguer, F., Rogez, G.: Posescript: 3d human poses from natural language. In: ECCV (2022) 4, 7, 9, 11, 20, 21, 22
8. Delmas, G., Weinzaepfel, P., Moreno-Noguer, F., Rogez, G.: Posefix: Correcting 3d human poses with natural language. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15018–15028 (2023) 4, 20

9. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: IEEE Computer Vision and Pattern Recognition Workshops (2019) 4
10. Du, Y., Wei, F., Zhang, H.: Anytool: Self-reflective, hierarchical agents for large-scale api calls. arXiv preprint arXiv:2402.04253 (2024) 4
11. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3D face model from in-the-wild images. vol. 40 (2021), https://doi.org/10.1145/3450626.3459936 2, 4, 20
12. Feng, Y., Lin, J., Dwivedi, S.K., Sun, Y., Patel, P., Black, M.J.: ChatPose: Chatting about 3d human pose. In: CVPR (2024) 4, 10, 11, 21
13. Feng, Y., Yang, J., Pollefeys, M., Black, M.J., Bolkart, T.: Capturing and animation of body and clothing from monocular video. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022) 2
14. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H.: Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 (2023) 4
15. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4D: Reconstructing and tracking humans with transformers. In: ICCV (2023) 2, 4, 6, 10, 11, 13, 20, 22
16. Han, S., Joo, H.: Chorus: Learning canonicalized 3d human-object spatial relations from unbounded synthesized images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15835–15846 (2023) 4
17. Hong, F., Zhang, M., Pan, L., Cai, Z., Yang, L., Liu, Z.: Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. arXiv preprint arXiv:2205.08535 (2022) 4
18. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv:2106.09685 (2021) 5, 9
19. IEEE: A 3D Face Model for Pose and Illumination Invariant Face Recognition (2009) 4
20. Joo, H., Neverova, N., Vedaldi, A.: Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation (2020) 4
21. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018) 2, 4
22. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023) 20
23. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: ICCV. pp. 2252–2261 (2019) 2, 4, 11
24. Kong, Y., Ruan, J., Chen, Y., Zhang, B., Bao, T., Shi, S., Du, G., Hu, X., Mao, H., Li, Z., 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) 4
25. LaViola Jr, J.J.: Bringing vr and spatial 3d interaction to the masses through video games. IEEE Computer Graphics and Applications **28**(5), 10–15 (2008) 2
26. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems **33**, 9459–9474 (2020) 4

27. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation. In: CVPR (2021) 4
28. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: 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), https://doi.org/10.1145/3130800.3130813 4, 5
29. Li, Z., Liu, J., Zhang, Z., Xu, S., Yan, Y.: CLIFF: Carrying location information in full frames into human pose and shape estimation. In: ECCV (2022) 2, 4, 13, 22
30. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1954–1963 (2021) 20
31. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014) 21
32. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023) 2, 6, 8, 9, 11, 12, 13, 20, 23
33. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023) 4
34. Liu, S., Cheng, H., Liu, H., Zhang, H., Li, F., Ren, T., Zou, X., Yang, J., Su, H., Zhu, J., Zhang, L., Gao, J., Li, C.: Llava-plus: Learning to use tools for creating multimodal agents (2023) 4
35. Liu, W., Qiu, Z., Feng, Y., Xiu, Y., Xue, Y., Yu, L., Feng, H., Liu, Z., Heo, J., Peng, S., et al.: Parameter-efficient orthogonal finetuning via butterfly factorization. In: International Conference on Learning Representations (2024) 9
36. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. In: ACM TOG (2015) 3, 5, 7, 22
37. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 9
38. von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In: ECCV (2018) 9, 10, 11, 13, 22
39. Matl, M.: Pyrender. https://github.com/mmatl/pyrender (2019) 22
40. Muller, L., Osman, A.A., Tang, S., Huang, C.H.P., Black, M.J.: On self-contact and human pose. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9990–9999 (2021) 2, 4
41. OpenAI: GPT-4 technical report. (2023) 5, 7, 11, 12, 21, 22
42. Pan, J., Chen, S., Shou, M.Z., Liu, Y., Shao, J., Li, H.: Actor-context-actor relation network for spatio-temporal action localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 464–474 (2021) 4
43. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: CVPR (2019) 3
44. Petrovich, M., Black, M.J., Varol, G.: TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In: International Conference on Computer Vision (ICCV) (2023) 20, 25
45. Qiu, Z., Liu, W., Feng, H., Xue, Y., Feng, Y., Liu, Z., Zhang, D., Weller, A., Schölkopf, B.: Controlling text-to-image diffusion by orthogonal finetuning. Advances in Neural Information Processing Systems **36**, 79320–79362 (2023) 9

46. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 9
47. Rajasegaran, J., Pavlakos, G., Kanazawa, A., Feichtenhofer, C., Malik, J.: On the benefits of 3d pose and tracking for human action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 640–649 (2023) 4
48. Rasley, J., Rajbhandari, S., Ruwase, O., He, Y.: Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. arXiv preprint arXiv:2002.11681 (2020) 9
49. Rasley, J., Rajbhandari, S., Ruwase, O., He, Y.: Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671 (2023) 2, 4, 5, 9, 10, 25, 27
50. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 20
51. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6) (Nov 2017) 4
52. Shen, Y., Song, K., Tan, X., Li, D., Lu, W., Zhuang, Y.: Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. arXiv preprint arXiv:2303.17580 (2023) 4
53. Shin, S., Kim, J., Halilaj, E., Black, M.J.: Wham: Reconstructing world-grounded humans with accurate 3d motion. In: CVPR (2024) 20
54. Su, H., Shi, W., Kasai, J., Wang, Y., Hu, Y., Ostendorf, M., Yih, W.t., Smith, N.A., Zettlemoyer, L., Yu, T.: One embedder, any task: Instruction-finetuned text embeddings. arXiv preprint arXiv:2212.09741 (2022) 6, 9, 25
55. Surís, D., Menon, S., Vondrick, C.: Vipergpt: Visual inference via python execution for reasoning. Proceedings of IEEE International Conference on Computer Vision (ICCV) (2023) 4
56. Sweller, J., Ayres, P., Kalyuga, S.: Cognitive Load Theory. Springer, New York, NY (2011) 3
57. Szolin, K., Kuss, D., Nuyens, F., Griffiths, M.: Gaming disorder: A systematic review exploring the user-avatar relationship in videogames. Computers in Human Behavior **128**, 107124 (2022) 2
58. Tewari, A., Zollhofer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., Theobalt, C.: Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 1274–1283 (2017) 4
59. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) 4
60. Tripathi, S., Chatterjee, A., Passy, J.C., Yi, H., Tzionas, D., Black, M.J.: Deco: Dense estimation of 3d human-scene contact in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8001–8013 (2023) 2, 4, 6, 9, 12, 20, 22, 23
61. Tripathi, S., Müller, L., Huang, C.H.P., Taheri, O., Black, M.J., Tzionas, D.: 3d human pose estimation via intuitive physics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4713–4725 (2023) 9, 13, 22, 24

62. Usman, M., Latif, S., Qadir, J.: Using deep autoencoders for facial expression recognition. In: 2017 13th International Conference on Emerging Technologies (ICET). pp. 1–6. IEEE (2017) 4
63. Wang, C., Luo, W., Chen, Q., Mai, H., Guo, J., Dong, S., Xuan, X.M., Li, Z., Ma, L., Gao, S.: Mllm-tool: A multimodal large language model for tool agent learning. arXiv preprint arXiv:2401.10727 (2024) 4
64. Wei, S.E., Saragih, J., Simon, T., Harley, A.W., Lombardi, S., Perdoch, M., Hypes, A., Wang, D., Badino, H., Sheikh, Y.: Vr facial animation via multiview image translation. ACM Transactions on Graphics (TOG) **38**(4), 1–16 (2019) 2
65. Xiang, D., Bagautdinov, T., Stuyck, T., Prada, F., Romero, J., Xu, W., Saito, S., Guo, J., Smith, B., Shiratori, T., et al.: Dressing avatars: Deep photorealistic appearance for physically simulated clothing. ACM Transactions on Graphics (TOG) **41**(6), 1–15 (2022) 2
66. Xu, H., Bazavan, E.G., Zanfir, A., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: GHUM & GHUML: Generative 3D human shape and articulated pose models. In: CVPR (2020) 3
67. Yang, R., Song, L., Li, Y., Zhao, S., Ge, Y., Li, X., Shan, Y.: Gpt4tools: Teaching large language model to use tools via self-instruction (2023) 9, 10, 27
68. Yang, R., Song, L., Li, Y., Zhao, S., Ge, Y., Li, X., Shan, Y.: GPT4Tools: Teaching llm to use tools via self-instruction. arXiv preprint arXiv:2305.18752 (2023) 2, 4, 5, 7, 20, 21, 25
69. Yang, Z., Liu, J., Han, Y., Chen, X., Huang, Z., Fu, B., Yu, G.: Appagent: Multimodal agents as smartphone users. arXiv preprint arXiv:2312.13771 (2023) 2, 4
70. Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., Wang, L.: Mm-react: Prompting chatgpt for multimodal reasoning and action. arXiv preprint arXiv:2303.11381 (2023) 4
71. Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., Sun, Z.: Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11446–11456 (2021) 4
72. Zhao, R., Chen, H., Wang, W., Jiao, F., Do, X.L., Qin, C., Ding, B., Guo, X., Li, M., Li, X., et al.: Retrieving multimodal information for augmented generation: A survey. arXiv preprint arXiv:2303.10868 (2023) 4

Appendix

A Additional Details

Here we describe in more detail how we train a multimodal LLM to use tools, discriminate, and integrate tool results to improve its ability to reason about humans. As mentioned in the main paper, here we provide additional information about the implementation, training process, evaluation methodology, and ablation study. Due to space limits in the main paper, we provide more qualitative results here that illustrate the performance of ChatHuman and its limitations.

A.1 Tool Details

We support 24 human-related tools in our tool pockets, including 9 perception tools, 8 generation tools, and 7 reasoning tools as listed in Table 9.

Perception	Reasoning	Generation
Body Pose Estimation [15]	Selective Person Pose Detection [15,32]	Text-to-Pose Generation [7]
Body Shape Measurement [1]	Specific Person Shape Measurement [1,32]	Speculative Pose Generation [7,32]
Hand Pose Estimation [30]	Targeted Hand Pose Estimation [30,32]	Text-to-Image Generation [50]
Face Reconstruction [11]	Described Person Face Reconstruction [11,32]	Text-based Pose Editing [8]
Human Segmentation [22]	Described Person Segmentation [22,32]	Remove Something From The Photo [22,32,50]
HOI Detection [60]	Selective Person Contact Estimation [32,60]	Replace Something From The Photo [22,32,50]
Pose Description [7]	Visual Question Answering [32]	Instruct Image Using Text [50]
Image Caption [32]		Text-to-Motion Generation [44]
Motion Capture [53]		

Table 9: List of tools used in ChatHuman. ChatHuman supports 24 human-related tools, including 9 perception tools, 8 generation tools, and 7 reasoning tools.

A.2 Evaluation Metric Details

Tool Usage. We use the evaluation metrics proposed in GPT4Tools [68] to measure the tool usage accuracy, including:

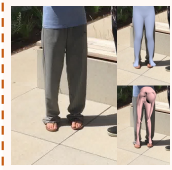
- **Successful Rate of Thought** ($\mathrm{SR_t}$), which measures the decision accuracy, calculated as $\mathrm{SR_t} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\tau_i)$, where N is the number of instructions and τ_i is a singular process. When the predicted thought is the same as the ground-truth thought, $\mathbb{I}(\tau_i)$ is equal to 1, and 0 otherwise.
- **Successful Rate of Action** ($\mathrm{SR_{act}}$), which measures the tool name prediction accuracy, calculated as $\mathrm{SR_{act}} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\alpha_i)$, where α_i is the matching process of the tool name. If the predicted tool name is correct, $\mathbb{I}(\alpha_i)$ is equal to 1, and 0 otherwise.
- **Successful Rate of Arguments** ($\mathrm{SR_{args}}$), which measures the tool arguments prediction accuracy, calculated as:

$$\mathrm{SR_{args}} = \frac{1}{N}\sum_{i=1}^{N}\eta_i, \quad \eta_i = \frac{1}{K}\sum_{j=1}^{K}\eta_{i,j}, \tag{5}$$

where K is the number of tool arguments. When the argument is a file name, $\eta_{i,j}$ equals 1 if the predicted file name is the same as the ground-truth file name, and 0 otherwise. When the argument is text, $\eta_{i,j}$ equals the BLEU score between the predicted and groud-truth text.

Human: Please compare the pose estimation results of Tool A (blue mesh) and Tool B (red mesh). Assess which one of the two meshes more accurately replicates the pose and shape of the human figure in the original image, taking into account their alignment and resemblance.

Assistant: Tool B's result is better.



Human: Please compare the pose estimation results of Tool A (blue mesh) and Tool B (red mesh). Assess which one of the two meshes more accurately replicates the pose and shape of the human figure in the original image, taking into account their alignment and resemblance.

Assistant: Tool A.

Fig. 8: Examples of the instruction-following data for discriminating between pose estimation results.

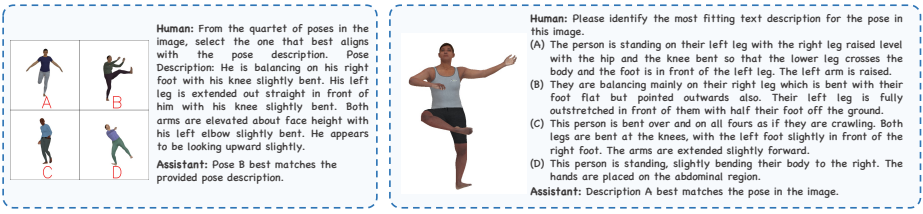
– **Interaction over Union** (IoU), which quantifies the percent overlap between the predicted text and ground-truth text.

Human Understanding. We use the following evaluation metrics to measure the performance of ChatHuman in human-related tasks:

– **Pose Estimation.** We adopt the same evaluation metrics as ChatPose [12] to evaluate the 3D pose estimation accuracy, including Mean Per-Joint Position Error (MPJPE), Mean Per-Joint Position Error after Procrustes alignment (PA-MPJPE), and Mean Per-Joint Rotation Error (MPJRE).
– **Pose Generation.** We use the evaluation metrics established in PoseScript [7], including the text-to-pose recall rate R^{P2T} and pose-to-text recall rate R^{T2P} of the retrieval models trained on real poses and evaluated on generated poses. We use the retrieval model from the jounal-version of PoseScript [7] and ChatPose [12] for the classical pose generation and speculative pose generation tasks, respectively.

A.3 Training Data Details

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 [68], we provide GPT-4 [41] with a textual description of an image from the COCO training set [31] and a tool-related prompt containing a description of tools. One of our key observations is that human-related tools often come with an academic paper containing rich background knowledge and varied applications, which are useful for the generation of user queries covering a wide range of application scenarios. Thus, we also incorporate the paper content into GPT-4 to generate the tool usage instruction-following data. To improve efficiency, we first prompt GPT-4 to summarize the paper content, rearticulate the tool functions and enumerate 50 potential user queries for tool activation. The details of the prompt are represented in Table 15. The summarized tool description and user queries will be fed to GPT-4 along with the image description to generate the instruction-following data about tool usage. Table 16 illustrates the prompt for the second step.
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 3D human datasets.



(a) Text-to-Pose Selection (b) Pose-to-Text Selection

Fig. 9: Examples of the instruction-following data for discriminating pose generation and pose description results.

– **Pose Estimation Results Discrimination.** To teach the LLM-based model to discriminate the pose estimation results from different tools, we build 17K pairs of instruction-following data based on the 3DPW [38] and MOYO [61] training sets. Specifically, we use HMR 2.0 [15] and CLIFF-SMPLify [2, 29] to predict the human mesh and calculate the reconstruction error between the predicted mesh and ground truth mesh. Based on MPVPE, we determine which tool is better for each image and construct instruction-following data as shown in Figure 8. Pose visualization results are rendered with Pyrender [39].

– **Pose Generation Results Discrimination.** The human pose generation tool, PoseScript [7], has multiple outcomes for each text input. Here we construct 44K pairs of instruction-following data to teach the multimodal LLM-based model to discriminate the multiple pose generation results. Specifically, we use PoseScript training data as the source and construct the data in two formats. The first one is about text-to-pose selection, as shown in Figure 9(a). Given a textual description, we visualize the corresponding pose and three other different poses from the training data and ask the agent to discriminate and choose the one that best aligns with the textual description. The second one is about pose-to-text matching, as shown in Figure 9(b). Given a 3D pose, we visualize it as an image by rendering the 3D body mesh in that pose. Then, we combine it with the corresponding text description and three other pose descriptions in the format of a multiple choice question. Finally, we ask the agent to choose the one that best describes the pose shown in the image.

– **Human Contact Detection Results Integration.** The outcome of the human contact prediction tool, DECO [60], is a vertex-wise contact prediction in a vector representation $y_c \in \mathbb{R}^{6890\times 1}$, which can not be directly used as input for our multimodal LLM baseline, LLaVA. To solve this problem, we transform the vertex-wise contact label of ground-truth and DECO's result into a textual description based on the vertex-to-part mapping dictionary of the SMPL model [36]. Subsequently, we feed the textual descriptions along with the RGB image from the DECO training set [60] into GPT-4V and prompt GPT4 [41] to generate instruction-following data about human-object interaction as shown in Figure 10. Notably, the transformed tool result

Human: Please describe the interaction between the person and any objects in as much detail as possible. The contact region detected by DECO are: right hand, left foot, right foot. Notably, the provided contact regions might not be complete and you should detect all the contact regions by yourself.

Assistant: The person is holding a tennis racket with his right hand, and his stance indicates that he may be in the action of swinging at a tennis ball. Both feet are planted on the ground, providing stability for the action he is performing.

Human: Are both feet of the person touching the ground or any object? The contact region detected by DECO are: right upper leg, left foot, back, right foot, left hand, right hand, left upper leg, hips. Notably, the provided contact regions might not be complete and you should detect all the contact regions by yourself.

Assistant: Yes, both the left and the right feet of the person are touching the ground.

Fig. 10: Instruction-following data about integrating results from human contact detection. The contact labels detected by DECO [60] are combined with the user query.

Human: what would be the appropriate clothes size for this person? Clues from tool: height: 1.65 m, weight: 65 kg, chest circumference: 1.03 m, waist circumference: 0.87 m, hips circumference: 1.24 m. Please note that these clues are provided as references and may not be entirely accurate.

Assistant: I would recommend a clothes size of L for the person.

Human: What is the body height and mass of the person in the image? Clues from tool: height: 1.82 m, weight: 80.2 kg, chest circumference: 1.03 m, waist circumference: 0.89 m, hips circumference: 1.01 m. Please note that these clues are provided as references and may not be entirely accurate.

Assistant: The person is about 1.98 m, 87.3 kg.

Fig. 11: Instruction-following data for integrating results from human body estimation. Measurements of the estimated body shape from BEDLAM-CLIFF [1] are added to the user query as clues.

is merged with the user query as a clue. The details of the prompt are shown in Table 17.

– **Body Shape Measurement Integration.** Similar to human contact prediction, the outcome of the body shape measurement tool is the SMPL body shape parameter $\beta \in \mathbb{R}^{10}$, which is also in a vector representation and can not be used by the LLM directly. Thus, we first convert the shape parameter into measurements based on the shape-to-measurement module from SHAPY [5] and represent it in a textual format. Subsequently, we feed the body measurement description along with attribute labels from the SHAPY training set into GPT-4 and prompt it to generate instruction-following data about human body shape as shown in Figure 11. Similarly, we merge the body measurement predicted by the tool with the user query as a clue. The prompt for GPT-4 is detailed in Table 18.

A.4 Benchmark Details

Tool Usage Benchmark. To evaluate the tool usage accuracy of our method, 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. Similar to the training data construction, we feed a textual description of an image from the COCO validation set, a tool description, and some examples summarized from the tool paper into GPT-4 and prompt it to generate instruction-following data about tool usage. We use the image description captioned by LLaVA [32] instead of the original image captions

to ensure a difference between training and test sets. Finally, we manually check the question-answering pairs to ensure the accuracy of the benchmark.

MixPose Benchmark. To validate whether the multimodal LLM-based agent can discriminate the pose estimation results from different tools, we build a new benchmark, MixPose. Considering that different tools excel in different scenarios, a benchmark covering diverse scenarios and corner cases is needed. To construct this benchmark, we selected 100 images featuring extreme camera angles from the MoYo [61] test set, 100 full-body images from the 3DPW test set, and 100 images with significant truncation also from the 3DPW test set. This approach ensures our benchmark includes typical in-the-wild images, shots taken from extreme views, and images with heavy truncation. This diverse distribution allows us to test whether the agent can accurately choose the right tool based on sceneries of the image and the tool performance. To get the truncated images from 3DPW, we resize the human bounding box by 2/3 and crop the human image based on the rescaled human bounding box.

B Additional Ablation Studies

Paper Components. To improve the tool usage accuracy, we propose a paper-based RAG mechanism. Here we conduct an ablation study to analyze the effects of each component of the paper for instructing tool usage. The baseline model is derived by removing the RAG operation and is trained with the instruction-following data constructed based on the manually defined tool descriptions and examples (T), without referring to paper content. We then add the tool descriptions and examples summarized from different paper components into the prompt and request GPT-4 to generate new instruction-following data. The paper components include the abstract (A), introduction (I), related work (R), method (M), and experiment section (E). The results are presented in Table 10. As shown, incorporating the paper content consistently improves the accuracy of tool usage. Interestingly, feeding the abstract and introduction of the paper achieves the best performance, and adding additional paper components like the experiments does not result in further improvement. This is expected since the abstract and introduction have covered the tool function and potential applications in most cases. We also find that examples generated based on paper content can be noisy in some cases, and thus we manually check the examples and remove the incorrect samples before feeding them to GPT-4. This data cleaning operation improves the accuracy.

Base Model for Value Extraction. During our body shape measurement and human contact detection experiments, the answer of ChatHuman is a sentence and thus could not be directly used to compute the evaluation metrics. Thus, we need to prompt an LLM to extract the value from the sentence and compare the extracted value with the ground truth label to calculate the metric. The prompts for body shape measurement and contact estimation are shown in Table 13. Here we use two different LLM models for value extraction and report the results in

T	Paper A	I	R	M	E	C	Seen Tools SR_t	SR_{act}	SR_{args}	SR	IoU	Unseen Tools SR_t	SR_{act}	SR_{args}	SR	IoU
✓						✓	1.0	0.97	0.93	0.96	0.96	0.95	0.89	0.78	0.82	0.87
✓	✓						1.0	0.97	0.95	0.97	0.97	0.99	0.94	0.85	0.90	0.93
✓	✓	✓					1.0	0.97	0.95	0.97	0.97	1.0	0.97	0.86	0.91	0.94
✓	✓	✓	✓				1.0	0.98	0.95	0.97	0.97	1.0	0.97	0.84	0.91	0.93
✓	✓	✓	✓	✓			1.0	0.98	0.94	0.97	0.97	0.99	0.95	0.82	0.87	0.92
✓	✓	✓	✓	✓	✓		1.0	0.98	0.95	0.97	0.97	1.0	0.96	0.86	0.91	0.94
✓	✓	✓				✓	1.0	0.97	0.95	0.97	0.98	1.0	0.97	0.89	0.95	0.95

Table 10: Ablation study on the impact of each paper component in the paper-based RAG mechanism. T denotes tool description, A, I, R, M, E are abstract, introduction, related work, method, experiment section from the paper, and C denotes the data after manually cleaning. Successful rate of thought (SR_t), action (SR_{act}), arguments (SR_{args}), execution (SR), and IoU are reported.

Evaluator	Precision ↑	Recall ↑	F1 Score ↑
GPT-3.5	0.67	0.67	0.63
GPT-4	0.69	0.69	0.64

(a) HOI Contact Detection.

Evaluator	Height ↓	Weight ↓	Chest ↓	Waist ↓	Hip ↓
GPT-3.5	6.7	10.4	6.1	13.0	6.4
GPT-4	6.7	10.4	6.1	13.0	6.4

(b) Body Shape Measurement.

Table 11: Ablation study on the base model for value extraction.

Table 11. As shown, the choice of the LLM model for value extraction does not introduce a significant difference.

Text Embedding Model in Retrieval. During the paper-based retrieval mechanism, we retrieve a relevant example by matching the text embedding of the query and those within the tool documents using a pretrained text embedding model [54]. Here we conduct an ablation study of the text embedding model to analyze the robustness of our RAG mechanism. We experiment with two models, i.e., instructor-xl and instructor-base. As shown in Table 12, the performance of our method is not greatly affected by the text embedding model, which demonstrates the robustness of ChatHuman.

Tool Use in Multi-turn Dialogue. In many real-world scenarios, the models need to correctly use the tool based on the context within a multi-turn dialogue. Here, we build a benchmark to evaluate the tool use accuracy within multi-turn conversations. We randomly select several single-turn question-answer pairs from the original multi-turn tool use benchmark and ask GPT-4 to merge them into a multi-turn dialogue. We then evaluate the performance of GPT4Tools [68], Visual ChatGPT [49], and ChatHuman on the built benchmark, which contains 1000 instruction-response pairs. As shown in Table 14, ChatHuman achieve a superior performance in the multi-turn setting, demonstrating its capacity to understand the comprehensive context information within the multi-turn dialogue and correctly use the tool to solve the problems.

C More Results

Motion-related Tools. We also integreta motion tools for tasks related to human motions like text-to-motion generation. Specifically, we utilize TMR [44]

Method	Seen Tools					Unseen Tools				
	SR_t	SR_{act}	SR_{args}	SR	IoU	SR_t	SR_{act}	SR_{args}	SR	IoU
instructor-base	1.000	**0.975**	0.947	**0.972**	0.974	0.997	0.950	0.884	0.949	0.949
instructor-xl	**1.000**	0.974	**0.950**	0.970	**0.975**	**0.999**	**0.967**	**0.893**	**0.954**	**0.953**

Table 12: Ablation study of text embedding model for RAG. Successful rate of thought (SR_t), action (SR_{act}), arguments (SR_{args}), execution (SR), and IoU are reported.

(a) You are an AI assistant. Your input will be a description of body measurements, including height, weight, chest circumference, hip chest circumference, and waist circumference. Your task is to extract the value of each attribute and return a result like:

"height: 1 m, weight: 1 kg, chest circumference: 1 m, waist circumference: 1 m, hip circumference: 1 m"

If there is no measurement value, return "There is no measurement value."

(b) You are an AI assistant. Your input will be a description of the human-object interaction information of a person. Your task is to extract the body parts that contact with objects and return a list. Consider the following possible body parts: right hand, right upper leg, left arm, left leg, left foot, back, left shoulder, right shoulder, right foot, head, right arm, left hand, right leg, left forearm, right forearm, neck, left upper leg, hips.

Table 13: Prompt to extract the target values from a sentence generated by ChatHuman for metric computation. Prompt (a) is used to request GPT-3.5 for body measurement values. Prompt (b) instructs GPT-3.5 to extract the body part names.

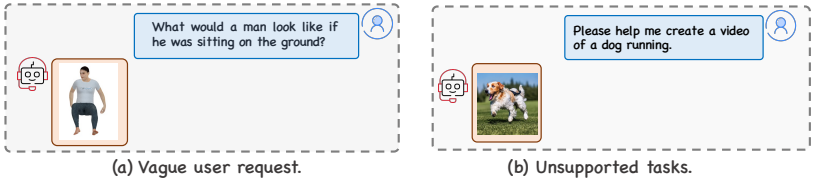


Fig. 12: Failure cases of ChatHuman. When the user inputs a vague query or requests a task that could not be finished by current tools, ChatHuman might fail.

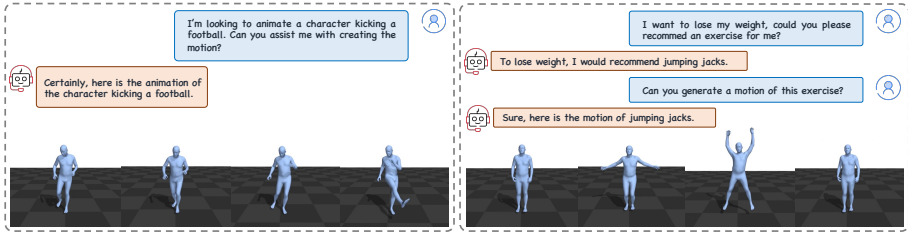


Fig. 13: Text-to-motion generation examples.

to generate 3D human motions from text input. Figure 13 shows two examples about synthesizing human motion from a textual description.

	SR_{args}	SR	IoU	SR_t	SR_{act}
GPT4Tools [67]	0.582	0.551	0.553	0.513	0.612
Visual ChatGPT-3.5 [49]	0.438	0.203	0.162	0.173	0.691
Visual ChatGPT-4 [49]	0.860	0.794	0.711	0.744	0.789
ChatHuman	**1.000**	**0.959**	**0.927**	**0.955**	**0.962**

Table 14: Comparison of tool usage accuracy within multi-turn dialogue. Successful rate of thought (SR_t), action (SR_{act}), arguments (SR_{args}), execution (SR), and IoU are reported.

Failure Cases and Limitations. There are mainly two kinds of failure cases of ChatHuman. Firstly, ChatHuman may fail in certain calling scenarios, particularly when the user request is vague, as also shown in Figure 7 in the main paper. Secondly, when the user query requires a task (e.g., video generation) that is not supported by the tools in the tool libraries, ChatHuman might incorrectly choose a tool to generate a response, instead of informing the user that the task cannot be completed and recommending the creation or introduction of a new tool. As shown in Figure 12(b), there is a need for tools for video generation. Future developments may include expanding to tools related to these scenarios.

You are an AI visual assistant tasked with analyzing a paper on a method in the field of 3D human modeling. Your goal is to extract key information about the method—its name, purpose, uses, and potential application scenarios. Based on this, you need to succinctly define the method in the following formats "Method name is a tool to do something. Useful when you want to do something. Like: user query."

Subsequently, craft 50 diverse, realistic user prompts that indirectly pertain to using this method. These queries should be framed as questions, demands, or scenarios from consumers who are unaware of the method's name but whose needs align with its capabilities. Assume that consumers have an image and seek assistance in achieving a task related to the image using this method. Each prompt should introduce the task of the user in an imperative tone. The prompt should specify and refer to the image.

Here is one example:

Method definition:
name="HMR2.",
description="HMR2.0 is a tool to estimate the 3D pose and shape of the person in the image. Useful when you want to detect poses of the humans in the image. Like: estimate the human poses in the image."

Possible queries:
1. Can you help me estimate the pose of the person in the photo?
2. Please reconstruct a 3D avatar for the person in the image.
3. Could you please estimate the SMPL parameter of the man in the picture?
4. Someone in the image is riding a bicycle, could you please help me estimate her pose?

Table 15: Prompt to request GPT-4 to summarize paper content, rearticulate tool functions, and enumerate possible user queries for tool activation.

Given an image whose image path is "example.jpg". Image caption: "{caption}". The image caption includes detail image description and each object paired with the bounding box [x1, y1, x2, y2]. For the bounding box, (x1, y1) refers to the top left, and (x2, y2) refers to the bottom right. x1 less than x2, and y1 less than y2.

Below are 22 visual tools. Each tool is defined as "tool name: usage scenario, and tool arguments".

Please generate 3 visual instructions for each tool, so you need to generate 66 visual instructions in total.

The generated instructions should follow the format of "instruction content, [tool name, tool arguments]". Each instruction must relate to the caption and can be solved by the tool.

You can not revise the "tool name", or add any other fake tools that is not defined. You must keep the correct "tool arguments".

Tools: {tool description}

Note that your generated visual instructions should be highly related to the image caption. Directly reply to me with the list, here are some examples: {examples}

Diversify the instructions to cover a wide range of possible user queries. Feel free to adapt and rephrase the examples provided to generate diverse, complex, and deceptive instructions as much as possible.

For example, you can also change the subject position or the person and pose description positions. Don't use too much imperative sentence, you should also use interrogative sentence.

Table 16: Prompt to request GPT-4 to generate instruction-following data about tool usage based on the image description, tool description, and tool usage examples summarized from tool paper.

You are an AI visual assistant, and you are seeing a single image and a sentence about the human-object contact regions of the person in the image. The sentence include the human-object contact body parts of the person. Notably, the provided contact regions might not be complete and you should detect all the contact regions by yourself.

Design a conversation between you and a person asking about the human-object contact information of the person. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question. Ask diverse questions and give corresponding answers.

Include questions asking about the person's human-object contact information, etc. Only include questions that have definite answers: (1) one can see the content in the image that the question asks about and can answer confidently; (2) one can determine confidently from the image that it is not in the image. Do not ask any questions that cannot be answered confidently.

Provide detailed answers when answering complex questions. In your answer, you should imitate as if you see the image and the contact regions are estimated by you. You should only ask questions about the human-object interaction. The answer should be as detailed as possible. Don't mention any other irrelevant information! Directly reply to me with a list, here are some examples:

1. Please help me detect the contact regions of the person in the image. [The person's contact region includes his feet, hands, and back. His feet touch the ground and his hands are holding a mobile phone.]

2. Does the person's hand contact any objects? [Yes, her right hand holds an umbrella.]

3. Describe the human-object interaction information of the person, as detailed as possible. [The person holds a phone with his left hand and stands on a skateboard with both their left and right feet.]

Notably, at least one question is to ask all the contact regions of the person. In your answer, you should distinguish and specify the left and right body parts. Notably, you should distinguish based on the body pose and orientation. If the person is facing the camera, the hand, foot, and ear on the left side of the image is the person's right hand, right foot, and right ear, and the one on the right side of the image is the person's left hand, left foot, and left ear. If the person has their back to the camera, the one on the left side of the image is the person's left body part, and the one on the right side of the image is the person's right body part.

Table 17: Prompt to request GPT-4V to generate instruction-following data about human-object interaction based on the textual contact description and RGB image.

You are an AI visual assistant, and you are seeing a single image. What you see are provided with a sentence, describing the body shape of the person in the image. Answer all questions as you are seeing the image.

The sentence includes information about the person's gender, body mass, height, chest circumference, waist circumference, and hip circumference. Besides, it includes 15 linguistic shape attributes scale from 1 (strongly disagree) to 5 (strongly agree).

Design a conversation between you and a person asking about the body shape of the person. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question.

Ask diverse questions and give corresponding answers.

Include questions asking about the visual content of the image, including the person's overall body fit, shape, height, mass, etc. Only include questions that have definite answers:

(1) one can see the content in the image that the question asks about and can answer confidently;

(2) one can determine confidently from the image that it is not in the image. Do not ask any questions that cannot be answered confidently.

Provide detailed answers when answering complex questions. When the question is about the measurement, provide an explicit and concrete metric number in the answer.

In your answer, you should imitate as if you see the image and the measurements and linguistic attributes are estimated by you. The linguistic attribute score is only used to help you understand and don't mention it in your answer.

Directly reply to me with a list, here are some examples:

1. How tall is the person in the image? [The person looks quite tall. He is about 1.85 m.]

2. Please help me estimate the body measurements of the man in the image. [The man is about 1.74 m and 60 kg. His chest circumference is about 0.9 m.]

3. What's the waist circumference of the person? [The chest circumference is about 0.95 m.]

Table 18: Prompt to request GPT-4 to generate instruction-following data about human body shape based on the textual description about human body.