SOLAMI: Social Vision-Language-Action Modeling for Immersive Interaction with 3D Autonomous Characters

Supplementary Material

A. Future Work

Our work, SOLAMI, represents a preliminary exploration for building 3D autonomous characters. While it has performed well in comparative experiments, there remains significant room for improvement on aspects as follows:

- **Input Modality:** For dyadic social interaction, using the user's body motion and speech as input is sufficient. However, when considering multi-person interaction or interaction involving the environment and objects, video [16, 47] or dynamic 3D scenes [31] might be a better choice;
- **Data Collection:** Our synthetic dataset, SynMSI, enables satisfactory user evaluation results. However, collecting real-time data of actual dyadic interaction could enable our model to generate more precise and natural body language and speech, while also supporting duplex streaming conversations, similar to [5, 46]. Compared to text and video modalities, the collection of embodied 3D data is undoubtedly challenging. Potential solutions include: capturing [9] or learning human behavioral data [6] from existing video datasets, building immersive interaction platforms [34] to gather data on human interactions, and using surrogate control to collect data from human interactions with 3D characters [14];
- **Cross Embodiment:** Using a unified SMPL-X [30] model to represent characters' motion inevitably introduces challenges in cross-embodiment for different characters. While some degree of error and misalignment may not hinder information exchange in social language interaction, such representations clearly lack generalizability for fine-grained tasks (*e.g.*, handshaking, object manipulation). The challenges of retargeting in 3D human-related tasks and cross-embodiment in robotics [47] share similarities, providing opportunities for mutual inspiration and methodological exchange;
- **Long-Short Term Design:** Although SOLAMI demonstrates effective modeling for real-time interactions, its architecture encounters challenges such as computational redundancy, forgetting, and training difficulties during extended social interactions. A promising direction [10, 15] to explore is integrating long-term memory, knowledge, and skills with short-term real-time interaction. This approach could ensure interaction quality while reducing computational overhead and simplifying the training process;
- **Efficient Learning Method:** Although our dataset, SynMSI, tries to collect large-scale motion data, the inherently long-tail distribution [45] of human motions results in some behaviors having very low occurrence frequencies [19, 21, 41]. In particular, the data volume for signature actions of 3D characters is inherently limited. While models like GPT-3 [8] have demonstrated remarkable few-shot learning capabilities, the data-intensive training required is currently unsustainable in the field of digital humans. Therefore, exploring effective learning methods is essential. Leveraging character-focused knowledge embedded in existing foundation models [40, 42] or incorporating human evaluators [28] to guide the model in learning new skills from a small number of samples are promising research directions.

B. More Details of Architecture Design

In this section, we discuss the input and output modalities of SOLAMI in Appendix B.1, compare the motion representation in Appendix B.2, and introduce details of our motion tokenizer and pre-training design in Appendix B.3.

B.1. Input and Output Modalities

Our ultimate goal is to establish a unified behavioral modeling system for any character, where input modalities include a wide range of sensory observations, including vision, audio, and haptics *etc.*, and output modalities represent actions in the finest possible granularity. However, currently, we need to balance the ideal with the constraints of existing data and devices to develop a model that provides an optimal user experience.

Regarding devices, we employ VR headsets instead of mobile phones or computers because VR headset enables a more immersive interactive experience by capturing and presenting richer information.

In terms of input modalities, while 3D scenes or videos could serve as input and have some foundational models [23, 31], collecting corresponding social interaction data is challenging. For instance, datasets like Ego4D [17] and Ego-Exo4D [18] capture first-person videos and motion data but include very limited social interaction content and no data involving character interaction. Within VR environments, the majority of incremental information a character can observe comes from user's behaviors that VR devices can capture. Consequently, we chose user motion and speech as the primary input for SOLAMI.

Similarly, for easy synthetic data generation and model training, we maintain the same types of output modalities for the character as for the user's input. This symmetry en-

Table 1. **Quantitative results of pre-training on text-to-motion task**. '↑'('↓') indicates that the values are better if the metrics are larger (smaller). The best results are in bold and the second best results are underlined.

ID	Body & Hand	Repre	Backbone	Token Interleaved	Metrics			
					FID↓	Diversity↑	PA-MPJPE↓	Pred Valid↑
1	bind	joints	GPT-2	-	**1.48**	9.03	148.00	<u>0.836</u>
2	bind	rotation	GPT-2	-	3.44	<u>12.94</u>	143.70	0.813
3	separate	rotation	GPT-2	Yes	3.00	11.64	117.26	0.676
4	separate	rotation	GPT-2	No	2.72	**14.05**	<u>112.53</u>	0.638
5	separate	rotation	Llama2	No	<u>1.82</u>	10.40	**110.23**	**0.999**

Table 2. **Quantitative results of Motion VQVAE**. '↑'('↓') indicates that the values are better if the metrics are larger (smaller). The best results are in bold.

ID	Body & Hand	Repre	Motion Metrics	
			PA-MPJPE↓	FID↓
1	separate	joints	87	**1.0**
2	bind	joints	**80**	1.3
3	separate	rotation	88	1.88
4	bind	rotation	113	2.34

sures alignment between what the model observes and what it produces, facilitating a more natural and precise interactive experience.

B.2. Motion Representation Comparison

Common representations of human motion are often based on 3D keypoints [19, 22, 26], which provide higher precision compared to methods based on joint rotations. However, this approach is inconsistent with the driving mechanism of 3D engines such as Unity Engine. When the model generates 3D keypoints, retargeting is necessary to derive the relative rotation of each joint with respect to its parent joint. Considering human motion priors, a typical approach [29] involves fitting an SMPL-X [30] model to the 3D keypoints using optimization strategies, and subsequently retargeting the fitted SMPL-X model to the character. However, this process has two main drawbacks:

1. **Time-Consuming Fitting Process:** The fitting step is computationally intensive. With optimized methods like SMPLify [29], achieving an adequate result requires about 1 second of iteration on a V100 GPU.
2. **Fitting Artifacts and Distortion:** Inevitable fitting errors can lead to biologically implausible joint rotations, significantly degrading visual quality.

In our experiments, we observed that while human motion representation based on 3D keypoints performs well in terms of motion metrics, as shown in Tab. 1 and Tab. 2, its visual fidelity is inferior to representation based on joint rotations. To address this, we adopted a cont6d representation for joint rotations, achieving improved visual outcomes.

B.3. Motion Tokenizer and Pre-training

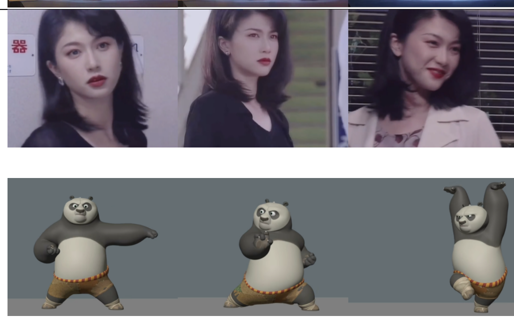
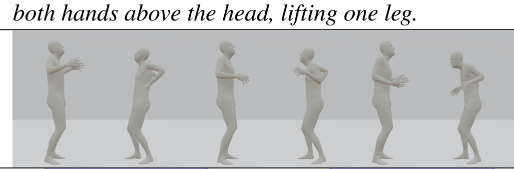
After processing as described in Appendix B.2, we obtained a 315-dimensional motion representation. When converting this motion representation into tokens via the tokenizers, several issues need to be discussed. Should body and hand motion features be represented separately? If so, how should their tokens be handled? Should the tokens for the body and hand motions be interleaved, or should they be input as independent sequences in the pre-training stage?

Considering our computational cost, we conducted ablation experiments on the text-to-motion task using the GPT-2 [33] backbone as the baseline model. Finally, we compared the models under the same settings using Llama2-7B [37] as the backbone.

As shown in Tab. 2 and Tab. 1, compared to unified representations of hand and body motion (marked as "bind"), the separate representation (marked as 'separate") achieves better performance, particularly with higher precision on the text-to-motion task (t2m). However, the trade-off is that the probability of GPT-2 [33] producing outputs that conform to the expected format (marked as "Pred Valid") decreases. However, this issue is mitigated in large part by using Llama2 [37] as the backbone model. We think this improvement is due to the differences in the language models: GPT-2, the relatively smaller language model, has weaker comprehension of textual instructions. In contrast, Llama2, trained on extensive corpora, demonstrates significantly stronger text understanding capabilities. Moreover, compared to interleaved tokens ("Yes" for "Token Interleaved"), separate sequence representations ("No" for "Token Interleaved") achieve better motion metrics. We hypothesize that this is because learning separate sequences reduces the overall complexity of the motion pre-training task, thereby improving performance.

Based on the above experimental evaluations, we ultimately select Llama2-7B [37] for its strong text comprehension capabilities as the LLM backbone. For processing motion representation, we employ separate motion tokenizers that convert the motion representation into noninterleaved token sequences. This configuration is used for the final instruction fine-tuning stage.

Table 3. **Methods of collecting multimodal interaction data.**

Methods	Input	Output
MoCap Human Motions from Internet Videos with SMPLer-X [9]		
Motion Captioning on Internet Videos with GPT-4o [27]		*1-3s: Turn head to the right and look straight ahead, with a neutral expression; 4-5s: Turn body and look sideways, with a serious expression, almost no movement; 6-8s: Turn to the left side, smiling while looking forward.* *1-2s: A panda in a combat stance, right hand raised in a fist, left hand extended, with a serious facial expression; 3s: Panda's body tilts to the left side, right hand clenched in a fist, left hand stretched forward, eyes looking to the right front; 4-5s: Panda raises both hands above the head, lifting one leg.*
Real Data Collection from VR Platforms		
Synthetic Data Generation from Existing Datasets		

C. More Details of Data Generation

In this section, we first discuss several methods for collecting multimodal social interaction data in Appendix C.1. Then, we introduce the technical details of SynMSI generation pipeline in Appendix C.2.

C.1. Comparison of Data Collection Methods

From the perspective of data sources, we discuss three sources: internet videos, Immersive VR platform, and existing incomplete motion capture datasets, as shown in Tab. 3.
Collecting from Internet Videos. The development of mobile devices has led to an explosion of video content, and researchers naturally expect the model to learn knowledge and capabilities from internet videos. Many works aim to implicitly learn human capabilities from videos [13, 39], but for our task, we anticipate obtaining explicit multi-modal interactive data through various tools [9, 27]. Human motions can be captured through video motion capture, but current video motion capture [9] faces challenges such as occlusion, temporal discontinuity, and long-tail problems, making it difficult to obtain high-quality motions. Understanding and annotating human behaviors in videos can be achieved using Vision-Language Models (VLM) [25, 27], and we find that with appropriate post-processing these annotations are usable. Additionally, there is another issue: the data obtained through this method lacks first-person view and is often fixed at a third-person view, which presents challenges in perspective transformation.

Collecting from VR Platforms. Building a VR interaction platform to directly collect user interaction data is the most straightforward method. However, two key problems arise: 1) Current VR devices' body tracking systems [38] cannot provide ground truth-level data. For instance, existing VR devices estimate lower body postures instead of capturing with wearable sensors, and tracking becomes unreliable when hands move beyond the sensor range of VR equipment. 2) Human interaction data differs from 3D character representations. Specifically, animated characters' movements tend to be more exaggerated compared to real human motions, which naturally introduces a data distribution gap.

Collecting from Existing Incomplete Datasets. Due to the novelty of our task, there is no dataset that perfectly suits our needs. Common open-source datasets [19, 24, 41] typically provide semantic annotations for motion sequences or co-speech gestures. The most cost-effective and convenient approach is to complete these datasets or use them to synthesize multimodal social interaction datasets. However, this faces several challenges: How can we ensure the diversity of dialogue content? How can we ensure that synthesized speech and motion are reasonable? Can synthetic data guarantee user satisfaction? We address these questions in Sec. 4 and Sec. 6.3 of the main paper. And we will introduce some technical details about data synthesizing latter.

In summary, obtaining data from the internet has high potential, but current video motion capture technology is insufficient to realize this potential, and it also involves perspective transformation challenges. Data collection from VR platforms is limited by hardware capabilities and faces difficulties in replicating character behaviors. Synthesizing data based on existing datasets represents an optimal choice when balancing cost and effectiveness.

C.2. Details of SynMSI Generation Pipeline

Motion Post-process Existing motion-text datasets [10, 41] primarily provide semantic-level text annotations, often overlooking behavioral details (such as sitting versus standing positions, orientations, *etc.*). Considering GPT-4o's capability [27] in understanding human behaviors in videos, as shown in Tab. 3, one approach would be to render all motions into videos and then use VLM for annotation. However, for a small research team, the cost of VLM API calls is relatively high. We propose a compromise strategy: combining multiple text annotations for a single motion and using GPT-4o [27] to generate a comprehensive, detailed description. In practice, we find this method to be quite effective.

Topics Collection. Without topic guidance, conversations with LLMs often converge to simple, generic content rather than character-specific, in-depth content [10, 35]. Using prompts to guide conversation is a common strategy. We collected topics from the following perspectives:

1. Character-related topics: These topics are difficult to collect in bulk from the internet and were generated through GPT-4o [27] brainstorming;
2. News-related topics: Google Trends [2] has compiled many news topics that people care about in daily life;
3. Daily life topics: Some community websites, such as Jike, specifically curate such topic content;
4. Topics people are curious about: Common Q&A websites (such as Quora, Zhihu [3]) specifically organize these topics.

(a) Samantha (b) K-VRC

(c) Batman (d) Banaya

Figure 1. Word cloud visualization of the keywords in the collected characters' topics.

After collecting these topics, we used LLMs to post-process them, filtering and organizing them into topics suitable for character conversation. Topic keywords are shown in Fig. 1.

Task Generation. Beyond daily conversation content, we also want SOLAMI to learn direct understanding of human body language and the ability to explicitly follow human instructions. For this purpose, when synthesizing data, we set up different tasks in the system prompt:

- **common:** daily conversation;
- **motion understanding:** requires users to generate motions with strong semantic information, and the character can clearly express understanding of body movements;
- **instruction following:** requires users to give clear motion instructions, and the character can output corresponding instructed movements;
- **imitation:** requires the character to imitate user's motion.

Script Generation Methods. Since we are using the chat version of LLMs, we experimented with and compared three script generation strategies:

1. Method 1: Round-by-Round completion: Using LLM to complete and refine the speech and motion text for each character round by round, which is the method mentioned in our main paper.
2. Method 2: Character Agent Dialogue: Similar to the SocioMind approach [10], using two LLMs to play two roles (User and Character), and alternately outputting speech and motion text, followed by refinement.
3. Method 3: One-shot generation: Generating the entire multi-turn dialogue script at once, then revising the script round by round based on retrieved motions.

According to our experimental results, Method 1 and Method 2 produce better results. Although Method 3 initially generates good scripts, the quality deteriorate after multiple rounds of modifications during motion-text database alignment. To produce SynMSI, we randomly alternate between Methods 1 and 2 to generate text scripts.

Interactive Motion. If we only use single-person motions, our model would lack the capability for two-person interaction. To address this issue, during script generation, when we retrieve a motion of one person in an interactive motion, we ask the LLM whether to use the motion of another person from the same interactive motion when generating the next round of motion text.

D. More Details of Experiments

D.1. LLM Selection

We chose Llama2-7B [37] because at the time of our experiments, end-to-end models with speech pre-training were scarce, with AnyGPT [43] being one of the few that performed well. Thus we selected the Llama2 series as the backbone for fair comparison in subsequent experiments. Readers aiming to achieve the best results can certainly choose state-of-the-art models as the backbone.

The Llama2-7B-chat model [37] tends to output increasingly longer dialogue content, which for *LLM+Speech* methods results in high inference latency from both LLM and TTS (sometimes exceeding 30 seconds). Therefore, through post-processing, we truncate the output content to a maximum of 3 sentences. While truncating output content somewhat affects user experience, the lower user latency generally results in a better overall experience.

D.2. Voice Cloning Comparison

Voice cloning / TTS has numerous available products and open-source models in both industry and academia, each with different focuses. We aim to achieve the best voice cloning effect in near real-time conditions. For this purpose, we compare these software and algorithms: ElevenLabs Instant Voice Cloning [1], ChatTTS + OpenVoice [4, 32], XTTS_v2 [12], MARS5 [11], and Bark [36]. Among them, MARS5 [11] uses a diffusion [20] framework and is relatively slow; ElevenLabs [1] produces the best results but has high API costs and tends to generate speech at a faster pace. XTTS_v2 [12] is a more suitable option, and can achieve a good balance between speed and quality.

When SOLAMI processes speech, we use the pre-trained SpeechTokenizer [44] and SoundStorm [7] from AnyGPT [43]. In SpeechTokenizer [44], one second of speech is encoded into 400 tokens across 8 layers. We only select tokens from the first semantic layer (50 tokens in total) to send to SOLAMI for processing. During SoundStorm [7] decoding, we choose 4 to 6 seconds of voice prompt based on the character and generate the speech with 4 iteration steps.

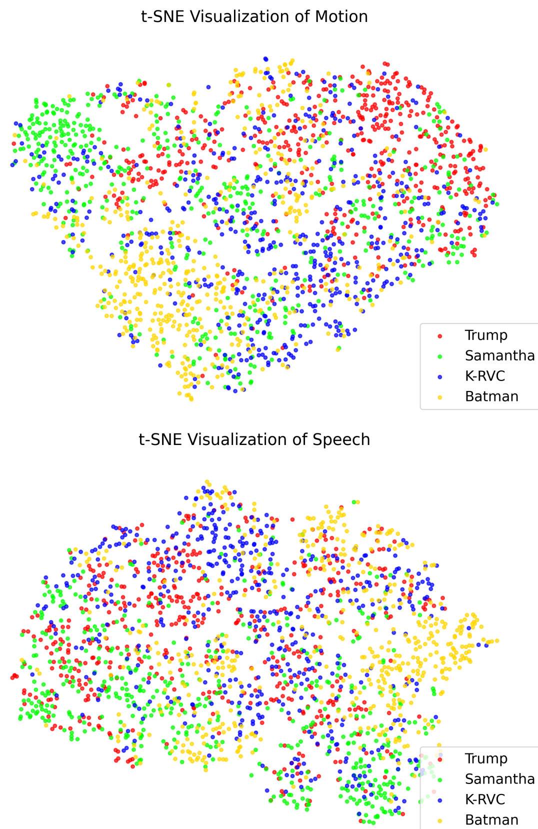


Figure 2. t-SNE visualization of generated motion and speech.

D.3. Additional Experimental Results

To visually demonstrate the diversity of motion and speech generated by SOLAMI, we used the speech and motion data stored on the server in the user study and performed t-SNE analysis with the features extracted by the encoder in the tokenizers. Results shown in Fig. 2 indicate the characters indeed have character-specific behaviors.

The average response latency of SOLAMI's VR demo with two H800s is 2.588 s. Specifically, the response process consists of: motion & speech tokenization (0.125s), LLM inference (1.926s), motion & speech decoder (0.187s), audio-to-face (0.353s), motion retargeting (0.032s), rendering (50 FPS).

References

[1] Elevenlabs. https://elevenlabs.io/. 5

[2] Google trends. https://trends.google.com/trends. 4

[3] Zhihu. https://www.zhihu.com. 4

[4] 2noise. Chattts: A generative speech model for daily dialogue. 2024. 5

[5] Tenglong Ao. Body of her: A preliminary study on end-

to-end humanoid agent. *arXiv preprint arXiv:2408.02879*, 2024. 1

[6] Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024. 1

[7] Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. Soundstorm: Efficient parallel audio generation. *arXiv preprint arXiv:2305.09636*, 2023. 5

[8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020. 1

[9] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. Smpler-x: Scaling up expressive human pose and shape estimation. In *NeurIPS*, 2023. 1, 3

[10] Zhongang Cai, Jianping Jiang, Zhongfei Qing, Xinying Guo, Mingyuan Zhang, Zhengyu Lin, Haiyi Mei, Chen Wei, Ruisi Wang, Wanqi Yin, Liang Pan, Xiangyu Fan, Han Du, Peng Gao, Zhitao Yang, Yang Gao, Jiaqi Li, Tianxiang Ren, Yukun Wei, Xiaogang Wang, Chen Change Loy, Lei Yang, and Ziwei Liu. Digital life project: Autonomous 3d characters with social intelligence. In *CVPR*, 2024. 1, 4

[11] CAMB.AI. Mars5: A novel speech model for insane prosody. 2024. 5

[12] Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, et al. Xtts: a massively multilingual zero-shot text-to-speech model. *arXiv preprint arXiv:2406.04904*, 2024. 5

[13] Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, Hanbo Zhang, and Minzhao Zhu. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation, 2024. 3

[14] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-television: Teleoperation with immersive active visual feedback. 2024. 1

[15] Konstantina Christakopoulou, Shibl Mourad, and Maja Matarić. Agents thinking fast and slow: A talker-reasoner architecture. *arXiv preprint arXiv:2410.08328*, 2024. 1

[16] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In *ICML*, 2023. 1

[17] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolár, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard A. Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3, 000 hours of egocentric video. In *CVPR*, 2022. 1

[18] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zachary Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, María Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Dutt Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J. Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina González, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbeláez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard A. Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shout, and Michael Wray. Ego-exo4d: Understanding skilled human activity

from first- and third-person perspectives. In *CVPR*, 2024. 1

[19] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, 2022. 1, 2, 4

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 5

[21] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE TPAMI*, 2014. 1

[22] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. In *NeurIPS*, 2023. 2

[23] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 1

[24] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. BEAT: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *ECCV*, 2022. 4

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 3

[26] Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Humantomato: Text-aligned whole-body motion generation. In *ICML*, 2024. 2

[27] OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 3, 4

[28] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022. 1

[29] 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

[30] 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. 1, 2

[31] Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction. In *ECCV*, 2024. 1

[32] Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. Openvoice: Versatile instant voice cloning. *arXiv preprint arXiv:2312.01479*, 2023. 5

[33] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 2019. 2

[34] David Saffo, Caglar Yildirim, Sara Di Bartolomeo, and Cody Dunne. Crowdsourcing virtual reality experiments using vrchat. In *CHI*, 2020. 1

[35] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-llm: A trainable agent for role-playing. In *EMNLP*, 2023. 4

[36] suno.ai. Chattts: A generative speech model for daily dialogue. 2023. 5

[37] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2, 5

[38] Alexander W. Winkler, Jungdam Won, and Yuting Ye. Questsim: Human motion tracking from sparse sensors with simulated avatars. In *SIGGRAPH Asia*, 2022. 3

[39] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *ICLR*, 2024. 3

[40] Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas J. Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v(ision) is a human-aligned evaluator for text-to-3d generation. In *CVPR*, 2024. 1

[41] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, Yunhui Liu, Wenjun Zeng, and Xiaokang Yang. Inter-x: Towards versatile human-human interaction analysis. In *CVPR*, 2024. 1, 4

[42] Wenhao Yu, Nimrod Gileadi, Chuyuan Fu, Sean Kirmani, Kuang-Huei Lee, Montserrat Gonzalez Arenas, Hao-Tien Lewis Chiang, Tom Erez, Leonard Hasenclever, Jan Humplik, Brian Ichter, Ted Xiao, Peng Xu, Andy Zeng, Tingnan Zhang, Nicolas Heess, Dorsa Sadigh, Jie Tan, Yuval Tassa, and Fei Xia. Language to rewards for robotic skill synthesis. In *CoRL*, 2023. 1

[43] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yu-Gang Jiang, and Xipeng Qiu. Anygpt: Unified multimodal LLM with discrete sequence modeling. In *ACL*, 2024. 5

[44] Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech large language models. *arXiv preprint arXiv:2308.16692*, 2023. 5

[45] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE TPAMI*, 2023. 1

[46] Zhipu. Glm-4-voice. 2024. 1

[47] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong T. Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: vision-language-action models transfer web knowledge to robotic control. In *CoRL*, 2023. 1