References

[1] Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[2] Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards understanding mixture of experts in deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[3] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 5

[4] CMU MOCAP. CMU Graphics Lab Motion Capture Database. https://http://mocap.cs.cmu.edu/, 2010. Accessed: 2023-10-25. 5

[5] Jieming Cui, Ziren Gong, Baoxiong Jia, Siyuan Huang, Zilong Zheng, Jianzhu Ma, and Yixin Zhu. Probio: A protocol-guided multimodal dataset for molecular biology lab. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2

[6] Erik Gärtner, Mykhaylo Andriluka, Erwin Coumans, and Cristian Sminchisescu. Differentiable dynamics for articulated 3d human motion reconstruction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[7] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 7, A1

[8] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[9] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. In *ACM SIGGRAPH Conference Proceedings*, 2023. 3

[10] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 2, 3

[11] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *ACM Transactions on Graphics (TOG)*, 41(4):1–19, 2022. 2, 3, 5, 7, A2

[12] Jiang Hua, Liangcai Zeng, Gongfa Li, and Zhaojie Ju. Learning for a robot: Deep reinforcement learning, imitation learning, transfer learning. *Sensors*, 21(4):1278, 2021. 2

[13] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[14] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2

[15] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[16] Jordan Juravsky, Yunrong Guo, Sanja Fidler, and Xue Bin Peng. Padl: Language-directed physics-based character control. In *ACM SIGGRAPH Conference Proceedings*, 2022. 2

[17] Sai Shashank Kalakonda, Shubh Maheshwari, and Ravi Kiran Sarvadevabhatla. Action-gpt: Leveraging large-scale language models for improved and generalized zero shot action generation. *arXiv preprint arXiv:2211.15603*, 2022. 2

[18] Charles Khazoom, Daniel Gonzalez-Diaz, Yanran Ding, and Sangbae Kim. Humanoid self-collision avoidance using whole-body control with control barrier functions. In *International Conference on Humanoid Robots (Humanoids)*, 2022. 2

[19] K Niranjan Kumar, Irfan Essa, and Sehoon Ha. Words into action: Learning diverse humanoid behaviors using language guided iterative motion refinement. In *2nd Workshop on Language and Robot Learning: Language as Grounding*, 2023. 2

[20] Alex X Lee, Anusha Nagabandi, Pieter Abbeel, and Sergey Levine. Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[21] Jiye Lee and Hanbyul Joo. Locomotion-action-manipulation: Synthesizing human-scene interactions in complex 3d environments. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[22] Junfan Lin, Jianlong Chang, Lingbo Liu, Guanbin Li, Liang Lin, Qi Tian, and Chang-wen Chen. Being comes from not-being: Open-vocabulary text-to-motion generation with wordless training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[23] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

[24] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021. 2

[25] Josh Merel, Leonard Hasenclever, Alexandre Galashov, Arun Ahuja, Vu Pham, Greg Wayne, Yee Whye Teh, and Nicolas Heess. Neural probabilistic motor primitives for humanoid control. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[26] Josh Merel, Saran Tunyasuvunakool, Arun Ahuja, Yuval Tassa, Leonard Hasenclever, Vu Pham, Tom Erez, Greg Wayne, and Nicolas Heess. Catch & carry: reusable neu-

ral controllers for vision-guided whole-body tasks. *ACM Transactions on Graphics (TOG)*, 39(4):39–1, 2020. 2, 3

[27] Marko Mihajlovic, Shunsuke Saito, Aayush Bansal, Michael Zollhoefer, and Siyu Tang. Coap: Compositional articulated occupancy of people. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[28] OpenAI. Introducing gpt-4. https://openai.com/blog/gpt-4, 2023. 7, A1

[29] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 2, 3

[30] Xue Bin Peng, Michael Chang, Grace Zhang, Pieter Abbeel, and Sergey Levine. Mcp: Learning composable hierarchical control with multiplicative compositional policies. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3

[31] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (TOG)*, 40(4):1–20, 2021. 2, 3

[32] Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Transactions on Graphics (TOG)*, 41(4):1–17, 2022. 2, 7, 8

[33] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 7, A1

[34] Siyuan Qi, Yixin Zhu, Siyuan Huang, Chenfanfu Jiang, and Song-Chun Zhu. Human-centric indoor scene synthesis using stochastic grammar. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 2

[36] Zhiyuan Ren, Zhihong Pan, Xin Zhou, and Le Kang. Diffusion motion: Generate text-guided 3d human motion by diffusion model. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. 2

[37] Juan Rocamonde, Victoriano Montesinos, Elvis Nava, Ethan Perez, and David Lindner. Vision-language models are zero-shot reward models for reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2023. 2, 3, 7, 8

[38] Tim Salzmann, Marco Pavone, and Markus Ryll. Motron: Multimodal probabilistic human motion forecasting. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[39] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 4

[40] SFU MOCAP. SFU Motion Capture Database. https://mocap.cs.sfu.ca/, 2023. Accessed: 2023-11-01. 5

[41] Jiaming Song, Hongyu Ren, Dorsa Sadigh, and Stefano Ermon. Multi-agent generative adversarial imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[42] Chen Tessler, Yoni Kasten, Yunrong Guo, Shie Mannor, Gal Chechik, and Xue Bin Peng. Calm: Conditional adversarial latent models for directable virtual characters. In *ACM SIGGRAPH Conference Proceedings*, 2023. 2, 3

[43] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 3, 5, 7, A2

[44] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*, 2022. 2

[45] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3d human mesh from monocular images: A survey. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023. 2

[46] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *International Conference on Intelligent Robots and Systems (IROS)*, 2012. 2

[47] Shashank Tripathi, Lea Müller, Chun-Hao P Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. 3d human pose estimation via intuitive physics. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[48] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[49] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[50] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[51] Shu Wang, Muzhi Han, Ziyuan Jiao, Zeyu Zhang, Ying Nian Wu, Song-Chun Zhu, and Hangxin Liu. Llmˆ 3: Large language model-based task and motion planning with motion failure reasoning. *arXiv preprint arXiv:2403.11552*, 2024. 2

[52] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning (ICML)*, 2020. 3

[53] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say, interact as you can: Language-guided human motion generation with scene affordance. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[54] Jungdam Won, Deepak Gopinath, and Jessica Hodgins. A scalable approach to control diverse behaviors for physically simulated characters. *ACM Transactions on Graphics (TOG)*, 39(4):33–1, 2020. 2

[55] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[56] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 7, 8

[57] Pei Xu, Xiumin Shang, Victor Zordan, and Ioannis Karamouzas. Composite motion learning with task control. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023. 3

[58] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[59] Shusheng Xu, Huaijie Wang, Jiaxuan Gao, Yutao Ouyang, Chao Yu, and Yi Wu. Language-guided generation of physically realistic robot motion and control. *arXiv preprint arXiv:2306.10518*, 2023. 2

[60] Heyuan Yao, Zhenhua Song, Yuyang Zhou, Tenglong Ao, Baoquan Chen, and Libin Liu. Moconvq: Unified physics-based motion control via scalable discrete representations. *arXiv preprint arXiv:2310.10198*, 2023. 2

[61] Hongwei Yi, Chun-Hao P Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J Black. Mime: Human-aware 3d scene generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[62] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[63] Haotian Zhang, Ye Yuan, Viktor Makoviychuk, Yunrong Guo, Sanja Fidler, Xue Bin Peng, and Kayvon Fatahalian. Learning physically simulated tennis skills from broadcast videos. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023. 3

[64] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[65] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[66] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. *arXiv preprint arXiv:2306.10900*, 2023. 2

[67] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[68] Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, and Siyu Tang. Synthesizing diverse human motions in 3d indoor scenes. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[69] Shuai Zhao, Xiaohan Wang, Linchao Zhu, and Yi Yang. Test-time adaptation with CLIP reward for zero-shot generalization in vision-language models. In *International Conference on Learning Representations (ICLR)*, 2024. 7, 8

A. Data

This section offers a detailed account of the data's origins and the methodologies employed for its processing.

A.1. Text Data

Text descriptions sourced from publicly available online datasets are often marked by redundancy, ambiguity, and insufficient detail. To address these issues, it is necessary to preprocess the descriptions to render them more practical and usable. For generating practical text descriptions, we implemented a three-tiered process leveraging GPT-4 [28]. This encompasses **filtering text** to discard non-essential details, **scoring text** for assessing utility, and **rewriting text** to improve clarity and applicability. Our goal is to identify text descriptions that significantly contribute to mastering open-vocabulary physical skills from a robust pre-existing dataset, and to standardize the collection of text instructions.

Filter text Initially, we compiled 89,910 text entries from HumanML3D [7] and Babel [33], discovering substantial repetition, including exact duplicates, descriptions of akin actions (*e.g.*, *"A person walks down a set of stairs"* vs. *"A person walks down stairs"*), frequency-related repetitions (*e.g.*, *"A person sways side to side multiple times"* vs. *"A person sways from side to side"*), and semantic duplicates (*e.g.*, *"The person is doing a waltz dance"* vs. *"A man waltzes backward in a circle"*).

To address this issue, we initiated a deduplication process, first eliminating descriptions that were overly brief (under three tokens) or excessively lengthy (over 77 tokens). We then utilized the LLAMA-2-7B MODEL with its 4096-dimensional embedding vector for further deduplication. By computing cosine similarities between each description pair and applying a 0.92 similarity threshold, descriptions exceeding this threshold were considered repetition. This procedure refined our dataset to 4,910 unique descriptions.

Scoring text After filtering out duplicates and semantically similar actions, we encountered issues like typographical errors, overly complex descriptions, and significant ambiguities in the remaining texts. These problems rendered the descriptions unsuitable for generating actionable human motion skills despite their uniqueness.

To further refine our text instructions, we evaluated the remaining descriptions for their suitability in model processing and practical motion generation. Our evaluation, detailed in Fig. A1, focused on fluency, conciseness, and the specificity of individual human poses within a brief sequence of frames. Descriptions that were direct and descriptive, containing clear verbs and nouns, were preferred over those with a sequential or ambiguous nature. Using a standardized scoring process, we ranked the action descriptions by their scores. After addressing issues in an initial round of scoring, a second evaluation was conducted to fine-tune our selection, as mentioned in Fig. A2. This led to the exclusion of descriptions within certain score ranges (0-0.92, 0.98-0.99), resulting in a curated dataset of 1,896 unique action descriptions optimized for model training.

Score Prompt I

You are a language expert. Please rate the following actions on a scale of 0 to 10 based on their use of language. The requirements are:

1. *The description should be fluent and concise.*
2. *The description should correspond to a single human pose, instead of a range of possible poses.*
3. *The description should describe a human pose at a short sequence of frames instead of a long sequence of frames (this requirement is not mandatory).*
4. *If the description contains sequential logic, rate it lower. "Walk in a circle" is a kind of sequential logic.*
5. *Except for the subject, the description should have only one verb and one noun.*
6. *If the description is vivid(like "dances like Michael Jackson"), rate it higher.*

Here are some examples you graded in the last round:

- *6 - A person is swimming with his arms.*
- *3 - Sway your hips from side to side.*
- *7 - A person smashed a tennis ball.*
- *4 - A person is in the process of sitting down.*
- *5 - A person brings up both hands to eye level.*
- *9 - A person dances like Michael Jackson.*
- *2 - A person packs food in the fridge.*
- *5 - A person flips both arms up and down.*
- *8 - Looks like disco dancing.*
- *3 - Kneeling person stands up.*
- *1 - A person does a gesture while doing kudo.*
- *6 - A person unzipping pants flyer.*
- *0 - then kneels on both knees on the floor.*
- *2 - A person is playing pitch and catch.*
- *1 - A person gesturing them walking backward.*
- *4 - A person seems confident and aggressive.*
- *1 - A person circles around with both arms out.*
- *5 - A person prepares to take a long jump.*
- *6 - A person jumps twice into the air.*
- *0 - Turning around and walking back.*

Now, please provide your actions in the format 'x - yyyy,' where 'x' is the score, and 'yyyy' is the original sentence. Please note that Do not change the original sentence.

Figure A1. **Score Prompt I**. This prompt focuses on filtering text descriptions for fluency, conciseness, and specificity, particularly targeting individual human poses within a short sequence of frames.

Score Prompt II

You are a language expert. Please rate the following actions on a scale of 0 to 10 based on the ambiguity of the description. Examine whether this action description corresponds to a unique action. If the description corresponds to fewer actions, like "wave with both arms", rate it higher. If the description corresponds to abundant actions, like "do yoga", rate it lower.

- *7 - grab items with their left hand.*
- *8 - hold onto a handrail.*
- *9 - do star jumps.*
- *5 - arms slightly curled go from right to left.*
- *3 - sit down on something.*
- *9 - kick with the right foot.*
- *7 - stand and put arms up.*
- *9 - cover the mouth with the hand.*
- *8 - stand and salute someone.*
- *2 - break dance.*
- *6 - spin body very fast.*
- *7 - open bottle and drink it.*
- *2 - do the cha-cha.*
- *5 - do sit-ups.*
- *4 - slowly stretch.*
- *6 - cross a high obstacle.*
- *7 - grab something and shake it.*
- *4 - lift weights to get buff.*
- *8 - move left hand upward.*
- *7 - walk forward swiftly.*

Now, please provide your actions in the format 'x - yyyy,' where 'x' is the score, and 'yyyy' is the original sentence. Please note that Do not change the original sentence.

Figure A2. **Score Prompt II**. This prompt selects for direct and richly detailed action descriptions, prioritizing clarity with a distinct verb and noun over descriptions based on sequential or complex logic.

Rewrite text In the final refinement phase, we address the specificity of action descriptions, crucial for accurately generating motions. Vague descriptions, such as *'jump rope'*, can lead to ambiguous interpretations and various motion realizations, challenging the model's training due to the similarity of rewards for different motions. This observation is consistent with other motion generation studies utilizing CLIP [11, 43].

To enhance the clarity and effectiveness of the reward calculation, we rephrase and detail the descriptions. For instance, *'jump rope'* is clarified to *'swinging a rope around your body'*, with further details like *'Raise both hands and shake them continuously while simultaneously jumping up with both feet, repeating this cycle'*. Additionally, we break down actions into more discrete moments, such as *'legs off the ground, wave hand'*, to improve the reward function's precision. Our methodology for this textual refinement is detailed in Fig. A3.

Rewrite Prompt

Describe an action of instruction for a humanoid agent. The description must satisfy the following conditions:

1. *The description should be concise.*
2. *The description should describe a human pose in a single frame instead of a sequence of frames.*
3. *The description should correspond to only one human pose, instead of a range of possible poses, minimize ambiguity.*
4. *The description should be less than 8 words.*
5. *The description should not contain a subject like "An agent", "A human".*
6. *The description should have less than two verbs and two nouns.*
7. *The description should not have any adjectives, adverbs, or any similar words like "with respect".*
8. *The description should not include details describing expressions or fingers and toes.*

For example, it's better to describe "take a bow" as "bow at a right angle."

Figure A3. **Rewrite Prompt**. This prompt is designed for rephrasing action descriptions to enhance clarity and incorporate additional details, aiming to improve the specificity and effectiveness of the generated motions.

A.2. Motion Data

For the study, we curated 93 motion clips, organizing them by movement type and style into a structured dataset. We delineated movements into three categories: *move_around*, *act_in_place*, and *combined*; and styles into five categories: *attack*, *crawl*, *jump*, *dance*, and *usual*. The clips were then classified into these eight categories, with a weighting system applied based on the inverse frequency of each category to enhance the representation of less common actions. For motions that spanned multiple categories, their weights were averaged based on their inverse frequency values. This approach aimed to ensure a balanced action distribution within the dataset, emphasizing the inclusion of rarer actions to avoid overrepresentation of any single action type. The categorization and its impact on the dataset distribution are illustrated in the diagram available in Fig. A10.

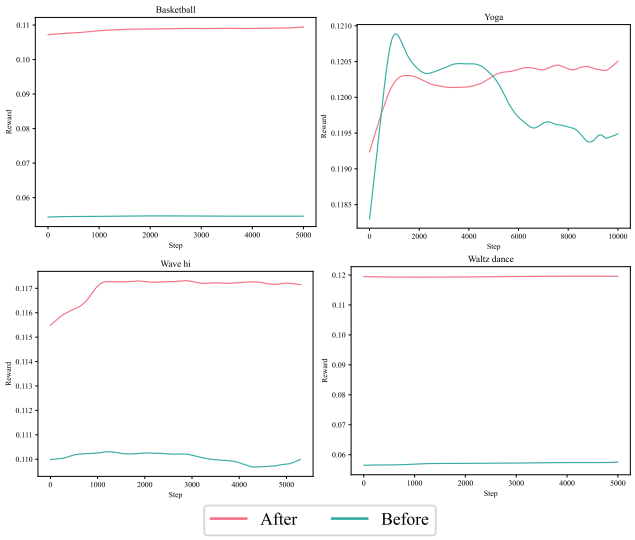


Figure A4. **Rewards before and after text enhancement.** The red curve depicts reward trends following text enhancement, contrasting with the pre-enhancement trends shown by the green curve.

B. Experiments

This supplementary section expands on the experimental analyses from Sec. 4, focusing on the text description. Beyond the quantitative metrics addressed in the main document, we explore the changes in reward function dynamics pre- and post-text refinement across various instructions. This includes a detailed comparison of CLIP similarity scores during training to critically evaluate the effectiveness and design of different reward functions.

B.1. Text Enhancement

Utilizing the text enhancement strategy described in Appendix A.1, we have refined action descriptions from existing open-source datasets, reducing ambiguity and enhancing clarity and applicability. To gauge the impact of these refined descriptions on training efficacy, we track and compare the reward feedback during the training phases.

Selecting four instructions at random from our dataset for illustration, we compare reward trends before and after text enhancements—represented by green and red curves, respectively, in our graphs. This comparison reveals that refined instructions consistently yield superior reward trajectories from the start, showing a swift and steady ascent to a performance plateau. This indicates that text enhancement notably improves policy training efficiency and convergence speed. Specifically, for intricate actions like *Yoga* (as shown in the top right figure of Fig. A4), refined instructions result in a more stable and gradual reward increase, signifying improved training stability and model performance.

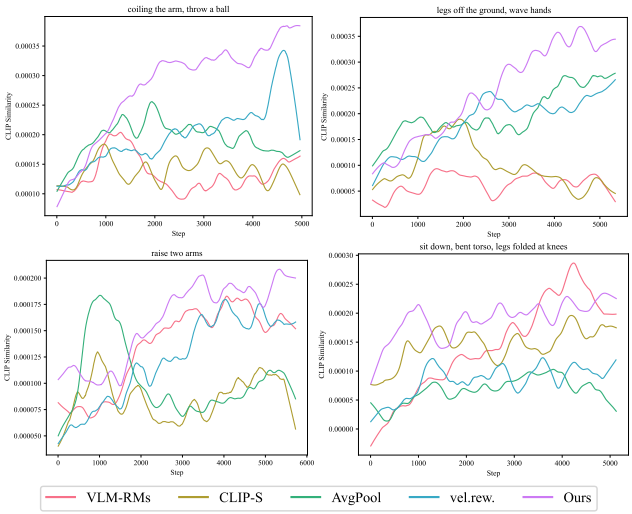


Figure A5. **The CLIP similarity calculated by different reward designs.**

B.2. Implementation Details

B.3. Reward Function Analysis

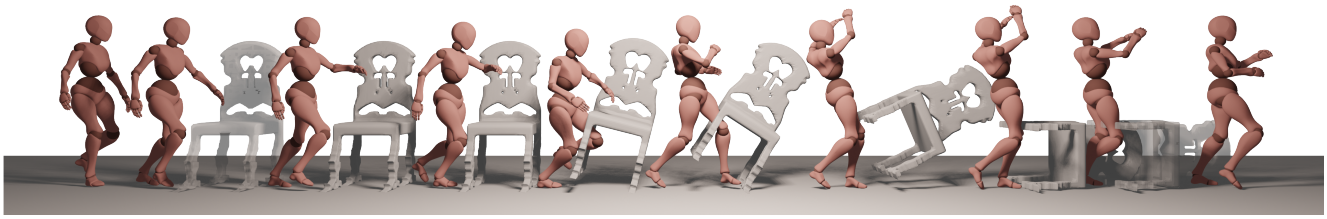
To evaluate and compare various reward function designs, we use cosine similarity between image and text features as a uniform metric, accommodating the differing numerical scales inherent to each reward design. As depicted in Fig. A5, we represent five reward functions using distinct colors, with our method marked in purple.

Aligning with discussions in the main text (Fig. 5), we examine four instructions from our user study for a detailed comparison. Our findings indicate that our method uniformly improves image-text alignment throughout training, achieving consistent convergence. While some methods exhibit comparable performance on select instructions, they generally show less consistency, with initial gains often receding over time. In contrast, our approach demonstrates robustness against the variabilities of open-vocabulary training, leading to stable and reliable performance improvements.

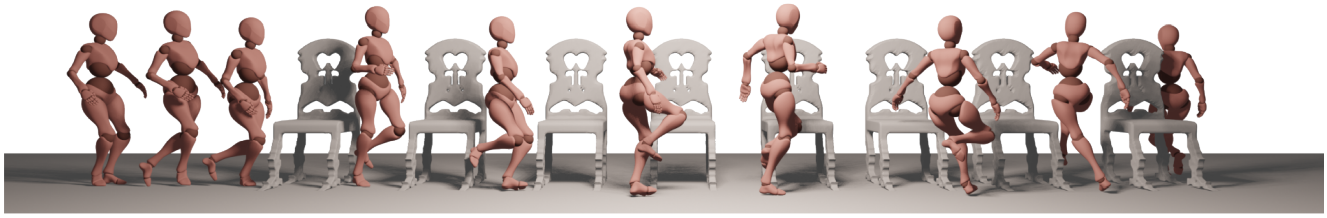
To assist readers in replicating our work, we have included a comprehensive breakdown of hyperparameter settings in Tabs. A1 and A2.

B.4. Interaction Motions

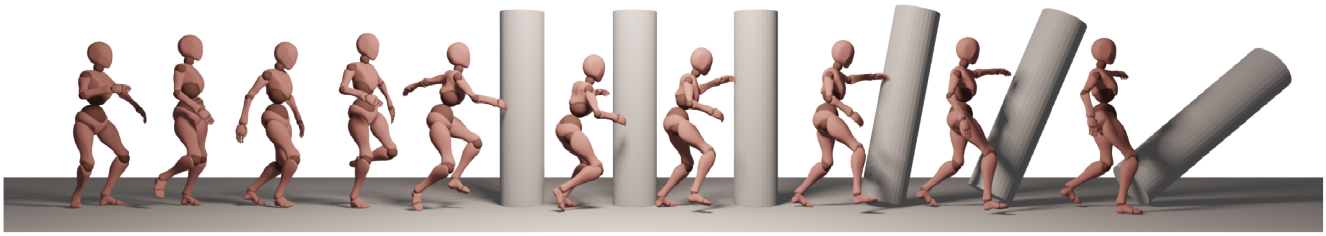
Within the main text, we highlighted `AnySkill`'s proficiency in mastering tasks involving interactions with diverse objects, underscoring its capability to adapt across a spectrum of interaction scenarios. For experimental validation, we deliberately chose a range of objects, both rigid (*e.g.*, pillars, balls) and articulated (*e.g.*, doors, chairs), to demonstrate the method's versatility. The quantitative analyses of these object interactions, as detailed in Appendix B.2, affirm the flexibility of our approach. Our system is shown to adeptly navigate a variety of action requirements, as speci-



(a) kick the white chair



(b) move around the white chair



(c) strike the pillar

Figure A6. **Additional results of interaction motions.**

Table A1. **Hyperparameters used for the training of low-level controller.**

Hyper-Parameters	**Values**
dim(Z) Latent Space Dimension	64
Encoder Align Loss Weight	1
Encoder Uniform Loss Weight	0.5
w gp Gradient Penalty Weight	5
Encoder Regularization Coefficient	0.1
Samples Per Update Iteration	131072
Policy/Value Function Minibatch Size	16384
Discriminators/Encoder Minibatch Size	4096
γ Discount	0.99
Learning Rate	2×10^{-5}
GAE(λ)	0.95
TD(λ)	0.95
PPO Clip Threshold	0.2
T Episode Length	300

fied by different text descriptions, maintaining efficacy even when faced with repetitive initial conditions or identical objects.

Table A2. **Hyperparameters used for the training of high-level controller.**

Hyper-Parameters	**Values**
w gp Gradient Penalty Weight	5
Encoder Regularization Coefficient	0.1
Samples Per Update Iteration	131072
Policy/Value Function Minibatch Size	16384
Discriminators/Encoder Minibatch Size	4096
γ Discount	0.99
Learning Rate	2×10^{-5}
GAE(λ)	0.95
TD(λ)	0.95
PPO Clip Threshold	0.2
T Episode Length	300

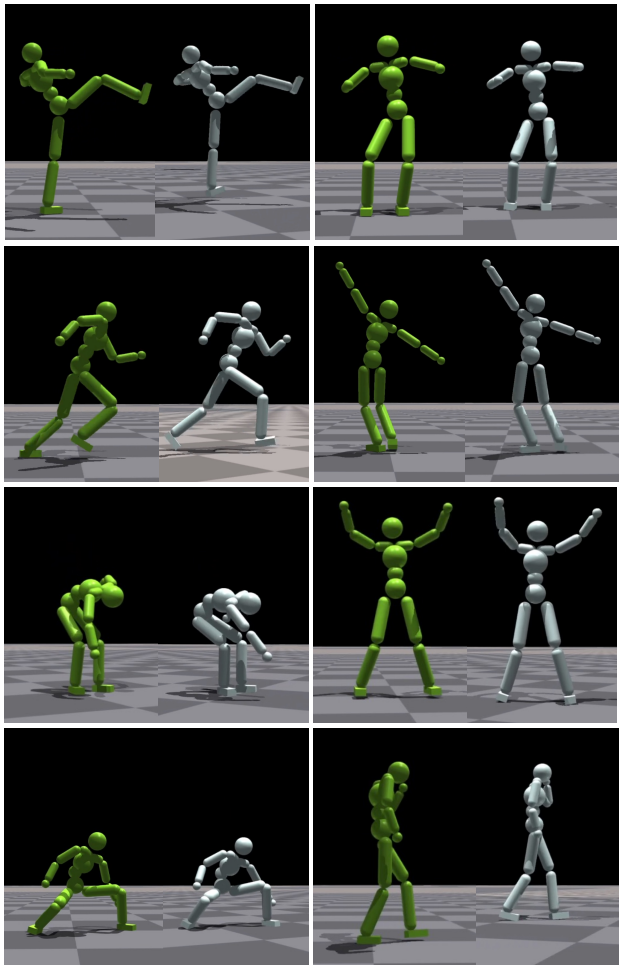


Figure A7. **Atomic actions from the trained low-level controller.** In each subfigure, the green agent shows the reference motion from the dataset, and the white agent shows our learned atomic action.

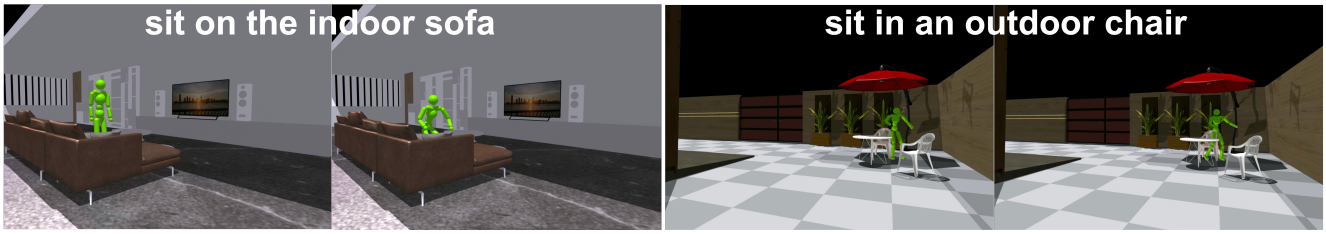
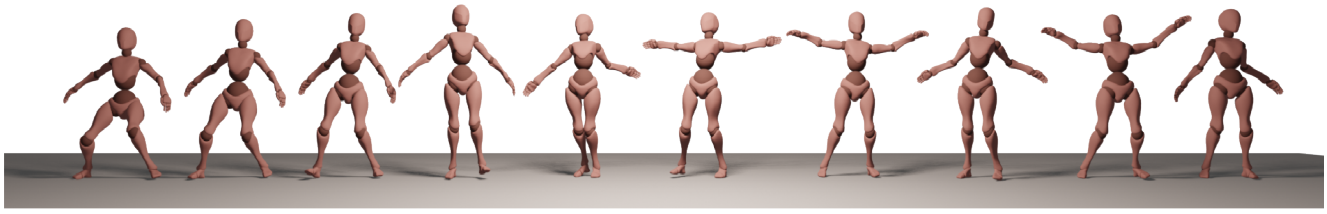
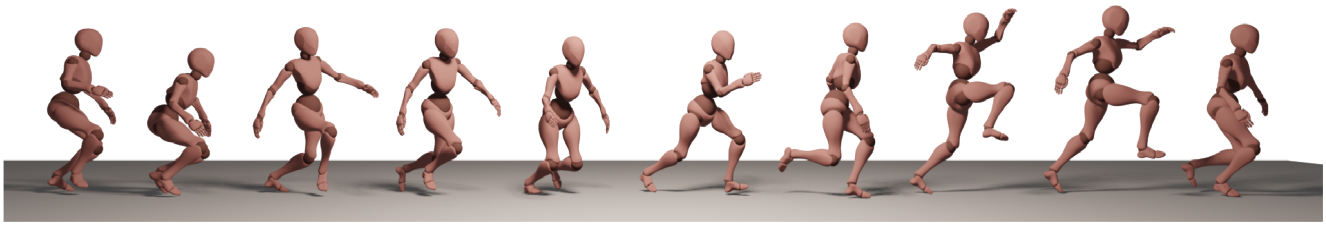


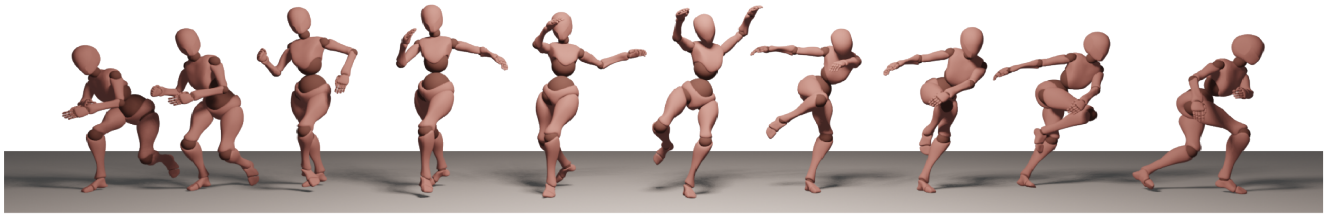
Figure A8. **Real-time scene interaction.** We employed both indoor and outdoor scenes within IsaacGYM. Throughout the training process, we conducted real-time rendering and obtained feedback on physical interactions.



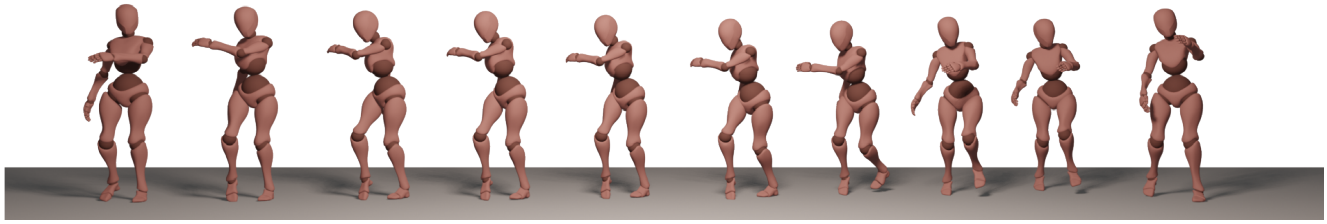
(a) wave hands up and down



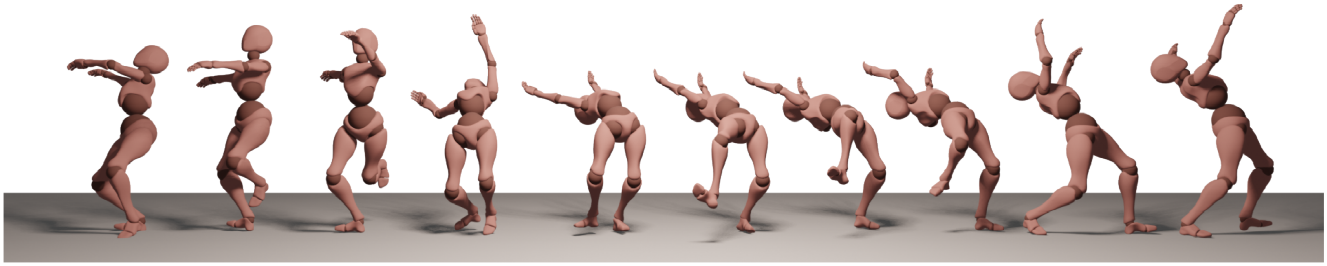
(b) jump high



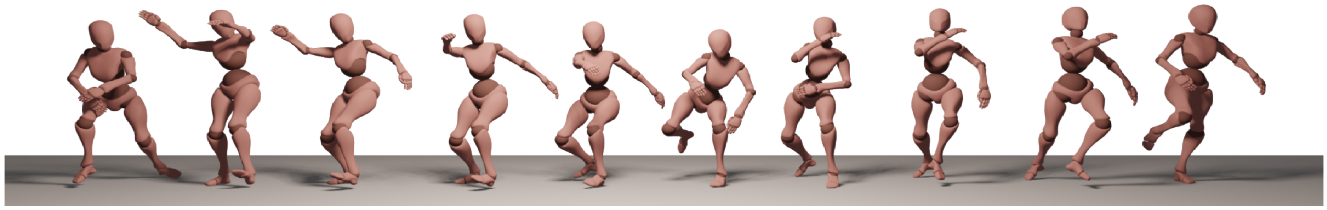
(c) left leg forward, right leg retreats



(d) raise one arm, put the other hand down



(e) raise hands above head, bend body



(f) hit a tennis smash with arm

Figure A9. **More results of open-vocabulary physical skills.**

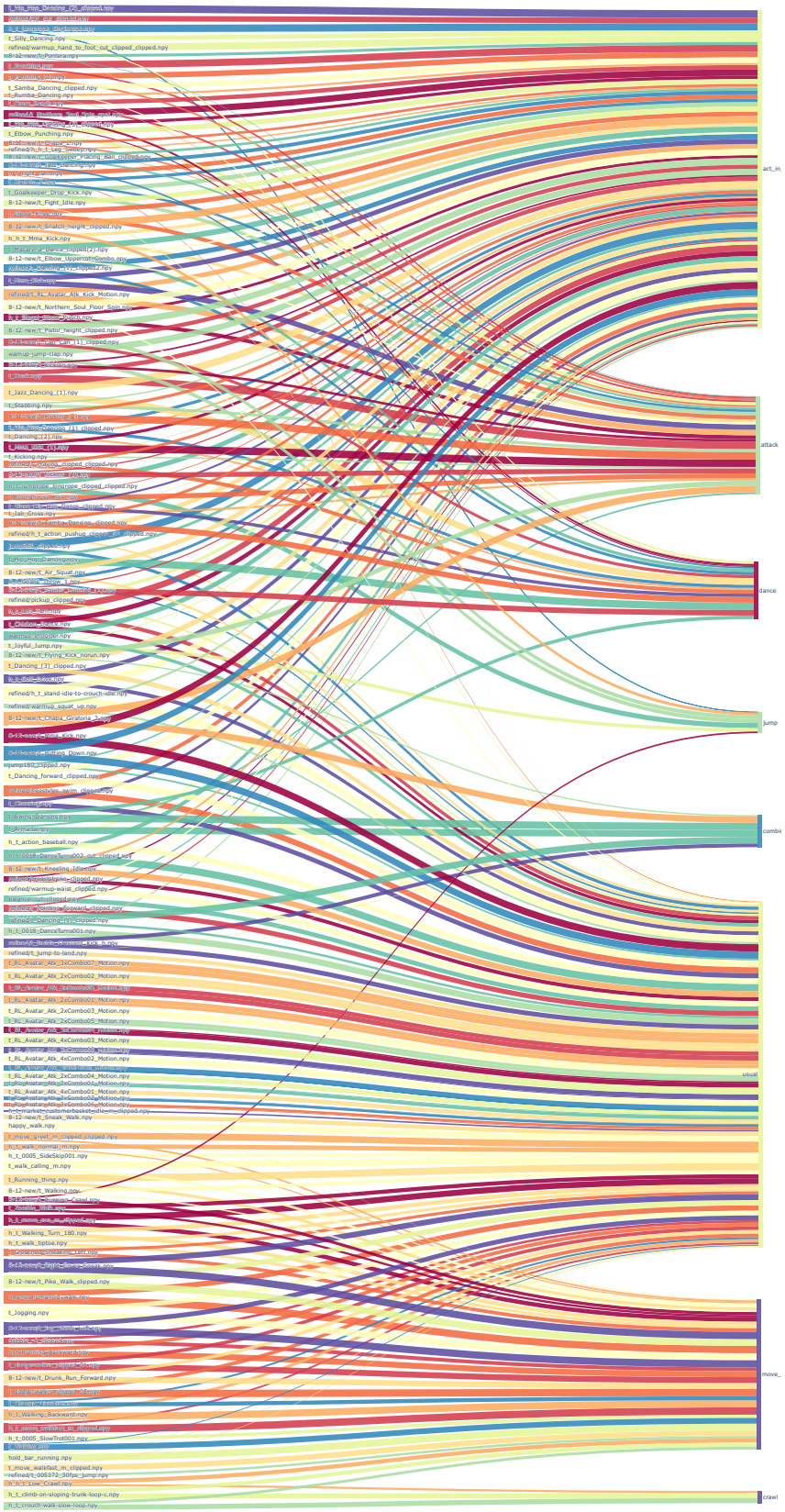


Figure A10. **The distribution of actions and their corresponding categories.**