A. Data

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

A.1. Text Data

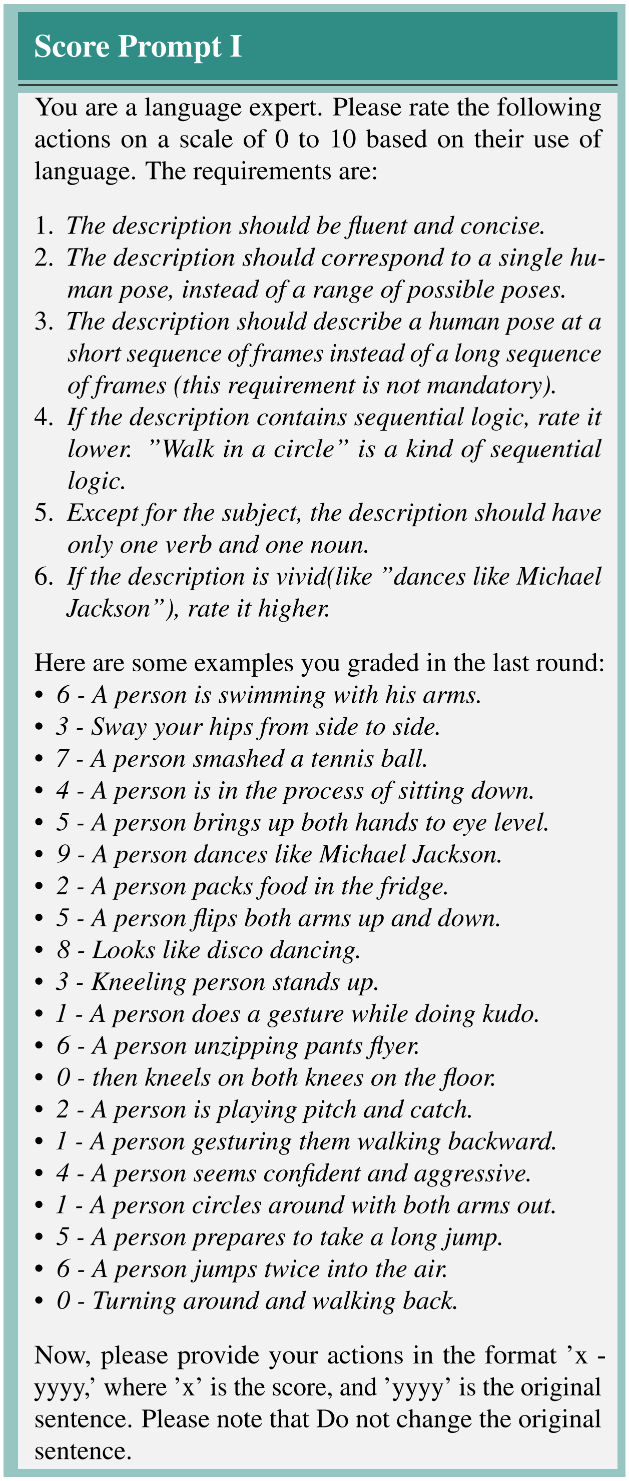
Text descriptions sourced from publicly available online datasets are often marked by redundancy, ambiguity, and insufficient detail. To address these issues, it is necessary to preprocess the descriptions to render them more practical and usable. For generating practical text descriptions, we implemented 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.

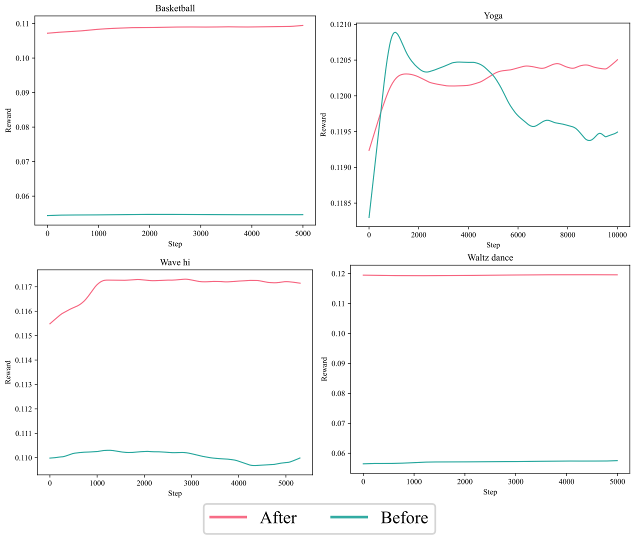


Figure A4. **Rewards before and after text enhancement.** The red curve depicts reward trends following text enhancement, contrasting with the pre-enhancement trends shown by the green curve.

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.

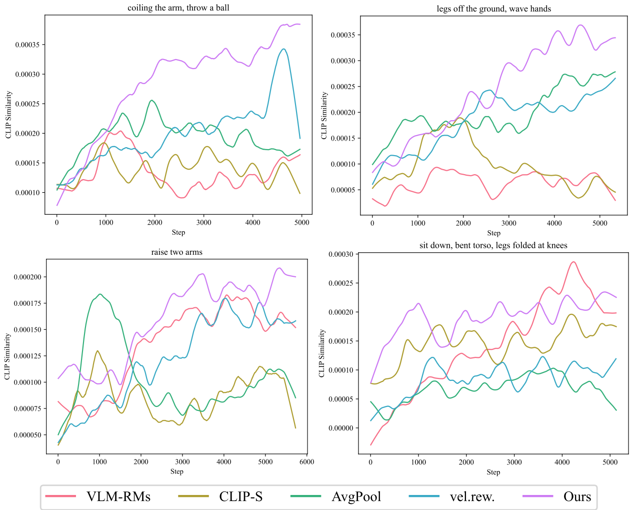


Figure A5. **The CLIP similarity calculated by different reward designs.**

B.2. Implementation Details

B.3. Reward Function Analysis

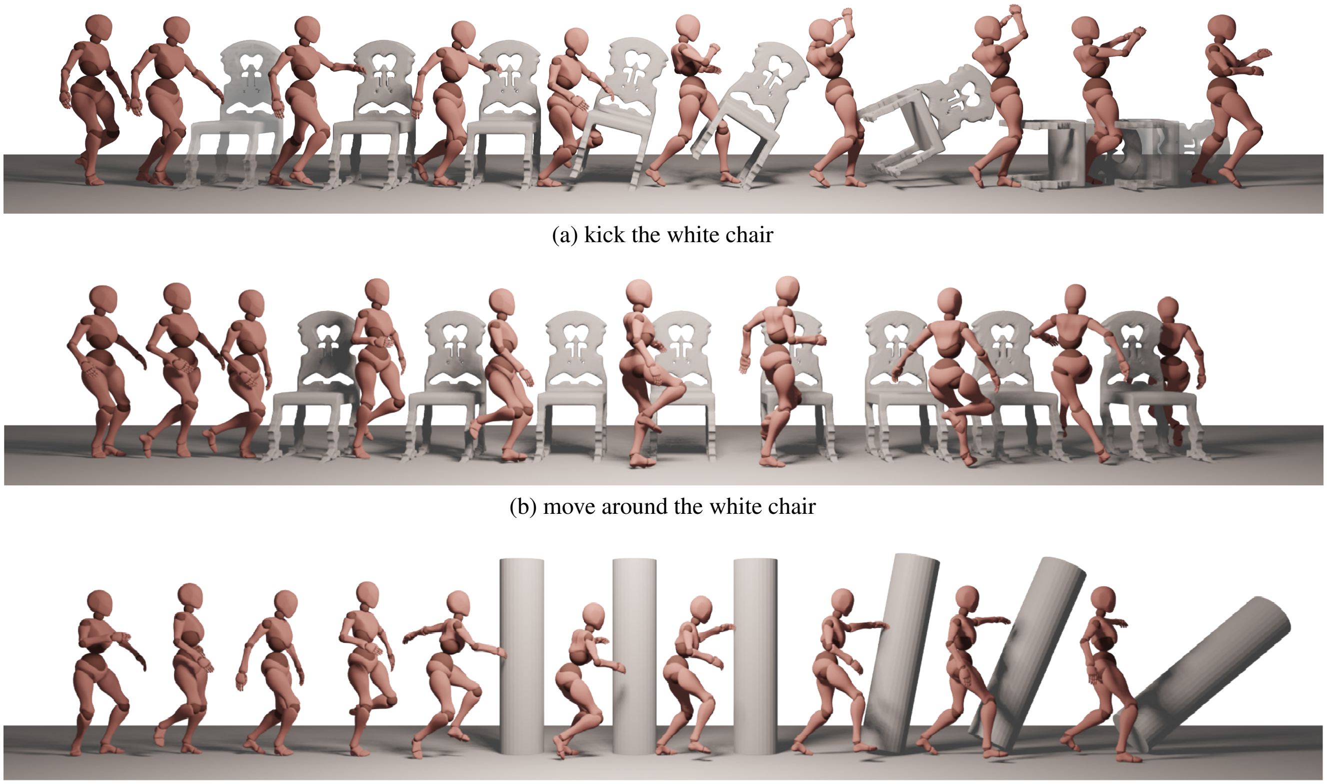
To evaluate and compare various reward function designs, we use cosine similarity between image and text features as a uniform metric, accommodating the differing numerical scales inherent to each reward design. As depicted in Fig. A5, we represent five reward functions using distinct colors, with our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

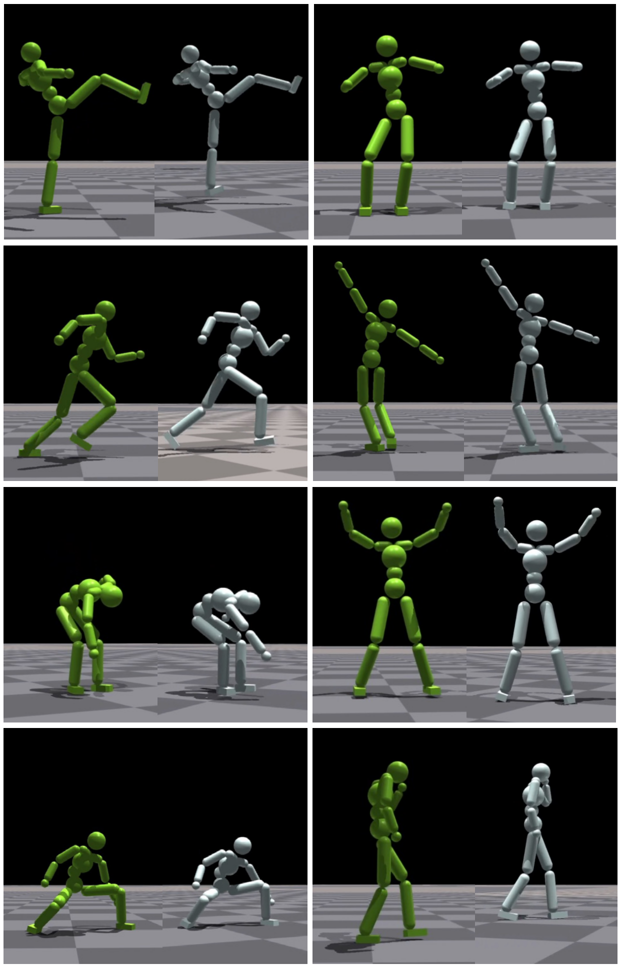


Figure A7. **Atomic actions from the trained low-level controller.** In each subfigure, the green agent shows the reference motion from the dataset, and the white agent shows our learned atomic action.

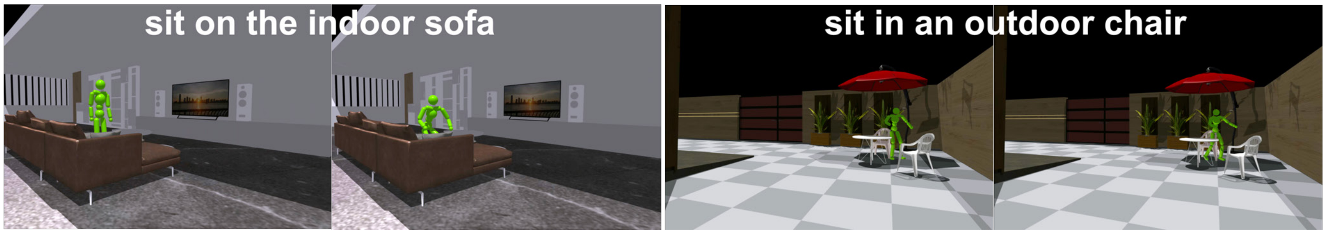
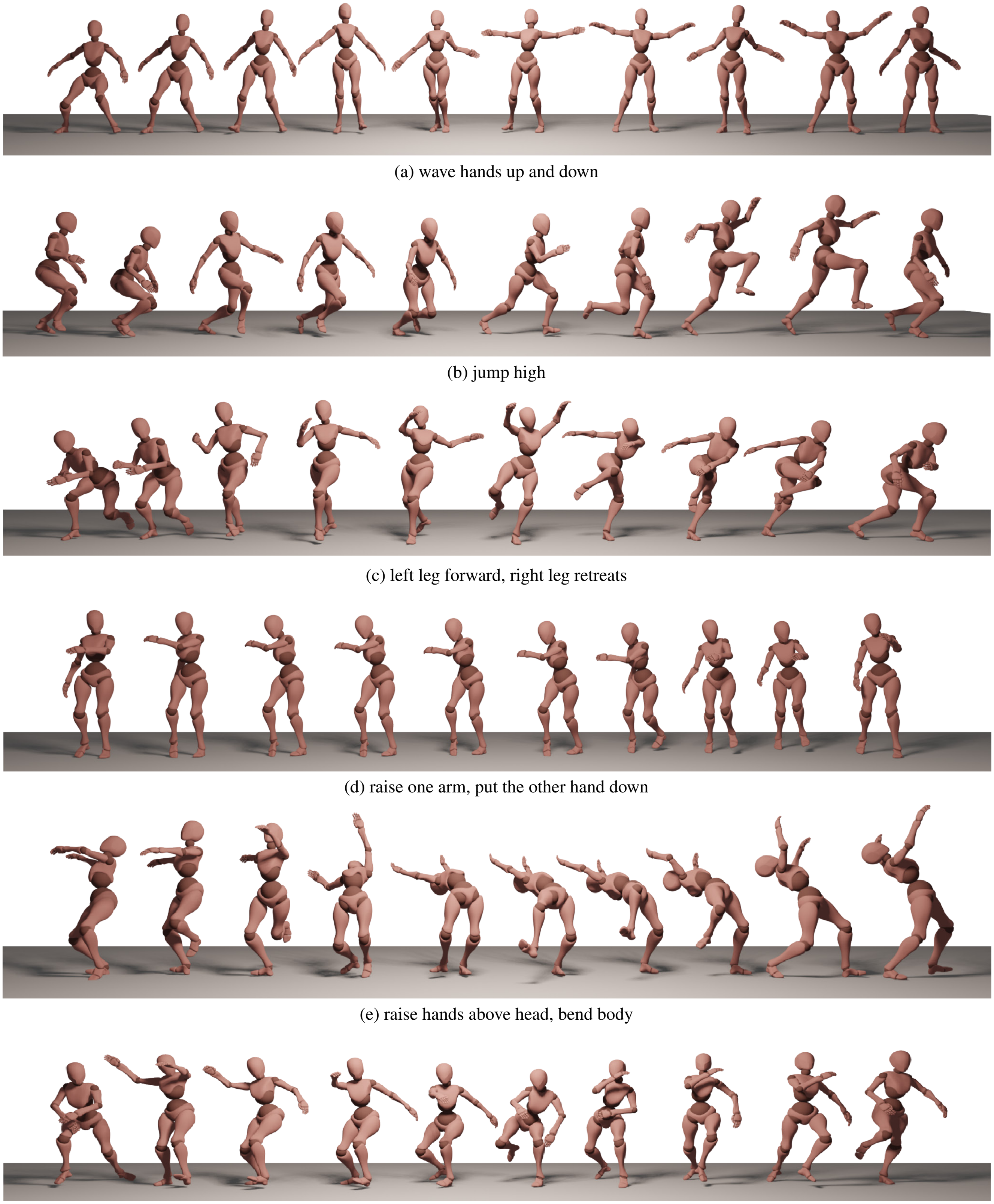


Figure A8. **Real-time scene interaction.** We employed both indoor and outdoor scenes within IsaacGYM. Throughout the training process, we conducted real-time rendering and obtained feedback on physical interactions.



(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.**

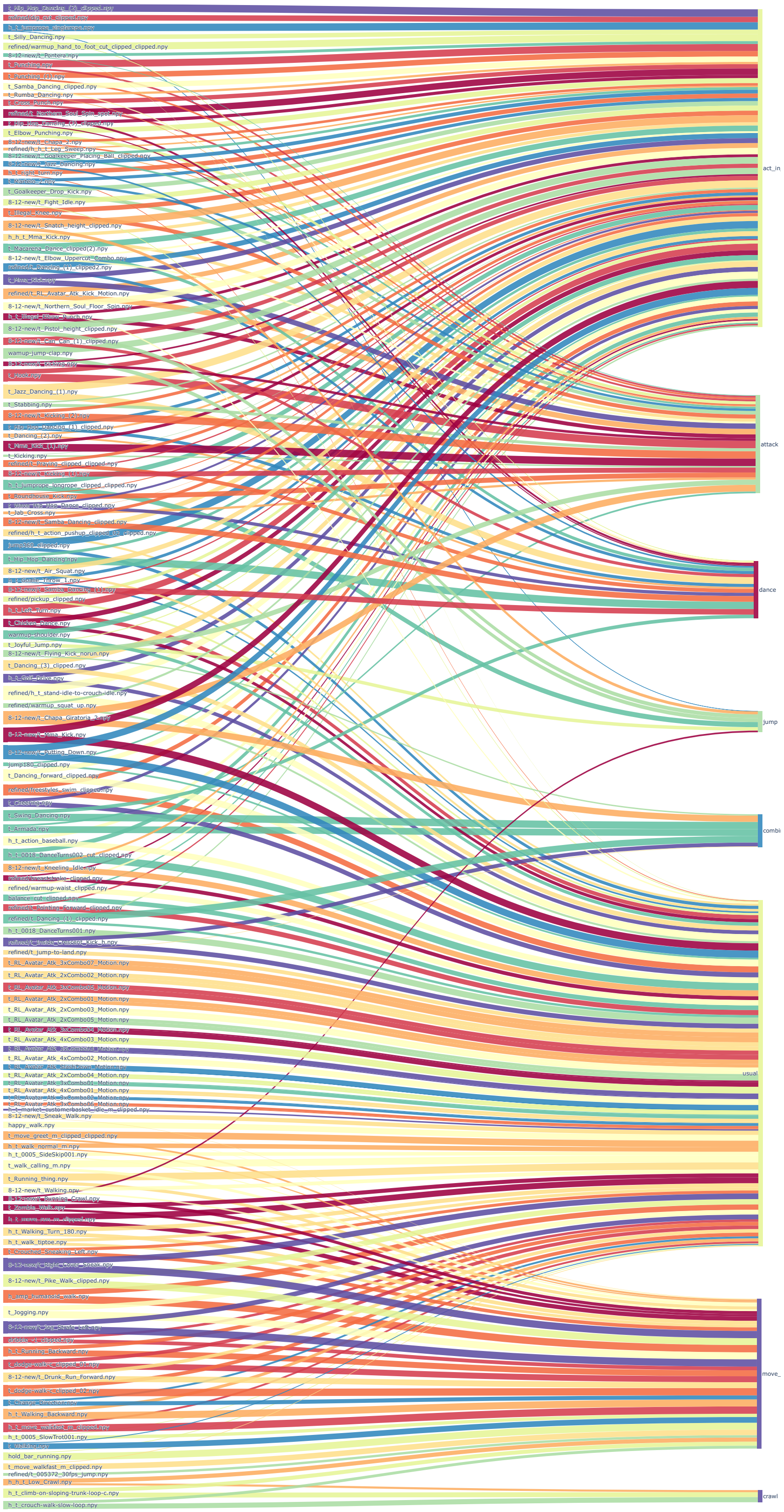


Figure A10. **The distribution of actions and their corresponding categories.**