PersonaBooth: Personalized Text-to-Motion Generation

Supplementary Material

A. PerMo Dataset Details

A.1. Motion Capture Environment

The dataset was captured in a studio shown in Fig. A with a floor area of 12m x 10m using 33 OptiTrack[1] cameras: 25 PrimeX 22 and 8 Prime17W. The PrimeX 22 cameras record at a resolution of 2048×1088 with a 79° field of view, while the Prime17W cameras record at a resolution of 1664×1088 with a 70° field of view. All videos were recorded at 120 fps.

A.2. Persona and Content Categories

PerMo was captured by five professional actors, including two women (Actors 1 and 2) and three men (Actors 3, 4, and 5), each with diverse body shapes and heights. The 34 style categories we captured are organized into six groups, as shown in Table A. Since each of the five actors performed all 34 styles, a total of 170 personas were created. Additionally, for each persona, we recorded motion data for 10 diverse contents to evenly represent full-body movement. These 10 contents fall into one of four action types: *leg actions*, *arm actions*, *ground locomotion*, and *leaping locomotion*, as illustrated in Table B. Note that actions that were not effective in representing the characteristics of each style were excluded. For instance, in the 'muddy-floor' style, the 'throw' action was omitted because it does not effectively convey the texture of a sticky floor.

Examples of the content motions are shown in Fig.J, and examples of each actor's personas are presented in Fig.K. Additional examples can be viewed in the attached video.

[1]https://www.optitrack.com

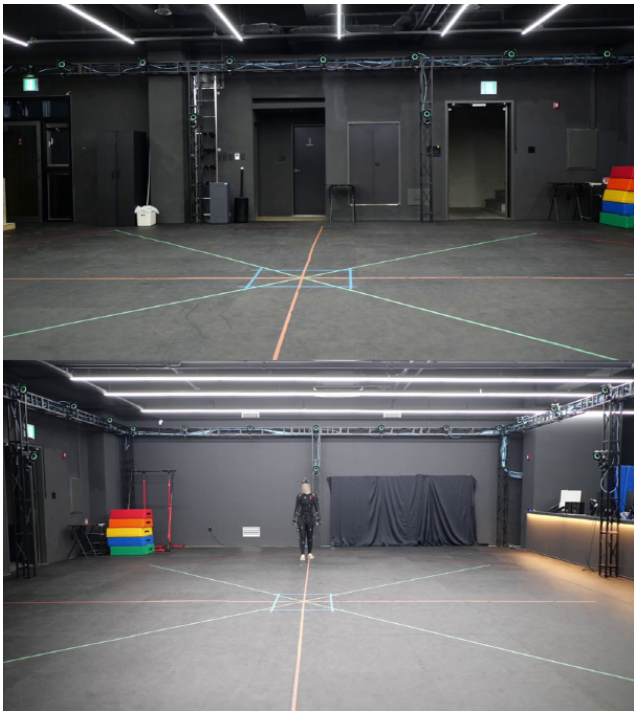


Figure A. Side and front views of a motion capture studio equipped with OptiTrack cameras

Table A. Style categories included in the PerMo dataset

Parent Category	Style Category
Age	Childish, Neutral, Old, Teenage
Character	Ballerina, Hulk, Monkey, Ninja, Penguin, Robot, SWAT, Waiter, Zombie
Condition	Arm-aching, Drunken, Exhausted, Head-aching, Healthy, Leg-aching, Text-necked
Emotion	Angry, Fearful, Happy, Sad, Strained, Surprising
Traits	Elegant, Shy, Silly, Uppity
Surroundings	Cold, Crowded, Muddy-floor, Unpleasant-floor

Table B. Content categories included in the PerMo dataset

Parent Category	Content Category
Leg Action	Kick, Kick Something
Arm Action	Punch, Throw, Wave
Ground Locomotion	Walk, Run, Transition
Leaping Locomotion	Jump, Hop

A.3. Motion Capture Instructions

We provided instructions to ensure that each actor could effectively embody their unique persona. First, we presented the style categories and gave each actor time to consider how they would express them. Actors were given as much freedom as possible to express their personas. Atomic motions (contents) within the same persona were performed consecutively to ensure a consistent persona was conveyed across those motions. When selecting the 34 styles, we took the actor's opinion into account and avoided choosing styles that could be similarly expressed in the motion, opting instead for distinct ones.

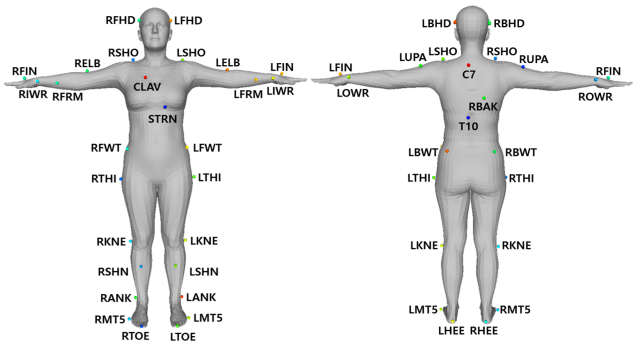


Figure B. Positions of the 41 markers

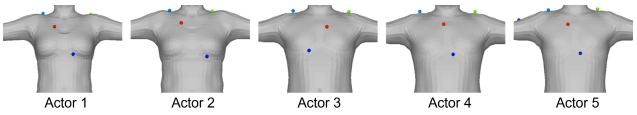


Figure C. Positions of the CLAV and STRN markers of each actor. The red dot on the upper chest represents CLAV, and the blue dot on the lower chest represents STRN

However, for the content, stricter guidelines were provided to ensure all actors performed actions within a common framework, allowing the differences between personas to be more clearly highlighted. Among the content types, *leg action* and *arm action* are stationary motions performed while standing, whereas *ground locomotion* and *leaping locomotion* involve movements across a wide range. For stationary motions, we marked the exact center of the studio and ensured the motions were performed at that location. For motions requiring movement, we placed a marker 4 meters away from the center. Actors started their motions from the edge of one side of the studio, and during post-processing, the sequence was cropped to begin when the actor stepped on the marker. Each motion was recorded in four to five takes for the same persona and content type.

We predefined the repetition count for each content type and cropped the data during post-processing to include only the specified number of repetitions. *Leg action* and *arm action* were repeated three times each, while *ground locomotion* and *leaping locomotion* were repeated five times each. All motions were recorded with a focus on the right side. For example, *leg action* and *arm action* were performed using the right arm and right leg. *Ground locomotion* started with stepping on the marker with the right foot. The 'Transition' action involved turning toward the right, and the 'Hop' action was performed using the right leg. For training, the data was augmented by flipping left and right, resulting in a total of 13,220 motion samples.

A.4. Data Format and Post-Processing

Each actor was equipped with 41 optical markers during the recording process. The positions of these markers are shown in Figure B. To distinguish between actors during data refinement, the positions of the CLAV and STRN markers vary slightly depending on the actor, as illustrated in Figure C. The marker sequences are stored in C3D files.

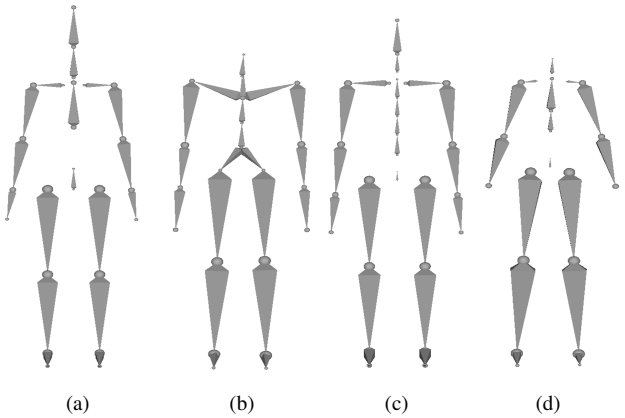


Figure D. The skeleton structure of each dataset. (a) PerMo (b) Xia [34] and BFA [1] (c) 100 STYLE [19] (d) BN [13]

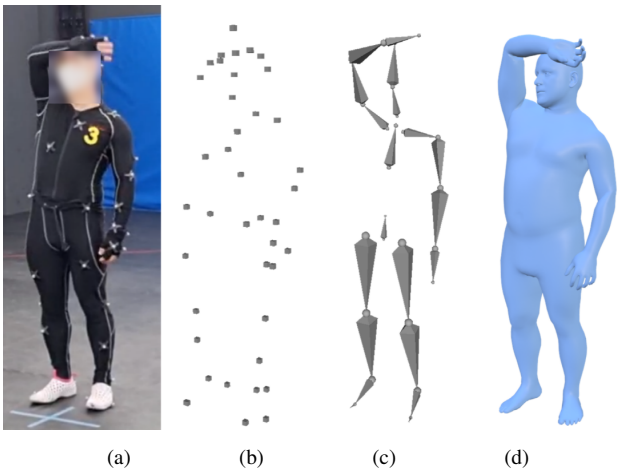


Figure E. Example of the data formats. (a) RGB image (b) Marker (c) Skeleton (d) Mesh

From the marker data, skeleton information comprising 20 bones is generated and saved in BVH files using OptiTrack software. The skeleton structure of PerMo is depicted in Figure D (a). Figures D (b), (c), and (d) show the skeleton structures of other motion style datasets. As shown, the skeleton structures vary across datasets, making it challenging to use them together for training. To address this, we convert the data into the standardized SMPL, a widely used 3D human mesh format. This unified format allows compatibility with other large-scale motion datasets such as AMASS [18] and HumanML3D [6].

We use MoSh++ [18] to convert optical marker data into SMPL-H format. In the first step of MoSh++, the shape pa-

Table C. Description of data format and naming conventions

Data Type	File Name	Description
Marker	[style]_[content]_[actor]_[take].c3d	41 markers for each frame
Skeleton	[style]_[content]_[actor]_[take].bvh	20 bones for each frame
Mesh	[style]_[content]_[actor]_[take].npz	SMPL-H pose data
Mesh Shape	shape_[style]_[content]_[actor].npz	SMPL-H shape data for each processing group
Rendered Mesh	[style]_[content]_[actor]_001.mp4	Rendered video of the mesh data

rameters are estimated. To achieve this, we grouped motions that share the same persona and content (4 or 5 takes), and then optimized the mesh for each group. Since the motion capture sessions were conducted over multiple days, slight variations in marker positions occurred between sessions. However, for motions within the same group, the marker positions remained consistent, enabling effective mesh optimization. A subset of frames from the motion sequences within each group was extracted, and the shape parameters were estimated from these frames. Subsequently, using the extracted shape parameters, the poses for each frame of the motion sequences were optimized.

To make it easier for users to check the motions, we provide rendered mesh videos for the first take of each group. Examples of the described data formats are shown in Fig. E, and the summary and naming conventions of the released data are presented in Table C. In addition, the folder structure of the PerMo dataset is shown in Fig. F.

A.5. Data Validation

We have prepared to release clean data by conducting a rigorous validation of the acquired motion data. The validation process considers the following four factors: (1) the cleanliness of the skeleton data (free from distortions), (2) the accuracy of motion cropping, (3) the presence of missing markers and bones, and (4) the synchronization between marker and skeleton data.

For (1), any twisted skeletons are manually corrected by experts during data validation. Regarding (2), each motion file is manually verified by at least two reviewers, who check whether the actor's starting position in *ground locomotion* is accurate and whether the correct number of repetitions is cropped. For (3) and (4), custom validation scripts were developed to perform automated checks.

A.6. Text Description

Examples of text descriptions included in the PerMo dataset are presented in Fig. G. To generate diverse descriptions, we first provided ChatGPT with detailed explanations of the motions. For example, for the 'Hop' motion, we used a description like: *"In the motion, one person hops forward on one leg. The person hops several times."* We then instructed

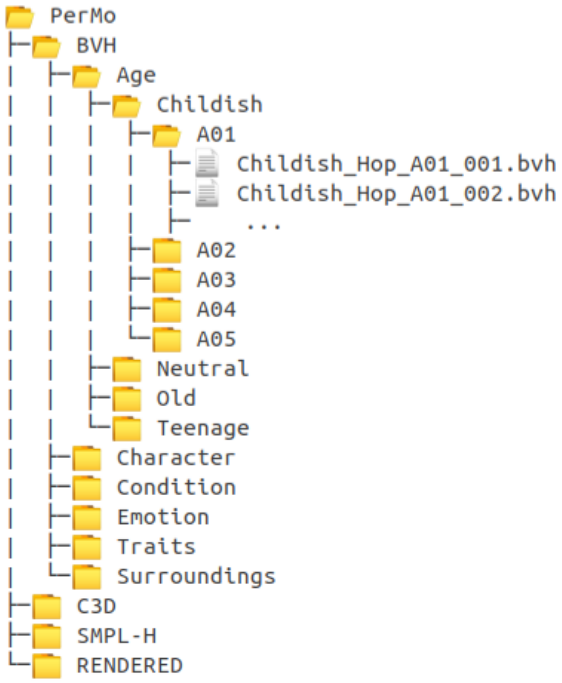


Figure F. Folder structure of the PerMo dataset

ChatGPT to create 20–30 variations of sentences, ranging from short and simple high-level descriptions to long and detailed low-level ones. As a result, we obtained descriptions of the 'Hop' motion at various levels of detail, as shown in Fig. G (a). Additionally, we represented the placeholder word [P*] within the sentences using *sks* [30]. When extracting sentence features, the word embedding corresponding to the *sks* index is replaced with P^*.

B. Ablation Study on Hyperparameters

We conduct an ablation study to examine the effects of hyperparameters s_t, s_v, g_t, g_v, and b, as shown in Fig. H. The experiments were performed using the single-input PersonaBooth. Fig. H (a) shows how the FID, R-precision (Top 3), and PRA metrics vary with the scaling factors s_t and s_v. As the scaling factor increases, the influence of personality features becomes more pronounced, leading to higher PRA

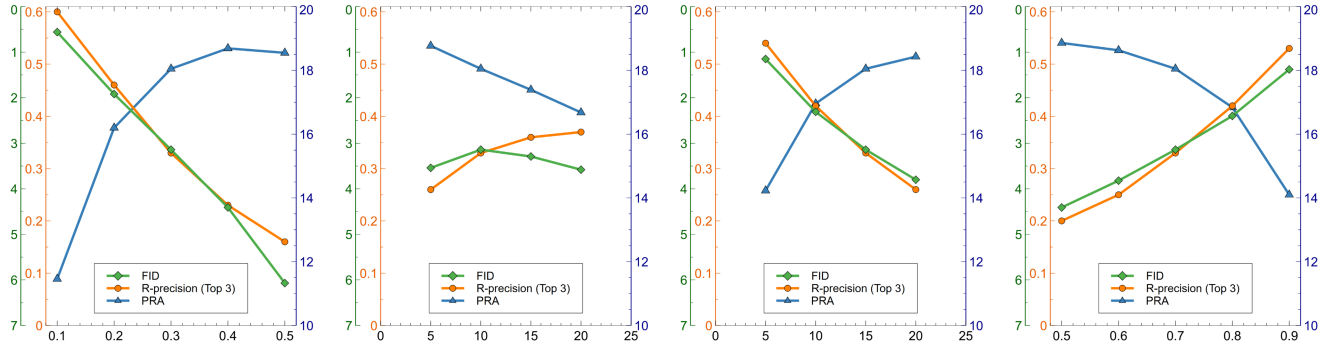
```
sks person hops ahead.
sks person is jumping ahead.
sks hops forward on one leg.
sks is hopping on a single foot.
sks person is hopping on one leg.
sks person jumps ahead on one foot.
sks hops forward several times on one leg.
sks person is bouncing ahead on one foot.
sks leaps forward, balancing on a single leg.
sks person hops forward, using only one leg.
sks person hops, moving steadily on one leg.
sks person leaps, using only one leg to move.
sks person hops several times, moving on one leg.
sks hops forward repeatedly, balancing on one leg.
Hopping on a single foot, sks person moves ahead.
sks person hops ahead, keeping balance on one leg.
sks person leaps, landing on the same leg each time.
sks person hops forward, using one leg for each leap.
sks person hops forward repeatedly on the same leg.
sks hops forward, pushing off with one leg each time.
sks person jumps several times, keeping balance on one leg.
sks person leaps forward several times, balancing on one leg.
sks person hops several times, using one leg to move forward.
sks person hops forward repeatedly, staying balanced on one leg.
sks person jumps forward on one leg, moving with steady balance.
sks hops forward, keeping balance on one leg with each movement.
sks person hops forward on one leg, repeating the motion several times.
```

(a) Hop

```
sks is jumping ahead.
sks person leaps ahead.
sks person leaps forward.
sks person bounds forward.
sks jumps ahead repeatedly.
sks person jumps repeatedly.
sks jumps forward several times.
sks takes several forward jumps.
sks makes continuous forward jumps.
sks person jumps forward repeatedly.
sks is making several forward leaps.
sks person jumps ahead multiple times.
sks person hops forward several times.
sks person repeatedly springs forward.
sks person hops forward multiple times.
sks person jumps forward several times.
sks person leaps forward several times.
sks person is taking several forward jumps.
sks person moves ahead with a series of jumps.
sks person moves forward through a series of jumps.
sks person propels themselves forward with several hops.
Jumping forward repeatedly, the sks person covers ground.
sks person pushes off the ground and jumps forward several times.
sks person pushes off the ground and performs several forward jumps.
```

(b) Jump

Figure G. Examples of text descriptions in PerMo dataset



(a) Scaling Factor (s_t, s_v) (b) Textual Guidance Scale (g_t) (c) Visual Guidance Scale (g_v) (d) Balancing Factor (b)

Figure H. Ablation study for hyperparameters s_t, s_v, g_t, g_v, and b. Higher positions on the graph indicate better performance across all three metrics, FID, R-precision (Top 3), and PRA.

values. However, PRA values reach a saturation point at scaling factors above 0.4. In contrast, FID and R-precision exhibit a trade-off with PRA, showing a decline in performance as the scaling factor increases.

Fig. H (b) shows the results for the textual guidance scale g_t. As g_t increases, the generated motions align more closely with the prompt, leading to improved R-precision. Conversely, PRA decreases, demonstrating a trade-off with R-precision. The FID metric performs best around $g_t = 10$. Fig. H (c) presents the results for the visual guidance scale g_v. As g_v increases, the generated motions more accurately reflect the features of the input motion, resulting in higher PRA values. Conversely, both FID and R-precision performance decrease. Fig. H (d) shows the results for the balancing factor b. Larger b values place more emphasis on the text, leading to a decrease in PRA but an improvement in both FID and R-precision. In selecting hyperparameters, we prioritized reflecting the persona, as the main objective of this task is personalization. Therefore, we aimed to maintain a high level of PRA while also achieving favorable FID and R-precision scores.

C. Ablation Study on the Number of Inputs

An ablation study on the number of inputs is conducted in the multiple input (MI) setting. As $|M_i|$ increases, all metrics demonstrate an increasing trend. This can be attributed to the higher likelihood of encountering motions that are more contextually aligned with the prompt as the diversity of input motions grows.

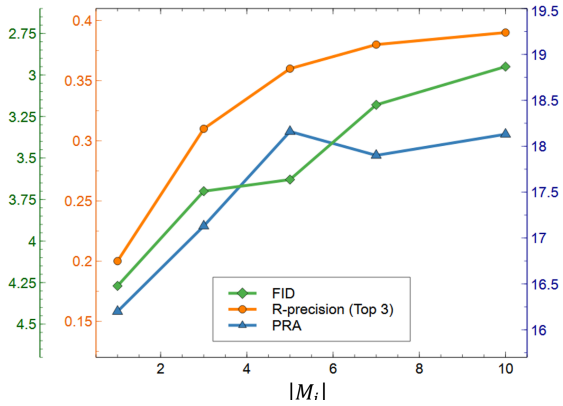


Figure I. Ablation study on the number of input motions. Higher positions on the graph indicate better performance.

Table D. Sup-parent categories for measuring the PRA metric

Sub-Parent Category	**Persona Category**
Age	Childish, Neutral, Old, Teenage
Character 1	Ballerina, Hulk, Monkey, Ninja
Character 2	Penguin, Robot, SWAT, Waiter, Zombie
Condition 1	Arm-aching, Drunken, Exhausted, Head-aching
Condition 2	Healthy, Leg-aching, Text-necked
Emotion 1	Angry, Fearful, Happy
Emotion 2	Sad, Strained, Surprising
Traits	Elegant, Shy, Silly, Uppity
Surroundings	Cold, Crowded, Muddy-floor, Unpleasant-floor

D. Persona Recognition Accuracy Details

For PerMo, PRA aims to classify 170 personas. To reduce the burden on a single classifier handling too many classes, we used nine separate classifiers, each responsible for a sub-parent category. Table D shows these sub-parent categories for PRA. For instance, the classifier for the ”Age” category handles 20 personas, with each style containing five different personas.

Each classifier is a 2-block transformer-based model, pretrained separately for each sub-parent category. The classifiers are trained using ground truth motions in PerMo paired with persona labels. These classifiers evaluate the generated motions, and their classification accuracy determines the PRA. Note that all 170 personas are combined into a single dataset during training. For 100Style, Style Recognition Accuracy (SRA) is evaluated across 40 style categories, excluding content-oriented ones. Only one classifier is used for 100Style.

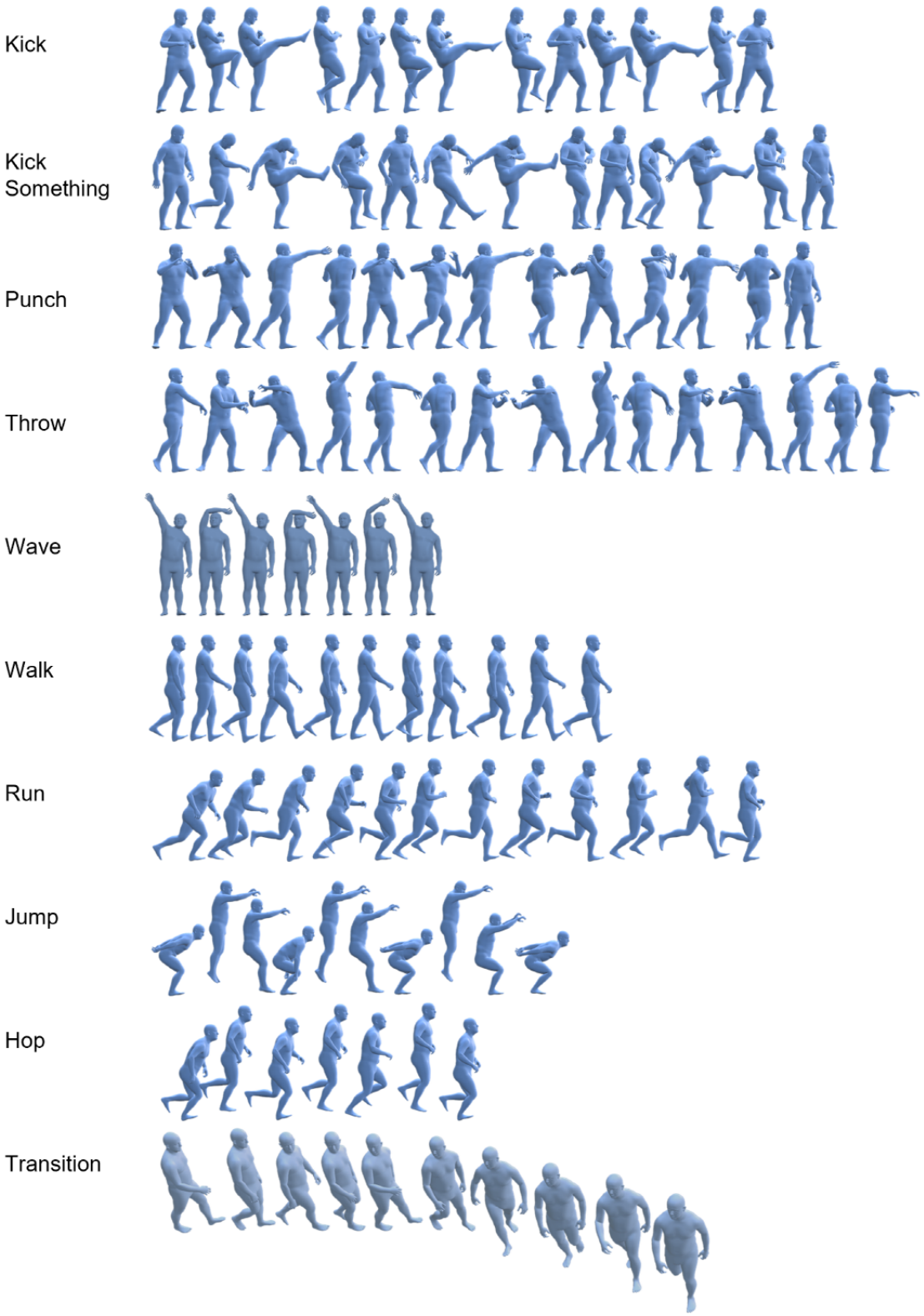


Figure J. Examples of content types in the PerMo dataset. Please refer to the attached video for a more detailed visualization

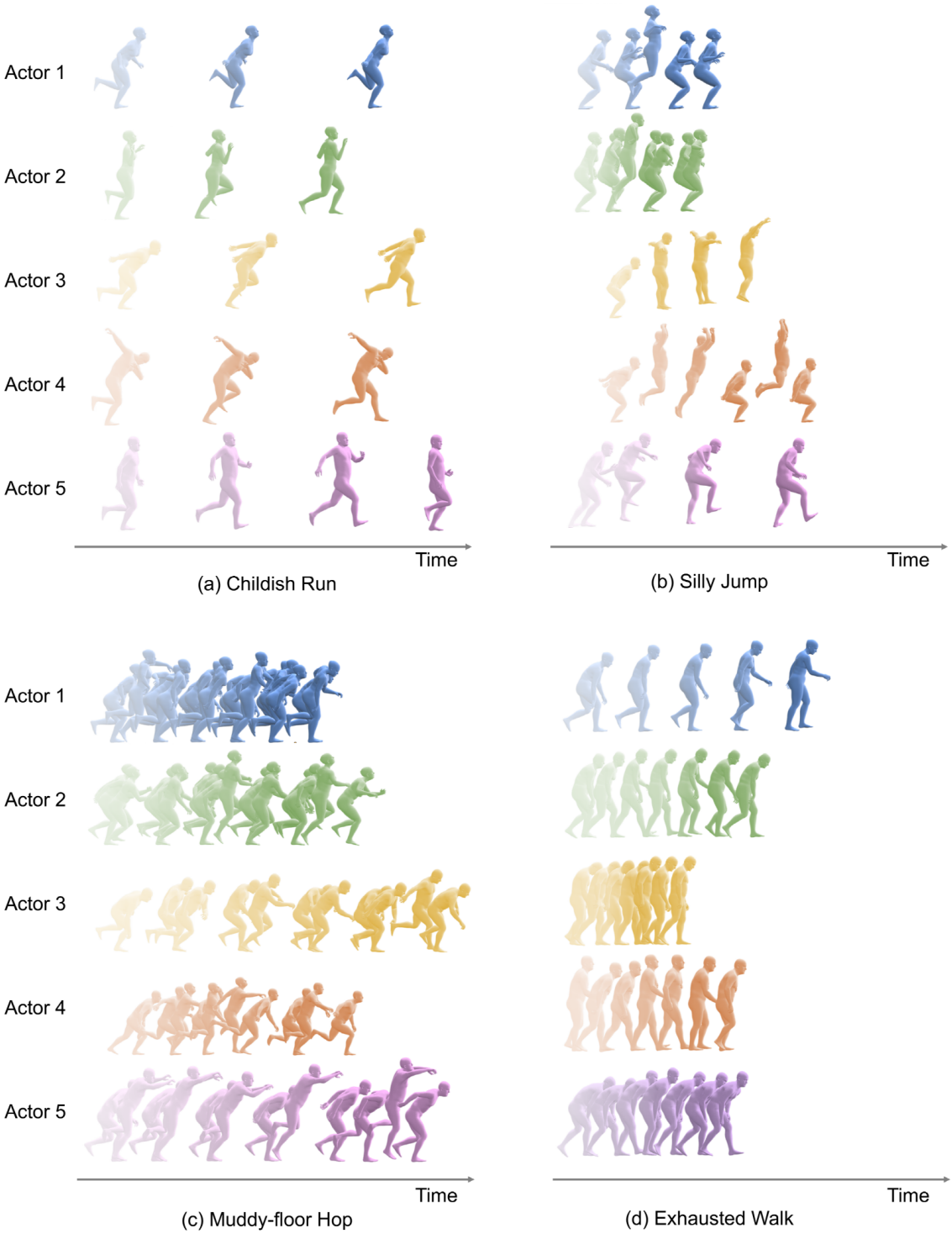


Figure K. Examples of personas in the PerMo dataset. Even for the same style and content, each actor portrays a different persona. Please refer to the attached video for a more detailed visualization