A. Dataset Analysis Details

A.1. Statistics

We provide a visualization of statistics of the objects that occur in our dataset and the sequence duration distribution in Fig. 11. A detailed comparison of actions, objects and scenes between existing dataset statistics is included in the supp. material.

A.2. Metrics

A.2.1. Metrics Overview

To quantitatively assess the quality of our HUMOTO dataset compares to others, we define the following metrics that capture different aspects of motion naturalness and interaction accuracy.

For human and object motion: **Foot sliding** measures unnatural horizontal movement during ground contact. For foot joints below a height threshold, we calculate horizontal displacement with a weighting function that decreases as joints lift from the ground. Lower values typically indicate more natural motion. **Jerk** quantifies motion smoothness by measuring the rate of change of acceleration. Lower jerk represents smoother motions. **Motion Signal-to-Noise Ratio (MSNR)** evaluates motion quality through the SNR of joint kinematics. Higher SNR indicates smoother motion, though overly smoothed signals may lose important details. **Coherence** quantifies motion consistency by measuring pose cluster compactness. Values approaching 1 indicate highly consistent movement patterns with minimal deviation. **Diversity** measures variety of motion patterns using normalized Shannon entropy across pose clusters. Higher values indicate a wider range of motion patterns, though this may potentially identify jitter as diversity.

For interaction quality: **Penetration** assesses the physical plausibility of human-object interactions by measuring object intrusion into the human mesh. Lower values indicate more physically plausible interactions. **Contact entropy** quantifies the diversity of interaction states and transitions. Higher values indicate more diverse and complex interactions with a balanced distribution of contact behaviors. **State consistency** measures the temporal stability of interaction states, rewarding smooth contacts while penalizing rapid fluctuations. Higher scores indicate more consistent interaction states with fewer changes.

Jerk is computed for both human and object motion. Foot sliding, MSNR, Coherence, and Diversity apply only to human motion. Penetration, contact entropy, and state consistency evaluate human-object interaction quality. These metrics are influenced by features of the dataset that do not necessarily represent quality issues. Therefore, they should be interpreted holistically rather than in isolation, as their values are influenced by multiple factors including motion and interaction complexity. A complete definition of metrics is provided in Appendix A.2.2.

A.2.2. Metrics Formulation

Foot sliding measures unnatural horizontal movement during ground contact. For each foot joint j (ankles and toes) with height below threshold H_j, we compute:

$$\text{Sliding}_j = N_f \sum_{t \in \mathcal{S}_j} \|\mathbf{p}_{j,t+1}^{xy} - \mathbf{p}_{j,t}^{xy}\|_2 \cdot (2 - 2^{(\mathbf{p}_{j,t}^{z}/H_j)}) \quad (1)$$

where $\mathbf{p}_{j,t}$ is the position of joint j at frame t, $\mathcal{S}_j$ are frames where $\mathbf{p}_{j,t}^{z} < H_j$, and N_f is the total frame count. The exponential weighting function gradually decreases influence as joints lift from the ground. The final metric averages across all four foot joints and is reported in centimeters. In a standard setting, the lower the foot sliding value, the more natural the motion.

Jerk quantifies motion smoothness by measuring the rate of change of acceleration. For a sequence of joint positions $\mathbf{p}$ with N_f frames, we compute:

$$\text{Jerk} = \frac{1}{N_f - 3} \sum_{t=1}^{N_f - 3} \|\mathbf{a}_{t+1} - \mathbf{a}_t\|_2, \quad (2)$$

where velocities and accelerations are calculated as finite differences. As indicated here, lower jerk represents more smooth motions.

Motion Signal-to-Noise Ratio (MSNR) quantifies motion quality through the SNR of joint kinematics, computed as:

$$\text{SNR} = 10 \log_{10} \left(\frac{P_{\text{signal}}}{P_{\text{noise}}} \right) = 10 \log_{10} \left(\frac{\mathbb{E}[\hat{v}^2]}{\mathbb{E}[|v - \hat{v}|^2]} \right), \quad (3)$$

where v represents the normalized local joint velocities, and $\hat{v}$ is the temporally smoothed version of v obtained through convolution with a kernel size of 3. This metric captures the relationship between meaningful motion patterns and undesirable jitter or noise. A higher SNR value indicates a smoother motion. However, we should note that an overly smoothed signal may lose important details or contain less informative action.

Coherence score quantifies motion consistency by measuring pose cluster compactness. We compute coherence as

$$C = 1 - \frac{\mu_d}{\max_d}, \quad (4)$$

where μ_d is the mean distance from poses to their cluster centroids, and $\max_d$ is the maximum observed distance. Values approaching 1 indicate highly consistent movement patterns with minimal deviation.

Diversity metrics, on the other hand, quantify the variety of motion patterns in a dataset. We compute motion diversity using normalized Shannon entropy across pose clusters.

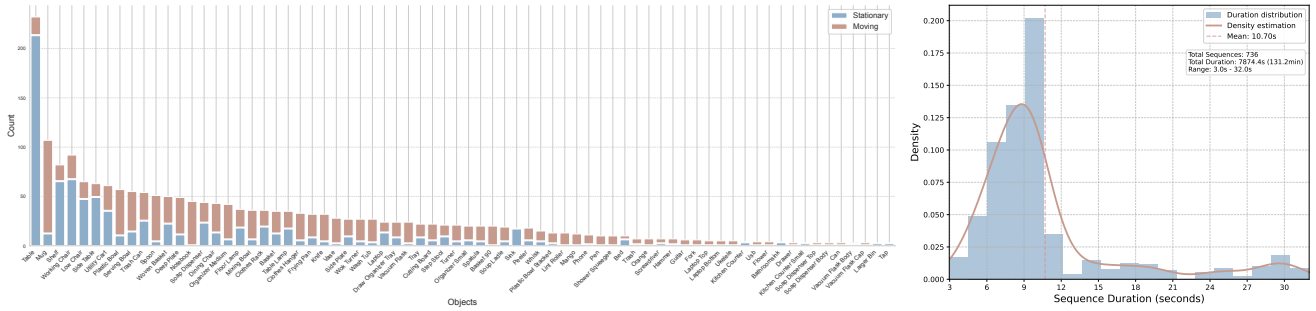


Figure 11. **Dataset statistics.** *Left:* Object occurrence frequency by motion type (stationary vs. moving). *Right:* Sequence duration distribution across the dataset.

After k-means clustering, we calculate

$$D = -\frac{\sum_{i=1}^{n} p_i \log_2 p_i}{\log_2 n}, \tag{5}$$

where p_i represents the proportion of frames in the i-th cluster. Higher diversity values indicate a wider range of motion patterns. However, this metric also identifies jittering or noise as diverse patterns.

Penetration quantifies the physical plausibility of human-object interactions by measuring object intrusion into the human mesh. For each frame, we sample points $\mathcal{P}_{obj}$ on object surfaces and compute the maximum penetration depth as:

$$\text{Penetration}(t) = \min_{p \in \mathcal{P}_{obj}} d(p, \mathcal{M}_h), \tag{6}$$

where $d(p, \mathcal{M}_h)$ is the signed distance from point p to the human mesh $\mathcal{M}_h$. Positive distances indicate interior points, with more positive values representing deeper penetration. We report the average maximum penetration across all frames, with lower values indicating more physically plausible interactions.

Contact entropy quantifies the diversity of interaction states and transitions during human-object interaction. For a sequence of interaction states discretized into categories (large penetration, contact, proximity, and distance), we compute:

$$\text{Entropy} = -\sum_{i,j} p(s_i \to s_j) \log_2 p(s_i \to s_j), \tag{7}$$

where $p(s_i \to s_j)$ is the probability of transitioning from state s_i to state s_j across all sampled points and frames. Higher entropy values indicate more diverse and complex interactions, with a balanced distribution of different types of contact and approach behaviors.

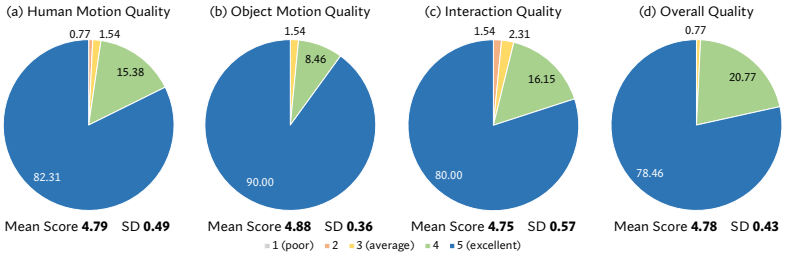
State consistency measures the temporal stability of interaction states, rewarding smooth and persistent contacts while penalizing rapid state fluctuations. For each sampled point, we calculate the average run length normalized by sequence length:

$$\text{Consistency} = \frac{1}{N_p} \sum_{p=1}^{N_p} \frac{\text{Avg. Run Length}_p}{\text{Sequence Length}}. \tag{8}$$

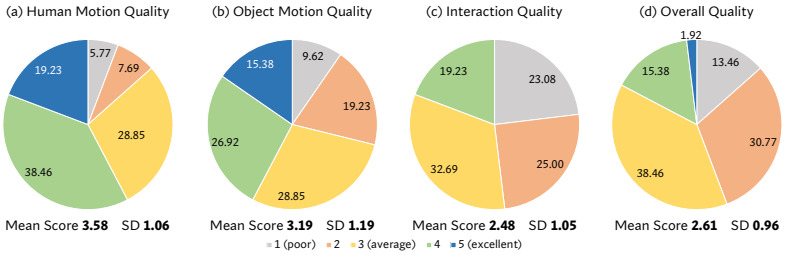
We additionally penalize points with large penetrations by applying a scaling factor based on large penetration duration. Higher consistency scores indicate a more consistent interaction state with fewer state changes.

A.3. Perceptual Evaluation Results

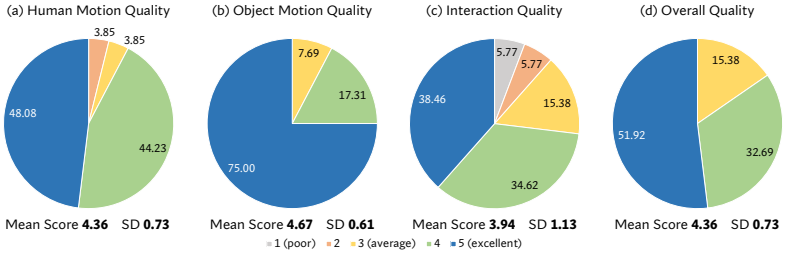
Following the details of our perceptual study setup provided in Sec. 4.2.2, we provide the detailed score distribution percentages of absolute quality evaluations in Fig. 12 and pairwise evaluations in Fig. 13.



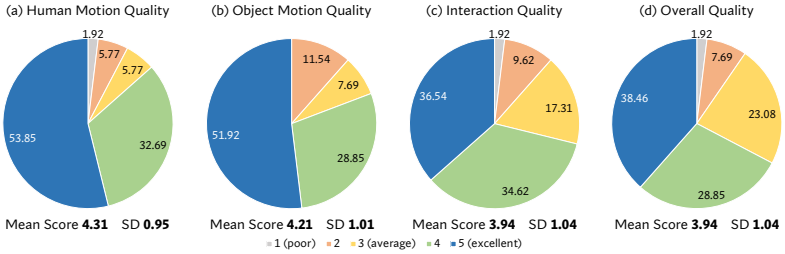
(a) HUMOTO



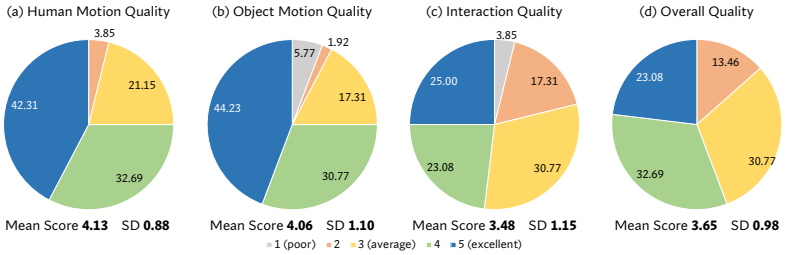
(b) BEHAVE



(c) OMOMO



(d) IMHD



(e) ParaHome

Figure 12. **Perceptual absolute quality ratings.** We show the aggregate percentages of absolute quality ratings on five-point Likert scales from our participants for HUMOTO, BEHAVE [1], OMOMO [39], IMHD [82], and ParaHome [35]. We assess the quality on four aspects: *(a) Human Motion Quality*, how plausible the human motions appear; *(b) Object Motion Quality*, how plausible the object motions appear; *(c) Interaction Quality*, how realistic the interactions between the humans and the objects appear; and *(d) Overall Quality*, how realistic the overall animations appear. We observe significant increases in ratings of 5 for HUMOTO in all four aspects.

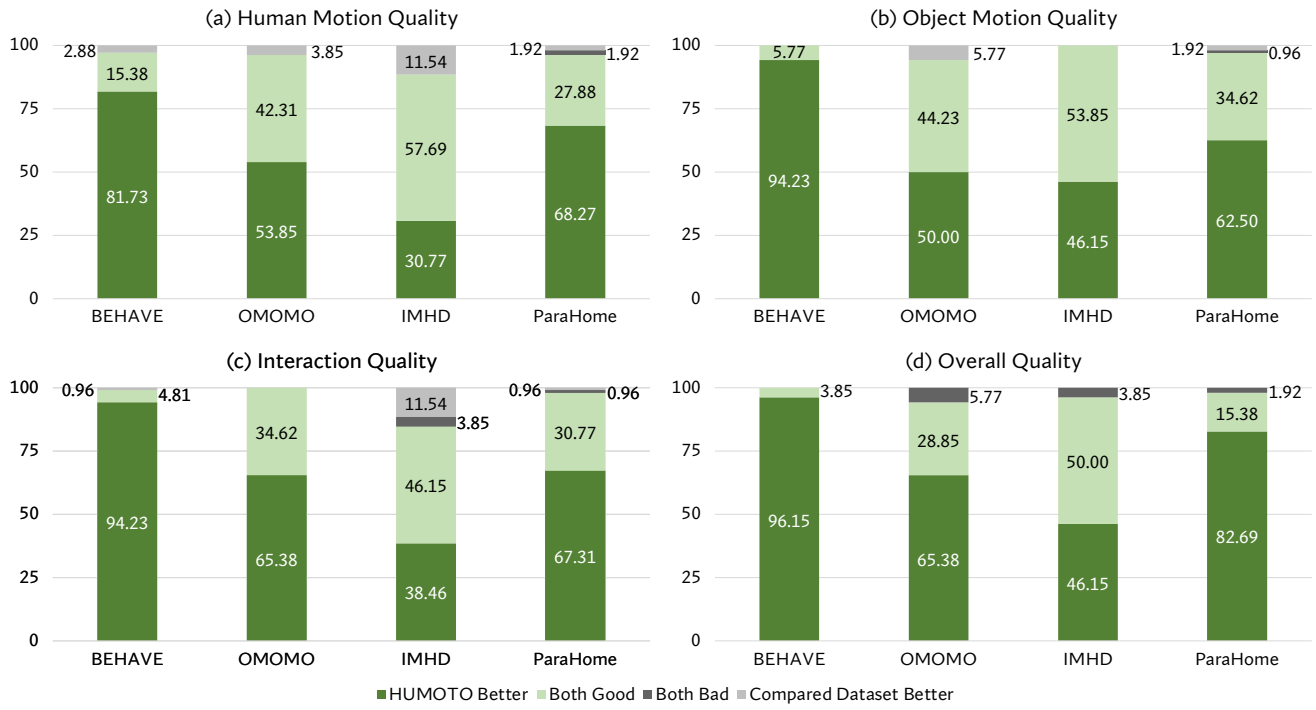


Figure 13. **Perceptual pairwise comparisons.** We show the aggregate percentages of pairwise comparison results from our participants, comparing side-by-side between HUMOTO and other datasets, including BEHAVE [1], OMOMO [39], IMHD [82], and ParaHome [35]. We assess the comparisons on four aspects: *(a) Human Motion Quality*, how plausible the human motions appear; *(b) Object Motion Quality*, how plausible the object motions appear; *(c) Interaction Quality*, how realistic the interactions between the humans and the objects appear; and *(d) Overall Quality*, how realistic the overall animations appear. After accounting for ties, we observe significant preferences for HUMOTO in all four aspects.

As an expert screenwriter and director, your task is to create scripts for a human-object interaction dataset. You'll be given a set of objects and a specific setting. Please follow these guidelines to create the script:
1.Object usage: Only use objects from the provided list. You may use all or a subset of these objects in each script. You cannot use any objects that not in the provided list.
2.Interaction focus: Each script should involve one person interacting with one or more objects.
3.Return format: Please return JSON output with format {"scene": <your_scene>, "short_script": <your_short_script>, "long_script": <your_long_script>, "objects": ["objects", "you", "used"]}.
Step 1: Set the scene: Envision a general setting where the given objects would naturally be found or follow the user provided setting.
Step 2: Write a concise script: Describe the interaction in 1-2 sentences, mentioning the objects involved.
Step 3: Write a longer script: Extend the interaction into a long paragraph. The goal is to create more accurate and detailed instructions that enable the actor to perform the intended actions precisely.
Step 4: List all objects you used in the scene, even those not provided originally.
setting: Your script should be based on a scene in a bedroom. You may select any combination of the provided objects. Your script should describe interaction lasts around 30 seconds.
objects: bed, clothes_hanger, clothes_rack, side_table, floor_lamp, mug, trash, trash_can, lint_roller, talbe_lamp, lower_chair.


```
{
  "scene": "Morning Outfit Selection",
  "short_script": "A person wakes up in bed, turns off the table lamp, walks to the clothes rack, and selects two clothes hangers with different outfits, holding each in front of themselves, examining their reflection in the mirror.",
  "long_script": "A person lying in bed stirs awake and slowly sits up, stretching and yawning. They reach over to the table lamp on the side table and touch the top of the lamp to turn it on. After a moment, they stand up from the bed and walk across the room toward a clothes rack filled with hanging garments. They pause in front of the rack, scanning the options, then carefully select a clothes hanger from the left side, lifting it off the rack with their right hand. They hold the hanger at arm's length, examining the garment. The person then walks to a mirror while holding the first hanger, positions themselves centered in front of the mirror, and holds the clothes on the hanger up against their body, tilting their head slightly while evaluating their appearance.",
  "objects": ["bed", "table_lamp", " side_table ", "clothes_rack", "clothes_hanger"]
}
```


As an expert screenwriter and director, your task is to create scripts for a human-object interaction dataset. Please follow the requirements and steps to write a script of human-object interaction based on the user given objects and its description:
1.Object usage: Only use the one provided object in your script.
2.Interaction focus: The interaction could only include one object and one performer. Your scripts should describe interactions lasts around 10 seconds.
3. Return format: Please return JSON output with format {"scene": <your_scene>, "short_script": <your_short_script>, "long_script": <your_long_script>, "objects": ["objects", "you", "used"]}.

<steps are the same as the scene-based scripts>

objects: mug: A ceramic mug with a comfortable handle, perfect for enjoying hot beverages like coffee or tea.


```
{
  "scene": "Mug Examination and Drinking",
  "short_script": "A person holds a mug, examines it, takes a sips, and sets it down again.",
  "long_script": "A person holds a ceramic mug by its handle, with their index finger threaded through the loop and their thumb resting on top. They lift the mug upward with a smooth motion, bringing it closer to their face and slightly rotating it to examine its contents. They then raise the mug to their lips, tilting it at approximately 30 degrees, and take a small sip. Finally, they lower the mug down to their waist level.",
  "objects": ["mug"]
}
```


Figure 14. **Examples of how we use LLMs to develop our human-object interaction scripts for capturing.** *Top:* We cluster objects into different scene types and create possible interactions within that scene. *Bottom:* For each individual object, we prompt LLMs on how one person would be possible to interact with the object.

Short script: The subject scoops ingredients using the spoon with left the hand from the deep plate. The subject adds ingredients with the left hand using the spoon to the mixing bowl. The subject mixes the content of the mixing bowl with the left hand.
Long script: The subject stand at the back of the table. The subject scoops ingredients inside deep plate with left hand using the spoon. The subject adds ingredients with left using spoon hand from deep plate to mixing bowl. The subject lifts the mixing bowl with the right hand. The subject inserts left hand into the mixing bowl. The subject mixes content with left hand inside the mixing bowl with.

Figure 15. **Motion generation results comparing our text-annotated dataset with MotionGPT [30].** *Top:* Generated motion sequence from short script input. *Middle:* Generated motion sequence from detailed long script input. *Bottom:* Ground truth motion sequence from our HUMOTO dataset. While MotionGPT can generate basic movements following general instructions, it struggles with the fine-grained hand-object interactions and precise manipulation sequences present in our dataset.