OpenT2M: No-frill Motion Generation with Open-source, Large-scale, High-quality Data

Supplementary Material

In this supplementary, we provide additional details of OpenT2M in Section 7. We provide the details about dividing human body into independent parts in Section 8. We also provide details of evaluation metrics in Section 9. We provide visualization examples of OpenT2M in Section 10.

7. Additional Analysis of OpenT2M

7.1. Data Distribution

Figure 5a shows the number distribution of motion sequences across different subsets in OpenT2M on a logarithmic scale, demonstrating variations in dataset sizes. OpenT2M integrates 21 curated subsets, amounting to a comprehensive collection of 1 million motion sequences. A substantial portion of motions in OpenT2M are extracted from web videos utilizing motion estimation models [32], such as Kinetics-700 [17], Internvid [38]. These motions undergo rigorous physically feasible validation and multi-granularity filtering. We set the number of visible keypoints to 8, while the whole body corresponds to 17 keypoints. Each motion sequence accounts for over 50% of the duration of the corresponding original video, ensuring temporal consistency and semantic validity. OpenT2M also integrates open-source human motion datasets [1, 2, 8], such as Motion-X [21]. Leveraging the proposed long-horizon motion curation pipeline, we construct 190K long-horizon motion sequences. The OpenT2M $_{\text{long}}$ comprises motions spliced from two, three, four, and five individual motion sequences. Figure 5b shows the average length distribution of OpenT2M across different subsets. We observe that the dataset with the shortest average sequence length is Postrack, comprising merely 16.12 frames, while 3DPW exhibits the longest average length, exceeding 500 frames. Following a meticulous curation process, OpenT2M exhibits a substantially longer average length compared with previous work [3].

7.2. Comparison of Long-horizon Datasets

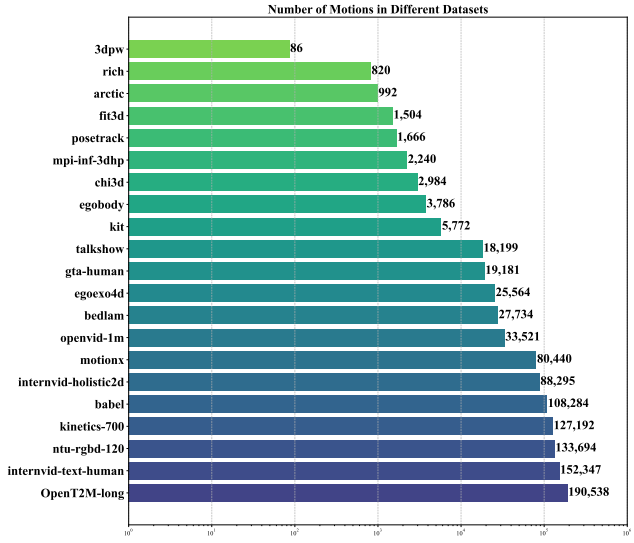
We first detail the pipeline for long-horizon motion curation. Two different motion sequences are initially aligned in orientation by rotating the initial frame of the second sequence to match the facing direction of the last frame in the first sequence. Subsequently, the entire second sequence is translated spatially to align its position with that of the last frame of the first sequence. Finally, a fixed transition duration is applied, during which spherical linear interpolation is performed between the last frame of the first motion and the initial frame of the second motion to ensure smooth kinematic continuity. To ensure that long-horizon motion sequences adhere to physical constraints, we utilize the concatenated motion sequence as reference poses for an RL policy, driving the avatar in the IsaacGym to track the reference motion. The resulting motion, refined through physical simulation, is adopted as the final long-horizon motion sequences.

Figure 6 shows the length distribution comparison between OpenT2M $_{\text{long}}$ and BABEL [30]. BABEL labels about 43 hours of mocap sequences from AMASS [26] with fine-grained action labels. BABEL exhibits a substantial variation in motion length, containing motion sequences from 5s to over 100s. In BABEL, 37.9% of motion sequences last 5s or less, which significantly limits its effectiveness for evaluating the long-horizon motion generation capability of T2M models. In contrast, OpenT2M $_{\text{long}}$ contains only 0.33% of motions within 5s. Furthermore, OpenT2M $_{\text{long}}$ contains 20 times motion sequences than BABEL. As a result, even intervals with relatively low proportions in OpenT2M $_{\text{long}}$ may contain a larger number of motions compared to BABEL. For instance, motions lasting from 35s to 40s only constitute 0.76% in OpenT2M $_{\text{long}}$, yet OpenT2M $_{\text{long}}$ contains 1,454 motion sequences from 35s to 40s. Meanwhile, although the same interval accounts for a higher proportion (0.9%) in BABEL, it represents merely 89 motions.

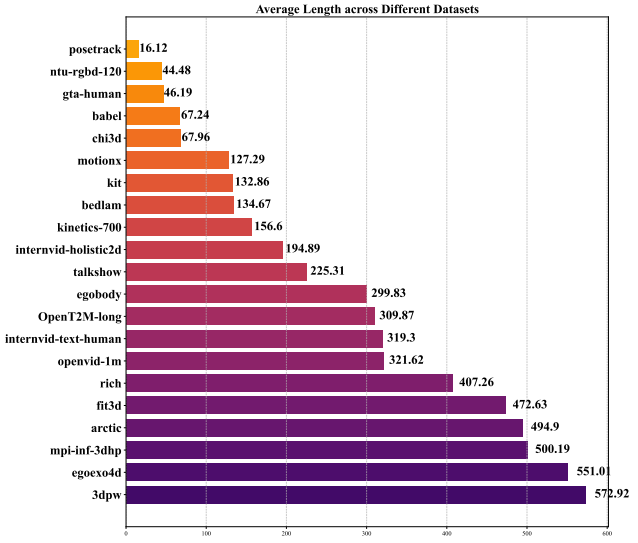
7.3. Second-wise Text Annotation

Previous works [3, 39] typically annotate motion sequences by directly feeding corresponding videos into Vision-Language Models (VLMs) to generate coarse textual descriptions. While this approach offers efficiency, it suffers from a critical limitation: motion sequences extracted from web videos often comprise complex and continuous motion clips. When VLMs are applied in an end-to-end manner to entire video clips, they tend to overlook fine-grained and crucial motion details. Such omissions impact the quality and utility of annotated texts, particularly for applications requiring high temporal precision or detailed kinematic analysis.

In this work, we design a second-wise annotation scheme as shown in Figure 7. The annotation task mainly contains second-wise captions and a general summary task. The annotation process begins by uniformly extracting video frames every 0.5s. Each second video frames are first annotated individually with second-wise descriptions. These second-wise captions are then summarized to form a precise caption for the entire video clip. In the annotation process, we deliberately exclude any descriptions of backgrounds, facial

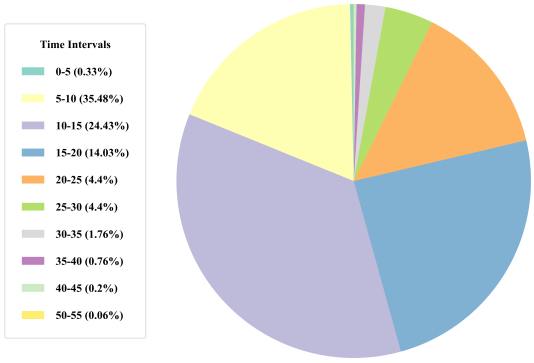


(a) Distribution of motion sequences across different subsets in OpenT2M.

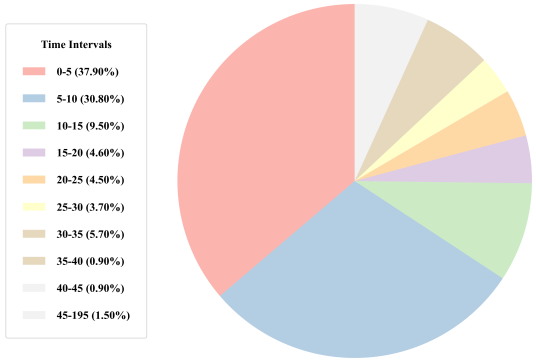


(b) Average length distribution of OpenT2M across different subsets.

Figure 5. Statistics of the OpenT2M dataset. (a) Motion sequence distribution (log scale). (b) Average motion length distribution.



(a) OpenT2M $_{\text{long}}$ Length Distribution



(b) BABEL Length Distribution

Figure 6. Length distribution comparison between OpenT2M $_{\text{long}}$ and BABEL datasets.

expressions, clothes, and other attributes that are irrelevant to human motion. We present annotated examples in Figure 8 to illustrate the precise alignment between text and motion.

8. Additional Details of 2D-PRQ

In this work, we propose 2D-PRQ, a tokenizer that divides the joints of the whole body into 5 parts, including:

- Left Hand: spine$_1$, spine$_2$, spine$_3$, left collar, left shoulder, left elbow, left wrist
- Right Hand: spine$_1$, spine$_2$, spine$_3$, right collar, right shoulder, right elbow, right wrist
- Left Leg: spine$_1$, spine$_2$, spine$_3$, left hip, left knee, left ankle, left foot
- Right Leg: spine$_1$, spine$_2$, spine$_3$, right hip, right knee, right ankle, right foot
- Torso: spine$_1$, spine$_2$, spine$_3$, neck, left collar, right collar, head

The pelvis, spine$_1$, spine$_2$, and spine$_3$ are shared across all parts, as they remain relatively stable during human motion. Each joint is represented by relative 6D rotations and redundant 3D positions, resulting in a dimensionality of 63+8 per part, including 4D root node and 4D foot contact information. When aggregating part features into motion features, we average the shared joints.

9. Evaluation Metrics

Text-to-motion. We adapt R-precision, MMDist, and FID to evaluate T2M model follow Guo et al. [11]. Each metric is illustrated as follows:

- **R-precision**: The retrieval metric is designed to evaluate the semantic consistency between text and generated motion. The R-precision is computed as the accuracy of its

Prompt Template

You are a video annotation AI that describes and analyzes ONLY the visible physical motion of the person within the given BBOX in the videos.

Input

You will receive a series of video frames in chronological order, with each frame sampled every0.5 seconds.

To analyze motion over each 1-second segment, group every three consecutive frames as follows:

- Second 1 : frames 0, 1, 2 (time 0.0s-1.0s)
- Second 2 : frames 2, 3, 4 (time 1.0s-2.0s)
- Second 3 : frames 4, 5, 6 (time 2.0s-3.0s)
- And so on.

In general, for the i-th second, analyze frames at indices (2i - 2), (2i - 1), and 2i, covering the time interval from (i - 1) to i seconds.

Task

Your task consists of 2 parts:

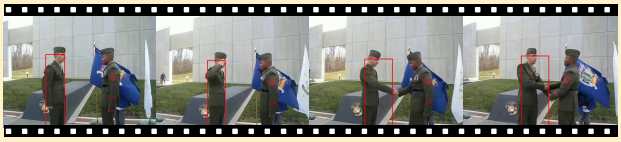
1. Second-wise Caption

For each 1-second segment of video, give one sentence to describe the physical motion of the person within the given BBOX. If the person within the given BBOX is not visible during that second, return null.

2. General Summary

After listing the second-wise results, give one sentence that summarizes the overall physical motion of the person within the given BBOX. This summary should:

- Describe the overall physical motion of the person in the BBOX to highlight common types of motion (e.g., walking, playing basketball, pivoting), but ONLY when 100% certain.
- Add some action details about the limb ONLY if clearly visible (e.g., left hand).



Text Annotation


```
{
  "second-wise": [
    {
      "second": 0,
      "start_frame": 0,
      "end_frame": 2,
      "text": "The person is standing still and looking to their right."
    },
    ......
    {
      "second": 2,
      "start_frame": 4,
      "end_frame": 6,
      "text": "The person raises their right hand to their forehead."
    },
    {
      "second": 3,
      "start_frame": 6,
      "end_frame": 8,
      "text": "The person continues to hold their right hand to their forehead."
    },
    {
      "second": 4,
      "start_frame": 8,
      "end_frame": 10,
      "text": "The person lowers their right hand."
    },
    ......
    {
      "second": 6,
      "start_frame": 12,
      "end_frame": 14,
      "text": "The person is shaking hands with their right hand."
    },
    ......
    {
      "second": 9,
      "start_frame": 18,
      "end_frame": 20,
      "text": "The person is shaking hands with their right hand."
    }
  ],
  "summary": "The person performs a salute and then shakes hands with another individual."
}
```


Figure 7. Prompt template for generating second-wise text annotations utilizing Gemini-2.5.

ground-truth text description being ranked Top-1 when retrieved by the generated motion from a text pool. Following Guo et al. [11], we set the size of the description pool to 32.

- **MMDist:** MultiModel Distance is computed as the average Euclidean distance between motion feature and corresponding text feature.
- **FID**: Frechet Inception Distance is designed to measure the similarity between the distribution of generated motions and ground-truth motion in the feature space. It is computed as the Fréchet distance between the feature distributions of the generated motion and ground-truth motion.

Motion Reconstruction. We adapt FID and MPJPE to evaluate motion tokenizers on the motion reconstruction task.

- **FID**: Similar to T2M, Frechet Inception Distance for motion reconstruction is computed as the Fréchet distance between the feature distributions of reconstruction motion and ground-truth motion.
- **MPJPE**: The metric is computed by averaging the L2 distances between all joints of reconstruction motion and ground-truth motion across all frames.

10. Visualization Examples

We provide visualization examples of `OpenT2M` in Figure 8. Visualization examples demonstrate that `OpenT2M` encompasses a diverse range of motion patterns and exhibits strong text-motion alignment, providing a high-quality data foundation for building large motion models.

Text Annotation: The person repeatedly lunges forward with their right arm extended and then retracts their arm while stepping back.

Text Annotation: The person is performing a series of dance moves, involving rotations, leans, and arm extensions.

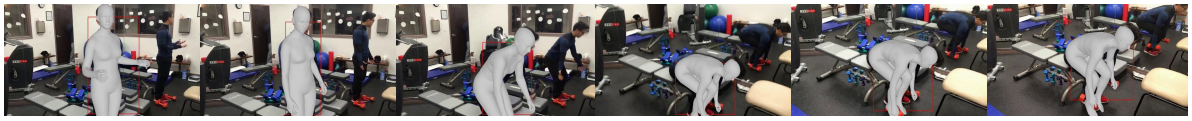
Text Annotation: The person is performing push-ups, moving their chest up and down towards and away from the floor.

Text Annotation: The person performs a series of slow, deliberate movements, characterized by shifting weight between legs, extending and retracting arms in a flowing motion.

Text Annotation: The person is a softball pitcher who performs a pitching motion including shifting weight, raising their arm, and releasing the ball, followed by recovery.

Text Annotation: The person is performing a weightlifting movement, starting from a deep squat and lifting the barbell overhead.

Text Annotation: The person performs ballistic side lunges with dumbbells, alternating lunges to the right and left while maintaining an upright posture.



Text Annotation: The person starts standing and talking, then bends down to pick up two dumbbells and continues to hold them in a bent-over position.

Figure 8. Visualization examples of OpenT2M, each example is annotated with precise text.