RobotMotion: A Multi-Source Humanoid Behavior Dataset for Physical AI

Abstract

RobotMotion is a multi-source dataset designed to support learning and evaluation of humanoid and physical AI systems. The dataset integrates egocentric human demonstrations, large-scale simulation data, and teleoperated humanoid robot executions, all organized under a unified skill–episode abstraction. RobotMotion emphasizes transparent data formats, synchronized multi-modal sensing, and compatibility with open-source robot learning pipelines, enabling research across imitation learning, reinforcement learning, and human-to-robot transfer.

I. Introduction

Robotic systems capable of operating in unstructured, real-world environments require large amounts of diverse, high-quality data to acquire robust and generalizable behaviors. In recent years, progress in robot learning has been driven in part by the availability of large-scale datasets capturing manipulation, navigation, and interaction tasks across a variety of settings. However, existing datasets often focus on a single data source, embodiment, or sensing configuration, limiting their applicability across different learning paradigms and robot platforms.

Human demonstrations provide rich behavioral priors and task structure, but collecting large-scale, well-organized human data that aligns with robotic learning requirements remains challenging. Simulation offers scalability and perfect state observability, yet transferring learned behaviors to real robots often suffers from domain mismatch. Teleoperated robot demonstrations provide direct embodiment alignment but are costly to collect and typically limited in scale. As a result, no single data source fully addresses the diverse needs of modern physical AI systems.

To address these challenges, we introduce RobotMotion, a multi-source humanoid behavior dataset that integrates egocentric human demonstrations, large-scale simulation data, and teleoperated humanoid robot executions under a unified representation. RobotMotion organizes all data using a shared skill–episode abstraction, enabling consistent indexing, synchronization, and analysis across heterogeneous data modalities and collection processes.

RobotMotion emphasizes flexible modality support rather than fixed sensor specifications. Visual observations, proprioceptive signals, audio, and optional tactile information are recorded when available, with explicit metadata describing modality presence, frequency, and synchronization. Data are stored in transparent, human-readable formats compatible with open-source robot learning pipelines, facilitating inspection, reuse, and extension.

II. Related Work

Recent research in robot learning has explored large-scale datasets, human demonstration-based learning, and simulation-driven approaches to enable robots to acquire diverse and generalizable skills.

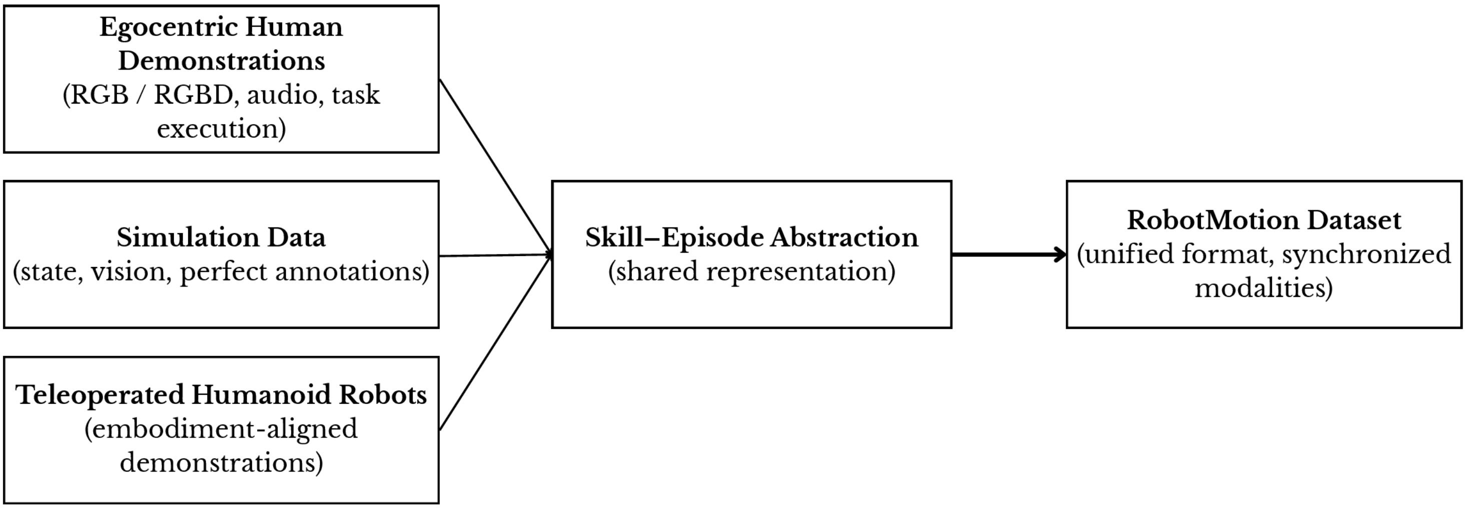

Egocentric Human Demonstrations
(RGB / RGBD, audio, task execution)
Simulation Data
(state, vision, perfect annotations)
Teleoperated Humanoid Robots
(embodiment-aligned demonstrations)
Skill–Episode Abstraction
(shared representation)
RobotMotion Dataset
(unified format, synchronized modalities)

Figure 1. Dataset overview. RobotMotion integrates three complementary data sources—egocentric human demonstrations, simulation data, and teleoperated humanoid robot executions—under a unified skill–episode abstraction. This shared representation enables consistent organization, synchronization, and analysis of heterogeneous data modalities across collection sources.

A. Large-Scale Robot Learning Datasets

Recent efforts have introduced large-scale datasets to support data-driven robot learning across diverse manipulation and interaction tasks. Datasets such as RH20T focus on contact-rich manipulation sequences with synchronized multi-modal sensing, while Open X-Embodiment aggregates data from a wide range of robot platforms and embodiments to enable general-purpose policy learning. More recent humanoid-focused datasets explore whole-body behaviors and embodiment-aligned demonstrations. While these datasets have significantly advanced robot learning, they typically emphasize a single data source or collection paradigm, motivating complementary approaches that integrate multiple forms of supervision.

B. Human Demonstration and Imitation Learning

Human demonstrations have long been used as an effective source of supervision for robot learning, providing rich behavioral priors and task structure that are difficult to specify manually. Imitation learning approaches leverage demonstrations to bootstrap policy learning, reduce exploration complexity, and improve sample efficiency, particularly for contact-rich manipulation and long-horizon tasks. As a result, human demonstration data has played a central role in enabling robots to acquire complex skills in real-world environments.

Recent work has explored a variety of demonstration modalities, including kinesthetic teaching, teleoperation, and video-based learning. Egocentric video demonstrations, in particular, have attracted growing interest due to their ability to capture task intent, object interactions, and temporal structure from a first-person perspective. However, large-scale human demonstration datasets often lack direct alignment with robot embodiments, sensing configurations, or control interfaces, limiting their direct applicability to robotic policy learning. Moreover, collecting and organizing human demonstrations at scale presents practical challenges related to annotation, synchronization across modalities, and consistency of task definitions. These limitations motivate approaches that preserve the richness of human demonstrations while enabling structured integration with other data sources and learning pipelines.

RobotMotion builds upon this line of work by incorporating egocentric human data within a unified representation that facilitates alignment with simulation and teleoperated robot data.

C. Simulation and Digital Twins for Robotics

Simulation has become a fundamental tool in robot learning due to its scalability, safety, and access to precise state information. Simulated environments enable the collection of large amounts of data at low cost and allow for controlled experimentation across diverse tasks, object configurations, and environmental conditions. As a result, simulation-based datasets have been widely used to train policies for manipulation, navigation, and whole-body control.

Despite these advantages, transferring policies learned in simulation to real robots remains challenging due to discrepancies in dynamics, sensing, and contact interactions. To mitigate this gap, prior work has explored techniques such as domain randomization, system identification, and digital twin modeling, which aim to align simulated environments more closely with real-world execution.

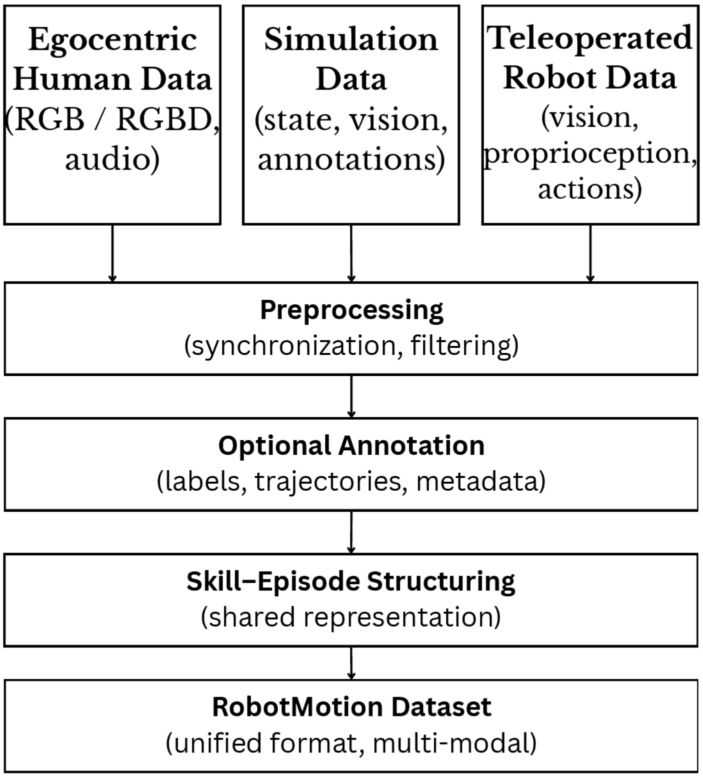
Digital twins aim to capture task-relevant aspects of the physical world within simulation while abstracting away unnecessary complexity.

While simulation and digital twin approaches offer scalability and structured supervision, they are most effective when complemented by real-world data. This observation has motivated hybrid strategies that combine simulation with human demonstrations or teleoperated robot data. RobotMotion follows this direction by integrating simulation data alongside egocentric human and teleoperated robot demonstrations within a shared representation, enabling consistent use of simulation-derived structure together with real-world behavioral data.

Together, these lines of work highlight the need for dataset designs that integrate multiple data sources under a unified and extensible representation.

III. RobotMotion Dataset

All data sources are processed through a shared pipeline that synchronizes heterogeneous observations and organizes them into a common skill–episode representation for consistent analysis and reuse.


Egocentric Human Data
(RGB / RGBD, audio)
Simulation Data
(state, vision, annotations)
Teleoperated Robot Data
(vision, proprioception, actions)
Preprocessing
(synchronization, filtering)
Optional Annotation
(labels, trajectories, metadata)
Skill–Episode Structuring
(shared representation)
RobotMotion Dataset
(unified format, multi-modal)

III.A. Data Sources

RobotMotion integrates three complementary data sources to capture diverse aspects of humanoid behavior: egocentric human demonstrations, simulation data, and teleoperated humanoid robot executions.

Each data source provides distinct supervision signals and inductive biases, and together they enable a unified representation that supports learning across heterogeneous modalities, embodiments, and environments.

III.A.1 Egocentric Human Demonstrations

Egocentric human demonstrations provide rich behavioral priors and natural task execution strategies grounded in real-world environments. In RobotMotion, human data is collected from a first-person perspective using wearable sensors, primarily RGB or RGBD cameras, and may optionally include audio signals. These demonstrations capture task-relevant visual context, object interactions, and temporal structure directly from human execution, without assuming a specific robot embodiment.

Human demonstrations are organized into task-specific episodes, each corresponding to a single execution of a skill under a particular environment configuration. This data source emphasizes realism, diversity, and natural variability in task execution, while remaining agnostic to downstream robot morphology. As such, egocentric human data in RobotMotion serves as a flexible foundation for imitation learning, representation learning, and cross-embodiment transfer.

III.A.2 Simulation Data

Simulation data provides scalable, controlled, and fully observable task executions that complement real-world demonstrations. RobotMotion incorporates simulation-generated episodes collected from physics-based simulators, enabling access to precise state information, perfect annotations, and large-scale data generation under diverse conditions.
Simulated data may include visual observations, proprioceptive states, object poses, contact information, and automatically generated annotations such as segmentation masks and bounding boxes. Domain randomization can be applied to object properties, scene layouts, lighting conditions, and dynamics to increase diversity and robustness. All simulation episodes are aligned with the same skill-episode abstraction used for real-world data, allowing simulated trajectories to be analyzed and utilized alongside human and robot demonstrations.

III.A.3 Teleoperated Humanoid Robot Data

Teleoperated humanoid robot demonstrations provide embodiment-aligned data that directly reflects the kinematic and dynamic constraints of real robotic systems. In RobotMotion, teleoperation is used to collect robot-centric executions of skills, capturing synchronized visual observations, proprioceptive signals, and control actions. These demonstrations bridge the gap between human intent and robot embodiment by grounding task execution in physical robot platforms. Compared to purely simulated data, teleoperated robot data captures real-world dynamics and sensor noise; compared to human demonstrations, it provides direct correspondence between observations and robot actions. While teleoperation is typically more costly and limited in scale, its inclusion in RobotMotion enables studies of human-to-robot transfer, policy fine-tuning, and embodiment-aware learning.

III.A.4 Complementarity of Data Sources

Each data source in RobotMotion addresses distinct limitations of the others. Egocentric human demonstrations offer realism and behavioral diversity, simulation data provides scale and precise supervision, and teleoperated robot executions ensure embodiment alignment. By integrating these sources under a unified skill–episode representation, RobotMotion enables consistent organization, synchronization, and analysis of heterogeneous data modalities, supporting flexible use across a wide range of robot learning paradigms.

III.B. Skill–Episode Representation

RobotMotion organizes all data using a unified skill–episode representation designed to provide a consistent abstraction across heterogeneous data sources, modalities, and embodiments. This representation serves as the fundamental organizational structure of the dataset, enabling alignment between egocentric human demonstrations, simulation-generated trajectories, and teleoperated humanoid robot executions while preserving task semantics and temporal structure.

III.B.1 Skills

A skill in RobotMotion corresponds to a task-level semantic unit, such as pick up an object, pour liquid, or carry an item. Each skill defines a consistent task specification independent of the data source or embodiment, allowing demonstrations collected from humans, simulators, or robots to be grouped under a shared semantic definition. Skills are identified by a unique identifier and accompanied by a concise, human-readable description that specifies the intended task and success criteria.

By organizing data at the skill level, RobotMotion enables aggregation of diverse executions that share common task intent while allowing variation in environment configuration, execution strategy, and sensing modality. This abstraction supports task-centric analysis and facilitates learning approaches that operate across multiple embodiments or data sources.

III.B.2 Episodes

An episode represents a single execution instance of a skill over a finite temporal horizon. Each episode consists of time-ordered observations and, when available, actions corresponding to one complete task execution. Episodes may vary in duration, sensing configuration, and data availability depending on the collection source, but all episodes adhere to a shared structural definition.

Within an episode, all recorded modalities—including visual observations, proprioceptive signals, audio, and optional tactile information—are temporally synchronized using frame indices and/or timestamps. This temporal alignment ensures consistent correspondence across modalities and enables downstream processing methods that rely on synchronized multi-modal inputs. Episodes constitute the primary unit for learning, evaluation, and replay within the RobotMotion dataset.

III.B.3 Metadata and Indexing

To ensure transparency, traceability, and ease of reuse, RobotMotion associates explicit metadata with both skills and episodes. Skill-level metadata captures task descriptions, identifiers, and aggregate statistics, while episode-level metadata specifies modality presence, recording parameters, synchronization information, and file paths to associated data.

All metadata is stored in human-readable formats that allow direct inspection without reliance on proprietary tools. Explicit indexing of skills and episodes enables consistent referencing across data sources and supports flexible dataset traversal. This design facilitates selective data access, filtering by modality or task, and integration with open-source robot learning pipelines.

III.C. Modalities and Synchronization

RobotMotion supports a flexible set of sensing modalities to accommodate heterogeneous data sources and evolving robot platforms. Rather than enforcing a fixed sensor configuration, the dataset explicitly records which modalities are present for each episode, together with their corresponding parameters and synchronization information. This design allows RobotMotion to support a wide range of learning methods while remaining agnostic to specific hardware setups.

III.C.1 Supported Modalities

Across different data sources, RobotMotion may include visual observations (RGB or RGBD), proprioceptive signals (e.g., joint states, IMU, odometry), audio recordings, and optional tactile or force-related measurements when available. Simulation data may additionally provide access to precise state information and automatically generated annotations such as segmentation masks or object poses. Not all modalities are present in every episode; instead, modality availability is explicitly encoded as part of the episode metadata.

This selective modality design reflects practical data collection constraints and enables researchers to choose appropriate subsets of data depending on the target task and learning paradigm. For example, visiononly learning can be conducted using egocentric RGB data, while embodiment-aware policy learning may leverage proprioceptive and action signals from teleoperated robot episodes.

III.C.2 Temporal Synchronization

All modalities within an episode are temporally aligned using either frame indices or timestamps, depending on the data source. Synchronization metadata specifies the temporal reference used, ensuring consistent correspondence across visual, proprioceptive, and action streams. This alignment is preserved throughout preprocessing and storage, allowing downstream methods to reliably access synchronized multi-modal observations.

For simulation data, synchronization is typically defined with respect to the simulator's control timestep, while real-world data sources rely on recorded timestamps or camera frame indices. By making synchronization explicit rather than implicit, RobotMotion enables reproducible data access and reduces ambiguity in multi-modal learning setups.

III.C.3 Design Considerations

The modality and synchronization design of RobotMotion prioritizes transparency and extensibility over strict uniformity. By avoiding hard constraints on sensing configurations, the dataset remains compatible with diverse robot embodiments and collection setups. At the same time, explicit metadata and synchronization guarantees provide sufficient structure for systematic analysis and learning.

This approach allows RobotMotion to serve as a common substrate for research spanning imitation learning, multi-modal representation learning, and human-to-robot transfer, while remaining adaptable to future sensing modalities and data sources.

IV. Scale and Statistics

RobotMotion is designed as a growing dataset that integrates multiple data sources under a unified representation. At the time of writing, data collection is ongoing, and the dataset scale is expected to expand across tasks, environments, and modalities.

IV.A Dataset Scale

The dataset consists of skill-organized episodes collected from egocentric human demonstrations, simulation environments, and teleoperated humanoid robot executions. Each episode corresponds to a single execution of a defined skill and may vary in duration and modality availability depending on the collection source. Aggregate statistics such as total number of episodes, total recording duration, and per-skill episode counts are maintained as part of the dataset metadata and will be released alongside the dataset.

IV.B Task and Environment Coverage

RobotMotion covers a diverse set of everyday manipulation and interaction tasks, including object picking, placing, pouring, and carrying. Data is collected across multiple environment configurations and object instances to capture variation in task context and execution strategy. Skills are defined at a semantic level to allow consistent grouping of episodes across different environments and data sources.

IV.C Modality Availability

Modality availability varies across episodes depending on the data source. Egocentric human demonstrations primarily provide visual and audio observations, simulation data may include additional state and annotation information, and teleoperated robot episodes provide embodiment-aligned proprioceptive and action signals. Explicit modality metadata enables filtering and analysis based on available sensing modalities without assuming uniform sensor configurations.

V. Limitations

While RobotMotion is designed to integrate complementary data sources under a unified representation, several limitations remain. First, modality availability is not uniform across all episodes. Egocentric human demonstrations, simulation data, and teleoperated robot executions naturally differ in sensing configurations and annotation richness, which may limit direct comparability across subsets without appropriate filtering.

Second, although simulation data enables scalable collection with perfect annotations, transferring policies learned in simulation to real-world robotic systems remains challenging due to domain discrepancies in dynamics, sensing noise, and contact interactions. RobotMotion does not eliminate this gap, but rather provides structured data to facilitate hybrid learning approaches that combine simulation and real-world observations.

Third, teleoperated humanoid robot demonstrations offer embodiment-aligned data but are more costly and time-consuming to collect than simulated trajectories, resulting in smaller-scale coverage compared to simulation-derived data. Finally, the dataset is under active development, and the current release represents an initial snapshot rather than a finalized corpus. Future expansions will further increase task diversity, environment coverage, and modality availability.

VI. Conclusion

We presented RobotMotion, a multi-source humanoid behavior dataset designed to support learning and evaluation of physical AI systems. By integrating egocentric human demonstrations, large-scale simulation data, and teleoperated humanoid robot executions under a unified skill–episode abstraction, RobotMotion enables consistent organization, synchronization, and analysis of heterogeneous behavioral data. The dataset emphasizes transparent data formats, flexible modality support, and compatibility with existing robot learning pipelines. We believe RobotMotion provides a practical foundation for research in imitation learning, reinforcement learning, and human-to-robot transfer, and hope it will support future progress toward generalizable humanoid robot behavior.