Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ30306

Limitations in Contemporary Robot Learning Datasets and A Unified Framework for Enhanced Generalization

Yukun Wang

Brunel London School, North China University of Technology, Beijing, China
13611162505@163.com

Abstract. Developing robust and generalizable embodied intelligence relies heavily on large-scale, multi-modal datasets. Despite the impressive scale of resources like Open X-Embodiment and AgiBot World, this paper identifies three critical limitations hindering further progress. Firstly, there is a predominant reliance on visual data while neglecting other sensory modalities such as force-torque and tactile feedback, which are crucial for contact-rich manipulation. Secondly, data bias and low fidelity often lead to shortcut learning, where models exploit spurious correlations rather than learning underlying physical principles. Lastly, the lack of standardization in data formats and evaluation protocols creates significant barriers to reproducibility and fair benchmarking. To solve these issues, this study proposes a single dataset that covers all of the data. The framework promotes a multi-modal approach that involves diverse sensory components, structured task hierarchies that promote compositional thinking, and compatibility across platforms through standardized data specifications. The findings demonstrate that prioritizing data quality and thoughtful structure over mere scale is essential for learning robust representations. This work concludes that such a systematic approach establishes a more reproducible foundation for the field, accelerates progress by enabling more meaningful comparisons and data reuse, and ultimately supports the development of reliable robotic systems for complex, real-world environments. Future research should focus on physically plausible data augmentation, improved failure annotation, and adaptive data collection protocols.

Keywords: Robot Learning Datasets, Embodied AI, Multimodal Alignment, Data Standardization, Shortcut Learning

1. Introduction

Advancements in embodied intelligence are increasingly driven by large-scale, multi-modal datasets such as Open X-Embodiment [1] and AgiBot World [2, 3], which facilitate the training of generalist policies across diverse tasks and platforms. These datasets contain a large amount of robotic path data from various sources, which is of great value for the data-driven learning. For instance, Open X-Embodiment [1] incorporates over 1 million real robot trajectories spanning 527 skills and 160,000 tasks across 22 distinct robot types. At the same time, AgiBot World [2, 3] boasts an even larger scale, with 10 times more long-horizon data and 100 times greater scene coverage than its predecessor.

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ30306

Nevertheless, despite these impressive achievements in scale, critical limitations persist in current dataset design approaches that hinder progress toward robust and generalizable embodied intelligence. This paper identifies three fundamental shortcomings in contemporary robot learning datasets: (1) an overreliance on visual modalities at the expense of other sensory streams; (2) persistent issues of data bias and low fidelity leading to shortcut learning [4]; and (3) a lack of standardization in data formats and evaluation protocols. To address these challenges, we propose a unified dataset framework that emphasizes multi-modal alignment, structured task hierarchies, and cross-platform compatibility. This approach aims to establish a more systematic and reproducible foundation for data-driven robot learning, promoting more efficient progress toward general-purpose robotic systems operating in complex, unstructured environments.

2. Limitations in contemporary robot learning datasets

2.1. Scarcity of multimodal sensory data

Contemporary robot learning datasets predominantly prioritize visual data while largely neglecting other critical sensory modalities. Despite its unprecedented scale, Open X-Embodiment [1] primarily contains visual information with limited integration of force-torque feedback, tactile sensing, or audio cues. The observed visual bias substantially limits the capacity of trained models to excel in contact-rich manipulation tasks necessitating an understanding of physical interaction, exemplified by intricate assembly procedures or the manipulation of pliable objects. The omission of haptic feedback is particularly problematic for fine manipulation capabilities. While humans rely heavily on tactile feedback when performing precision tasks, most current robotic datasets provide only visual information, creating a fundamental gap in what learning models can perceive about physical interactions. This limitation becomes evident in tasks requiring careful force modulation, such as inserting components without damage or handling fragile objects. Some recent initiatives, like the HEART project, have begun integrating force feedback technologies such as SenseGlove haptic gloves [5] with robotic systems to capture richer interaction data.

Table 1. Comparison of sensory modalities across major robotic datasets

Dataset	Visual	Tactile	Force-Torque	Audio	Proprioceptive
Open X-Embodiment	✓	✗	Limited	✗	✓
AgiBot World	✓	Partial	Partial	✗	✓
ARIO	✓	✗	✗	✗	✓
RoboMIND	✓	✗	Limited	✗	✓
LIBERO	✓	✗	✗	✗	✓

The consequences of this sensory imbalance are particularly evident in real-world manipulation scenarios. For example, tasks like knot-tying, material stiffness assessment, or liquid containment all require rich tactile and force feedback that current datasets largely lack. Without these critical sensory streams, learning models must rely exclusively on visual approximations of physical interactions, resulting in policies that struggle with fine manipulation and contact-rich tasks.

DOI: 10.54254/2755-2721/2026.TJ30306

2.2. Data bias and low-fidelity issues

Current robot learning datasets suffer from significant data bias and fidelity issues that lead to "shortcut learning" - where models exploit superficial correlations in the data rather than learning robust underlying concepts [4]. The technology research team discovered that robots trained on large datasets like Open X-Embodiment [1] often develop unexpected failure modes when encountering novel environments or tasks. For example, when instructed to "place the spoon on the towel," a model might persistently attempt to grasp a soda can simply because canned drinks frequently co-occurred with towel surfaces in the training data.

Two fundamental issues contribute to this problem: insufficient diversity within individual sub-datasets and excessive fragmentation between data sources. While aggregate datasets may appear diverse, individual sub-datasets (often collected by specific labs) typically contain limited environmental conditions, object appearances, or task configuration variations. Furthermore, different data sources frequently employ dramatically different collection protocols, creating substantial integration challenges and reinforcing spurious correlations.

The sim-to-real gap presents another significant source of data bias. Many datasets incorporate simulated data to augment real-world recordings, but physical inaccuracies in simulation engines often lead to policies that fail when deployed on physical systems. A general lack of annotated failure cases in most datasets compounds this problem. Learning models cannot develop robust error-handling capabilities essential for real-world deployment without examples of failures and recovery strategies.

Recent research has revealed that some of the most performant robotic systems, including $\pi 0.5$ and Gemini Robotics, intentionally avoid using heterogeneous datasets like Open X-Embodiment [1] in favor of carefully controlled, highly consistent data collections [4]. This approach suggests that current methods for aggregating disparate data sources may introduce more problems than they solve, highlighting the need for more sophisticated dataset design principles.

2.3. Lack of data standardization

The lack of standardized data formats, action representations, and evaluation metrics presents substantial obstacles to data integration, model reproducibility, and equitable benchmarking. Open X-Embodiment [1] attempted to address this by converting all source datasets into RLDS (Robot Learning Data Standard) format, but substantial heterogeneity persisted beneath this surface-level standardization. Different datasets employ incompatible action representations (joint positions vs. end-effector commands, absolute vs. relative values, etc.), varying camera configurations, and disparate observation spaces that resist straightforward integration.

This standardization problem manifests particularly acutely in action representations. The same fundamental operation (e.g., "grasp bottle") might be encoded dramatically differently across platforms, depending on the specific robot's kinematics, control interface, and data collection protocol. Without standardized action representations, models trained on aggregated data must learn to translate between these different encodings, adding substantial complexity to the learning problem.

Evaluation protocols similarly lack standardization across the field. Different datasets employ success metrics, task variants, and testing procedures, making comparing approaches challenging. For example, some datasets measure success based on the outcome, while others incorporate intermediate milestones or quality measures. This heterogeneity impedes progress by making it

DOI: 10.54254/2755-2721/2026.TJ30306

difficult to identify which techniques advance the state of the art versus those that benefit from favorable evaluation protocols.

The lack of standardization also affects sensory data. Cameras with different specifications (resolution, field of view, color calibration) are used across datasets, and positional variations in camera mounting create further inconsistencies. While Open X-Embodiment [1] attempted to address this by resizing all images to 320×256 pixels and selecting a "canonical" viewpoint from multi-view setups, such approaches represent partial solutions that inevitably discard potentially valuable information.

3. A unified dataset framework

3.1. Multimodal alignment

To address the sensory limitations of current datasets, we propose a unified framework that emphasizes comprehensive multimodal alignment. This approach integrates visual, tactile, force-torque, and audio data into a synchronized representation that provides a complete picture of robotic interactions. Rather than treating non-visual modalities as auxiliary signals, we position them as first-class data streams essential for understanding physical interactions.

The framework incorporates hardware-agnostic data collection protocols that ensure consistent recording of all sensory streams regardless of the specific robot platform. This includes standardized calibration procedures for force-torque sensors, tactile arrays, and audio recording equipment to ensure comparable measurements across different collection setups. For tactile data, we propose a normalized representation that captures essential contact information (pressure distribution, vibration, temperature) independent of specific sensor hardware.

To address the inherent difficulties of having multiple modes of synchronization, our framework includes precise time alignment across all data streams and specialized techniques for augmenting data that maintain the physical consistency of the data across different modes. For example, when applying visual augmentations like color jitter or noise, corresponding adjustments are made to tactile and force expectations to maintain cross-modal consistency.

A key innovation in our framework is using cross-modal self-supervision techniques that leverage the natural redundancy between sensory streams. For instance, visual information can help interpret tactile signals, while force readings can help ground visual predictions. These self-supervision signals enable more data-efficient learning and encourage the development of representations that capture underlying physical principles rather than superficial correlations.

3.2. Structured task hierarchies

To mitigate data bias and shortcut learning [4], our framework employs a structured task hierarchy to guide data collection, explicitly encoding compositional reasoning challenges. Rather than treating tasks as flat collections of demonstrations, we decompose complex skills into hierarchical structures that highlight shared subtasks and variations across applications.

The hierarchy begins with foundational motor primitives (grasping, pushing, rotating, etc.) that form the building blocks of more complex skills. These primitives are combined into compound skills (pour liquid, fasten screw, fold cloth) which in turn are organized into complete task solutions (prepare meal, assemble furniture). This hierarchical organization enables more efficient learning by identifying transferable components across different tasks and environments.

DOI: 10.54254/2755-2721/2026.TJ30306

Our framework incorporates explicit variations at each level of the hierarchy to ensure diversity and discourage shortcut learning. For object variations, we include systematic changes to appearance, size, material properties, and weight. For environmental variations, we modify lighting conditions, camera angles, background scenes, and workspace configurations. Crucially, we ensure that these variations are applied orthogonally to task factors, preventing models from relying on spurious correlations between task requirements and environmental contexts.

To address the lack of failure annotations in existing datasets, our framework includes deliberate collection of error demonstrations and recovery strategies. For each task variant, we record not only successful executions but also common failure modes and appropriate correction procedures. This approach provides learning models with crucial error detection and recovery information for robust real-world performance.

The structured task hierarchy also informs our evaluation protocol, which assesses performance at multiple levels of abstraction. Models are evaluated not only on overall task success but also on the quality of individual components and their ability to adapt to novel combinations of familiar skills. This multi-level evaluation provides clearer insights into where and why methods succeed or fail.

3.3. Cross-platform compatibility

This framework addresses the standardization challenges through a hardware-agnostic data specification that supports cross-platform compatibility while preserving the unique capabilities of different robot systems. The specification includes standardized representations for actions, observations, and tasks that can be mapped to various robot platforms through platform-specific adapters. For action representation, it propose a universal coordinate system that combines the task-level intention with the specific platform. actions are represented as commands that end up with the desired results in a common space, these can then be converted to joint commands or torques via the platform's specific inverse kinematics. This approach maintains the high-level semantic meaning of actions while accommodating mechanical differences between robots.

The observation schema similarly separates fundamental perceptual information from sensor-specific characteristics. Visual observations are accompanied by detailed camera calibration parameters, enabling models to understand and compensate for perspective distortions. Proprioceptive data is normalized relative to each robot's physical limits, allowing direct comparison of manipulation strategies across different platforms.

To facilitate reproducibility and fair benchmarking, our framework includes a standardized evaluation protocol with clearly defined success metrics for various task categories. The protocol specifies testing procedures, baseline implementations, and reporting formats to ensure consistent evaluation across different research groups. Additionally, it provide implementations of several cutting-edge methods [6,7] that are standardizable, this decreases the difficulty of entry for new scientists. The framework also supports efficient sim-to-real transfer by maintaining consistent data representations across simulated and physical environments. Simulation environments configured to match our data specification can generate synthetic training data that aligns closely with real-world recordings, reducing the domain gap that typically plagues transfer learning approaches.

4. Conclusion

This study has identified three major limitations in current robot learning datasets: insufficient sensory information, data bias that promotes shortcut learning [4], and a lack of standardization.

DOI: 10.54254/2755-2721/2026.TJ30306

Based on an analysis of existing datasets [1-3,8] and recent research [4], we have illustrated how these issues hinder the development of robust and generalizable embodied intelligence.

To tackle these challenges, we propose a unified framework centered on multimodal alignment, structured task hierarchies, and cross-platform compatibility. By prioritizing high-quality data and well-designed structures over sheer volume, this approach facilitates more efficient learning of robust representations grounded in physical principles rather than superficial patterns.

This work's primary contribution offers a more systematic and reproducible foundation for data-driven robot learning. Standardized data specifications [9, 10] and evaluation protocols can accelerate research by enabling more precise comparisons between methods and more effective data reuse across groups. Moreover, by intentionally mitigating shortcut learning through improved dataset design, our framework promotes the creation of robotic systems capable of reliable operation in complex, real-world settings.

For future work, we suggest several promising directions: developing more advanced data augmentation techniques that preserve physical consistency across modalities, creating better methods for injecting and annotating failures to improve error recovery learning, and designing adaptive data collection protocols that automatically identify and address gaps in existing datasets. Data curation will be critical for the development of general-purpose robotic systems as robots move from controlled environments to real-world applications.

References

[1] Chebotar, Y., et al. (Open X-Embodiment Collaboration). (2023). Open X-Embodiment: Robotic Learning Datasets and RT-X Models. arXiv.
[2] Zhiyuan Robot. (2024). AgiBot World: Million-robot dataset opens new frontiers for embodied AI. NetEase News.
[3] AgiBot World: Million-Robot Dataset Opens New Frontiers for Embodied AI. (2024).
[4] Xing, Y., et al. (2025). Does big data actually make robot learning worse? Revealing the dark side of datasets. Proceedings of the Conference on Robot Learning (CoRL).
[5] SenseGlove. (2025). SenseGlove tactile feedback gloves: Breakthroughs in automation and training. Company White Paper.
[6] Brohan, A., et al. (2022). RT-1: Robotics transformer for real-world control at scale. arXiv: 2212.06817.
[7] Brohan, A., et al. (2023). RT-2: Vision-language-action models transfer web knowledge to robot control. arXiv: 2307.15818.
[8] Ye, H., et al. (2024). ARIO: A new standard and unified dataset for versatile, general-purpose embodied agents. arXiv: 2408.10899.
[9] OASIS. (2025). Akoma Ntoso: Powering legal AI with open data standard. OASIS Standard.
[10] Author. (2025). Data middle platform construction methodology: OneData system detailed explanation. CSDN Blog.