TeachAnything: A Multimodal Crowdsourcing Platform for Training Embodied AI Agents in Symmetrical Reality

Category: Research

ABSTRACT

Symmetrical Reality (SR) is emerging as future trend for human–agent coexistence, placing higher demands on agents to acquire human-like intelligence. It calls for richer and more diverse human guidance. We introduce a three-stage demonstration paradigm integrating multimodal demonstration signals. Building on this paradigm, we developed **TeachAnything**, a cloud-based, crowdsourcing-oriented demonstration platform with physics simulation capable of collecting diverse demonstration data across varied scenes, tasks, and embodiments. By unifying virtual and physical interactions through both methodological design and physics simulation, the system serves as a practical foundation for developing embodied agents aligned with Symmetrical Reality.

Index Terms: Crowdsourcing platform, Human–Robot Interaction, Agent Training, Symmetric Reality.

1 INTRODUCTION

Symmetrical Reality (SR) is increasingly regarded as an inevitable developmental trend in embodied AI, envisioning a future where physical and virtual worlds integrate seamlessly and intelligent agents interact coherently across both domains. [4] Achieving such embodied intelligence requires agents to develop unified perception and action capabilities across realities, which fundamentally depends on large-scale, diverse, and semantically aligned demonstrations. [3] Moreover, modern human in the loop learning frameworks, including VLA models and policy learning systems, also depend on large scale multimodal data to bridge high level intent, perceptual grounding, and low level control. [5] However, existing demonstration pipelines are misaligned with the needs of this SR-driven future. They are often restricted to fixed scenarios, predefined tasks, single embodiments, or single-modality inputs, limiting the richness and variability of supervision they can provide. [1] Complex real-world tasks commonly require multimodal teaching signals, span a wide variety of environments and goals, and often involve the generation of complex demonstration data, such as manipulation tasks that produce continuous action trajectories, which poses higher technical challenges for existing demonstration methods. As a result, there comes a substantial gap between the data required to train SR-capable agents and what existing demonstration-collection systems can supply.

To address these limitations, we introduce a three-stage demonstration paradigm that explicitly targets the key challenges of SR-oriented data collection. By decomposing human teaching into semantic, perceptual, and embodied channels, the paradigm enables multimodal supervision for complex tasks, supports open-ended demonstrations across diverse scenes and tasks, and provides dedicated mechanisms for generating fine-grained continuous action data. Specifically, language demonstrations capture high-level intent and task structure beyond fixed templates, video demonstrations ground task execution by providing rich perceptual evidence from diverse scenes and embodiments, and teleoperation-based demonstrations produce precise continuous control trajectories for manipulation-intensive skills. These stages form a coherent and scalable framework that bridges the gap between the rich demonstration data required by SR-capable agents and the limitations of existing demonstration pipelines.

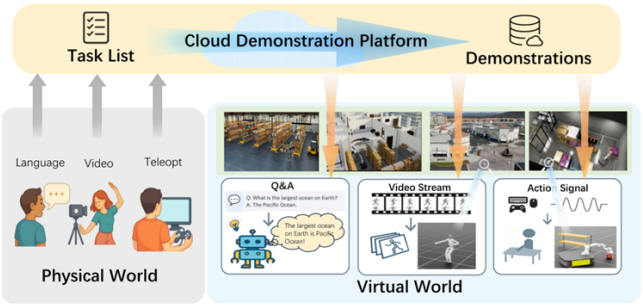

Figure 1: **Overview of TeachAnything**, a cloud-based crowdsourcing demonstration platform that enables users to teach anytime and anywhere through multimodal demonstrations. The platform supports both predefined and user-defined tasks within rich virtual scenes, and converts heterogeneous inputs into structured data for training embodied agents.

Building on this paradigm, we develop **TeachAnything**, a cloud-based demonstration platform that unifies all three stages within a single environment. The system supports configurable scenes, multiple interaction channels, and diverse robot embodiments, with a Gradio interface providing real-time visualization, teleoperation panels, and integrated recording. An Isaac Sim backend powered by PhysX provides high-fidelity physical interactions for embodiments such as the Franka arm and Unitree G1, while WebSocket streaming synchronizes scenes and commands, and Flask microservices enable camera input and HaMeR-based [2] gesture control. All language, video, and teleoperation data are consolidated into a unified structured format for embodied-agent training and virtual–physical integration in Symmetrical Reality.

2 METHODS

Our methodology centers on a three-stage paradigm that organizes human guidance by modality and interaction channel. Each stage serves as an information-carrying pathway for human demonstration, independent of specific sensors, embodiments, or task settings. By decoupling the demonstration interface from the underlying agent or environment, this structure unifies heterogeneous demonstration sources and enables consistent interpretation across tasks, robot embodiments, and virtual–physical domains.

- **Language-based demonstration**: Text or speech inputs that describe task execution, articulate high-level goals, or provide semantic annotations of demonstration content. This modality not only conveys procedural intent but also supplies contextual cues—such as object relations, constraints, or rationale for certain actions—that may not be visually observable.

- **Video-based demonstration**: Uploaded or recorded videos illustrating complete task executions or annotated visual content from any embodiment, including human demonstrations, robot executions, and simulation renderings. Video demonstrations capture temporal dynamics, motion patterns, object interactions, and environmental context with high perceptual

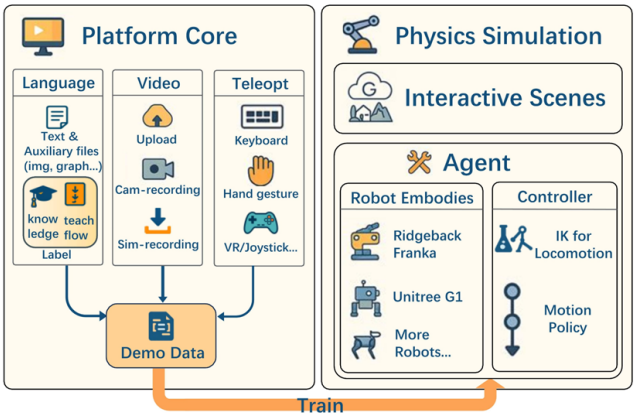

Figure 2: **Overview of the TeachAnything**, which integrates language, video, and teleoperation channels with real-time services and physics simulation to generate unified training data for SR-aligned embodied agents.

fidelity. These visual sequences provide dense behavioral supervision that complements linguistic abstractions and allow agents to learn spatial reasoning, motion understanding, and visual affordances directly from observation.

- **Teleoperation-based demonstration**: Real-time human control of an embodied agent in simulation, generating continuous action trajectories through all kinds of interfaces such as keyboard–mouse, VR controllers, or gesture-recognition systems. This modality delivers fine-grained motor-level supervision, enabling the platform to collect detailed manipulation strategies, corrective refinements, and personalized action preferences. Because teleoperation is embodied and temporally aligned with the robot's control loop, it maintains strong physical consistency and is well suited for training low-level policies that transfer effectively from simulation to real-world execution.

3 System Demostration

To realize the proposed paradigm, we implement a cloud-based crowdsourcing platform that enables users to initiate demonstrations anytime and anywhere for both predefined and user-defined tasks shown in Fig 2. The platform provides unified interfaces for collecting language, video, and teleoperation demonstrations within a single environment.

The platform core coordinates multimodal teaching from distributed users. Language demonstrations allow users to freely describe tasks, procedures, or declarative knowledge without relying on fixed templates, and are automatically categorized and enriched with semantic metadata.

The video demonstration pipeline supports uploads, local camera recording, and simulation-based capture. Users can reset, re-record, and download videos, enabling iterative refinement. All videos are collected with structured metadata such as timestamps, sources, and scene context, facilitating alignment with other modalities.

For embodied supervision, the platform supports teleoperation through keyboard–mouse input and gesture-based control. Continuous action commands and synchronized simulation states are transmitted via a WebSocket layer to ensure low-latency interaction. Gesture-based teleoperation is implemented through a Flask service integrating hand pose recognition as illustrated in Fig 3.

All demonstrations are executed within a physics-based simulation backend supporting diverse robot embodiments and interactive scenes. Robots such as Ridgeback–Franka and Unitree G1 operate under dedicated control stacks combining inverse kinematics and

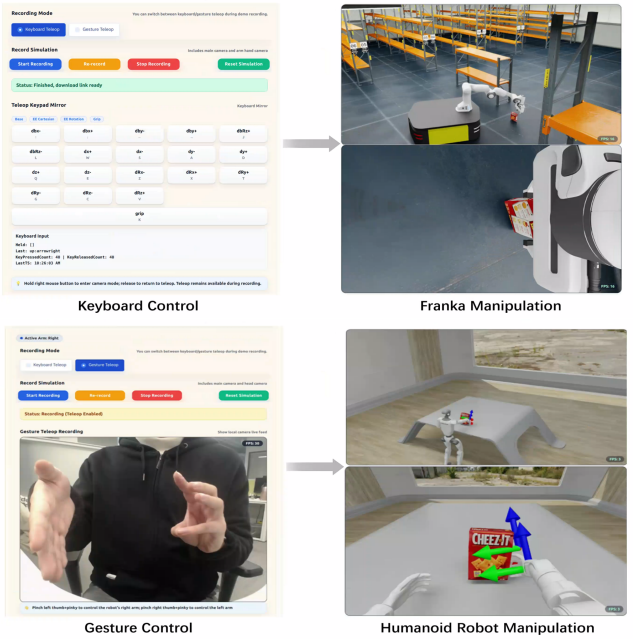

Figure 3: **Demonstrations on platform**, showing keyboard-based teleoperation of a Franka arm and hand-gesture-based teleoperation of a Unitree G1 robot, with videos and action trajectories recorded as demonstration data.

optional learned motion policies, enabling users to generate physically consistent demonstrations across varied tasks and embodiments.

4 Conclusion And Future Work

We present a three-stage demonstration paradigm and a cloud-based crowdsourcing platform that support scalable, multimodal supervision for training embodied agents in Symmetrical Reality. By unifying language, video, and teleoperation demonstrations within a physics-based simulation environment, the platform enables flexible and open-ended teaching across tasks, scenes, and embodiments. Although the platform is still evolving, additional scenes, robot embodiments, interaction methods, and a full data-to-training pipeline as future works will further strengthen its role as a foundation for scalable supervision and embodied learning in Symmetrical Reality.

References

[1] A. Mandlekar, Y. Zhu, A. Garg, J. Booher, M. Spero, A. Tung, J. Gao, J. Emmons, A. Gupta, E. Orbay, et al. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In *Conference on Robot Learning*, pp. 879–893. PMLR, 2018. 1

[2] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9826–9836, 2024. 1

[3] C. Yifan, M. Wei, X. Wang, Y. Liu, J. Wang, H. Song, L. Ma, D. Di, C. Sun, K. Liu, et al. Embodied ai: A survey on the evolution from perceptive to behavioral intelligence. *SmartBot*, 1(3):e70003, 2025. 1

[4] Z. Zhang, Z. Zhang, Z. Jiao, Y. Su, H. Liu, W. Wang, and S.-C. Zhu. On the emergence of symmetrical reality. In *Proceedings of the IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 639–649. IEEE, 2024. 1

[5] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023. 1