Data Infrastructure for Embodied AI

Multimodal physical perception, world modeling and industrial reuse

Linchao Zhu
Research Professor, Zhejiang University & Jinglianwen

Figure and OpenAI are investing in robotics data systems

Figure

Index: human behavior data for Helix

16M+ uploaded videos reported at launch

A dedicated pipeline for filtering, deduplication and annotation

OpenAI

Robotics data infrastructure roles

Public roles cover collection operations and distributed training-data systems.

Selection, automated labels and loaders

Industry signal: data production is becoming a core robotics capability.

Figure: Index, 25 Aug 2026. OpenAI: public robotics role descriptions, accessed 15 Oct 2026.

Figure reports a large gain from human-data pretraining

9%

Without Index pretraining

56%

With Index pretraining

Whole-task success across 30 unseen homes

Figure reports matched architecture, task data, training and evaluation.
Tasks: room tidying, towel folding and bed making.

Company-reported result, Helix 2.5, 17 Sep 2026. This is not an industry-wide benchmark.

Deployment requires more than recognizing the scene

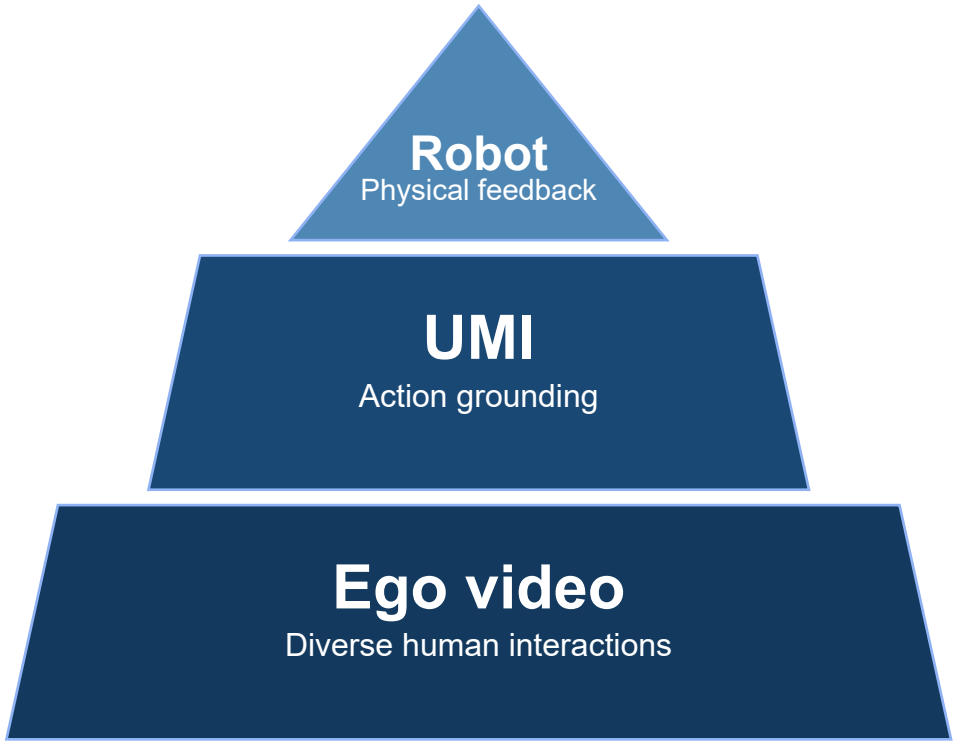
The same task must work across changing physical conditions.

Deployment challenge	Data the robot needs
New objects and settings	Diverse interactions and transferable action examples
Uncertain physical state	Geometry, motion, contact signals and robot state aligned with actions
Disturbances and failures	**Timely feedback, recovery attempts and task outcomes**

The data system must connect what is seen, what is done and what happens next.

Four data sources address different deployment gaps

Ego, UMI and robot data provide anchors for scalable simulation.



Real-robot interaction: physical grounding

Align vision / depth, robot state and actions with force, touch and other available sensors.

UMI demonstrations: action grounding

Connect visual observations with manipulation trajectories collected through handheld interfaces.

Ego video: interaction coverage

Learn diverse object interactions and visual dynamics from human first-person observations.

Simulation expands variation across the pyramid; real data calibrates physical fidelity.

Conceptual hierarchy, not dataset volumes. Sensor labels depend on the collection setup.

Ego video broadens coverage of human interactions



What it contributes

Objects, scenes and ways of handling them
Visual cues around contact and motion

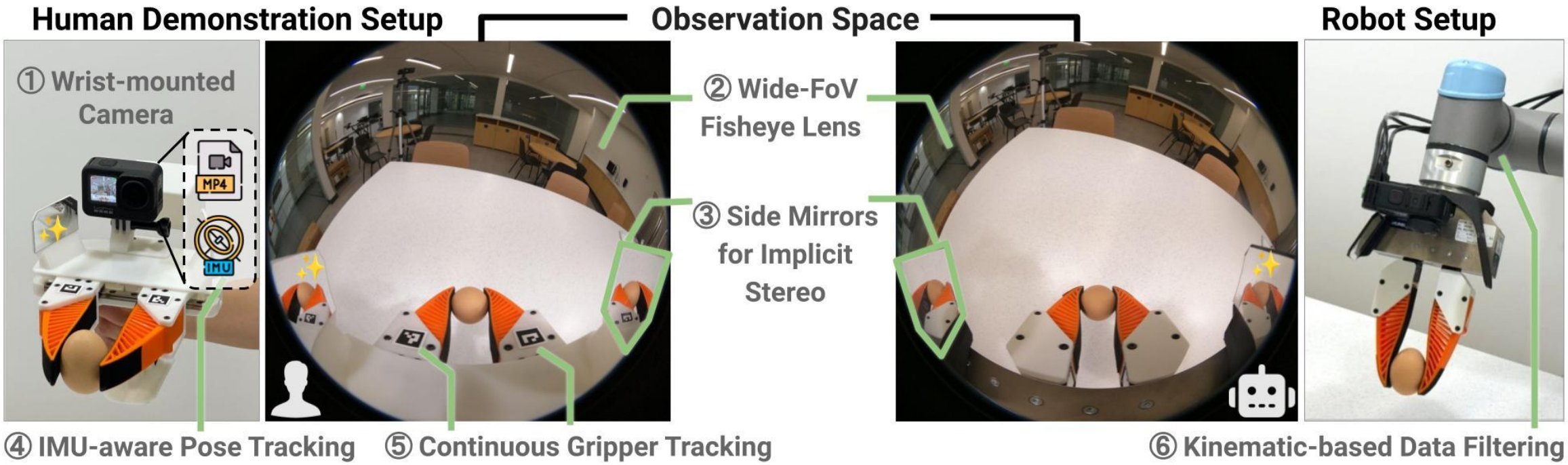
What needs additional supervision

Robot actions, 3D geometry and measured physical responses

Its role: broad interaction coverage before task-specific calibration.

Ego = first-person video. Illustration adapted from the supplied presentation.

UMI adds action supervision to visual experience



Handheld collection

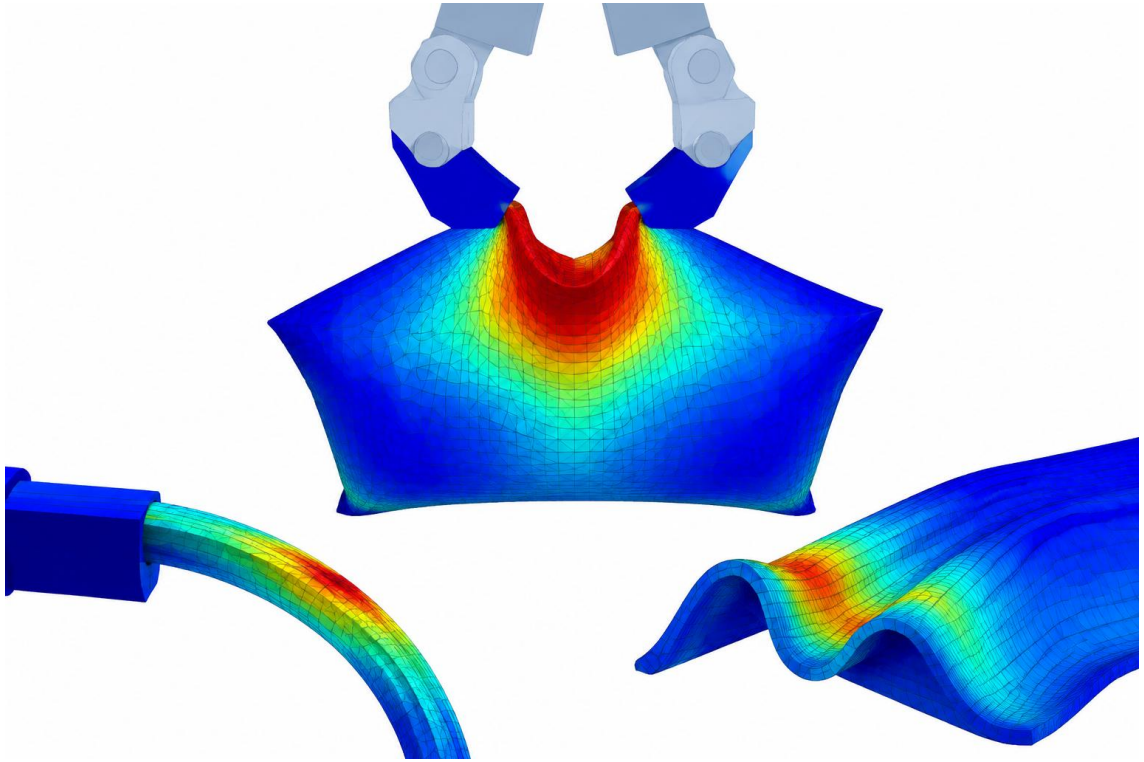
Record manipulation motion with portable interfaces.

Action supervision

Align observations with trajectories and gripper behavior.

UMI: Chi et al., RSS 2024. Hardware illustration: umi-gripper.github.io

Real robots supply measured physical feedback



Observe the full physical interaction

Align vision / depth, joint state, motion, force and touch.

Include failures and recoveries, not only successful trials.

Collect synchronized signals where physical uncertainty limits the task.

Illustrative physical response. Sensor channels depend on the collection system.

Simulation expands scenarios around real experience

Real examples → Simulated variations → Real-world validation

Industrial value	What simulation contributes
Expand coverage	Vary objects, layouts, lighting and physical parameters
Explore failures	Generate rare conditions, perturbations and recovery trials
Repeat at scale	**Produce action trajectories and repeatable test scenarios**

Industry example: NVIDIA GR00T-Mimic expands demonstrations into synthetic trajectories.

Simulate broadly; calibrate dynamics and sensor responses on real robots.

Sources: NVIDIA GR00T-Mimic blueprint; Tobin et al., Domain Randomization (2017).

Align the sources into usable interaction records

Different sources describe the same interaction in different ways.

Mismatch	What must align
View and embodiment	Human hands, handheld tools and different robots
Time and geometry	Time, coordinates, units, sensor calibration and physics parameters
Supervision	Measured vs. simulated signals, inferred labels and missing channels
Task and outcome	**What counts as completion, failure and recovery**

Reusable data depends on consistent meaning across sources.

Physical perception combines complementary sensor signals

No single modality captures the full state of an interaction.

Signal	What it reveals
Vision and depth	Objects, spatial geometry, motion and visible deformation
Force / torque and touch	Loads, contact, pressure distribution and slip cues
Proprioception	Joint positions, velocities, effort and end-effector state
Audio, where relevant	**Impact, vibration and changes in machine operation**

Align these signals with actions and outcomes to learn physical dynamics.

Conceptual sensor coverage. Available channels and reliability depend on the hardware and task.

World models predict how actions change the physical state

“If the robot acts, how will the scene and interaction change?”

Observe

Combine scene geometry, motion, robot state and contact signals.

Predict

Given an action, estimate future states, physical responses and uncertainty.

Act and update

Use predictions to choose actions; compare observed outcomes and correct.

Research direction: multimodal physical world modeling, grounded in measured interactions.

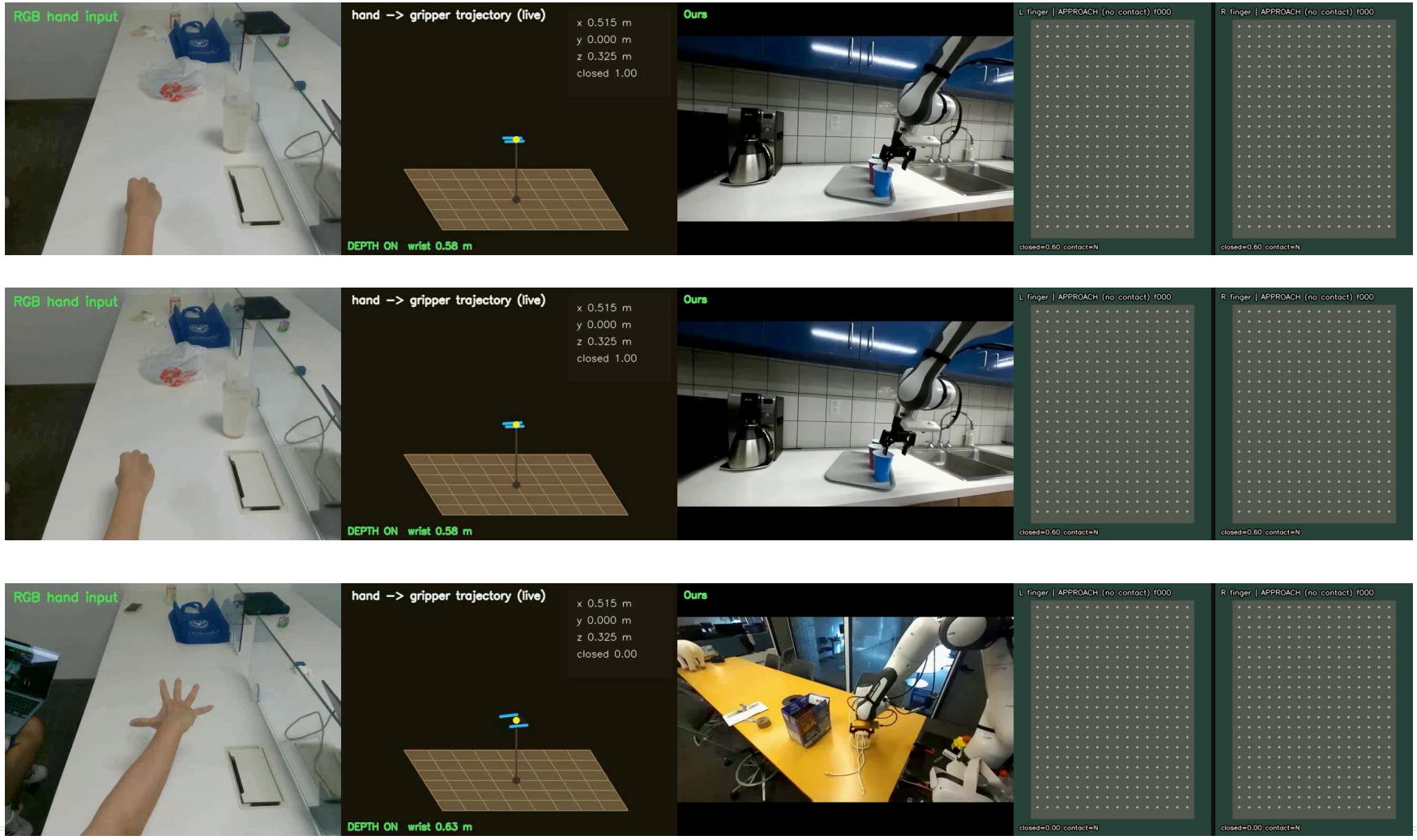
A multimodal prediction example: motion, vision and touch

One model predicts hand motion, future video and tactile signals from a visual input.

Input video | **Hand motion** | **Predicted video** | **Predicted tactile signals**

Visual input

Predicted interaction feedback



Deployment closes the loop by revealing missing experience

01

Observe failures

Capture the conditions that existing data missed.

02

Diagnose

Separate coverage gaps from hardware and control faults.

03

Improve data

Recreate gaps in simulation; collect targeted real feedback.

04

Verify transfer

Test new settings and protect existing capabilities.

Deployment access matters when it produces reusable learning.

Data reuse should lower the cost of the next deployment

Useful pretraining should reduce the need to recollect every task in every setting.

Upfront investment

Collection, simulation assets, training and physical validation

Value of reuse

Reusable scenarios and less task-specific data
for each additional deployment

Economic hypothesis to test across tasks and embodiments, not a quantified cost forecast.

Industry value spans the entire data-to-deployment chain

Different businesses can build complementary capabilities.

Capability	Where it creates value
Collection and simulation	Capture difficult interactions and expand relevant variation
Data infrastructure	Align, curate and trace observations, actions and outcomes
Models and robot systems	Turn experience into transferable, responsive behavior
Deployment and evaluation	**Expose new gaps and verify gains in operating conditions**

Defensibility comes from making accumulated experience useful in new deployments.

Measure data value through transfer and operating outcomes

A larger dataset is useful only when the added experience changes behavior.

Question	Evidence to collect
Does it transfer?	Success on held-out objects, tasks and environments
Does feedback help?	Recovery rate, retries and contact-related failures
Does deployment get easier?	Task-specific collection effort and operator intervention
Is the gain attributable?	**Matched comparisons with and without the added data**

Proposed evaluation framework. No production improvement or cost reduction is asserted.

Common interfaces enable data reuse across the industry

Proposed priorities for the industry and ITU

Shared building block	Why it matters
Interaction records	Consistent modality, action and outcome definitions
Timing and calibration	Comparable physical signals and response latency
Provenance and permitted use	Traceable collection, transformations and usage rights
Evaluation protocols	**Comparable transfer, physical feedback and recovery tests**

Common interfaces can support reuse without requiring all data to be public.

Build reusable physical experience for scalable deployment

Combine complementary data

Ego for coverage; UMI for actions; robots for physical signals; simulation for variation.

Turn experience into feedback

Fuse physical signals, predict action consequences and correct robot behavior.

Make each deployment improve the next

Feed failures back into data and verify transfer, reliability and adaptation cost.

Research focus: multimodal physical perception and world modeling