SOCIAL DEMOS, EPISODE 9

AI has read everything on the internet. Now it needs to watch how humans live, work and move.

The enterprise guide to physical AI training data: egocentric video, sensor streams and annotation at scale

Physical AI is in production. Humanoids are on warehouse floors. Autonomous systems run factory lines. The binding constraint is not the robot. It is the training data.

$176.8B
enterprise AI deployment spending, 2026

5.4x
year-on-year growth in enterprise AI spend

150+
environments in Corpshore AI physical data footprint

35+
languages covered by Corpshore AI physical programs

EGOCENTRIC VIDEO SENSOR STREAMS ACTION ANNOTATION



PHYSICAL AI TRAINING DATA: COLLECTION, SENSOR FUSION AND ANNOTATION

Jwuma by Corpshore AI | jwuma.com | info@corpshore.ai | +1 866 531 5727 | Part of Corpshore Solutions Corporation, Toronto, Canada | showcase.corpshore.ai | deck.corpshore.ai

Every video frame, annotation output, programme specification, quality metric and coverage figure in this document is a synthetic simulation produced by Corpshore AI. Nothing here is client data or live sensor material. The simulations illustrate physical AI data programme operations as Corpshore AI designs and runs them.

SECTION 1

Why physical AI is a data problem

Large language models learned from text already on the internet. Physical AI models have no such corpus. Every hour of training data for a robot must be collected, often in the exact environment the robot will work in.

Text AI learned from the web. Image AI learned from public datasets. Physical AI has no equivalent corpus. The internet does not contain footage of a person picking an item from a warehouse shelf from the first-person perspective of a robot arm. That footage must be collected from scratch.

The gap between a robot that works in a lab and a robot that works in a warehouse is almost never the model architecture. It is the training data the model never saw. Distribution shift from the controlled training environment to the real deployment environment is the leading cause of commercial robotics programme delays.

Corpshore AI and Jwuma solve the data problem. We collect egocentric video, real-world action demonstrations and synchronised multimodal sensor data across 18+ countries and 150+ environment types. The collection infrastructure exists. The annotation capability exists. What is missing is the programme.

THE FIVE PHYSICAL AI DATA TYPES

Egocentric Video — 100,000+

The core format. First-person POV captures how the world looks to the agent that must navigate it.

Real-World Action Demonstrations — 240+

Humans demonstrate. Robots learn. The gap between simulation and reality closes one demonstration at a time.

Multimodal Sensor Streams — 6

Robots do not just see. They feel, hear and sense. Training data must reflect all of it.

Human-Robot Interaction Data — 35+

Robots that work alongside humans need to understand human intent, not just objects.

Scene and Environment Diversity — 150+

A robot trained in one warehouse fails in another. Scene diversity is not optional.



EGOCENTRIC VIDEO: THE FOUNDATIONAL FORMAT FOR PHYSICAL AI TRAINING

Every video frame, annotation output, programme specification, quality metric and coverage figure in this document is a synthetic simulation produced by Corpshore AI. Nothing here is client data or live sensor material. The simulations illustrate physical AI data programme operations as Corpshore AI designs and runs them.

SECTION 2

Egocentric video and real-world action demonstrations

Humanoids learn grasping, navigation and manipulation from first-person demonstrations. Policy models trained on third-person video fail to generalise to embodied agents.

EGOCENTRIC VIDEO

Video recorded from a head-mounted or wearable camera, capturing the participant's first-person point of view: exactly what a humanoid or mobile robot camera would see.

- Indoor environments: kitchen, bathroom, office, retail, medical
- Outdoor environments: street, agricultural, construction, logistics dock
- Industrial: warehouse aisles, assembly stations, sorting lines, clean rooms
- Multi-sensor sync: RGB, depth (LiDAR, structured light), IMU where requested

ACTION DEMONSTRATIONS

Scripted or semi-scripted human demonstrations of target tasks, captured in the environments where the robot will actually operate. Participants are briefed, perform naturally and the full action sequ

- Household tasks: cooking, cleaning, object organisation, door and drawer operation
- Warehouse tasks: pick-and-place, bin sorting, label scanning, pallet building
- Factory tasks: assembly, quality inspection, tool handoff, maintenance
- Healthcare-adjacent: instrument handling, supply retrieval (non-clinical)

ANNOTATION DELIVERABLES: EGOCENTRIC VIDEO

- Frame-level action labels (pick, place, grasp, navigate, inspect, assemble)
- Object bounding boxes and instance masks per frame
- Quality flags: motion blur, occlusion, lighting drop
- Temporal phase segmentation (approach, contact, manipulation, release)
- Spatial relationship tags (above, inside, adjacent, gripping)

THE ANNOTATION ARCHITECTURE

The annotation layer for physical AI data exists at two time scales. Frame-level annotation applies per-frame labels for dense classification tasks. Segment-level annotation defines temporal spans for task segmentation and skill learning. Corpshore AI delivers both layers simultaneously, with segmen

- Frame-level labels at configurable intervals (every frame or every Nth frame)
- Segment labels with start and end timestamps (ms precision)
- Temporal alignment between frame and segment layers
- Phase ontology: approach, manipulation, transport, placement, error, recovery



EGOCENTRIC COLLECTION ENVIRONMENT: WAREHOUSE PICK-AND-PLACE STATION

Every video frame, annotation output, programme specification, quality metric and coverage figure in this document is a synthetic simulation produced by Corpshore AI. Nothing here is client data or live sensor material. The simulations illustrate physical AI data programme operations as Corpshore AI designs and runs them.

SECTION 3

Multimodal sensor streams and human-robot interaction

Robots do not just see. They feel, hear and sense. Vision-only models fail in occluded, low-light or texture-uniform environments. Training data must reflect the full sensor stack.

MULTIMODAL SENSOR STREAMS

Synchronised multi-channel data streams that go beyond RGB video: depth, audio, tactile, IMU and proprioceptive signals captured simultaneously and time-stamped to microsecond precision.

+ RGB video: 1080p and 4K, 30-60fps primary visual signal
+ Depth (LiDAR): point cloud for 3D environment mapping
+ Depth (structured light): depth map for close-range manipulation
+ IMU: 6-DoF at 200Hz for motion and orientation

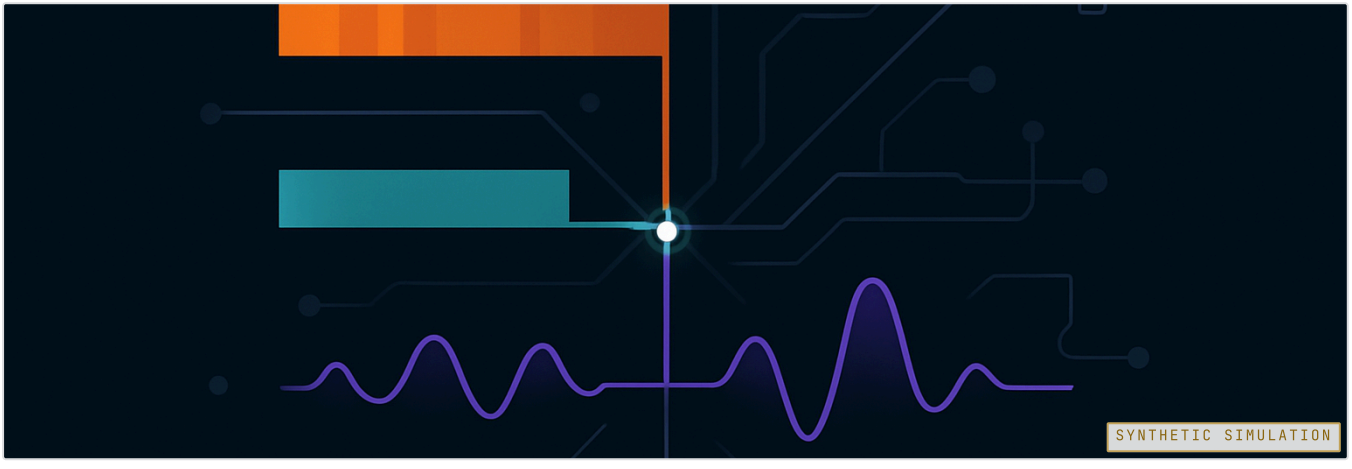
HUMAN-ROBOT INTERACTION

Recordings of natural human-robot interaction scenarios: humans giving verbal instructions, correcting robot actions, handing off objects and operating in shared workspaces.

+ Verbal instruction sequences in 35+ languages
+ Correction and re-instruction events
+ Object handoff sequences: human to robot and robot to human
+ Shared-workspace navigation with dynamic obstacles

SENSOR CONFIGURATION

SENSOR	SPEC	USE CASE
RGB Camera	1080p / 4K, 30-60fps	Primary visual stream
Depth (LiDAR)	Point cloud	3D scene mapping
Depth (structured)	Depth map	Close manipulation
IMU	6-DoF at 200Hz	Motion, orientation
Audio	48kHz stereo	Speech and env sounds
Tactile	Pressure grid	Contact and grip



MULTIMODAL SENSOR COLLECTION: SYNCHRONISED TO MICROSECOND PRECISION

Every video frame, annotation output, programme specification, quality metric and coverage figure in this document is a synthetic simulation produced by Corpshore AI. Nothing here is client data or live sensor material. The simulations illustrate physical AI data programme operations as Corpshore AI designs and runs them.

SECTION 4

Five dimensions of physical AI data quality

Enterprise buyers of physical AI training data are acquiring a liability as much as an asset. Consent failures, demographic gaps, sensor miscalibration and annotation inconsistency all trace back to the data vendor.

The question a physical AI buyer asks is not "can you collect this?" It is "can you prove to our data governance team that it was collected correctly, with proper consent, by calibrated sensors, with consistent annotation and in a format our pipeline will accept?" The answer requires systematic quality management across five layers.

01. Consent and Rights Infrastructure

Physical AI data collection involves recording real people in real environments. Rights and consent are non-negotiable for enterprise buyers.

+ Explicit informed consent per participant per session
+ Clear rights grant for AI training use
+ Participant withdrawal and data deletion process

02. Demographic and Environmental Diversity

A dataset's diversity is auditable. Corpshore AI delivers diversity metadata with every programme.

+ Participant demographic matrix: age, gender, body size distribution
+ Environment diversity log: lighting, floor surface, object density, cultural context
+ Geographic spread: collection locations per session logged

03. Annotation Accuracy and Inter-Annotator Agreement

Action labels are only as good as the annotators who apply them.

+ Domain-specialist annotators for each task category
+ Double-annotation with IAA calculation per batch
+ Target IAA >= 0.85 (Cohen's Kappa) for action labels

04. Sensor Calibration and Synchronisation

Multimodal datasets are only usable if sensor streams are calibrated and aligned.

+ Per-session calibration check for all sensors (RGB, depth, IMU)
+ Timestamp alignment validation: RGB-depth delta < 5ms
+ Calibration metadata delivered per session

05. Format and Integration Readiness

Data must arrive in formats that plug into the buyer's training pipeline on day one.

+ Output formats: HDF5, TFRecord, Parquet, COCO JSON, custom schema on request
+ Metadata standard: ROSBAG-compatible where requested
+ Delivery: secure cloud storage with API access or encrypted physical delivery

Every video frame, annotation output, programme specification, quality metric and coverage figure in this document is a synthetic simulation produced by Corpshore AI. Nothing here is client data or live sensor material. The simulations illustrate physical AI data programme operations as Corpshore AI designs and runs them.

SECTION 5

150+ environments. 18+ countries. Scene diversity at scale.

Most commercial robotics failures trace to distribution shift between the training environment and the deployment environment. Corpshore AI's 150+ environment footprint is designed to close that gap.

Cultural variation in workspace layout requires geographically diverse collection. A kitchen in Manila is laid out differently from a kitchen in Warsaw or Accra. A warehouse in Karachi operates with different lighting, floor surfaces and object densities than one in Dusseldorf. Training data collected in one geography does not transfer cleanly to another.

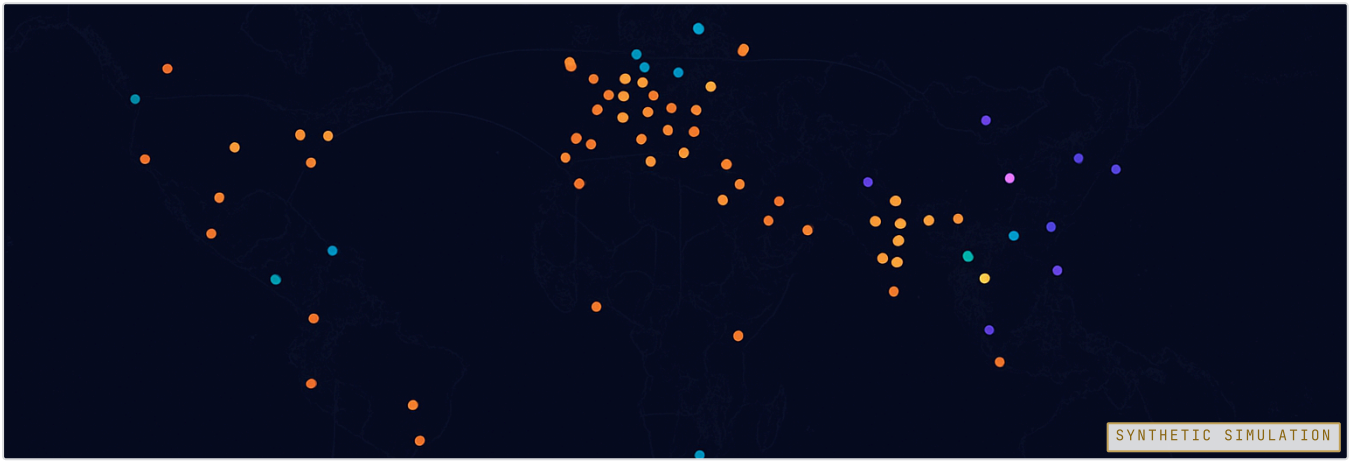
Corpshore AI collects in the region the robot will work in, using participants from the population the robot will work alongside.

ACTIVE AND ON-REQUEST COVERAGE

United States ACTIVE English Warehouse, Household	**United Kingdom** ACTIVE English Office, Retail	**Germany** ACTIVE German Manufacturing, Warehouse
France ACTIVE French Hospitality, Retail	**Philippines** ACTIVE English, Filipino Household, Office	**Ghana** ACTIVE English, Twi Agricultural, Household
Kenya ACTIVE English, Swahili Agricultural, Transport	**Nigeria** ACTIVE English, Hausa Household, Agricultural	**South Africa** ACTIVE English, Afrikaans Mining-adjacent, Manufacturing

ON-REQUEST AND JWUMA CONTRIBUTOR COUNTRIES

Japan · South Korea · Brazil · India · Australia · Netherlands · Uganda · Dominican Republic · Turkiye · Canada



PHYSICAL AI COLLECTION COVERAGE: 18+ ACTIVE COUNTRIES, 150+ ENVIRONMENT TYPES

Every video frame, annotation output, programme specification, quality metric and coverage figure in this document is a synthetic simulation produced by Corpshore AI. Nothing here is client data or live sensor material. The simulations illustrate physical AI data programme operations as Corpshore AI designs and runs them.

SECTION 6

Annotation architecture: IAA, formats and delivery

For robotics ML engineers and data infrastructure leads. Physical AI annotation is more complex than image labelling: the temporal dimension, ego-motion and contact-state labelling require a specialist protocol.

IAA PROTOCOL FOR ACTION LABELS

Action labelling in physical AI data is subjective. Two annotators watching the same clip may disagree on where one phase ends and another begins. Corpshore AI applies a structured IAA protocol with taxonomy alignment, d

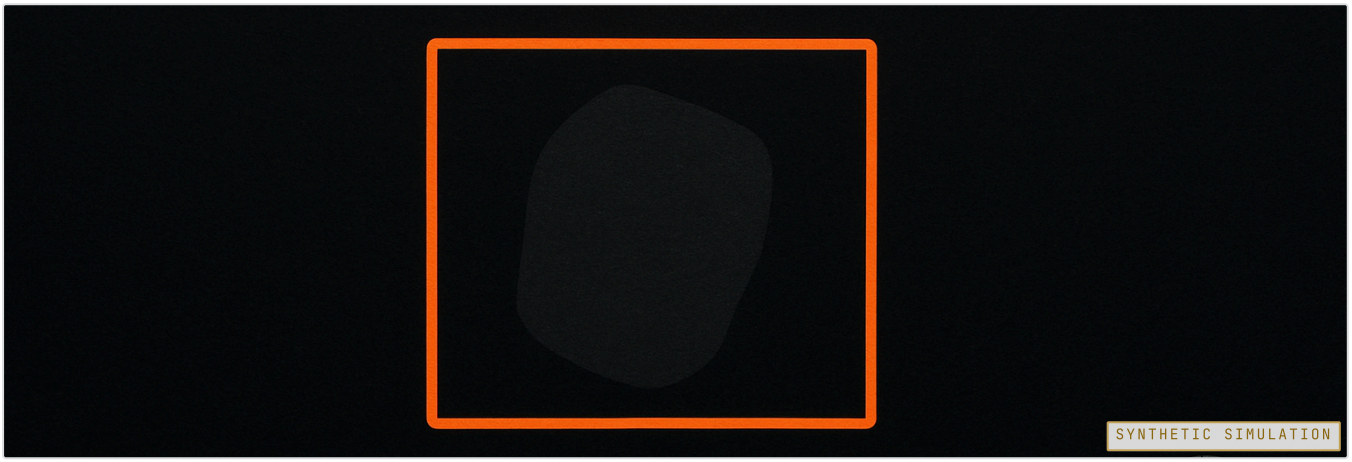
- Label taxonomy alignment session before collection: 50 gold-standard clips calibrated
- Double-annotation: 100% of clips annotated independently by two annotators
- Cohen's Kappa calculated per action class per batch
- Target: Kappa >= 0.75 (Basic), >= 0.82 (Standard), >= 0.88 (Premium)
- Arbitration by domain-specialist lead annotator for Kappa < threshold

IAA TARGETS BY TIER

TIER	IAA TARGET (COHEN'S KAPPA)	TYPICAL USE
Basic	⩾ 0.75	Pre-training, exploratory datasets
Standard	⩾ 0.82	Production training pipelines
Premium	⩾ 0.88	Safety-critical or fine-tuning

OUTPUT FORMATS

- HDF5: multimodal datasets and ML framework compatibility
- Parquet and Feather: tabular annotation metadata
- ROSBAG: ROS-based robotics pipelines
- TFRecord: TensorFlow training pipelines
- COCO JSON: object detection and segmentation tasks
- Custom schema: agreed at programme scoping



FRAME-LEVEL ANNOTATION WITH TEMPORAL PHASE AND CONTACT STATE LABELS

Every video frame, annotation output, programme specification, quality metric and coverage figure in this document is a synthetic simulation produced by Corpshore AI. Nothing here is client data or live sensor material. The simulations illustrate physical AI data programme operations as Corpshore AI designs and runs them.

SECTION 7

Physical AI is blocked by data, not models. We solve the data problem.

Corpshore AI and Jwuma collect egocentric video, real-world action demonstrations and multimodal sensor data across 18+ countries and 150+ environments. Our contributor pools are pre-qualified for safety compliance, consent documentation and demographic diversity. The collection infrastructure exists. The annotation capability exists. What is missing is a programme scoping conversation.

FREQUENTLY ASKED QUESTIONS

What is egocentric video and why is it important for physical AI training?

Egocentric video is recorded from a first-person, head-mounted or wearable camera: capturing exactly what a humanoid or mobile robot camera would see. It is the foundational training data format for embodied AI because policy models trained on third-person video fail to generalise to agents operatin...

How does Corpshore AI collect physical AI training data at scale?

Corpshore AI collects physical AI training data through its Jwuma contributor platform, which coordinates participant recruitment, consent management, equipment deployment and session logistics across 18+ countries and 150+ environment types. Each programme is scoped to match the buyer's target depl...

What annotation does Corpshore AI provide for robotics and physical AI datasets?

Corpshore AI provides frame-level and segment-level annotation for physical AI datasets, including action labels, temporal phase segmentation, object bounding boxes and instance masks, spatial relationship tags, hand-object contact state labels and quality flags. All annotation is performed by domai...

What multimodal sensor data can Corpshore AI collect?

Corpshore AI supports synchronised multi-channel data collection including RGB video, depth (LiDAR and structured light), inertial measurement unit (IMU) data, stereo audio and tactile or force data where partner hardware is provided. All sensor streams are timestamped to a common epoch and delivere...

Contact	**info@corpshore.ai**
Contributor platform	**jwuma.com**
Capability deck	**deck.corpshore.ai**
Interactive demo	**showcase.corpshore.ai/demos/physical-ai-robotics-data**

Jwuma by Corpshore AI | jwuma.com | info@corpshore.ai | +1 866 531 5727 | Part of Corpshore Solutions Corporation, Toronto, Canada | showcase.corpshore.ai | deck.corpshore.ai

Every video frame, annotation output, programme specification, quality metric and coverage figure in this document is a synthetic simulation produced by Corpshore AI. Nothing here is client data or live sensor material. The simulations illustrate physical AI data programme operations as Corpshore AI designs and runs them.