Causal Data Tiers for World-Action Models: A Survey of Egocentric Data, Model Design, and Evaluation

Benjamin Tay & Jaymari Chua
School of Computer Science and Engineering
The University of New South Wales
Sydney, NSW 2052, Australia

Abstract

World action models predict how environments change under action, but their accuracy depends on training data that records agent-environment causality. Third-person datasets capture state; egocentric data captures the action that causes it. This survey maps egocentric data sources for world model training. We organise existing data by causal completeness: uncurated video records outcomes without actions; laboratory telemetry links physical actions to states but at limited scale; synthetic environments generate full counterfactual trajectories. We then trace how this determines how models are constructed, and how they are designed to take advantage of the various strengths of the different data types. This is followed by an analysis on the state of the egocentric data economy, outlining the saturation of data collection businesses and highlighting the need to own dynamic data infrastructure and processing pipelines as opposed to just a raw data supplier. We conclude with a discussion on the ongoing problem of accurately evaluating the effectiveness of egocentric datasets in improving model performance.

Contents

1 Introduction **2**

2 Why Egocentric Data Matters **3**

2.1 Defining Egocentric Data 3

2.2 The Case For Egocentric Data 4

2.3 Curating Egocentric Data for World Models 4

3 A New Taxonomy for Egocentric Data **5**

3.1 Table Overview of the Data Taxonomy 5

3.2 Tier 1: In-the-Wild Egocentric Video 6

3.3 Tier 2: Controlled Laboratory and Robot Telemetry 7

3.4 Tier 3: Synthetic and Gameplay Environments 8

4 World Action Models as Data Consumers **8**

4.1 Design Philosophies Shaped by Data Tiers 8

4.2 Render-and-Decode WAMs 9

4.3 Latent-only WAMs 10
4.4 Video-Generation-Free WAMs 11
4.5 Pretraining and Post-training – A Short Summary 11

5 The Egocentric Data Economy **12**
5.1 The Egocentric Data Collection Movement 12
5.2 Structural Challenges 13
5.3 Data Is No Longer The Bottleneck 14

6 Evaluating Egocentric Data **15**
6.1 Scaling In The Dark 15
6.2 Why Egocentric Data Is So Hard To Benchmark 16
6.3 Research Directions Into Effective Evaluation 18

7 Conclusion **19**

1 Introduction

A world action model learns to predict future states conditioned on actions. Its function is $p(s_{t+1} \mid s_t, a_t)$, a conditional dynamics objective that underlies model-based control, latent imagination, and embodied planning (Ha & Schmidhuber, 2018; Hafner et al., 2020). This conditional defines the boundary between passive perception and embodied intelligence: a model that observes without acting learns correlation, while a model that acts learns causation.

The accuracy of this prediction depends on the data used for training. If the data records state transitions without the actions that produced them, the model must infer a hidden cause from observed effects. Third-person video datasets capture this ambiguity. They record what happened, but not what the observer did to make it happen.

A key bottleneck in world model research is therefore data source selection. In particular, egocentric observation is widely regarded as one of the most effective data modalities for training world models, as it provides an action-dependent stream which couples viewpoint changes, body movement, and interaction with the environment; in other words, it captures the causal loop between where to look, what becomes visible, and how the next action changes (Kerr et al., 2025).

However existing surveys of egocentric vision organise the field by task: action recognition, temporal anticipation, localisation, or gaze estimation (Chen et al., 2025). They treat egocentric data as a geometric variant of third-person data from mounted cameras, and this framing misses the causal structure that distinguishes world model training from video understanding. In contrast, we instead evaluate egocentric data by the completeness of its action-state record.

We propose that egocentric data sources should be classified by causal completeness, not by collection environment. The first tier contains in-the-wild egocentric video, where datasets such as Ego4D and EPIC-KITCHENS record outcomes, hands, objects, and long-horizon routines, but usually lack the action labels that caused the observed transitions (Grauman et al., 2022; Damen et al., 2022; 2021). The second tier contains laboratory and robot datasets with synchronised telemetry. Here, demonstrations, robotic actions, and sensor states are aligned, as in robot learning-from-demonstration and large-scale robot datasets, but hardware constraints limit scene and behavior diversity (Argall et al., 2009; Dasari et al., 2019; Ebert et al., 2021; Open X-Embodiment Collaboration, 2023; Khazatsky et al., 2024). The third tier contains synthetic environments and deterministic game replays, which generate full counterfactual trajectories: the same state can be rerun with different actions, yielding exact labels for what each action causes (Kolve et al., 2017;

Savva et al., 2019; Shen et al., 2021; Zhu et al., 2020; Reka Labs, 2026; Wu et al., 2026). This taxonomy explains why no single dataset solves the problem. Scale trades against causal density, and realism trades against controllability.

We then examine how the availability and causal structure of training data shape different world-action model architectures. We divide data into in-the-wild egocentric video, synchronised robot telemetry, and synthetic environments, showing how each supplies complementary visual priors, action grounding, and counterfactual supervision. These trade-offs produce three broad design philosophies—render-and-decode, latent-only, and video-generation-free models—with different computational costs and dependencies on each data tier. Across these architectures, scalable egocentric and synthetic data provide cost-effective world-model pretraining (Ma et al., 2026b), while scarce, embodiment-specific robot demonstrations remain the principal bottleneck for converting predictive representations into deployable actions.

Ultimately, the rapid commercialisation of egocentric data is creating an emergent data economy for embodied AI. More than 50 collection companies have reportedly entered the market, while providers such as Lightwheel advertise hundreds of thousands of delivered recording hours (Lightwheel AI, 2026b). This growth is supported by record robotics investment (Sternstein, 2026) and demand from venture-backed foundation-model and humanoid-robotics companies. Yet raw footage is becoming increasingly commoditised, and its scale does not guarantee usefulness for training world-action models. We examine the market's dependence on downstream robotics funding and argue that durable value will shift from collecting hours toward processing, action extraction, simulation, curation, and standardised evaluation of control-relevant data.

The final part of the paper examines why the rapid scaling of egocentric datasets has outpaced the field's ability to evaluate their usefulness for world-action models. Reported hours, tasks, and environments are inconsistently defined (Khazatsky et al., 2024; Grauman et al., 2022; Open X-Embodiment Collaboration, 2023; Lightwheel AI, 2026b; Build AI, 2025) and reveal little about whether footage provides transferable, control-relevant supervision. Unlike language data, egocentric video contains no native robot actions, so its value depends on human-to-robot conversion, target embodiment, policy architecture, and closed-loop deployment conditions. The section therefore outlines emerging alternatives to scale-based reporting.

2 Why Egocentric Data Matters

2.1 Defining Egocentric Data

Egocentric data consists of sensory streams captured from a first-person perspective, typically via head-mounted, chest-mounted, or wrist-mounted cameras. Unlike traditional third-person (exocentric) datasets, which record an event from a static, external vantage point, egocentric data records the environment exactly as the agent experiences it.

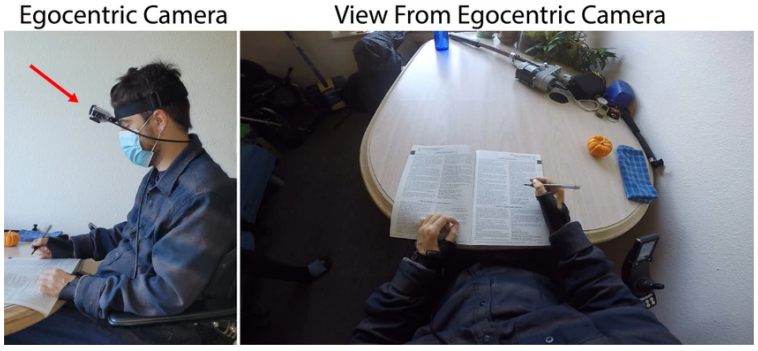

Figure 1: A head-mounted egocentric camera (left) and the corresponding first-person egocentric view it captures (right) (Objectways, 2026).

The defining visual characteristic of egocentric manipulation is that the agent's hands, tools, or grippers dominate the foreground. Though it excels in capturing the real-world vision of the agent, a challenge is that it creates severe causal occlusion; its own embodiment blocks the camera's view of the contact point, e.g. through its hands blocking its view of its own action.

2.2 The Case For Egocentric Data

A world model is defined as an internal predictive representation of a particular environment, designed to allow an agent to simulate how the world is affected by its actions.

A world model is therefore useful when it can answer "what would happen if the agent acted differently?" In embodied settings, that answer is not exhausted by pixels.

In almost all instances, tackling this question requires understanding and knowledge of: geometry, object identity, contact state, task intent, reward and latent physical variables. A robot deciding whether to grasp an object, open a door, align a tool, or recover from failure needs to be able to accurately predict different real-world potential outcomes as opposed to merely generating visually plausible continuations.

Thus, egocentric data matters because it puts prediction inside the agent's own information state. From this perspective, the camera records what the agent could plausibly use at decision time; hands, grippers, tools, and body motion create the same occlusions that will appear at deployment; and, when actions or telemetry are synchronised, interventions can be linked to outcomes. Third-person footage can describe what occurred, but egocentric action-state data can show how an embodied agent made it occur. That difference makes egocentric data a supervision regime rather than merely a camera geometry.

In other words, egocentric data is the only format that perfectly bridges the gap between passive observation and embodied agency. Ma et al. (2026b) show that, under matched pretraining budgets, embodied foundation models pretrained on egocentric human video rather than teleoperated real-robot trajectories achieve a 24% lower validation loss on real-robot action prediction, along with 52.5% and 90% higher success rates on in-distribution and out-of-distribution real-robot task execution, respectively. Similarly, Wang et al. (2026b) found that a policy trained on just 30 minutes of human egocentric video per task reaches 92.5% average success across four real-world tasks, outperforming matched-time robot teleoperation by 41%.

2.3 Curating Egocentric Data for World Models

Existing egocentric-vision surveys usually organise the field around recognition, localisation, anticipation, or human activity understanding (Chen et al., 2025).

For world models, however, the organising question changes. As their goal is to understand consequences and outcome, the relevant test is whether a dataset supports action-conditioned prediction and downstream control, not simply whether it supports classification or captioning.

In summary, all egocentric data should be causally structured. On a high level, dataset curation needs to address three major requirements.

Pseudo Action Labelling: World Models need to know what intervened in the scene, and the pipeline must therefore recover the missing causal link.

Causal Filtering: The pipeline needs to verify the camera is stable, that a meaningful physical interaction occurs, and that the acting agent is fully visible throughout the process of the interaction.

Deduplication: Raw datasets often suffer from immense redundancy, with hundreds of clips that are identical or of extremely similar quality. To ensure the model doesn't overfit to a specific background, we need to deduplicate the dataset. However, a key thing to note is that deduplication must be interaction-aware, because if they have the exact same background but different physical motions, both must be retained.

Fulfilling these three requirements is fundamental to maintaining an effective dataset for training World Models. However, not all data is made equal; different data types have various inherent advantages/disadvantages that contribute to their overall usefulness, as outlined in the following section.

3 A New Taxonomy for Egocentric Data

We propose a new taxonomy for egocentric data, that is fundamentally ordered around causal completeness as opposed to collection environment, which is a far more significant metric for application to world model policy training.

Table 1 summarises the three data tiers most relevant to egocentric world model training. In-the-wild video supplies semantic breadth; laboratory and robot telemetry supplies physical grounding; synthetic and gameplay environments supply controllable counterfactuals.

It is important to note, therefore, that the tiers are complementary rather than hierarchical, as each tier specialises in providing a certain necessary component.

3.1 Table Overview of the Data Taxonomy

Table 1: Representative datasets across the three egocentric data tiers, compared by environment, scale, perspective, action labeling, proprioception, contact/physics grounding, and counterfactual coverage.

Dataset	Tier	Environment	Scale (Hrs)	Persp.	Action Labels	Proprio.	Contact/ Physics	Counterf.
Tier 1: In-the-Wild Egocentric Video (high semantic breadth, low causal completeness)								
Ego4D (Grauman et al., 2022)	1	Open-world daily	3,670	Ego	None (Narrations)	No	No	No
EPIC-KITCHENS (Damen et al., 2022)	1	Real kitchens	100	Ego	None (Narrations)	No	No	No
EgoDex (Hoque et al., 2025)	1	Tabletop	829	Ego	Pseudo (Hand Pose)	No	No	No
HumanNet (Ego) (Deng & Zhou, 2026)	1	Open-world daily	800,000	Ego	Pseudo (Hand Pose)	No	No	No
EgoSuite-Open100K (Lightwheel AI, 2026a)	1	Multi-domain (7 types)	100,000	Ego (+Wrist)	Pseudo (Hand/Body Pose)	No	No	No
Tier 2: Controlled Laboratory and Robot Telemetry (high causal completeness, low semantic breadth)								
RoboNet (Dasari et al., 2019)	2	Lab tabletop	45	Exo	Motor Commands	Yes (Joints)	No	No
RT-1 (Brohan et al., 2023)	2	Office kitchens	130	Exo	Action Tokens	Yes (Gripper)	No	No
DROID (Khazatsky et al., 2024)	2	In-the-wild robot	350	Exo	Continuous Vectors	Yes (Joints/EE)	Depth (Partial)	No
Open X-Embodiment (Open X-Embodiment Collaboration, 2023)	2	Multi-robot	3,000	Ego/Exo	Continuous Vectors	Yes (Varies)	Task Success	No
AgiBot World (AgiBot-World-Contributors et al., 2025)	2	Domestic/Retail	2,976	Ego/Exo	Continuous Vectors	Yes (Full body)	No	No
Tier 3: Synthetic and Gameplay Environments (high counterfactual coverage, domain gap to reality)								
AI2-THOR (Kolve et al., 2017)	3	Simulated homes	∞^*	Ego/Exo	Discrete API	Yes (Virtual)	Full State	Yes
Habitat (Savva et al., 2019)	3	Scanned scenes	∞^*	Ego	Continuous Control	Yes (Virtual)	Full State	Yes
CS2-10k (Reka Labs, 2026)	3	Gameplay	10,000	Ego	Keyboard/Mouse	No	Game State	Yes
EgoCS-400K (Wu et al., 2026)	3	Gameplay	10,000	Ego	Keyboard/Mouse	No	Game State	Yes

*Procedurally generated simulators support effectively unbounded interaction hours, not a fixed corpus size.

3.2 Tier 1: In-the-Wild Egocentric Video

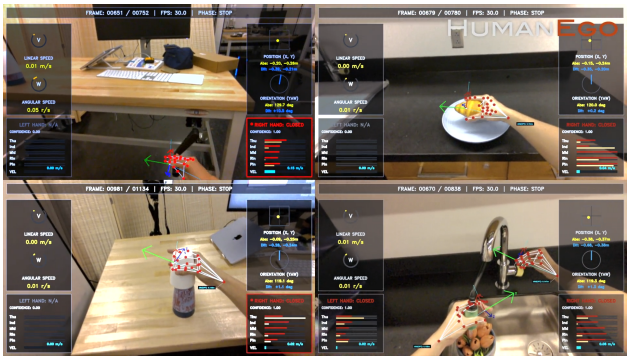

Figure 2: HumanEgo egocentric dataset (Wang et al., 2026b)

'In-the-wild' egocentric video is defined as first-person footage, usually of people, performing ordinary unscripted activities in their contextual environments. This is seen through datasets such as Ego4D and EPIC-KITCHENS-100, which hands, objects, routines, gaze-driven framing, and long-horizon activity in natural environments (Grauman et al., 2022; Damen et al., 2022; 2021). For instance, Ego4D contains 3670 hours of video recorded by 931 unique participants, and the footage spans hundreds of everyday scenarios such as household chores, workplace tasks, leisure and outdoor work. In contrast EPIC-KITCHENS-100 focuses more narrowly within kitchen environments, where participants wore a head-mounted GoPro Hero 7 while doing cooking, cleaning and food preparation tasks.

Such data is usually collected through head-mounted wearable cameras, with long continuous recordings that capture full routines. The primary signal is RGB video, and there is no other modalities such as contact forces, joint angles, proprioception. They offer breadth without direct intervention labels, at a scale and diversity that instrumented robot collections cannot match.

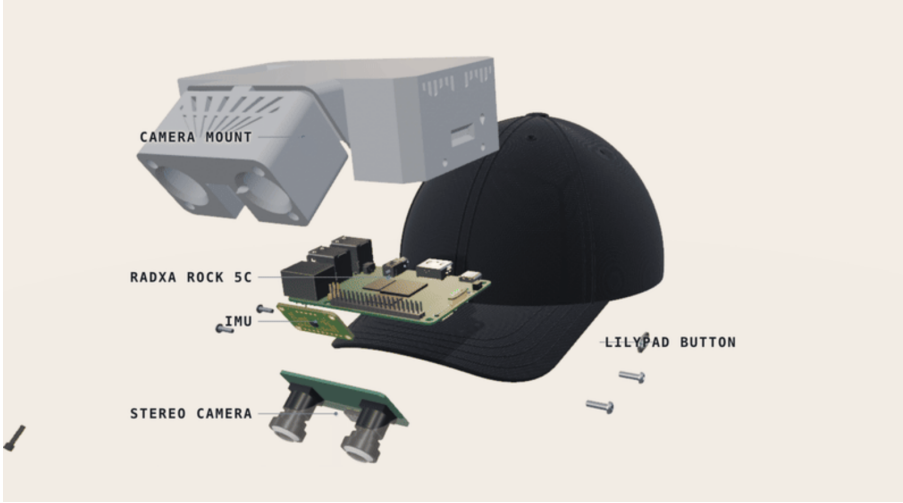

Figure 3: Ego-OSCAR: an open-source egocentric capture system (Paul et al., 2026).

However, because the data records only visual outcomes and not the physical actions, core critical signals are missing, such as logged motor commands, torques, object poses and tactile measurements. This means that any world model trained solely on tier 1 video must infer affordances and causal structure purely from

visual correlations; this is much less effective and therefore needs to be supplemented by further grounding through robot or laboratory data to supply missing action-state pairs.

Ultimately, though tier 1 data is unmatched in terms of scale, realism and diversity, it is so far still fundamentally incomplete for learning the conditional dynamics that define a world model.

3.3 Tier 2: Controlled Laboratory and Robot Telemetry

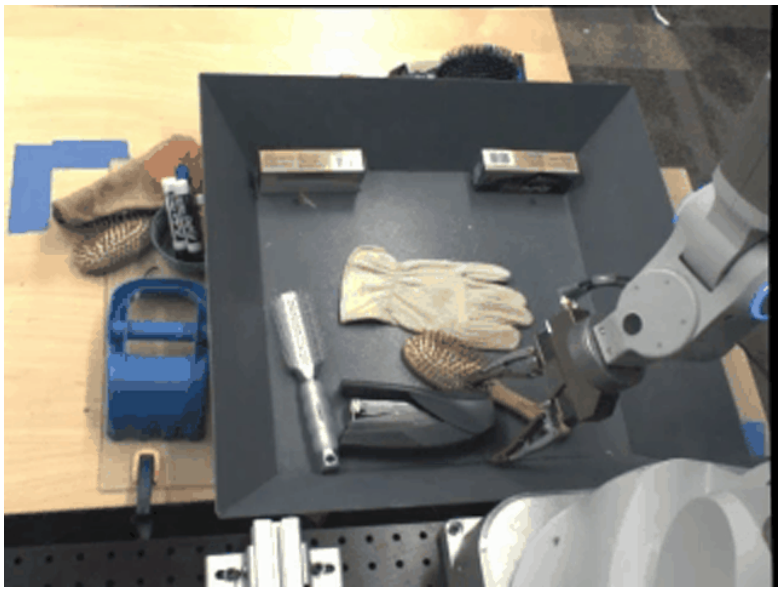

Figure 4: The RoboNet robotic dataset featuring a Google R3 (Dasari et al., 2019).

Controlled laboratory and robot teleoperated egocentric data consists of first-person video and sensor streams recorded while a human operator teleoperates a robot, or while a robot executes logged actions while inside a structured environment. A core difference to Tier 1 data is that every state transition is paired with an explicit action label.

In particular, robot and laboratory datasets such as RoboNet, RT-1, Open X-Embodiment, and DROID provide synchronised interventions, proprioception, gripper states, task outcomes, and sensor streams (Dasari et al., 2019; Brohan et al., 2023; Open X-Embodiment Collaboration, 2023; Khazatsky et al., 2024). Because the action is logged at the moment it is executed, the model can learn how a command changes the observed state. This precision makes the tier essential for physical grounding, as you have the true actions that produce each state transition. There is also a very high signal-to-noise ratio for contact-rich manipulation.

Its weakness is distributional: laboratory scenes, hardware constraints, and scripted tasks cover less of the world than human wearable video. This is due to the more involved nature of obtaining such data; it requires manual teleoperation, which is a much more time-consuming process as opposed to capturing raw video like Tier 1. It is also a relatively expensive procedure and it is impractical to scale.

Overall Tier 2 data is the strongest source of causally correct supervision for training world models, however, it cannot by itself provide the breadth and scale needed for a truly generalisable embodied intelligence.

3.4 Tier 3: Synthetic and Gameplay Environments

Figure 5: CS2-10k: a large-scale egocentric Counter-Strike 2 dataset (Reka Labs, 2026).

The third tier is best understood as a counterfactual laboratory. This refers to counterfactual control, which is the evaluation of hypothetical intervention scenarios, allowing the robot to test alternate action paths before executing any physical movement, which is significant for ensuring safety particularly in high-stakes deployments such as surgery or space station maintenance.

Simulators and deterministic gameplay corpora such as AI2-THOR, Habitat, iGibson, robosuite, CS2-10k, and EgoCS-400K can expose privileged state, exact action histories, rare events, and alternative rollouts from the same initial condition (Kolve et al., 2017; Savva et al., 2019; Shen et al., 2021; Zhu et al., 2020; Reka Labs, 2026; Wu et al., 2026).

This density is difficult to obtain in the real world, where repeated resets and exhaustive action branching are expensive or unsafe. At the same time, simulation and gameplay introduce rendering, physics, and behavior gaps. They are most useful not as replacements for reality, but as sources of counterfactual supervision, ablation, and stress testing before transfer is measured on real egocentric data. It is important to note that while tier 3 data maximises causal completeness and controllability, it comes at the sacrifice of realism and scale of natural variation.

4 World Action Models as Data Consumers

4.1 Design Philosophies Shaped by Data Tiers

The different types of designs for world models are directly influenced by the nature of the data available, as researchers seek to find the optimal mix of tradeoffs and advantages of each data type, whether it be scalability or compute capacity. However, at the core of this, world models are designed to predict causality and through the deliberate arrangement of egocentric data into 'tiers' of causality, and Figure 6 shows how each different tier influences a unique world model design.

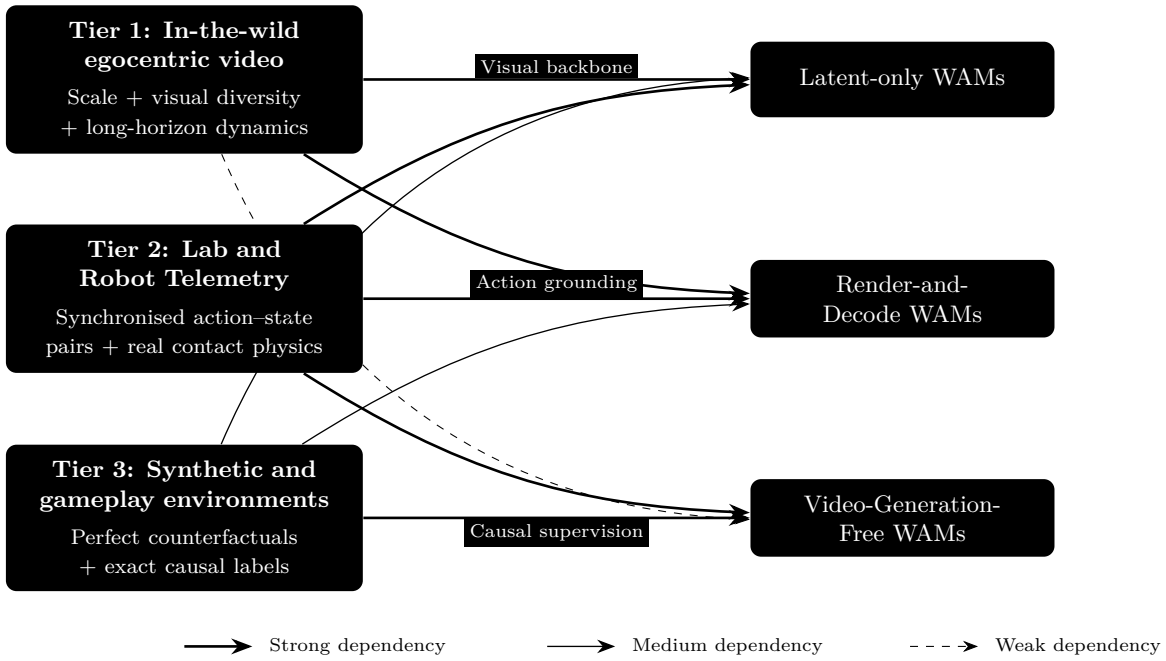

Figure 6: Data tiers contribute complementary strengths to different World Action Model (WAM) families. Tier 1 anchors visual, appearance-driven priors; Tier 2 supplies the synchronised action-state grounding every WAM family ultimately depends on; Tier 3 supplies dense causal and counterfactual supervision. Line weight encodes the strength of each dependency.

4.2 Render-and-Decode WAMs

A Render-and-Decode model (e.g., UniPi (Du et al., 2023), GR-1 (Wu et al., 2023), Gen2Act (Bharadhwaj et al., 2024)) does the following process at inference time:

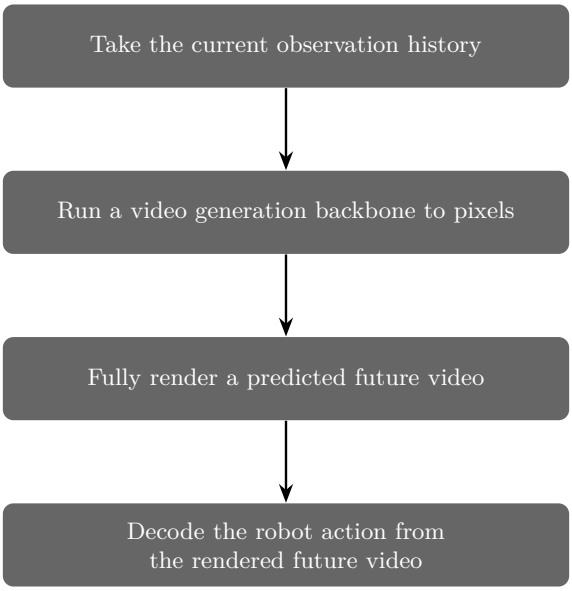

Figure 7: Inference-time pipeline of a Render-and-Decode WAM.

The core design philosophy is that the fully rendered video is worth producing as it preserves the complete visual prior learned by the video model, and the action module thus is able to read the inspectable and contextually rich future prediction. It is the earliest style of WAM designed, and is still widely used.

Render-and-Decode WAMs are dependent on the complementary strengths of the three different data tiers.

For Tier 1 data (in-the-wild egocentric video), this is the primary training source for the video generation and rendering backbone. As Tier 1 datasets are much larger, diverse and long-horizon, it is more effective in helping the model learn broad visual statistics of common objects, interactions and motions. However, tier 1 data is still insufficient by itself because it has no action labels, and therefore cannot supervise the model's action decoder.

Therefore, Tier 2 data supplies the missing observation and action pairs, and as it has action labels, it is used to fine-tune the video backbone onto robot embodiments and train the action decoder.

Finally, Tier 3 provides the unique advantages of exact counterfactual trajectories, and is therefore ideal for large-scale supervised training of the full pipeline – as through generating hard negative futures we are able to test if it has effectively learnt the causal dynamics of the task.

Ultimately as seen in Figure 6, Render-and-Decode WAMs have a strong dependency on all types of causal data, meaning that we require a relatively quantity of data to operate this model. In particular, it depends heavily on Tier 2 data, which is difficult to scale at a large amount.

4.3 Latent-only WAMs

Latent-only WAMs, in contrast, keep the powerful predictive prior learned by a video or world model, however, unlike a render and decode world action model, it never decodes to pixel level during inference time. Instead, it stops at an intermediate latent representation, and decodes the robot action directly from the latent space.

Concretely, what counts as a "latent representation" varies across implementations: it can be a compressed pixel-space tensor, an intermediate feature map from a video or vision-language backbone, an explicit geometric field, or an affordance map, each of which hands the action decoder a different kind of control signal. Table 2 summarises the four substrates most common in current Latent-only WAMs.

Table 2: Common latent substrates used by Latent-only WAMs.

Latent substrate	Technical form	Control signal	Representative paper
Pixel latent	VAE/VQ spatiotemporal tensor $z_{t:t+H}$	Decodable appearance and dynamics	LaWAM (Chen et al., 2026)
Feature latent	Intermediate DiT states or teacher embeddings $h_{t:t+H}$	Semantic, action-relevant dynamics	JEPA-WAM (Lin et al., 2026)
Geometric latent	2D/3D flow, tracks, depth or correspondences	Explicit object/scene motion	GeoSem-WAM (Ma et al., 2026a)
Affordance latent	Masks, value maps, contact or progress fields	Feasible interactions and task state	DreamWAM (Yuan et al., 2026)

As mentioned above, full pixel rendering under render-and-decode is computationally expensive, and often the generated pixel detail is irrelevant for the actual decision. Latent-only WAMs retain the useful predictive signal, while discarding the unnecessary rendering step, allowing it to be faster and more practical.

Like pixel-based models, Latent-only WAMs also draw on all three tiers of causal egocentric data, but the mixture differs sharply in how much Tier 2 data is actually required. The clearest evidence comes from JEPA-style architectures: V-JEPA 2 is pretrained self-supervised on over one million hours of large-scale internet video, and its action-conditioned variant, V-JEPA 2-AC, is then post-trained on fewer than 62 hours of teleoperated robot interaction from the DROID dataset, after which it plans zero-shot in unseen labs at 65–80% success (Assran et al., 2025).

4.4 Video-Generation-Free WAMs

These WAMs forgo the video generator entirely, predicting entirely in embedding, token, or geometric spaces (e.g., FLARE (Zheng et al., 2025), PointWorld (Huang et al., 2026), Audio-WM (Zhang & Gienger, 2025)). Because there is no video backbone running, these models are faster and cheaper at inference than both other types of WAMs.

Because these predict structured, non-visual futures, Tier 3 is the only data source which can supply these signals at scale with perfect accuracy. In a simulator, you have exact 3D object poses, velocities, and full information of every action and physical interaction within the space.

Furthermore, this type of WAM does not need massive visual diversity by virtue of it not relying on a video generator. Instead, it can focus on learning accurate dynamics and geometry rather than spending compute on rendering, making simulation a more effective match.

4.5 Pretraining and Post-training – A Short Summary

WAMs are typically built in two distinct stages; a large-scale pretraining of a predictive world backbone, and then followed by post-training which couples the backbone to robot actions.

Within pretraining, the goal is to learn a generalisable predictive representation of the world. As such, the egocentric data used is mostly of Tier 1 and Tier 3, with minimal or no Tier 2 data. This is a deliberate design choice due to the scarcity of Tier 2 data, with its quantity insufficient for the diversity and breadth needed to train a backbone.

Within post-training, the goal is to turn the pretrained world model into an actionable policy that outputs real robot actions for a target embodiment and task distribution. Here, this is dominated by Tier 2 data, as we require specific demonstrations of the task at hand to anchor the model.

Ultimately, this shows that scaling Tier 1 / 3 remains the most cost effective way to improve the foundation, and scaling Tier 2 is the bottleneck for making the robot deployments perform as well as possible. Figure 8 summarises this pipeline and the resulting scaling asymmetry between the two stages.

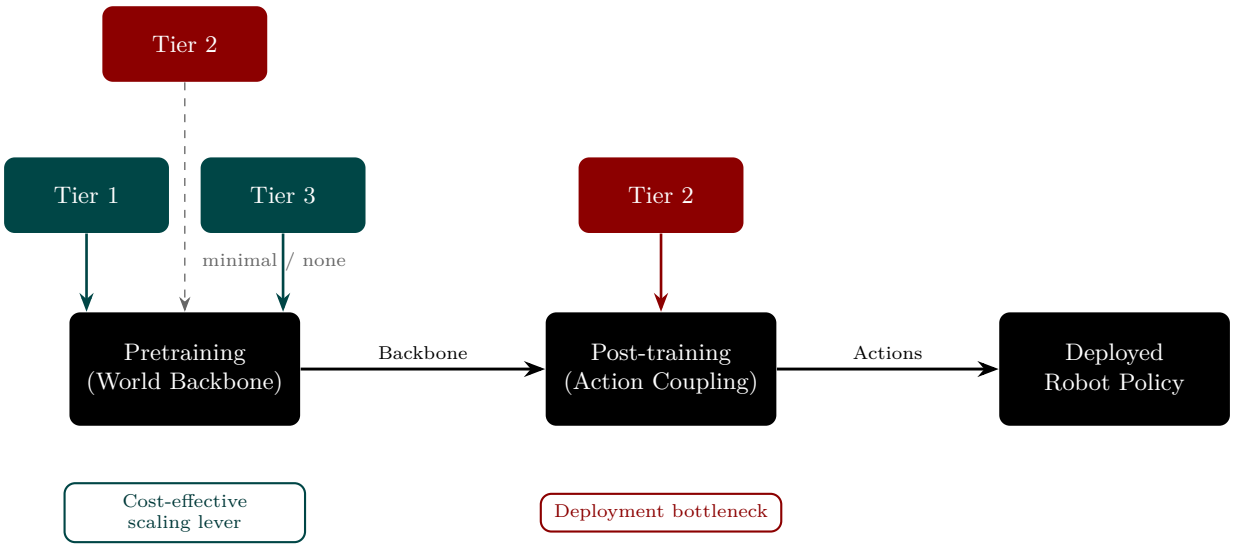

Figure 8: Data-tier composition across the two stages of a WAM pipeline. Pretraining draws almost entirely on abundant Tier 1 and Tier 3 data, while post-training is dominated by scarce, task-specific Tier 2 demonstrations – making Tier 1/3 the cheap lever for improving the backbone, and Tier 2 the bottleneck on final deployment quality.

5 The Egocentric Data Economy

5.1 The Egocentric Data Collection Movement

The first half of 2026 has seen a rise in egocentric data collection companies.

Driven by a relatively low barrier to entry, a sudden industry surge in robotics funding and the recognition that internet video is mostly third-person and insufficient for embodied AI, the egocentric data collection space has gone from research datasets to a commercial industry with > 50 players in a very short window, with almost all the companies having formed in the last $\sim$18 months against a backdrop of projected demand for more and more hours of training data.

This funding surge is not merely anecdotal. Robotics venture activity reached an all-time high in Q1 2026, with roughly \$16B invested across just under 500 deals in the quarter, a pace roughly 2$\times$ larger by deal count and 4.5$\times$ larger by deal value than the preceding 2021–2025 stretch (Sternstein, 2026). Figure 9 shows this trajectory. Because egocentric capture hardware and annotation pipelines are themselves capital- and labor-intensive, this broader capital rotation toward physical AI is a direct tailwind for the egocentric data economy: it is the same investment wave that is funding the $>$50 data-collection startups described above.

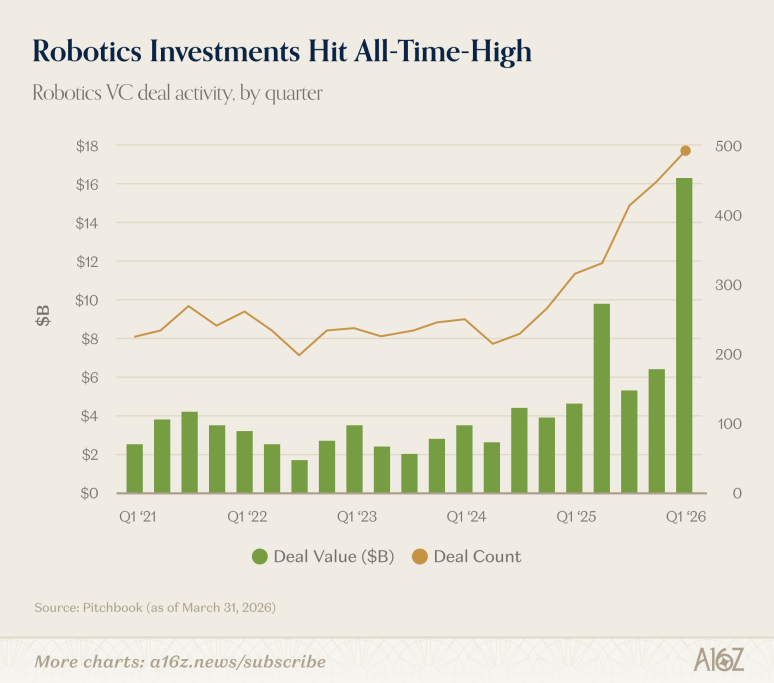

Figure 9: Robotics VC deal activity by quarter, Q1 2021–Q1 2026. Both deal value (bars) and deal count (line) reached record highs in Q1 2026. Source: PitchBook, as of March 31, 2026, reproduced from Sternstein (2026).

One of the most eminent data collection companies is Lightwheel, which runs EgoSuite, a large-scale egocentric human data operation for embodied AI and world models (Lightwheel AI, 2026b). They claim to have over 300,000 hours of egocentric data already delivered to robotics / VLA / WAM teams, 20,000+ hours per week of ongoing collection, and operate across 7 countries (Lightwheel AI, 2026b). They have also recently open-sourced the EgoSuite-Open100K dataset, which spans 100,000 hours of annotated human data, one of the largest public releases in the category to date (Lightwheel AI, 2026a).

5.2 Structural Challenges

Generalist AI's breakthrough GEN-1.5 model builds on a pretraining corpus of over half a million hours of real-world data captured from low-cost wearable devices on humans performing everyday activities; notably, this pretraining set contains no robot data at all, with downstream adaptation to a new task requiring only around one hour of robot-specific data (Generalist AI, 2026). Ego2Robot demonstrates a complementary strategy for stretching a comparatively small egocentric corpus: it converts roughly 1,940 hours of egocentric human video, pooled from four sources (ANT, EgoDex, ViTRA, and EgoVerse), into 18,561 hours of synthetic robot training data spanning 15 robot morphologies via action retargeting and visual synthesis, which is then co-trained alongside roughly 6,565 hours of real robot demonstrations to improve out-of-distribution generalisation (Wang et al., 2026a).

Therefore, such proven demand for data and the volume of data collection startups that have emerged over the past year, coupled with the amount of funding that is continually going into the space, it is a safe assumption to make that the egocentric data space will become increasingly more commoditised, as raw ego pricing – Tier 1 data – will increasingly drop as supply keeps growing.

However, beyond the hype, there is a key unresolved structural challenge: namely, that buyers and suppliers of data are often funded by the same pool of capital.

Indeed, most of the customers of data collection companies: foundation model teams, robotics startups or AI-transitioning traditional robot manufacturers. They are mostly all at a pre-revenue or early commercialisation stage, meaning their data budgets are funded by venture financing rather than actual operating cash flow.

In this sense, the market for egocentric data collection is highly correlated with the success of downstream robotics companies. Within the emergent space of robotic deployments, if they fail to reach successful commercial impact, this could lead to a rapid slowdown of data demand.

Table 3 makes this concrete: across six of the most visible humanoid and robot-foundation-model companies, funding and reported valuations run into the hundreds of millions to tens of billions of dollars, while disclosed revenue and commercial deployment remain comparatively small. Data-collection spend at this scale is therefore, almost by construction, venture capital chasing a promised future market rather than cash flow from an existing one.

Table 3: Selected humanoid and robot-foundation-model companies illustrate the circularity between data supply and demand: most named customers of egocentric and robot data providers are themselves pre-revenue or early-commercialisation companies funded by venture capital rather than operating cash flow.

Company	Focus	Funding & Valuation	Traction	Data-Spend Context
Figure AI	Full-stack humanoid	∼$2.3B raised; $39B Series C (Sep 2025) (Figure AI, 2025)	Revenue undisclosed; ∼150 units shipped in 2025; 2026 revenue estimated in the low tens of millions	Highest-profile humanoid; data spend cannot come from operating profit at this scale
Physical Intelligence	Foundation models (π_0 / $\pi_{0.5}$)	$2.4B → $5.6B valuation; $11B+ round reported, unclosed (The AI Insider, 2026)	Pre-product model lab; buys and self-collects egocentric/robot data (Levin, 2025)	Classic pre-revenue model lab funded entirely by venture rounds
Generalist	Cross-robot foundation software (GEN-1.5)	$540M+ raised; $3B valuation reported (Aug 2026) (TechCrunch, 2026)	Scale AI data customer; building generalist manipulation policies (Levin, 2025)	Venture-backed model company, not a revenue-generating OEM
1X Technologies	Home humanoid (NEO)	∼$137M raised; $1B round in talks at $10B+ valuation (Tech Startups, 2025)	Early enterprise / teleoperated deployments; consumer NEO launch still pre-commercial	Fundraising far outpaces disclosed revenue, consistent with the sector-wide pattern
Dyna Robotics	Manipulation foundation model (DYNA-1)	$120M Series A (Sep 2025) (Dyna Robotics, 2025)	Early commercial pilots (e.g. laundry folding)	Research-to-product company still in the pilot stage, not cash-flow positive
Cobot (Collaborative Robotics)	Warehouse / hospital mobile manipulator (Proxie)	$140M+ raised (Series B) (Collaborative Robotics, 2024)	∼30 units in trials at Maersk and Mayo Clinic; 5,000+ operating hours (Interesting Engineering, 2024)	Named Scale AI data customer; deployments still early pilots, not yet a mature revenue stream (Levin, 2025)

5.3 Data Is No Longer The Bottleneck

Data volume is no longer the bottleneck for robotics training. Instead, like an oil industry for a post-GPT world, companies must build robust processing pipelines that turn raw data (Tier 1) into actual usable signal. Therefore, the new moat lies in dynamic data infrastructure, not the data itself.

Simulation is one such layer. Jensen Huang has framed the Physical AI era as one that will be "built in simulation first," and NVIDIA has productised that claim across Omniverse, Isaac Sim / Isaac Lab, Cosmos world models, and OpenUSD. Cosmos 3 is positioned as an open physical-AI foundation model that combines scene understanding, world generation, and action prediction; NVIDIA reports it ranking first on several open world-generation and robot-policy benches (NVIDIA et al., 2026). The practical implication for data companies is that synthetic and real-to-sim pipelines now compete with field collection. In relation to the taxonomy, it means that companies must take control over the production of Tier 1 and Tier 3 data, both of which are highly in demand.

Data processing is another layer. Embodied data is multimodal through RGB, wrist cameras, depth, IMU, tactile, proprioception, language, and arrives in inconsistent schemas. Encord's February 2026 Series C is the cleanest public proxy for this market: $60 million led by Wellington Management, $110 million total raised, platform volume growing from 1 petabyte to more than 5 petabytes in twelve months, and physical-AI revenue up 10× over the same period, across 300+ AI teams (Stanciuc, 2026). Encord explicitly sells curation, annotation, evaluation, and alignment, not raw capture. Scale AI occupies the adjacent generalist slot: it reported 150,000+ hours of physical-AI data delivered in 2025, a claimed pace of 1,000+ demonstration hours uploaded per day (Jadhav, 2026), and a partnership with Universal Robots to collect force-feedback data on production cobot arms rather than only on research cells (Levin & Dadabhoy, 2026).

Ultimately, long-term winners in this field will be firms that establish tight control over the simulation and processing stack, as opposed to collecting the most hours.

6 Evaluating Egocentric Data

6.1 Scaling In The Dark

Press releases and papers for robotics data companies' datasets feature taglines of the number of collection hours, task diversity, or magnitude of environments it spans. However, the simpler question of "is this data good?" still remains in the air, as there is currently no accepted or standardised metric that takes a raw dataset as its input and returns a number quantifying the effectiveness of said dataset onto improving a model at doing a task. In other words, the field has still no accepted function of

$$f(\text{raw dataset}) \rightarrow \text{quality score}$$

that is model-agnostic, embodiment-agnostic and predictive of closed-loop success.

In particular, quantifying by hours, tasks or scenes are weak proxies due to a lack of consistent standardisation. For instance, '10,000' tasks could mean 10,000 duplicates of pick-and-place movements, or 200 unique skills. There lacks a shared task ontology across datasets across Ego4D (Grauman et al., 2022), DROID (Khazatsky et al., 2024), Build AI (Build AI, 2025), EgoSuite (Lightwheel AI, 2026b) or Open X-Embodiment (Open X-Embodiment Collaboration, 2023).

Hours, task counts, and scene counts are not commensurable across releases because each corpus defines the unit differently. DROID (Khazatsky et al., 2024) explicitly counts a "task" as a unique verb extracted from language instructions—not a verb–object pair—and a "scene" only when the robot's workspace changes substantially (a new room or kitchen corner), not when objects or a tablecloth are rearranged. On that definition it reports 86 tasks / 564 scenes / 52 buildings / 350 hours / 76k trajectories. The same paper notes that verb counts, verb–object counts, and manually curated taxonomies produce different rankings of "diversity," which is why they had to devote a section to the definition at all.

Open X-Embodiment (Open X-Embodiment Collaboration, 2023) uses a third scheme. It advertises 527 skills and 160,266 tasks over 1M+ trajectories and 22 embodiments, but those "tasks" are language instruction instances pooled from 60 pre-existing datasets, and the authors themselves observe that most skills still collapse into the pick-and-place family. A buyer comparing "86 DROID tasks" to "160k OXE tasks" is not comparing the same object: one is a verb inventory on a single Franka; the other is a cross-lab instruction count with mixed action spaces and control rates (roughly 3–10 Hz). Ego4D (Grauman et al., 2022) is not a manipulation skill catalogue at all. The headline is 3,670 hours, 931 wearers, 74 locations, 9 countries. Internally it mixes 136 scenarios, a Moments taxonomy of 110 high-level activities, narrations with 1,772 unique verbs and 4,336 nouns, and later goal/step overlays (e.g. 319 goals / 34 scenarios, or 86 kitchen goals / 514 steps on a subset). "Cooking" is a scenario, "setting the table" is a moment, and "pick up" is a forecasting action—three granularities that would each yield a different "task" number if a vendor copied the marketing template.

Commercial and industrial dumps make the mismatch worse. Lightwheel's EgoSuite (Lightwheel AI, 2026b) cites 10,000+ tasks across 500+ environments without publishing a verb list, a scene-change rule, or an inter-annotator protocol. Build AI's Egocentric-10K / 100K / 1M series (Build AI, 2025) is sold in hours and frames (100K hours, 10.8B frames, 14k factory workers) rather than a skill ontology; "assembly / sort / pack / machine" is a process domain, not a closed task set. Those figures cannot be lined up with DROID's 86 verbs or OXE's 527 skills without an unstated mapping.

Figure 10: A promotional frame from Lightwheel's EgoSuite-Open100K dataset page, overlaying "15,000+ Scenes" and "15,000+ Tasks" on an egocentric grocery-shopping clip. The headline numbers are presented with no accompanying scene-change rule or task/verb definition, illustrating the marketing-unit problem described above. (Lightwheel AI, 2026a)

Even hours are inconsistent. DROID's (Khazatsky et al., 2024) 350 hours are teleoperated robot interaction at 15 Hz with wrist and exo cameras. Ego4D (Grauman et al., 2022) hours include unscripted daily life, much of it with no manipulation and no action labels. Gig-worker ego capture bills wall-clock recording time, including idle, transit, and talking-head footage. AgiBot-style (AgiBot-World-Contributors et al., 2025) robot dumps sometimes keep failed trajectories; hour-maximising vendors often drop them—yet failure-then-recovery is frequently the more useful supervision. Comparing "300,000 hours delivered" (Lightwheel AI, 2026b) to "350 hours of DROID" is therefore a unit error as well as a quality error. Scenes have the same inflation problem. One thousand layouts in a single housing stock, lighting regime, and object catalogue are highly correlated; DROID's (Khazatsky et al., 2024) rule (workspace move, not object shuffle) and Ego4D's (Grauman et al., 2022) 74 worldwide locations are stricter than a vendor counting every apartment as unique. Without a shared scene predicate—and without a distance over objects, viewpoints, and dynamics—scene count is a uniqueness claim the buyer cannot audit.

The practical consequence is that scale statistics are not a sufficient statistic for policy improvement, and are incomparable marketing units. Until corpora share a task ontology, a scene-change rule, and an hour definition that excludes idle and specifies whether actions are labelled, "10,000 tasks / 500 environments / N hours" cannot be interpreted as evidence that the data is genuinely effective and useful.

6.2 Why Egocentric Data Is So Hard To Benchmark

In order to understand the challenges of benchmarking the effectiveness of egocentric data, we must first examine how large language model evaluation (LLMs) work. Fundamentally, evaluating LLMs is a simpler task because training data, model inputs, and test items occupy the same discrete space that is text tokens.

There, a tokeniser maps language into a finite alphabet that is public, stable, and shared across pretraining, post-training, and evaluation. The string the model consumes is the same kind of object as the string it emits and the same kind of object as the items on a benchmark. There is no intervening body, no camera-to-actuator map, and no need to invent a supervision variable that the files do not already contain. Likelihood on held-out text is therefore a well-defined number as it scores the model in the units in which the model was trained. Exact-match, pass@k (Chen et al., 2021), and rubric-graded generation are likewise scores of the output itself, not of a later physical process that the output is hoped to control.

Because the test distribution can be frozen as text, evaluation is cheap, repeatable, and only weakly coupled to any particular laboratory. A benchmark is a file plus a prompt template plus decoding rules. It can be versioned, mirrored, and rerun on every new checkpoint without building an environment whose dynamics drift. The same suite can be applied to a base model, a mid-trained mixture, and a filtered crawl, which is why it is fine to treat "better data" as the mixture that moves a public score. Contamination (Xu et al., 2024) complicates the reading, but it does not change the ontology of the test: the worry is that the items leaked into the training stream, not that the items live in a different physical currency from the training stream.

Even the parts of LLM evaluation that look subjective remain inside the same modality. Preference labels, tool traces, and model-as-judge protocols (Zheng et al., 2023) are still comparisons among strings. They introduce noise and circularity of their own, yet they do not require choosing an embodiment, aligning clocks across sensors, or deciding whether a predicted future remains executable once it is converted into motor commands. The evaluation object is the behaviour under test.

What it does make them is LLM evaluation structurally convenient. Train, test, and output share a representation; the target functional can be computed offline; and a data-ablation can be interpreted against a common leaderboard.

This is the convenience egocentric supervision for world-action models does not have. The files are not in the robot's problem form, the quantity of interest is a closed-loop treatment effect rather than held-out likelihood, and the states that would tell you whether the data were sufficient appear only after the policy has already begun to act.

For WAMs, the first core problem is representational. What arrives from collection is almost never what a world-action model consumes at training time. Human first-person video is not a robot observation stream and does not contain a robot action. Usefulness therefore depends on an adapter that sits outside the corpus: hand and object tracking, retargeting into joints or end-effector twists, inpainting, and pseudo-action labels inferred from vision (Wang et al., 2026a). A set that is strong as video can be weak as supervision once those maps are applied; a smaller set with accurate contact-rich trajectories can be the reverse. Any score computed on raw frames either ignores the map, and so answers the wrong question, or conditions on a particular extractor version, and so is no longer a property of the data. Camera geometry is part of the same problem. Field of view, shutter, sensor synchronisation, motion blur, and whether the wrist stays in frame changes the inferred actions, which change the policy.

The second is that the failure mode that matters is not visible in the dataset. A corpus is useful to the extent that its state coverage includes the states the policy will visit after it begins to err, and those states exist only after training and rolling out. Offline likelihood on held-out frames, caption quality, and hand–object detection can tell you whether the video is parseable. They cannot tell you whether predicted futures remain executable when turned into motor commands, or whether a small error in an action boundary produces an unsafe recovery. Contact events, grasp stability, and temporally locked multimodal annotation are the supervision that actually trains control, but frame level inspection does not see them.

The third is that there is no shared experimental skeleton on which two corpora could be compared. There is no common action variable: dumps mix unlabelled RGB, retargeted hands, end-effector twists, joint commands, and discrete skills. Choosing one representation to score against immediately privileges an embodiment. There is no common held-out world either. Simulation suites concentrate on a handful of tabletop distributions and saturate (Liu et al., 2023); rank orderings reverse when distractors, perturbations, horizon, or safety constraints are added (Fei et al., 2025); the homes and factory lines that would settle transfer are private. Mixing two clean sets can still hurt as negative-influence trajectories exist, camera conventions clash, action scales disagree, so quality is neither additive nor monotonic in volume.

What the field does instead is evaluate models, then treat the dataset as an implicit factor in that evaluation. The default public ritual is a simulation leaderboard. A policy is trained or fine-tuned and scored on LIBERO-style task suites (Liu et al., 2023), sometimes on broader arenas that add distractors, extrapolation, and long horizon. Those numbers are cheap and comparable across papers. They are also brittle. Success collapses under modest perturbation (Fei et al., 2025); binary success inflates reported capability; high visual

plausibility of a world model's predicted video does not imply that the predicted video can be executed (Li et al., 2026b). Real-robot evaluation is the next layer and is worse as a market instrument: protocols are lab-specific, trial counts are often too small for a confidence interval (Snyder et al., 2025), and the result does not travel with the dataset to the next robot.

6.3 Research Directions Into Effective Evaluation

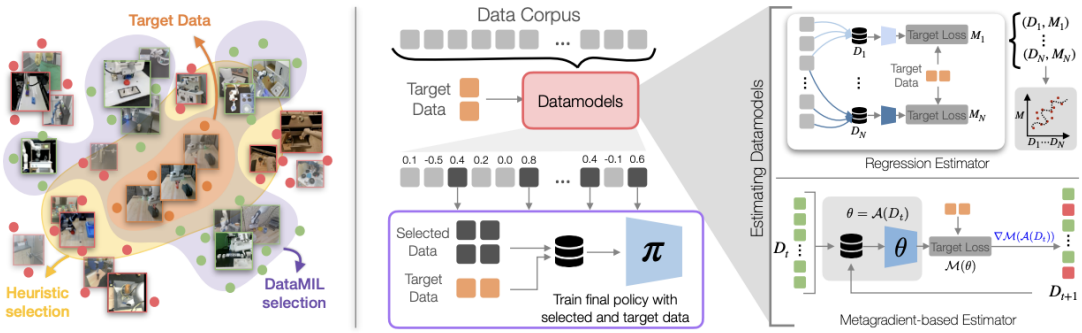

Figure 11: The DataMIL method for policy-performance-driven data evaluation (Dass et al., 2025). Rather than scoring data by heuristics, DataMIL uses datamodels to estimate the marginal effect of each candidate data point on a target policy's loss.

A promising direction is to evaluate egocentric datasets through their marginal contribution to downstream world-action models, rather than through hours, frames, or nominal task counts. Methods such as DataMIL (Dass et al., 2025) and CUPID (Agia et al., 2025) could be adapted to estimate how individual egocentric clips, activities, or dataset subsets affect closed-loop policy performance after human-to-robot conversion. This would produce a conditional measure of quality, $f(D, M, A, T) \to \Delta P$, where usefulness depends on the dataset D, world-action model M, human-to-robot adapter A, and target task distribution T. Although not fully model-agnostic, this formulation would distinguish useful supervision from redundant, unparseable, or negatively influential footage.

A second direction is to develop control-relevant coverage metrics specifically for egocentric video. Existing measures of visual, scene, or language diversity do not reveal whether footage captures the states required for successful action. Inspired by the action-sensitive coverage proposed in It's Not Just More Demos (D'urso et al., 2026), egocentric datasets could instead be scored by their coverage of contact transitions, alternative action choices, failure-and-recovery sequences, object-state changes, and long-horizon dependencies. This would distinguish datasets containing many visually diverse but behaviorally equivalent clips from datasets that expose models to consequential decision points and out-of-distribution recovery states.

Finally, another way is to benchmark the complete egocentric-to-robot conversion pipeline. H2R-Bench (Rong et al., 2026) evaluates human-to-robot manipulation-video generation, while KineBench (Liu et al., 2026) tests whether embodied world-model predictions remain kinematically grounded without tying evaluation to a particular inverse-dynamics model. EgoWAM (Li et al., 2026a) further investigates separating transferable content—objects, scenes and task semantics—from human-specific embodiment, and Human-Scale (Ma et al., 2026b) evaluates egocentric video through downstream embodied pretraining. Together, these approaches move benchmarking beyond visual realism toward physical consistency, action recoverability, controllability, and transfer to robot policies. The resulting benchmark should therefore represent egocentric-data quality as a multidimensional profile, rather than a universal scalar:

$$Q(D) = [\,\Delta\text{closed-loop success},\ \text{action-sensitive coverage},\ \text{human-to-robot transfer},\ \text{physical consistency},\ \text{recovery coverage},\ \text{utility per hour}\,].$$

Measured using shared adapters, policy families, embodiments, and closed-loop test environments, this profile would make comparisons between egocentric datasets auditable. More importantly, it would change collection incentives: datasets would be rewarded for capturing visible hands, temporally precise contact, object-state transitions, failures, corrections, and transferable action structure, not just accumulating recording hours.

7 Conclusion

Egocentric data is becoming foundational to world-action models, but its value lies not in perspective or scale alone: it lies in the causal information connecting actions to changes in the world. This survey introduced a three-tier taxonomy that distinguishes the semantic breadth of in-the-wild video, the action–state grounding of robot telemetry, and the counterfactual control of synthetic environments, showing how their complementary strengths shape different model architectures and stages of training. As raw collection becomes commoditised, progress will increasingly depend on combining these tiers through effective processing, action extraction, simulation, and curation. The field must therefore move beyond reporting hours, tasks, and scenes toward a more standardised evaluation that involves downstream transfer, physical consistency, recovery coverage, and closed-loop policy improvement.

References

Christopher Agia, Rohan Sinha, Jingyun Yang, Rika Antonova, Marco Pavone, Haruki Nishimura, Masha Itkina, and Jeannette Bohg. CUPID: Curating data your robot loves with influence functions. *arXiv preprint arXiv:2506.19121*, 2025. URL `https://arxiv.org/abs/2506.19121`.

AgiBot-World-Contributors, Qingwen Bu, Jisong Cai, Li Chen, et al. AgiBot World Colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025. URL `https://arxiv.org/abs/2503.06669`.

Brenna D. Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5):469–483, 2009.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. URL `https://arxiv.org/abs/2506.09985`.

Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2Act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024. URL `https://arxiv.org/abs/2409.16283`.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. RT-1: Robotics transformer for real-world control at scale. In *Robotics: Science and Systems*, 2023. URL `https://arxiv.org/abs/2212.06817`.

Build AI. Egocentric-100K. `https://huggingface.co/datasets/builddotai/Egocentric-100K`, 2025. Dataset, Hugging Face.

J. Chen et al. Challenges and trends in egocentric vision: A survey. *arXiv preprint arXiv:2503.15275*, 2025. URL `https://arxiv.org/abs/2503.15275`.

Jialei Chen, Kai Wang, Kang Chen, Shuaihang Chen, Feng Gao, Wenhao Tang, Zhiyuan Li, Weilin Liu, Zhuyu Yao, Boxun Li, Yuanbo Xu, and Chao Yu. Lawam: Latent world action models for efficient dynamics-aware robot policies. *arXiv preprint arXiv:2606.15768*, 2026. URL `https://arxiv.org/abs/2606.15768`.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. URL https://arxiv.org/abs/2107.03374.

Collaborative Robotics. Collaborative robotics raises $100 million in series b funding. https://www.prnewswire.com/news-releases/collaborative-robotics-raises-100-million-in-series-b-funding-302112670.html, 2024. Press release, published April 10, 2024; brings total funding to over $140 million.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. EPIC-KITCHENS: A dataset for egocentric vision in kitchens. *International Journal of Computer Vision*, 2021.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, et al. Rescaling egocentric vision. *International Journal of Computer Vision*, 2022. URL https://arxiv.org/abs/2006.13256.

Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. In *Conference on Robot Learning*, 2019.

Shivin Dass, Alaa Khaddaj, Logan Engstrom, Aleksander Mądry, Andrew Ilyas, and Roberto Martín-Martín. DataMIL: Selecting data for robot imitation learning with datamodels. *arXiv preprint arXiv:2505.09603*, 2025. URL https://arxiv.org/abs/2505.09603.

Yufan Deng and Daquan Zhou. Humannet: Scaling human-centric video learning to one million hours. *arXiv preprint arXiv:2605.06747*, 2026. URL https://arxiv.org/abs/2605.06747.

Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *arXiv preprint arXiv:2302.00111*, 2023. URL https://arxiv.org/abs/2302.00111.

Giovanni D'urso, Kaushik Roy, Nicholas Lawrance, and Brendan Tidd. It's not just more demos: Counterfactual action sensitivity coverage for data-efficient robust robot imitation. *arXiv preprint arXiv:2607.27261*, 2026. URL https://arxiv.org/abs/2607.27261.

Dyna Robotics. Dyna robotics raises $120 million to advance robotic foundation models on the path to physical artificial general intelligence. https://www.therobotreport.com/dyna-robotics-closes-120m-funding-round-to-scale-robotics-foundation-model/, 2025. Press release, published September 15, 2025.

Frederik Ebert, Yifeng Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. In *Robotics: Science and Systems Workshop*, 2021.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. LIBERO-Plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025. URL https://arxiv.org/abs/2510.13626.

Figure AI. Figure exceeds $1b in series c funding at $39b post-money valuation. https://www.figure.ai/news/series-c, 2025. Company press release, published September 16, 2025.

Generalist AI. GEN-1.5: Embodied foundation models are one-shot learners. https://generalistai.com/blog/gen-1.5, 2026. Company blog post, published August 19, 2026.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022. URL https://arxiv.org/abs/2110.07058.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025. URL https://arxiv.org/abs/2505.11709.

Wenlong Huang, Yu-Wei Chao, Arsalan Mousavian, Ming-Yu Liu, Dieter Fox, Kaichun Mo, and Li Fei-Fei. PointWorld: Scaling 3D world models for in-the-wild robotic manipulation. *arXiv preprint arXiv:2601.03782*, 2026. URL https://arxiv.org/abs/2601.03782.

Interesting Engineering. New proxie robot beats costly humanoids with AI-powered efficiency. https://interestingengineering.com/innovation/proxie-cobot-robotic-automation, 2024. Published November 25, 2024.

Abhishek Jadhav. 5 physical AI infrastructure platforms shaping robotics in 2026. https://www.therobotreport.com/5-physical-ai-infrastructure-platforms-shaping-robotics-in-2026/, 2026. The Robot Report, published July 30, 2026.

Justin Kerr, Kush Hari, Ethan Weber, Chung Min Kim, Brent Yi, Tyler Bonnen, Ken Goldberg, and Angjoo Kanazawa. Eye, robot: Learning to look to act with a BC-RL perception-action loop. In *Conference on Robot Learning*, 2025. URL https://arxiv.org/abs/2506.10968.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Kenny Lifshitz, Jesse Thomason, et al. DROID: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. AI2-THOR: An interactive 3d environment for visual AI. In *arXiv preprint arXiv:1712.05474*, 2017.

Ben Levin. Expanding our data engine for physical AI. https://scale.com/blog/physical-ai, 2025. Scale AI company blog, published September 24, 2025.

Ben Levin and Natasha Z. Dadabhoy. Scale AI and Universal Robots: Enabling physical AI for real-world deployment. https://scale.com/blog/scale-ai-universal-robots-physical-ai, 2026. Scale AI company blog, published March 16, 2026.

Baoyu Li, Xinchen Yin, Mengying Lin, Yixin Zhang, and Danfei Xu. EgoWAM: World action models beyond pixels with in-the-wild egocentric human data. *arXiv preprint arXiv:2607.08436*, 2026a. URL https://arxiv.org/abs/2607.08436.

Huiqiong Li, Jiayu Wang, Zhiting Mei, Anirudha Majumdar, Jingjing Chen, and Bin Zhu. RoboTrust-Bench: Benchmarking the trustworthiness of video world models for robotic manipulation. *arXiv preprint arXiv:2606.01600*, 2026b. URL https://arxiv.org/abs/2606.01600.

Lightwheel AI. EgoSuite-Open100K: The largest fully-annotated open egocentric human dataset. https://egosuite100k.lightwheel.ai/, 2026a. Dataset available at https://huggingface.co/collections/LightwheelAI/egosuite-open100k.

Lightwheel AI. EgoSuite: Global-scale egocentric human data collection. https://lightwheel.ai/egosuite, 2026b. Company product page. Accessed 2026-08-26.

Yihan Lin, Jiawei He, Shifeng Bao, Chen Zhao, Yang Li, Xiaobo Wang, Yan Wang, Cheng Chi, and Jing Zhang. JEPA-WAM: Learning vision-language-action policies with joint-embedding world modeling. *arXiv preprint arXiv:2608.09381*, 2026. URL https://arxiv.org/abs/2608.09381.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. *arXiv preprint arXiv:2306.03310*, 2023. URL https://arxiv.org/abs/2306.03310.

Zeyu Liu, Zhangzhe Zhu, Yang Zhang, Chenyou Fan, Chenjia Bai, and Xuelong Li. KineBench: Benchmarking embodied world models via IDM-free kinematic grounding. *arXiv preprint arXiv:2607.19875*, 2026. URL https://arxiv.org/abs/2607.19875.

Fulong Ma, Daojie Peng, Wenjun Yue, Jiahang Cao, Bintao Wang, Qiang Zhang, and Jun Ma. GeoSemWAM: Geometry- and semantic-aware world action models. *arXiv preprint arXiv:2606.03188*, 2026a. URL https://arxiv.org/abs/2606.03188.

Juncheng Ma, Jianxin Bi, Yufan Deng, Xuanran Zhai, Kewei Zhang, Ye Huang, Bo Liang, Shukai Gong, Jiankai Tu, Xiaotian Tang, et al. Humanscale: Egocentric human video can outperform real-robot data for embodied pretraining. *arXiv preprint arXiv:2606.20521*, 2026b. URL https://arxiv.org/abs/2606.20521.

NVIDIA, Yin Cui, et al. Cosmos 3: Omnimodal world models for physical AI. *arXiv preprint arXiv:2606.02800*, 2026. URL https://arxiv.org/abs/2606.02800.

Objectways. Training robots in real life with egocentric data. https://objectways.com/blog/training-robots-in-real-life-with-egocentric-data/, 2026. Blog post.

Open X-Embodiment Collaboration. Open x-embodiment: Robotic learning datasets and RT-X models. *arXiv preprint arXiv:2310.08864*, 2023.

Gunjan Paul, Senthil Palanisamy, Satpal Singh Rathore, Pratyush Kumar Patnaik, Shubhanshu Khatana, and Abhishek Anand. Ego-OSCAR: Egocentric open source stereo capture system. *arXiv preprint arXiv:2608.08285*, 2026. URL https://arxiv.org/abs/2608.08285.

Reka Labs. Cs2-10k: A large-scale dataset for action-conditioned world models from professional gameplay. https://huggingface.co/datasets/reka-labs/CS2-10k, 2026. Dataset release.

Dingyi Rong, Yue Shi, Chaofan Ma, Jiezhang Cao, Zongrui Wang, Zeyu Zhang, Yao Mu, Guangtao Zhai, and Ning Liu. H2R-Bench: Benchmarking human-to-robot manipulation video generation in world models. *arXiv preprint arXiv:2608.13049*, 2026. URL https://arxiv.org/abs/2608.13049.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied AI research. In *IEEE/CVF International Conference on Computer Vision*, 2019.

Bokui Shen, Fei Xia, Chengshu Li, Roberto Martín-Martín, Linxi Fan, Guanzhi Wang, Shyamal Buch, Claudia D'Arpino, Sanjana Srivastava, Lyne P. Tchapmi, et al. iGibson 1.0: A simulation environment for interactive tasks in large realistic scenes. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021.

David Snyder, Asher James Hancock, Apurva Badithela, Emma Dixon, Patrick Miller, Rares Andrei Ambrus, Anirudha Majumdar, Masha Itkina, and Haruki Nishimura. Is your imitation learning policy better than mine? policy comparison with near-optimal stopping. *arXiv preprint arXiv:2503.10966*, 2025. URL https://arxiv.org/abs/2503.10966.

Ana-Maria Stanciuc. Encord raises €50m to build the data layer for physical AI. https://thenextweb.com/news/encord-raises-e50m-to-build-the-data-layer-for-physical-ai, 2026. The Next Web, published February 27, 2026.

Moses Sternstein. Charts of the week: Cycles, different but the same. a16z Newsletter (Substack), 2026. URL https://www.a16z.news/p/charts-of-the-week-cycles-different. Published June 27, 2026.

Tech Startups. Norway's 1X raising $1b at $10b valuation to bring humanoid robot NEO into homes. https://techstartups.com/2025/09/24/norways-1x-raising-1b-at-10b-valuation-to-bring-humanoid-robot-neo-into-homes/, 2025. Published September 24, 2025.

TechCrunch. Robotics startup generalist reaches $3b valuation, sources say. https://techcrunch.com/2026/08/25/robotics-startup-generalist-reaches-3b-valuation-sources-say/, 2026. Published August 25, 2026.

The AI Insider. Report: Physical intelligence to raise $1b with valuation north of $11b. https://theaiinsider.tech/2026/03/30/report-physical-intelligence-to-raise-1b-with-valuation-north-of-11b/, 2026. Published March 30, 2026.

Ye Wang, Pei Lin, Xiong-Hui Chen, Haoqi Yuan, Zhixuan Liang, Yiyang Huang, Anzhe Chen, Zixing Lei, Jie Zhang, Tao Zhang, Haoyang Li, Tong Zhang, Chenxi Xiao, Ziyuan Jiao, and Qin Jin. Ego2robot: Scalable robot data synthesis from egocentric human data. *arXiv preprint arXiv:2608.02580*, 2026a. URL https://arxiv.org/abs/2608.02580.

Zhi Wang, Botao He, Kelin Yu, Seungjae Lee, Ruohan Gao, Furong Huang, and Yiannis Aloimonos. Humanego: Zero-shot robot learning from minutes of human egocentric videos. *arXiv preprint arXiv:2605.24934*, 2026b. URL https://arxiv.org/abs/2605.24934.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. *arXiv preprint arXiv:2312.13139*, 2023. URL https://arxiv.org/abs/2312.13139.

Zhuoyuan Wu, Jun Gao, et al. Egocs-400k: An egocentric gameplay dataset for world models. *arXiv preprint arXiv:2606.18180*, 2026. URL https://arxiv.org/abs/2606.18180.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*, 2024. URL https://arxiv.org/abs/2406.04244.

Shanglin Yuan, Weiheng Zhao, Xin Shi, Haoyi Jiang, Xianda Guo, Liu Liu, Wenyu Liu, Wei Sui, and Xinggang Wang. DreamWAM: Beyond RGB future prediction for world action models. *arXiv preprint arXiv:2608.04996*, 2026. URL https://arxiv.org/abs/2608.04996.

Fan Zhang and Michael Gienger. Learning robot manipulation from audio world models. *arXiv preprint arXiv:2512.08405*, 2025. URL https://arxiv.org/abs/2512.08405.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023. URL https://arxiv.org/abs/2306.05685.

Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loic Magne, Avnish Narayan, You Liang Tan, Guanzhi Wang, Qi Wang, Jiannan Xiang, Yinzhen Xu, Seonghyeon Ye, Jan Kautz, Furong Huang, Yuke Zhu, and Linxi Fan. FLARE: Robot learning with implicit world modeling. *arXiv preprint arXiv:2505.15659*, 2025. URL https://arxiv.org/abs/2505.15659.

Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning. In *arXiv preprint arXiv:2009.12293*, 2020.