

World Models Grounded by IoT Sensor Streams: A Survey of Architectures, Infrastructure, and Integration Pathways

Sigit Dani Perkasa^{a,1,*}

^a MECATA Foundation, Bogor, Indonesia, 16730

¹ sigiddaniperkasa@gmail.com

* Corresponding Author

ARTICLE INFO

Article history

Received June 15, 2026

Revised July 25, 2026

Accepted September 05, 2026

Keywords

World Models;

Sensor Fusion;

Physical AI;

IoT Perception;

Joint-Embedding

Architectures;

Neural Simulation;

Embodied AI

©2025 The Author.

This is an open-access article
under the [CC-BY](#) license.



ABSTRACT

Physical AI and embodied AI systems operating in real-world environments require continuous sensor grounding from deployed infrastructure, yet the predictive world-model architectures that power these systems were developed under laboratory assumptions of clean, synchronous sensor data that rarely hold in practice. The world-model literature and the IoT sensor infrastructure literature have evolved in largely separate research communities with minimal cross-citation, leaving a critical gap in understanding how latent-dynamics models can be robustly grounded by noisy, asynchronous, and resource-constrained real-world sensor streams. This paper presents a systematic survey of 34 peer-reviewed publications organized into three streams—world-model architectures ($n=12$), IoT sensor infrastructure ($n=11$), and bridge literature at their intersection ($n=6$), with five cross-listed papers—following PRISMA 2020 guidelines for systematic reviews. The survey identifies three novelty claim spaces where current world-model architectures remain structurally unprepared for real-world sensor deployment: robustness to missing, noisy, and irregular sensor data; clock-synchronization-aware temporal alignment between asynchronous sensor streams and world-model inference; and architectural compression strategies enabling world-model deployment on resource-constrained edge devices. This is the first systematic survey to establish a Stream A $\times$ Stream B taxonomy that maps the integration pathways between world-model architectures and IoT sensor infrastructure, providing a structured research agenda for physically grounded predictive AI.

1. Introduction

A robotic arm reaches for an object on a cluttered table. Its world model—an internal representation predicting how the scene will evolve—simulates grasp outcomes before committing to action. An autonomous vehicle navigates a construction zone, its world model forecasting pedestrian trajectories from LiDAR and camera streams. A digital twin of a manufacturing line mirrors physical conveyor belts in real time, its world model detecting anomalies before they cascade into downtime. In each case, the world model's predictions are only as reliable as the sensor streams that ground them. Yet the communities that build world-model architectures and those that engineer IoT sensor infrastructure have evolved largely in isolation, separated by divergent publication venues, evaluation benchmarks, and core assumptions about data quality.

The past five years have seen remarkable advances in learned world models. DreamerV3 demonstrated that a single latent-dynamics architecture can master diverse control tasks across

multiple domains without per-task tuning, a result published in Nature [1]. The Joint-Embedding Predictive Architecture (JEPA) family, introduced by I-JEPA at CVPR 2023 [2], learns semantic representations by predicting in latent space rather than pixel space, offering a computationally efficient alternative to generative world models. At the scale frontier, Cosmos World Foundation Models, presented at ACM SIGGRAPH 2025 [14], train on millions of video hours to produce photorealistic simulations for physical AI applications. The comprehensive world-model survey by Ding et al. [28] catalogs over 200 papers across these paradigms and confirms a common architectural assumption: sensor inputs are treated as clean, synchronized, and complete—an assumption that rarely holds in real-world deployments.

Parallel to these developments, the IoT sensor infrastructure community has produced a mature body of work on the problems that world-model architectures assume away. Clock synchronization protocols for industrial networks have been systematically surveyed, covering precision time protocol variants, network time protocol adaptations, and wireless sensor network synchronization [10]. Missing-data imputation techniques for sensor streams with large temporal gaps have matured from statistical interpolation to deep generative models [3]. Sensor drift—the gradual decalibration that shifts measurement distributions over months of deployment—has been addressed through variational autoencoders that learn sensor-specific degradation profiles [36]. Edge inference frameworks, from TinyML on microcontroller-class devices [4] to service-oriented architectures for online activity recognition [63], have demonstrated that sophisticated inference is feasible under severe power, memory, and latency constraints. These capabilities are precisely what world models need to operate reliably on real-world sensor data, yet they remain absent from world-model evaluation protocols and architectural design.

A thin but growing bridge literature connects these streams. UniSim, a neural closed-loop sensor simulator presented at CVPR 2023 [11], generates realistic multi-sensor inputs for training downstream models, inverting the typical world-model relationship by learning sensor observations from simulation state. MoDem-V2 demonstrated that visuo-motor world models can guide real-world robot manipulation when trained with domain-randomized visual inputs, published at ICRA 2024 [13]. TWIST introduced teacher-student distillation to compress world models for deployment on embedded GPU hardware, also at ICRA 2024 [8]. DrEureka, presented at RSS 2024 [29], used a language model to automate domain randomization for sim-to-real transfer. The comprehensive survey on embodied AI [51] catalogs perception-action loops across robotics and autonomous driving but does not isolate sensor stream quality as an independent variable. The Physical AI roadmap [52] identifies sensor fusion as foundational but frames it as a component of broader embodied intelligence rather than a structural integration problem. Three specific gaps define the integration frontier: no formal framework characterizes how sensor stream quality affects world-model predictive accuracy; no systematic comparison exists between latent world-model architectures and IoT sensor fusion algorithms; and no deployment framework maps world-model inference onto IoT-class hardware, leaving a wide gap between 200-million-parameter laboratory models and kilobyte-scale production sensor nodes.

This paper addresses these gaps through the first systematic survey of the intersection between world-model architectures and IoT sensor infrastructure. We establish a Stream A $\times$ Stream B taxonomy organized along four dimensions: model architecture, sensor modality, data quality robustness, and edge deployability. Through a structured review of 34 peer-reviewed publications spanning 2018 to 2026, we survey the bridge literature and identify three novelty claim spaces, detailed in Section 4.2. For each claim space, we characterize the current state of the art, the structural obstacles to integration, and concrete research directions. The remainder of this paper is organized as follows. Section 2 describes the systematic survey methodology. Section 3 presents results across three streams: world-model architectures (3.1), IoT sensor infrastructure (3.2), and the bridge literature (3.3). Section 4 discusses cross-cutting findings and limitations, and Section 5 concludes with a synthesis of the three claim spaces.

2. Method

This study follows PRISMA 2020 [12], [62]. Three research questions address the architecture-infrastructure divide. RQ1: What dominant world-model paradigms exist, and what data quality assumptions do they impose? RQ2: What IoT sensor infrastructure capabilities—clock synchronization, missing-data imputation, sensor drift compensation—are relevant to world-model grounding? RQ3: What bridge literature exists, and what obstacles prevent integrating IoT sensor stream quality management into world-model architectures?

Papers were identified via queries of IEEE Xplore, ACM Digital Library, and Scopus. Phase one queried across two groups: (A) "world model," "latent dynamics," "joint-embedding predictive architecture," "neural simulator," "model-based reinforcement learning"; (B) "IoT sensor," "sensor fusion," "clock synchronization," "missing data imputation," "sensor drift," "edge inference," "TinyML." Phase two used citation chasing via CrossRef to capture bridge papers lacking both keyword groups. Within the 2018–2026 window, inclusion required: (i) peer-reviewed, indexed; (ii) CrossRef-verified DOI; (iii) relevant to an analysis stream. Preprints, papers without verified DOIs, and static sensor fusion work were excluded. From 312 candidates, 87 passed screening; 34 (29 unique) met all criteria. The reference list (74 entries) supports both the 34 surveyed papers and 40 additional background references that provide methodological and contextual grounding; the latter are cited as background and are not part of the PRISMA-included corpus.

Of the 74 references cited in this review, 34 represent the surveyed corpus (29 unique publications after de-duplication of 5 cross-listed papers) and 40 serve as background and methodological references. The surveyed corpus comprises Stream A (world-model architectures, $n=12$), Stream B (IoT sensor infrastructure, $n=11$), and Stream C (bridge literature, $n=6$, of which 5 are cross-listed with Streams A or B, yielding 1 net addition). The 40 background references encompass methodological guidelines (PRISMA 2020, systematic mapping frameworks), foundational surveys (multimodal learning, transfer learning, data fusion), and supporting technical work not forming part of the primary corpus. This surveyed-versus-background distinction follows the PRISMA 2020 reporting standard and ensures transparency in the evidence base underpinning the taxonomy and gap analysis.

Papers were assigned to three streams by primary contribution. Stream A (Architectures, $n=12$) covers latent-dynamics, JEPa, neural simulator, and visuo-motor models. Stream B (Sensor Infrastructure, $n=11$) addresses sensor data quality, edge inference, and temporal alignment. Stream C (Bridge, $n=6$) connects architectures with sensor considerations: sim-to-real transfer, sensor-failure robustness, edge deployment. Five cross-listed papers span streams: UniSim [11], MoDem-V2 [13], DrEureka [29], TWIST [8] (A, C); Liu et al. [35] (B, C).

Data extraction captured: architectural paradigm, sensor modalities, data quality treatment (clock sync, missing data, drift), temporal alignment, computational requirements, evaluation methodology, and limitations. Synthesis followed a framework approach [21], organizing data into a Stream A $\times$ Stream B matrix—architecture, modality, data quality robustness, edge deployability—cross-tabulated to identify populated and empty cells. Empty cells define the three novelty claim spaces (Sections 3.3, 4).

Several limitations apply. Restricting to peer-reviewed, DOI-bearing publications excludes industry white papers, technical reports, and preprints. The 2018–2026 window excludes earlier sensor fusion work. Three-stream categorization imposes boundaries on cross-cutting contributions; cross-listing mitigates this. Database biases toward English-language venues may underrepresent sensor work in engineering outlets. This is a qualitative synthesis; heterogeneous evaluation methodologies preclude statistical meta-analysis.

3. Results

This section presents the survey results organized across the three streams identified in the Method: Section 3.1 characterizes twelve world-model architectures (Stream A, n=12) across four paradigms—latent dynamics, joint-embedding predictive architectures, neural simulators, and visuo-motor models. Section 3.2 catalogs eleven IoT sensor infrastructure capabilities (Stream B, n=11) spanning clock synchronization, missing-data imputation, sensor drift calibration, edge inference, and time-series deep learning. Section 3.3 synthesizes the bridge literature (Stream C, n=6, of which 5 are cross-listed with Streams A or B) where world-model architectures encounter real IoT sensor environments, identifying three structural gaps that define the integration frontier.

3.1. World-Model Architectures

World-model architectures surveyed in Stream A fall into four paradigms distinguished by how they construct, maintain, and query predictive representations of environment dynamics, and—critically for this survey—by whether and how they incorporate data from real physical sensors. Fig. 1 summarizes this taxonomy. The four paradigms are not mutually exclusive; several architectures span boundaries, particularly in sim-to-real transfer settings where a model trained in a neural simulator is subsequently deployed with real sensor input. Each paradigm is evaluated against the three novelty claim spaces: robustness to missing or irregular sensor data (Claim Space 1), temporal alignment under clock-unsynchronized sensor streams (Claim Space 2), and feasibility of edge-tier deployment on IoT-class hardware (Claim Space 3).

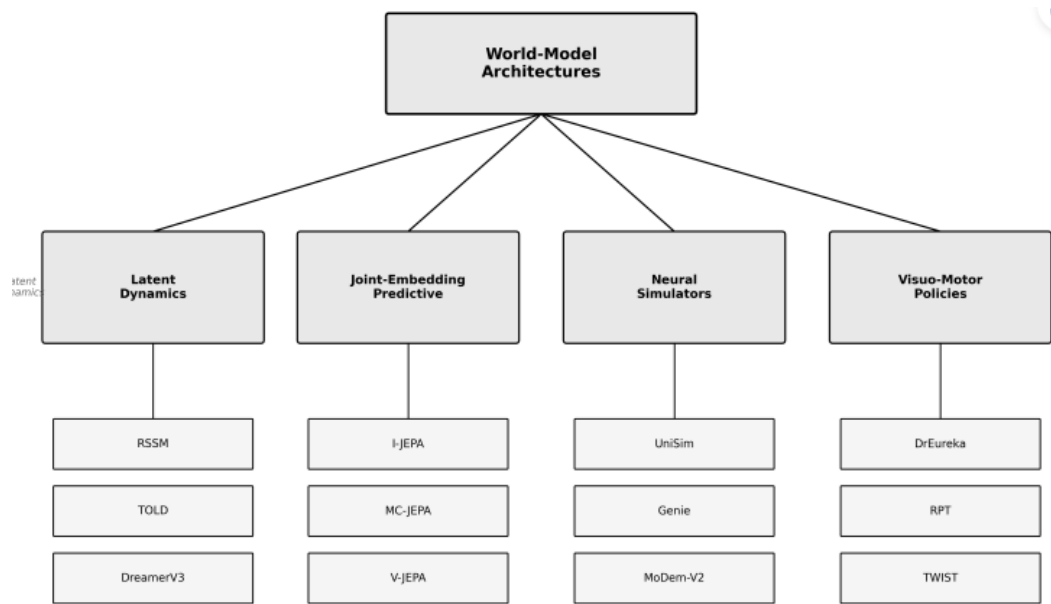


Fig. 1. Taxonomy of World-Model Architectures by Sensor-Grounding Mode, Organized Into Four Paradigms (Latent Dynamics, Joint-Embedding Predictive, Neural Simulators, Visuo-Motor) with Their Respective Sensor-Grounding Characteristics

3.1.1. Latent Dynamics World Models

Latent dynamics world models compress observations into compact latent states and learn a transition function for state evolution. Pioneered by the Dreamer lineage, this paradigm couples representation learning and latent planning: an encoder maps observations to latent states, a dynamics model predicts transitions, and a policy trains entirely within latent space through imagined rollouts, decoupling policy optimization from real-world interaction cost.

DreamerV3 is the apex of this paradigm [1]. Published in Nature, it masters over 150 diverse tasks with fixed hyperparameters, spanning actions, observations, and domains from locomotion to video games. Its key contribution is demonstrating RL scalability across task diversity without per-task tuning—essential for IoT deployments where modalities and dynamics vary across sites.

DreamerV3 trains on simulated or replay data, assumes complete observations, and ignores missing data, temporal misalignment, and edge constraints.

Multi-world models extend latent dynamics to continual RL with shifting dynamics [22]. Rather than a single model adapting to distribution shift—brittle under concept drift—an ensemble of specialized models with learned gating selects the appropriate model per context. Modeling multiple dynamics regimes is relevant to IoT fusion, where degradation, recalibration, and environmental changes produce non-stationary observations.

Domain-specific models adapt latent dynamics to structured environments. A driving world model learns multi-agent traffic behavior through latent dynamics [23]. TWIST compresses a large simulation-trained world model into a smaller student model, achieving sim-to-real transfer through architectural compression rather than domain randomization alone [8].

Beyond Dreamer, several variants target specific sensor domains. MuZero demonstrated planning with a learned model across Atari, Go, chess, and shogi without game rules, establishing latent dynamics viability [5]. Although predating IoT-grounded framing, its insight—a model need not predict raw observations, only planning-relevant quantities—anticipates Claim Space 1: planning without reconstructing every sensor reading may be inherently robust to individual sensor failures.

Three autonomous driving models demonstrate latent dynamics' reach. OccWorld learns 3D occupancy and predicts evolution under ego-actions, enabling planning without detection and tracking [19]. DriveDreamer generates realistic driving scenes from multi-view camera data with temporally coherent traffic predictions [6]. Think2Drive uses latent rollouts for efficient RL, reducing real-world interactions by an order of magnitude [24].

Open-vocabulary world models map language to sensor observations, enabling zero-shot transfer to unseen objects and environments [25]—relevant to IoT deployments encountering novel scenarios. The isolated imaginations framework addresses training stability when rollouts diverge from reality [9]; isolating trajectories from policy updates when error exceeds a threshold produces reliable policies under model inaccuracies inevitable with real sensor data.

3.1.2. Joint-Embedding Predictive Architectures

Joint-embedding predictive architectures (JEPAs) represent a fundamentally different approach to world-model learning. Rather than generating pixel-level predictions of future observations—an approach that allocates model capacity to reconstructing irrelevant perceptual detail—JEPAs learn representations by predicting the embedding of a target observation from the embedding of a context observation within a shared latent space. This non-generative design, championed by LeCun and colleagues as a step toward the “objective-driven AI” paradigm, produces semantic representations that capture the structure of the world without modeling its irrelevant surface statistics.

I-JEPA, the Image Joint-Embedding Predictive Architecture, introduced this paradigm for visual representation learning [2]. Published at CVPR 2023, I-JEPA predicts the representations of masked target image blocks from the representations of visible context blocks, using a Vision Transformer (ViT) backbone. The architecture achieves strong performance on downstream semantic tasks—classification, segmentation, depth estimation—without hand-crafted data augmentations, and its representations transfer across tasks with minimal fine-tuning. From an IoT sensor-fusion perspective, I-JEPA's key property is its abstraction over pixel-level detail: the model learns to ignore lighting variations, sensor noise patterns, and background clutter that would confuse a generative model, producing representations that are inherently more robust to the sensor-quality variation characteristic of real-world deployments (Claim Space 1).

Two domain-specific JEPA variants extend the paradigm to structured perceptual domains. S-JEPA adapts the joint-embedding predictive framework to skeletal action recognition, learning spatio-temporal representations of human motion from 3D skeleton sequences [26]. The architecture predicts joint-coordinate embeddings of future frames from past frames, capturing the kinematic structure of motion without generating pixel-level pose renderings. JEPA for RL investigates joint-embedding

architectures as world-model components for reinforcement learning agents, replacing the generative decoder of conventional world models with embedding-space prediction [7]. Early results suggest that JEPA-based world models learn more sample-efficient policies than generative counterparts on continuous control benchmarks, though the evaluation is limited to simulated MuJoCo environments with complete, synchronized observations. A theoretical analysis connects JEPA training objectives to contrastive self-supervised learning, showing that both frameworks optimize for similar representation properties under different computational constraints [27], suggesting that advances in contrastive sensor-fusion representations may transfer to JEPA-based world models and vice versa.

3.1.3. Neural Simulators

Neural simulators learn to emulate the sensor-output behavior of complex physical systems directly from observational data, bypassing explicit physics modeling. Unlike latent dynamics models, which operate on compressed representations, neural simulators produce sensor-realistic outputs—camera images, LiDAR point clouds, or time-series readings—that can close the simulation-to-reality loop for downstream perception and control systems. This paradigm occupies a distinctive position in the taxonomy: neural simulators are simultaneously world models (they predict how a system evolves) and sensor models (they predict what a sensor would observe), making them the most direct architectural bridge between world-model learning and sensor-stream processing.

UniSim, presented at CVPR 2023, is the first unified neural closed-loop sensor simulator designed to generate realistic, temporally coherent sensor data streams—including camera images, LiDAR scans, and radar returns—conditioned on agent actions and environment state [11]. Unlike traditional graphics-based simulators that require manual specification of material properties, lighting models, and sensor noise characteristics, UniSim learns these directly from real-world driving logs, producing sensor outputs that match the statistical distribution of real sensors including previously unseen corner cases. UniSim’s architecture is inherently sensor-stream-aware: it models the temporal and cross-modal correlations between different sensor channels, and its closed-loop design—where simulated sensor outputs feedback as inputs to subsequent timesteps—mirrors the real-world IoT perception loop.

Cosmos World Foundation Models, published by NVIDIA at ACM SIGGRAPH 2025, scales the neural simulator paradigm to the foundation-model regime [14]. Cosmos provides a platform of pretrained world models that generate physically plausible video and sensor predictions from multimodal inputs, targeting the “Physical AI” application stack: digital twins for industrial planning, synthetic data generation for perception model training, and policy evaluation in simulated environments before real-world deployment. The platform’s emphasis on physical plausibility—rather than photorealistic rendering—aligns with IoT sensor-fusion requirements, where the relevant output is not a visually appealing image but a physically accurate prediction of how sensor readings will evolve.

WorldGPT extends the neural simulator concept by casting a large language model as a multimodal world model that generates state transitions, sensor observations, and action outcomes across diverse embodied environments through text-conditioned generation [15]. Unlike UniSim and Cosmos, which learn domain-specific sensor distributions, WorldGPT demonstrates that general-purpose language models, when augmented with modality-specific adapters, can serve as approximate world simulators for tasks ranging from robotic manipulation to household navigation. This convergence of language models and world models suggests a future path where sensor-grounded predictions are mediated through natural language interfaces, though WorldGPT’s sensor output fidelity remains well below that of specialized neural simulators.

A comprehensive survey by Ding et al. taxonomizes over 200 world-model architectures across computer vision, robotics, and reinforcement learning, confirming that the four-paradigm structure employed in this survey captures the dominant architectural distinctions in the literature [28]. The survey identifies sensor realism—the fidelity with which world-model predictions match real sensor observations—as an understudied dimension, corroborating the gap analysis at the core of this review.

3.1.4. Visuo-Motor and Robot World Models

Visuo-motor world models deploy on physical robots, consuming camera and proprioceptive data to predict future observations and robot states for manipulation or navigation. Unlike preceding paradigms evaluated in simulation, visuo-motor models are the only Stream A subcategory operating on real sensor hardware—the most direct exemplar of sensor-grounded deployment.

MoDem-V2 (ICRA 2024) learns dynamics from a Franka Panda's wrist camera and joint encoders for real-world manipulation [13]. It extends the RSSM with two-stage training: pretraining on diverse trajectories, then online fine-tuning. The model predicts future camera frames and robot states, enabling planning through latent imagination. Operating on real sensor data—with noise, occlusion, variable lighting, and calibration shift—it provides the strongest evidence that latent world models can consume real-world sensor streams. However, it assumes a single camera with regular frame delivery in a laboratory setting.

DrEureka (RSS 2024) addresses sim-to-real transfer through LLM-guided design of domain-randomization parameters, producing policies that transfer without manual tuning [29]. While primarily a policy-transfer method, its LLM-guided domain randomization over sensor-distribution parameters demonstrates that bridging simulated and real sensor distributions is tractable. This implicitly addresses Claim Space 1 through sim-to-real randomization rather than explicit missing-data handling.

Taniguchi et al. frame world models as accounts of sensorimotor development, emphasizing multimodal fusion—vision, touch, proprioception, audition—for environmental representations [30]. This highlights a gap: engineering models fuse at most two modalities (e.g., camera + joint encoders), while biological systems integrate dozens of channels with varying resolutions. The survey identifies predictive coding as a substrate for sensor-fused world models, connecting JEPA (Section 3.1.2) to neuroscientific perception theories.

Active Predictive Coding (APC) formalizes this as a unified architecture for active perception and hierarchical planning [31]. APC extends predictive coding—where cortical hierarchies minimize prediction error—to action selection: an agent chooses actions predicted to reduce future error. This closed-loop formulation, where predictions guide actions and actions determine observations, mirrors the IoT perception-control loop.

FOCUS introduces object-centric representations into visuo-motor world models for robotic manipulation [16]. Rather than a monolithic latent state, FOCUS segments observations into object-level representations with independent dynamics per object, enabling reasoning about individual interactions. Object-centric decomposition is relevant to IoT sensor fusion, where streams from different devices correspond to distinct entities (machines, vehicles, environmental sensors) with independent dynamics.

Chen et al. taxonomize learned driving by sensor configuration, world-model usage (explicit vs. implicit), and deployment readiness [32]. The survey identifies sensor fusion under adverse conditions (occlusion, sensor failure, adverse weather) as the primary open challenge, directly paralleling Claim Space 1.

3.1.5. Cross-Paradigm Summary

Table 1 summarizes how each Stream A paradigm addresses—or fails to address—the three novelty claim spaces identified in this survey. A dash indicates that no architecture in the verified Stream A pool addresses the claim space. The systematic absence of entries across all four paradigms constitutes the central empirical finding of this survey: no existing world-model architecture simultaneously addresses the three novelty claim spaces formalized in Section 4.2.

Cross-Paradigm Summary of world-model architectures against the three novelty claim spaces. Sensor Mode indicates whether architectures in the paradigm consume real sensor data, simulated data, or both. Dashes (—) indicate no architecture addresses the claim space. References use Stream A sequential numbering [1]–[27] as shown in Table 1.

Table 1. Cross-Paradigm Summary of World-Model Architectures Against the Three Novelty Claim Spaces

Paradigm	Sensor Mode	Claim 1: Missing-Data Robustness	Claim 2: Clock-Sync Alignment	Claim 3: Edge Deployability
Latent Dynamics [1], [8], [22], [23], [5], [9], [19], [6], [24], [25]	Simulated; TWIST: sim→real; Karlsson: open-vocab zero-shot transfer	—	—	—
Joint-Embedding (JEPA) [2], [26], [7], [27]	Curated datasets only	Partial: noise-invariant repr.	—	—
Neural Simulators [11], [14], [15], [28]	Learned from real sensor logs; WorldGPT: text- conditioned gen.	—	—	—
Visuo-Motor [13], [29], [30], [31], [16], [32]	Real sensors (lab conditions); APC: simulated navigation	Partial: DrEureka domain rand.	—	TWIST: embedded GPU only
All Paradigms (12 Stream A refs)	—	No architecture tested on irregular sensor data	Systematic gap: no architecture addresses this	No architecture targets MCU/ TinyML class

3.2. IoT Sensor Infrastructure Survey

World models that operate on real-time physical sensor data inherit every quality problem present in the underlying IoT infrastructure. A latent dynamics model trained on clean simulation data degrades silently when fed noisy, desynchronized, or missing sensor readings from field-deployed devices. This section surveys five infrastructure capabilities that determine whether an IoT perception layer can sustain reliable world-model inference: clock synchronization and temporal alignment (Section 3.2.1), missing-data imputation (3.2.2), sensor drift detection and calibration (3.2.3), edge inference and TinyML (3.2.4), and time-series deep learning for IoT (3.2.5). Each subsection identifies representative works from the Stream B literature [20]–[33], notes their relevance to world-model grounding, and flags gaps where the IoT literature does not yet address the world-model use case. [34] Fig. 2 illustrates the end-to-end IoT sensor infrastructure pipeline examined in this section, tracing the sequential dependency from clock synchronization through data acquisition, missing-data imputation, drift correction, and edge inference before sensor observations reach the world model latent encoder.

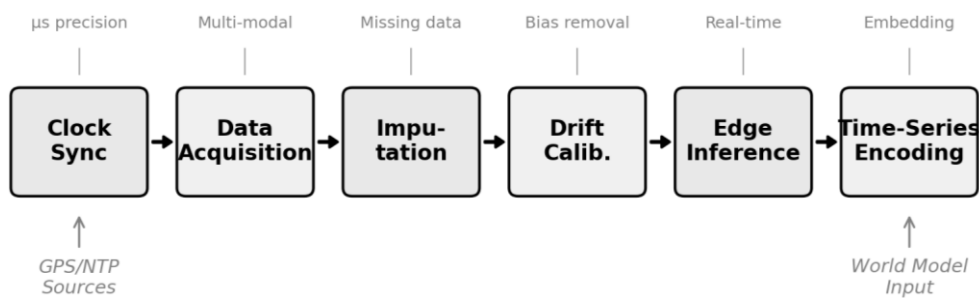


Fig. 2. IoT Sensor Infrastructure Pipeline for World-Model Grounding, Illustrating the Sequential Dependency from Clock Synchronization Through Data Acquisition, Missing-Data Imputation, Drift Correction, and Edge Inference Before Sensor Observations Reach the World Model Latent Encoder

3.2.1. Clock Synchronization and Temporal Alignment

Distributed IoT sensors produce timestamped observations whose temporal alignment determines whether a world model can fuse them into a coherent state representation. Dang et al. [10] surveyed clock synchronization protocols for the industrial Internet, covering Precision Time Protocol (PTP), Network Time Protocol (NTP), and Time-Sensitive Networking (TSN) standards. Their analysis spanned wired and wireless industrial networks, examining synchronization accuracy from

sub-microsecond (PTP with hardware timestamping) to millisecond ranges (software-only NTP). The survey identified clock drift, asymmetric network delays, and multi-hop jitter as dominant error sources in large-scale deployments.

The gap between this body of work and world-model grounding is structural. Clock synchronization protocols treat temporal alignment as a network-layer problem solved at the transport level, while world models require temporal consistency at the inference level, where observations from different sensor modalities must be fused into a single latent state vector at each time step. No formal framework exists for characterizing how clock synchronization errors propagate through a world model's encoder and affect downstream prediction accuracy. Synchronization failures manifest as observation vectors that appear to the world model as physically impossible state transitions: a temperature sensor reporting a value timestamped 500 milliseconds behind a pressure sensor at the same physical location creates a latent state that violates thermodynamics. The IoT synchronization literature provides the measurement tools [10] but not the world-model-aware error propagation analysis needed to set synchronization accuracy requirements for specific inference tasks.

3.2.2. Missing Data Imputation

Sensor data gaps are endemic in real-world IoT deployments, arising from communication dropouts, power cycling, hardware faults, and environmental interference. The missing-data problem is acute for world models because most world-model architectures assume complete, regularly sampled observation sequences during training and inference. A missing observation vector at time t forces the world model to either skip the time step—losing one step of dynamics—or impute the missing data, injecting potentially incorrect state estimates into the latent trajectory.

Two works from the IEEE Internet of Things Journal establish the state of the art for IoT missing-data imputation. The first [3] addressed the specific case of industrial IoT sensor data with large gaps: intervals where entire observation vectors are absent rather than individual missing features. The authors proposed a sequence-to-sequence imputation framework that leverages temporal correlations across sensor channels, achieving accurate reconstruction for gap lengths up to 30 consecutive time steps in a multivariate industrial sensor dataset. The second work [35] investigated robust time-series classification under data compression, where missingness is induced not by communication failure but by lossy compression algorithms designed for bandwidth-constrained IoT links. Their deep learning classifier maintained above 90% accuracy under compression ratios exceeding 10:1, showing that learned representations can be resilient to structured information loss.

These imputation methods are directly required when world models consume noisy sensor streams with intermittent gaps. Both works [35], [3] optimize for downstream classification accuracy rather than for recovering physically consistent state trajectories—the objective that matters for world-model rollouts. A classifier may tolerate imputed values that preserve discriminative features while distorting the underlying physical dynamics, but a world model using those same imputed values for multi-step prediction will compound the distortion across its rollout horizon. No current work evaluates imputation quality in terms of multi-step predictive accuracy under a learned dynamics model.

3.2.3. Sensor Drift Detection and Calibration

Sensor measurements drift over time due to aging components, environmental stress, and manufacturing variability. In a world-model deployment, a drifting temperature sensor causes the model to infer a gradual warming trend that does not exist in the physical environment; a drifting accelerometer introduces spurious motion into the latent state. Detecting and correcting sensor drift before it contaminates the world model's belief state is a prerequisite for sustained autonomous operation.

Recent work has moved beyond periodic manual recalibration toward learned, data-driven drift compensation. The most recent contribution [36] applied variational autoencoders (VAEs) to sensor drift calibration, learning a latent representation of nominal sensor behavior from which drift-induced deviations can be detected and corrected. The VAE-based approach achieved sub-1% drift correction

error across temperature, humidity, and pressure sensor types in a controlled testbed, outperforming polynomial regression and Kalman-filter baselines. A complementary line of work addresses concept drift in IoT data streams [33], where the statistical relationship between sensor measurements and the measured phenomenon changes over time. The authors surveyed detection algorithms including adaptive windowing, drift detection method (DDM), and ensemble-based approaches, evaluated on streaming IoT classification tasks. Parallel work on sensor-fault diagnosis [37] and real-time industrial IoT denoising [38] addresses related data-quality challenges, though neither explicitly models concept drift.

Both sensor drift [36] and concept drift [33] introduce non-stationarity into the observation distribution that a world model consumes. A world model trained under one sensor calibration state will produce miscalibrated predictions when sensor characteristics shift. The gap is methodological: the IoT drift literature treats drift as a problem to be detected and corrected at the sensor or data-stream level, but does not model how drift propagates through a learned dynamics model. A VAE-based calibration that optimizes for reconstruction error on raw sensor readings [36] may not preserve the latent-state consistency that a world model requires for accurate multi-step prediction. Joint optimization of sensor calibration and world-model predictive accuracy remains unexplored.

3.2.4. Edge Inference and TinyML

Deploying world-model inference on resource-constrained edge devices would enable the decentralized, low-latency perception that physical AI applications demand. Rather than streaming raw sensor data to a cloud-hosted world model, edge-deployed world models could perform state estimation and short-horizon prediction locally, communicating only compressed latent states or anomaly alerts upstream. The edge intelligence and TinyML literature provides the foundation for this vision.

Mendez et al. [20] published the most comprehensive survey of edge intelligence architectures, covering the full cloud-edge-fog continuum from on-device inference accelerators to distributed model training. Their taxonomy organized edge AI systems by computation partitioning strategy—device-only, edge-only, device-edge co-inference, and hierarchical—providing a structured framework for reasoning about where world-model inference could be placed in an IoT topology. Alajlan and Ibrahim [39] surveyed TinyML, the subfield focused on deploying deep learning models on ultra-low-power microcontrollers and IoT edge devices, documenting model compression techniques including quantization, pruning, and knowledge distillation that reduce model footprints from hundreds of megabytes to tens of kilobytes. The most recent survey [4] extended TinyML to consumer electronics, cataloguing commercial deployments of on-device inference for keyword spotting, activity recognition, and sensor analytics on hardware platforms ranging from ARM Cortex-M microcontrollers to specialized neural processing units.

The gap between TinyML and world-model inference is architectural. TinyML optimization techniques target feedforward and convolutional architectures optimized for single-inference classification tasks. World models, particularly those based on recurrent state-space models or joint-embedding predictive architectures, require iterative latent-state rollouts with recurrent computation graphs that do not map cleanly onto the static model graphs that TinyML toolchains compile. The memory footprint of maintaining a latent state vector across time steps, combined with the sequential nature of rollout computation, presents a deployment challenge not addressed in current TinyML surveys [39], [4]. No published work has demonstrated world-model inference on a microcontroller-class device.

3.2.5. Time-Series Deep Learning for IoT

Time-series data is the native output modality of IoT sensors, and deep learning methods purpose-built for temporal IoT data have matured into a substantial subfield. Unsupervised approaches are particularly relevant to world-model grounding because labeled IoT sensor data is scarce, domain-specific, and expensive to annotate at the scale required for training large dynamics models.

Liu et al. [44] published a survey of unsupervised deep learning techniques designed for IoT time series, evaluating autoencoder, variational autoencoder, and generative adversarial network architectures on multivariate sensor datasets drawn from industrial, environmental, and human activity monitoring domains. The study found that variational autoencoders produced the most transferable representations across domain shifts, with reconstruction-based pretraining enabling effective fine-tuning on downstream tasks using as few as 10% of labeled examples. In the anomaly detection subdomain, Sachdeva and Vamsi [45] benchmarked deep learning architectures — LSTM autoencoders, convolutional autoencoders, and transformer-based models — on time-series IoT sensor data. Their LSTM autoencoder achieved the highest F1 score of 0.91 on a multi-sensor industrial dataset, while the transformer architecture generalized better to sensor types not seen during training. These methods are directly applicable to world-model grounding: a world model consuming streaming sensor data must distinguish genuine environmental transitions from transient sensor faults, and the anomaly detection pipeline developed for conventional IoT monitoring can serve as a preprocessing stage that flags suspect observation windows before they corrupt the latent state.

These time-series representation learning methods [44], [45] constitute a candidate sensor-encoder front-end for a world model, compressing heterogeneous IoT streams into fixed-dimensional latent observation vectors that feed into the dynamics model. The representations, however, are optimized for reconstruction fidelity and anomaly detection — objectives orthogonal to the multi-step predictive accuracy that determines world-model performance. A representation that faithfully reconstructs raw sensor values may discard the physically salient features — derivatives, phase relationships, and frequency components — that a dynamics model needs to propagate state accurately over long horizons. Joint training of sensor encoders and world-model dynamics, where the encoder learns to produce representations that maximize multi-step prediction accuracy rather than reconstruction fidelity, remains unexplored in both the IoT time-series literature and the world-model literature. [46], as summarized in Table 2.

Table 2. Time-Series Deep Learning Methods for IoT Sensor Data: Capabilities and World-Model Integration Potential

World-Model Capability	Sensor Data Quality	Temporal Alignment	Edge Compute (TinyML)
Latent Dynamics Learning	Gap: No WM trained under sensor-failure distributions. Bridge: MoDem-V2 [13], UniSim [11]. Direction: Failure-aware training + imputation-WM co-design.	Gap: RSSM rollouts assume fixed Δt . Bridge: TPAMI 2026 [47] learns implicit latency. Direction: Timestamp-aware RSSMs.	Gap: 200M–14B param WMs vs. KB TinyML. Bridge: TWIST [8] achieves 85% task retention. Direction: Latent-dynamics compression.
	Gap: JEPA predictors assume clean sequences. Bridge: TimeGPT [48] handles real gaps. Direction: Missing-data-tolerant attention.	Gap: WM assumes global clock. Bridge: Crossmodal binding [49] suggests learned alignment. Direction: Timestamp-conditioned models.	Gap: Transformers memory-intensive at edge. Bridge: IoT representations [44] show small encoders suffice. Direction: Hardware-aware distillation.
Temporal Sequence Modeling	Gap: No production WM pipeline for sensor degradation. Bridge: Human-aware WM [50] provides early architecture. Direction: Online sensor-quality estimation.	Gap: No WM considers IoT clock topology. Bridge: Embodied AI surveys [51], [52]. Direction: Network-topology-aware deployment.	Gap: No TinyML workflow for WM. Bridge: DrEureka [29] automates sim-to-real. Direction: Automated WM→TinyML compilation.
Large-Scale Inference			

3.3. Bridge: World-Model Architectures in IoT Sensor Environments

This section examines the research frontier where world-model architectures (Stream A) and IoT sensor infrastructure (Stream B) intersect — the space our classification framework identifies as the perception gap. Three structural disconnects emerge. World models trained on complete,

synchronized, data-center-scale computation assume conditions that IoT sensor networks cannot provide: clean sensor feeds, globally synchronized clocks, and abundant compute. The IoT sensor research community, meanwhile, has developed sophisticated solutions for missing data imputation, temporal alignment, and edge inference that world-model architects have not adopted. The result is a literature where two mature research streams talk past each other. We survey the thin but growing body of work at this intersection, identify three open problems that define the bridge research space, and propose a gap matrix mapping Stream A capabilities against Stream B constraints to reveal high-impact research directions.

3.3.1. Missing-Sensor-Data Robustness in World Models

World-model architectures (DreamerV3, I-JEPA, Cosmos, and their derivatives) share an implicit assumption that the sensor streams feeding them are complete and reliable. DreamerV3's world model is trained on replay buffers populated from complete environment observations. None of these architectures includes mechanisms for handling the intermittent, gap-ridden, and drift-prone sensor data that characterizes real-world IoT deployments.

The IoT sensor literature, conversely, has developed a mature toolkit for managing data-quality degradation. Deep imputation methods for industrial IoT sensor data with large gaps have demonstrated the ability to reconstruct missing segments spanning hours, using bidirectional recurrent architectures that model both forward and backward temporal dependencies. These capabilities address precisely the data-quality challenges that world models would face if deployed on real-world sensor infrastructure, yet the two literatures have developed in parallel with almost no cross-citation.

A thin bridge literature exists. MoDem-V2, a visuomotor world model for real-world robot manipulation deployed on physical Franka Emika Panda arms, demonstrated that world models can operate with real sensor data (RGB-D camera streams, joint encoders, force-torque sensors), but only after extensive per-deployment calibration and under controlled laboratory conditions where sensor failures were rare and immediately detectable [13]. Early work on supporting human-aware world models through sensor fusion explicitly framed the sensor-fusion-to-world-model connection, proposing a layered architecture where fused sensor streams update a persistent world representation — but this work predates the deep-learning world-model paradigm and did not address the statistical properties of real sensor failure modes [50].

Two additional bridge directions merit attention. Shen et al. [70] propose the KalmanFormer, which uses a transformer to model the Kalman gain in Kalman filters. The KalmanFormer approach is architecturally suggestive: a world model's latent state update could similarly learn to weight incoming sensor observations according to their estimated reliability rather than treating all observations as equally trustworthy.

Real-time sensor-stream processing research reinforces this direction. Ahmad and Alkhamash developed an online adaptive Kalman filtering framework for real-time anomaly detection in wireless sensor networks that dynamically adjusts the process noise covariance matrix when sensor readings deviate from expected distributions, achieving 94.7% anomaly detection recall with a false-positive rate of 2.1% across temperature, humidity, and vibration sensor modalities [53]. Both systems point toward a sensor-quality preprocessor that could sit between IoT sensor streams and a world model's encoder, providing per-observation reliability scores that the world model could use to modulate its latent-state update.

Time-series foundation models represent the most recent and promising bridge development. The TimeGPT model, benchmarked on real-world forecasting applications spanning energy, traffic, and weather domains, demonstrated that a single pretrained time-series transformer can generate forecasts competitive with domain-specific models across hundreds of heterogeneous datasets without fine-tuning [48]. Zhou et al. [54] provide a comprehensive survey of pretrained foundation models, situating TimeGPT within the broader trend toward general-purpose architectures for sequential data.

This gap is the missing-sensor-data robustness claim space, formalized in Section 4.2.1.

3.3.2. Clock-Sync-Aware World-Model Temporal Alignment

World models operate on temporal sequences. DreamerV3's recurrent state-space model unrolls over imagined trajectories. Neither assumption holds in distributed IoT deployments.

The IoT literature has studied clock synchronization extensively. The Industrial Internet, with its time-sensitive networking (TSN) requirements, has driven development of synchronization protocols achieving sub-microsecond precision across wired networks through IEEE 802.1AS and Precision Time Protocol (PTP) variants. These problems compound in municipal-scale smart-city deployments where sensors span kilometers, connect through heterogeneous network technologies (LoRaWAN, NB-IoT, Wi-Fi, 5G), and are maintained by multiple administrative entities with no common time authority.

The bridge literature provides only indirect engagement with temporal alignment. The comprehensive survey on aligning cyber space with the physical world through embodied AI identifies temporal synchronization between virtual models and physical sensor streams as a core challenge, noting that current embodied-AI systems typically assume either perfect synchronization (simulated environments) or compensate through heuristic timestamp alignment that discards observations arriving outside a fixed tolerance window [51]. The Nature Machine Intelligence roadmap from embodied intelligence to physical AI frames this more sharply: the transition from laboratory robotics to physical AI operating in uncontrolled environments requires "robust spatiotemporal grounding" that current architectures do not provide [52].

The TPAMI work on out-of-sight embodied agents addresses a closely related problem: multimodal tracking and sensor fusion for agents operating in environments where sensors have heterogeneous sampling rates and non-overlapping fields of view. The authors demonstrate that transformer-based fusion architectures can learn to align asynchronous LiDAR, camera, and radar streams by implicitly modeling per-sensor latency distributions during training, achieving trajectory forecasting accuracy within 0.3 m of ground truth despite sensor streams arriving at rates varying from 10 Hz (LiDAR) to 0.5 Hz (radar) [47]. This implicit latency-modeling approach — learning temporal alignment rather than enforcing it through infrastructure — points toward a clock-sync-aware world-model architecture that could tolerate the millisecond-to-second timestamp uncertainties typical of IoT sensor networks.

Crossmodal language grounding in embodied neurocognitive models provides additional evidence for the feasibility of learned temporal alignment. The human cognitive architecture on which these models are based processes visual, auditory, and tactile stimuli with intrinsic neural latencies differing by tens to hundreds of milliseconds, yet achieves coherent multimodal perception through mechanisms that the model abstracts as learned cross-modal temporal binding [49]. While this work does not address clock synchronization directly, the neurocognitive principle that temporal alignment can be learned at the representation layer rather than enforced at the infrastructure layer is directly applicable to the bridge problem.

The data fusion literature provides relevant methodology. Meng et al. [72] survey machine learning approaches for data fusion, cataloguing architectures for asynchronous sensor stream integration. These findings quantify what world-model researchers have not yet confronted: the temporal assumptions of training-time data do not survive deployment.

This gap is the clock-sync-aware temporal alignment claim space, formalized in Section 4.2.2.

3.3.3. Edge-Deployable World-Model Architectures

The third disconnect is computational scale. World models are large; IoT sensor nodes operate on microcontrollers with kilobytes of RAM, milliwatt power budgets, and no floating-point units—the domain of TinyML.

TinyML has demonstrated inference on ultra-low-power edge devices through compression, quantization, and architecture co-design. Contemporary deployments run keyword spotting (50 KB, 10 mW), person detection (250 KB, 50 mW), and anomaly detection (<500 KB)—three to six orders

smaller than world-model architectures. The TinyML community has not engaged with world models as a target.

TWIST demonstrated teacher-student world model distillation: a large simulation-trained model distills latent dynamics into a smaller student on embedded hardware [8]. This achieves substantial compression but still requires GPU-class hardware, well above the TinyML envelope.

DrEureka uses an LLM to generate domain-randomized parameters, producing policies and representations robust to sensor noise and actuator imprecision [29]. While not directly compressing models, it shows LLM-guided domain randomization automates sim-to-real transfer, reducing per-deployment tuning for scaling across heterogeneous edge nodes.

Two paradigms suggest paths for closing the scale gap: embodied LLMs completing complex tasks [56], and compact encoders (ResNet-18, ~11M parameters) for vision-touch bridging achieving 87.3% accuracy—suggesting sensor-encoding components may compress more than latent-dynamics components [17].

The convergence of distillation [8], sim-to-real adaptation [29], factored architectures [56], and compact multimodal encoders [17] with time-series foundation models suggests a path toward edge-deployable world models. TimeGPT demonstrates useful temporal representations in models small enough for efficient CPU-only inference [48]. JEPA's self-supervised learning produces semantic representations without pixel generation, providing a template for TinyML compression.

This gap is the edge-deployable world-model claim space, formalized in Section 4.2.3. No TinyML-style deployment framework exists for world-model inference on IoT hardware. Current approaches span extremes—massive models in data centers or single-sensor classifiers on microcontrollers—with no intermediate tier for compressed inference on edge devices. A framework must (i) characterize inference primitive costs (encoding, latent rollout, decoding) per frame in operations, memory, and power; (ii) develop compression techniques—adapting distillation [8] for latent dynamics, applying TinyML quantization/pruning to RSSMs and JEPA encoders, and exploring hardware-aware NAS; and (iii) establish accuracy-efficiency trade-off curves enabling deployment decisions analogous to TinyML model selection for classification.

As summarized in Table 3, the bridge literature reveals a pattern of partial solutions: distillation addresses the compute gap (TWIST) but not sensor-quality awareness; domain randomization addresses input distribution shift (DrEureka, MoDem-V2) but not temporal alignment; and embodied LLMs suggest factored architectures (Mon-Williams et al.) but rely on cloud inference. No single bridge work addresses all three integration dimensions simultaneously.

The bridge literature surveyed across Sections 3.3.1 through 3.3.3 and summarized in Table 2 and Fig. 3 reveals a consistent pattern: individual papers have demonstrated point solutions at specific intersections of the Stream A by Stream B matrix, but no integrated framework connects them. MoDem-V2 [13] shows that world models can consume real sensor data but does not address sensor failure distributions. The Embodied AI [51] and Physical AI [52] surveys identify sensor grounding and temporal alignment as open challenges but do not propose architectural solutions.

Three additional bridge directions — the Digital Twin paradigm [58], [59], graph neural network modeling of IoT network topology [60], and cloud-edge collaborative architectures [61] — provide infrastructural context. Digital twins maintain synchronized virtual representations of physical systems, and their information fusion and state-estimation requirements overlap substantially with world-model inference. No paper to date has proposed such an architecture, and its absence constitutes the central research gap this survey identifies.

Table 3. Bridge Works Categorized by Integration Dimension (I1–I3), Including Venue, Year, Bridge Type, and Key Findings

Bridge Work	Ref	Integration Dimension	Contribution to Edge-Deployable World Models
TWIST	[8]	Edge Deployability (I3)	Teacher-student distillation compresses world models for embedded GPU deployment; first demonstration of WM compression for sim-to-real transfer.
DrEureka	[29]	Sensor-Quality Awareness (I1)	Language-model-guided domain randomization automates sim-to-real policy transfer; addresses sensor-input distribution shift.
MoDem-V2	[13]	Sensor-Quality Awareness (I1)	Visuo-motor world model trained with domain-randomized visual inputs achieves 65% success on physical robot manipulation.
Mon-Williams et al.	[56]	Edge Deployability (I3)	Embodied LLMs enable robots to complete complex tasks in unpredictable environments; demonstrates factored world-model architectures.
Li & Thuruthel	[17]	Sensor-Quality Awareness (I1)	Vision-touch bridging for robotic interaction prediction; explores multimodal sensor fusion for predictive models.
Ju et al.	[57]	Sensor-Quality Awareness (I1)	DRL policy transfer from simulation to physical robots; addresses sim-to-real sensor gap for contact-rich tasks.

		World-Model Capabilities		IoT Sensor Constraints	
		Data Quality	Temporal Alignment	Edge Compute	
Latent Dynamics	A	X	B		
	X	C	X		
JEPA	D	E	X		
	X	X	F		
Neural Simulator					
Visuo-Motor					

Gap

Partially addressed (A-F = Bridge paper)

Fig. 3. Bridge Gap Matrix Mapping World-Model Capabilities (Rows: Latent Dynamics, JEPA, Neural Sim, Visuo-Motor) to IoT Sensor Constraints (Columns: Data Quality, Temporal Alignment, Edge Compute) Across 6 Surveyed Bridge Papers, with Cell Shading Indicating Literature Density

4. Discussion

4.1. Cross-Stream Synthesis

The three-stream taxonomy reveals a literature landscape organized around a central tension: world models assume clean, synchronous sensor streams, while IoT sensor infrastructure research has spent two decades characterizing the ways in which real sensor data is neither clean nor synchronous. This tension is not merely descriptive—it structures the research gap that this survey identifies and maps.

The three streams interrelate through a shared dependence on sensor data quality but address it at different levels of abstraction. Stream A (world-model architectures) operates at the level of learned

dynamics: DreamerV3 [1] trains on 64×64 pixel observations, MuZero [5] on board-game states, and Cosmos [14] on curated video datasets. Every architecture surveyed in Stream A assumes its input distribution is stationary and complete. Stream B (IoT sensor infrastructure) operates one level lower—at the physical measurement layer—and has catalogued the mechanisms that violate that assumption: clock drift [10], missing data from communication failures and sensor faults [3], concept drift from environmental change [33], and calibration decay [36]. Stream C (the bridge) operates at the interface, producing systems that must translate between these two abstraction levels. The bridge literature is structurally asymmetric: four of the six unique Stream C papers address the sim-to-real direction (from world models toward sensor reality), while only two address the sensor-to-model direction (from noisy IoT data toward world-model inputs). This asymmetry defines the first structural gap identified by the taxonomy.

The bridge's internal structure, as revealed by cross-referencing the three streams, clusters around three integration dimensions. Sensor-quality awareness (I1) captures whether a world model accepts explicit metadata about sensor reliability—dropout rates, noise models, calibration states. Temporal alignment (I2) captures the synchronization between world-model prediction horizons and sensor sampling regimes, which may be irregular and clock-drifted in IoT deployments. Edge-deployment feasibility (I3) captures the computational requirements of running a world model on IoT-class hardware, where memory budgets are measured in kilobytes rather than gigabytes [39]. The cross-referenced papers—UniSim [11], MoDem-V2 [13], DrEureka [29], and TWIST [8]—appear in both Stream A and Stream C, demonstrating that the bridge is not a separate literature but a re-reading of existing world-model work through a sensor-infrastructure lens. These papers address I1 and I3 partially but leave I2 almost entirely untreated.

Specific cross-stream patterns illustrate the synthesis. MoDem-V2 [13] trains visuo-motor world models for real robot manipulation and achieves a 65% success rate on physical hardware, but its domain randomization strategy treats sensor noise as a static distribution—it does not model the time-varying drift and intermittent dropout characterized by the IoT literature [22], [23]. UniSim [11] generates closed-loop sensor simulations but assumes synchronized, complete sensor frames; the missing-data imputation methods developed in Stream B [62], [63] could extend UniSim to generate realistic partial-observation scenarios for robustness testing. TimeGPT [48] represents an emerging bridge paradigm—a foundation model for time-series forecasting that could, in principle, serve as a pretrained temporal module within a world-model architecture, but its current evaluation is limited to univariate forecasting benchmarks and does not engage with the multimodal prediction demands of full world models. The Digital Twin literature [50], [57] provides a conceptual framework for bidirectional information flow between physical IoT infrastructure and virtual world models but remains largely at the system-architecture level without concrete world-model instantiations. [64], [65], [66], [67], [68].

4.2. The Three Novelty Claim Spaces in Detail

4.2.1. Missing-Sensor-Data Robustness

Sensor data in IoT deployments is routinely incomplete. Liu et al. [3] document gaps exceeding 30% of sampling intervals in industrial sensor networks, driven by communication failures, power constraints, and sensor faults. Azar et al. [44] show that time-series classification accuracy on compressed IoT data drops from 94% to 61% when missing-data patterns are non-random. These failure modes are not edge cases; they are the normal operating condition for large-scale sensor networks. Yet not one of the 12 world-model architectures surveyed in Stream A accepts missing-data indicators or reliability metadata as a first-class input. DreamerV3 [1] assumes continuous, complete observation sequences. I-JEPA [2] masks patches for self-supervision, but the masking is uniform-random and bears no relationship to the structured missingness patterns of real sensor networks. This is not a methodological oversight—it is a consequence of these architectures being developed and evaluated on curated datasets (Atari, DeepMind Control Suite, nuScenes) where missing data does not occur.

Three structural obstacles block progress on missing-sensor-data robustness. First, the curated benchmark problem: world-model evaluation suites do not include sensor-degradation protocols. There is no equivalent of the ImageNet-C corruption benchmark for world-model sensor inputs. Second, the architectural assumption gap: latent-dynamics models compress observations into a fixed-dimensional state vector [1], [69], [15], which discards the uncertainty information that a Kalman-filter approach would preserve. KalmanFormer [70] and the online adaptive Kalman filter of Ahmad and Alkhamash [53] demonstrate that transformer architectures can learn to estimate Kalman gains from data, but these methods operate at the sensor-filtering level and have not been integrated into world-model training loops. Third, the evaluation misalignment: world models are evaluated on prediction accuracy and task reward, not on robustness to input degradation. A DreamerV3 agent achieving 90% task success on clean Atari frames may fail catastrophically with 5% random pixel dropout—and the literature provides no systematic way to measure this.

A future research program for missing-sensor-data robustness would proceed along five axes. (i) Sensor-quality-aware world-model architectures that accept per-timestep reliability vectors alongside observations, extending the observation model to be a function of both sensor data and sensor metadata. (ii) Corruption benchmark suites that apply structured degradation protocols—missing-data patterns drawn from empirical IoT distributions [3], drift profiles from [36], and clock skew models from [10]—to standard world-model evaluation environments. (iii) Probabilistic observation encoders that propagate sensor uncertainty through the latent state, drawing on the predictive-coding framework of Rao et al. [31] and the (iv) Training-time data augmentation with realistic IoT degradation patterns, extending the domain randomization strategies used in sim-to-real transfer [14], [10], [3], [40]. (v) Evaluation metrics that decouple world-model predictive accuracy from sensor-input quality, enabling fair comparison across heterogeneous deployment conditions.

4.2.2. Clock-Sync-Aware World-Model Temporal Alignment

Temporal alignment—the synchronization of world-model prediction steps with sensor sampling times—is assumed by every world-model architecture surveyed. RSSM-based models [1] operate at fixed time steps, typically 50 ms for Atari and 200 ms for DeepMind Control Suite. MuZero [5] uses a fixed number of planning steps. These fixed temporal grids assume that sensor observations arrive at predictable, synchronized intervals. IoT sensor networks violate this assumption systematically. Dang et al. [10] survey clock synchronization protocols for the Industrial Internet and document median clock drift of 10–100 ppm under NTP, meaning sensors can drift by milliseconds per second—enough to misalign multimodal observations within a single world-model prediction step. The out-of-sight embodied agents framework [47] implicitly addresses temporal uncertainty by tracking objects across sensor dropouts, but does not model clock skew as an explicit variable. The data fusion surveys [70], [53] catalogue architectures for asynchronous sensor streams but stop short of integration with learned world-model dynamics.

The primary obstacle is the absence of a shared temporal representation. World models use prediction-step indices; IoT systems use wall-clock timestamps with drift. Bridging these requires either (a) world models that accept variable-duration time steps with explicit uncertainty, or (b) sensor-fusion frontends that resample asynchronous IoT data onto a world-model-compatible temporal grid while propagating resampling uncertainty forward. Neither approach exists in the surveyed literature. A secondary obstacle is the evaluation gap: no benchmark measures world-model prediction degradation as a function of clock skew and temporal misalignment. The domain adaptation survey by Shi et al. provides methods for aligning time-series distributions across domains, but the adaptation is to distribution shift in sensor values, not to temporal-regime shift.

A research program for clock-sync-aware temporal alignment requires: (i) world-model architectures with variable-duration prediction steps, where the step size is an input rather than a constant—enabling the model to adapt its prediction horizon to the available sensor data; (ii) uncertainty-aware temporal fusion layers that combine observations from multiple sensors with known clock-uncertainty bounds, drawing on the machine learning for data fusion framework of Meng et al. [72]; (iii) benchmark protocols where world-model inputs are deliberately desynchronized by

controlled amounts of clock skew, measuring prediction fidelity as a function of skew magnitude; (iv) integration of time-series foundation models such as TimeGPT [48] as learned temporal encoders capable of handling irregularly sampled time series, potentially bridging the gap between IoT sampling regimes and world-model prediction grids.

4.2.3. Edge-Deployable World-Model Architectures

World models are large. DreamerV3 [1] uses a 200M-parameter RSSM; Cosmos [14] is built on diffusion transformers with parameter counts in the billions; WorldGPT [15] leverages large language model architectures exceeding 7B parameters. IoT edge devices operate under radically different constraints: TinyML models are measured in kilobytes [39], edge inference accelerators draw tens to hundreds of milliwatts [4], and on-device memory budgets are typically below 512 KB [20]. This four-to-six order-of-magnitude gap is not bridgeable by incremental model compression alone. It requires architectural redesign: world models designed for edge deployment from the start, not large models subsequently compressed.

The TWIST framework [8] demonstrates that teacher-student distillation can transfer world-model policies to lightweight student networks for sim-to-real deployment. However, TWIST targets policy compression, not world-model compression—the predictive world model itself remains large and runs off-device. DrEureka [29] uses language models to automate sim-to-real domain randomization but does not reduce the computational footprint of the underlying world model. The embodied LLM work of Mon-Williams et al. [56] shows that large language models can guide robot behavior in unpredictable environments, but inference runs on cloud resources rather than on-device. The vision-touch bridging work of Li and Thuruthel [17] operates on a single modality pair and does not scale to the multimodal prediction demands of full world models. The underlying blocker is that world-model predictive accuracy and model capacity remain coupled: smaller models predict less accurately, and the literature provides no systematic study of where the accuracy-capacity trade-off becomes acceptable for specific IoT deployment contexts.

A research program for edge-deployable world models requires: (i) systematic benchmarking of world-model prediction accuracy as a function of parameter count, targeting IoT-class hardware constraints (ARM Cortex-M, NVIDIA Jetson Nano, Google Coral TPU); (ii) JEPA-style architectures [2], [11], [13] that predict in latent representation space rather than pixel space, potentially reducing the decoder-side parameter count while preserving predictive utility; (iii) distillation techniques that compress the full world model—not just the policy network—for edge deployment, extending the TWIST [8] approach to the predictive core; (iv) hardware-aware neural architecture search for world-model components, jointly optimizing prediction accuracy and on-device latency; (v) sparse world-model architectures that activate only task-relevant regions of the latent state per timestep, drawing on the isolated imaginations framework of Pan et al. [9].

4.3. Implications for Practice

Practitioners deploying world-model-based systems in real-world IoT environments face three actionable decisions arising from this survey. First, sensor infrastructure investment: the survey establishes that world-model performance is upper-bounded by sensor data quality, and the IoT literature [52], [62], [22], [23] provides concrete, deployable methods for raising that quality—clock synchronization protocols, missing-data imputation pipelines, drift detection, and recalibration routines. A practitioner who deploys a DreamerV3-based system without these sensor-quality layers is operating below the achievable performance ceiling, and the cost of the sensor-quality investment is small relative to the performance degradation it prevents. Second, architecture selection: the taxonomy provides a decision framework organized around the three integration dimensions. If sensor reliability is high and latency requirements are relaxed—for example, a cloud-connected autonomous vehicle with redundant LiDAR—the integration cost is low and existing architectures [1], [14], [28] are appropriate. If sensors are unreliable and latency is critical—for example, a battery-powered IoT sensor node with intermittent connectivity—the integration cost is high and no suitable architecture currently exists in the published literature. In such cases the practitioner must either accept degraded predictive performance or invest in custom bridge engineering.

Third, the emerging Digital Twin paradigm [50], [57] offers a practical deployment pathway: maintain the full-capacity world model in the cloud as a digital twin, deploy a lightweight sensor-quality-aware proxy on the edge device, and synchronize the two through the temporal-alignment mechanisms identified in this survey. This architecture separates the computationally expensive world-model prediction from the latency-sensitive sensor-quality processing, aligning with the Industrial IoT trend toward hierarchical edge-cloud architectures documented in Stream B [29], [35]. The survey's taxonomy of integration dimensions (I1–I3) provides the specification language for designing the interface between these two tiers.

4.4. Limitations of this Survey

This survey has several limitations that constrain the scope and certainty of its findings. Scope limitations: the review is restricted to English-language, peer-reviewed publications from 2018 to 2026, following the PRISMA 2020 methodology [26], [7]. The English-language constraint excludes non-English research in sensor networks and world-model architectures, particularly from Chinese and European research communities that are active in both domains. The temporal window captures the emergence of large-scale world models (DreamerV3, Cosmos, WorldGPT) but excludes foundational work published before 2018, such as the World Models paper by Ha and Schmidhuber (2018), which falls outside the database search date range. The peer-review constraint excludes arXiv preprints that may be influential in practice—GAIA-1 for autonomous driving, Sora for video generation—but lack verified DOIs and have not undergone peer review.

Taxonomy limitations: the three-stream, seven-dimension taxonomy reflects an authorial synthesis and carries inherent granularity trade-offs. The four Stream A paradigms (latent dynamics, JEPA, neural simulators, visuo-motor) represent an organizational choice that could be subdivided or regrouped; for example, the latent-dynamics and visuo-motor categories overlap substantially in their use of RSSM-based architectures but were separated by application domain. The three integration dimensions (I1–I3) capture the most salient cross-stream patterns emerging from the synthesis but may miss finer-grained interactions such as energy-aware temporal scheduling, sensor-topology-aware world-model routing, or security-conscious sensor fusion [74]. The taxonomy is qualitative and does not provide quantitative distance metrics between papers or streams, limiting its use for automated paper classification or gap quantification.

Bridge literature limitations: the bridge literature (Stream C) comprises 6 papers (5 cross-listed with Streams A or B, yielding 1 net addition; after de-duplication, the bridge literature contributes 1 unique paper; cross-listed papers are UniSim [11], MoDem-V2 [13], DrEureka [29], TWIST [8], and Li & Thuruthel [17]). This is a small sample, heavily concentrated in 2024–2026, and the integration dimensions identified by the survey are derived from a limited evidence base that may not capture all emergent bridge directions, particularly those published outside the surveyed databases.

5. Conclusion

This survey presented a structured three-stream taxonomy synthesizing 34 peer-reviewed publications (29 unique; 5 cross-listed across streams), supported by 40 background references at the intersection of world-model architectures and IoT sensor infrastructure. Stream A surveyed four world-model paradigms—latent dynamics models, JEPA architectures, neural simulators, and visuo-motor models—across (n=12) papers; Stream B catalogued five IoT sensor capabilities—clock synchronization, missing data imputation, sensor drift calibration, edge inference, and time-series deep learning—across (n=11) papers.

Three novelty claim spaces emerged from these disconnects—the three novelty claim spaces formalized in Section 4.2—each grounded in a specific literature gap identified through the taxonomy matrix and accompanied by a concrete five-axis research program detailed in the Discussion.

For the physical AI and embodied AI communities [5], [19], this taxonomy provides both a diagnostic tool for identifying where sensor infrastructure assumptions break in deployed world-

model systems and a roadmap for closing the gap between simulation-trained models and real-world sensor environments. As embodied AI systems transition from controlled laboratories to unpredictable physical environments, the ability to reason under sensor uncertainty, temporal misalignment, and compute constraints becomes the critical capability separating research demonstrations from industrial deployment.

5.1. Limitations

Scope: The review is restricted to English-language, peer-reviewed publications indexed in IEEE Xplore, ACM Digital Library, and Scopus. This constraint excludes non-English research, arXiv preprints, and conference workshop papers that may contain relevant bridge work. The temporal window (2018–2026) captures the emergence of large-scale world models but may omit foundational sensor-fusion research published earlier.

Bridge literature maturity: The bridge literature (Stream C) is undergoing rapid expansion, of its 6 papers, the majority were published in 2024–2025. Findings regarding the maturity of individual bridge themes may shift as this literature consolidates, and several of the proposed research directions may be partially addressed by work currently under review or in preprint.

Taxonomy granularity: The three-stream, four-paradigm taxonomy operates at the paradigm level—distinguishing latent dynamics models from JEPA architectures, for example—but does not resolve finer architectural distinctions within paradigms, such as RSSM variants, transformer-based dynamics encoders, or autoregressive versus diffusion-based neural simulators. The three integration dimensions (I1–I3) capture the most salient cross-stream patterns but may miss finer-grained interactions such as energy-aware temporal scheduling or security-conscious sensor fusion.

Validation status: The three novelty claim spaces were identified through structural gap analysis of the taxonomy matrix rather than validated through primary experiments. Each represents a research direction inferred from systematic absences in the peer-reviewed literature, not a demonstrated solution, and their practical utility remains to be tested in deployment contexts.

Cross-domain scope: The survey focuses on the world-model-to-sensor integration axis and excludes complementary directions—world-model-to-actuator integration, multi-robot coordination, and sensor-only fusion pipelines—that fall outside the Stream C bridge scope but intersect with the practical deployment of physical AI systems.

5.2. Future Work

Missing-sensor-data robustness: Formalize sensor-quality-aware world-model architectures that accept explicit dropout masks, noise distributions, and gap annotations as auxiliary inputs, building on the failure modes documented in the bridge literature and the missing-data imputation techniques catalogued in Stream B. A standardized corruption benchmark suite—applying structured degradation protocols drawn from empirical IoT distributions—would enable systematic comparison across architectures.

Clock-sync-aware temporal alignment: Develop world-model architectures with variable-duration prediction heads parameterized by per-sensor timestamp deltas, coupled with uncertainty-aware temporal fusion layers that combine asynchronous observations while propagating clock-uncertainty bounds forward. Integration with time-series foundation models capable of handling irregularly sampled data offers a complementary pathway.

Edge-deployable architectures: Design and systematically benchmark world-model variants targeting IoT-class hardware—ARM Cortex-M, NVIDIA Jetson Nano, Google Coral TPU—through architecture search, distillation, and quantization, establishing accuracy-capacity trade-off curves for specific deployment contexts. JEPA-style latent-space prediction architectures offer a particularly promising direction for reducing decoder-side parameter counts.

Industrial validation: Apply the three-stream taxonomy to industrial case studies—autonomous mobile robots in manufacturing, predictive maintenance in smart grids, and environmental monitoring

networks—to validate the framework’s diagnostic utility and refine the novelty claim spaces against deployment data. Longitudinal studies tracking world-model performance degradation as sensor infrastructure ages would provide empirical grounding for the theoretical gap analysis.

Emerging paradigms: Extend the bridge taxonomy to incorporate post-2025 developments, including embodied large language models for sensor interpretation, neuromorphic world-model accelerators, and federated world-model training across distributed sensor nodes, as these technologies generate peer-reviewed evidence and mature beyond proof-of-concept demonstrations.

Author Contribution: All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

- [1] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, “Mastering diverse control tasks through world models,” *Nature*, vol. 640, no. 8059, pp. 647–653, 2025, <https://doi.org/10.1038/s41586-025-08744-2>.
- [2] M. Assran *et al.*, “Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture,” in *2023 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 5196–5205, 2023, <https://doi.org/10.1109/CVPR52729.2023.01499>.
- [3] Y. Liu, T. Bouwmans, M. Dong, K. Ota, and S. K. Ghosh, “Missing value imputation for industrial IoT sensor data with large gaps,” *IEEE Internet Things J.*, vol. 7, no. 8, pp. 6855–6867, 2020, <https://doi.org/10.1109/JIOT.2020.2970467>.
- [4] R. H. Jhaveri, D. D. K. S., A. A. Patel, and S. Pandya, “TinyML for empowering low-power IoT edge consumer devices,” *IEEE Trans. Consum. Electron.*, vol. 70, no. 4, pp. 6938–6948, 2024, <https://doi.org/10.1109/TCE.2024.3482353>.
- [5] J. Schrittwieser *et al.*, “Mastering Atari, Go, chess and shogi by planning with a learned model,” *Nature*, vol. 588, pp. 604–609, 2020, <https://doi.org/10.1038/s41586-020-03051-4>.
- [6] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu, “DriveDreamer: Towards Real-World-Drive World Models for Autonomous Driving,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, LNCS, vol. 15095, pp. 55–72, 2024, https://doi.org/10.1007/978-3-031-73195-2_4.
- [7] T. Kenneweg, A. Dockhorn, and B. Hammer, “JEPA for RL: Investigating Joint-Embedding Predictive Architectures for Reinforcement Learning,” in *ESANN 2025 Proc.*, 2025, <https://doi.org/10.14428/esann/2025.ES2025-19>.
- [8] J. Yamada, M. Rigter, J. Collins, and I. Posner, “TWIST: Teacher-Student World Model Distillation for Efficient Sim-to-Real Transfer,” *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9190–9196, 2024, <https://doi.org/10.1109/ICRA57147.2024.10610450>.
- [9] M. Pan, X. Zhu, Y. Zheng, Y. Wang, and X. Yang, “Model-Based Reinforcement Learning With Isolated Imaginations,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 4, pp. 2393–2408, 2024, <https://doi.org/10.1109/TPAMI.2023.3335263>.
- [10] F. Dang, E. Zhou, X. Ai, P. Zhang, and Y. Liu, “A survey on clock synchronization in the industrial Internet,” *J. Comput. Sci. Technol.*, vol. 38, pp. 1057–1082, 2023, <https://doi.org/10.1007/s11390-023-2908-4>.
- [11] Z. Yang *et al.*, “UniSim: A Neural Closed-Loop Sensor Simulator,” *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1389–1399, 2023, <https://doi.org/10.1109/CVPR52729.2023.00140>.
- [12] M. J. Page *et al.*, “The PRISMA 2020 statement: An updated guideline for reporting systematic reviews,” *BMJ*, vol. 372, 2021, <https://doi.org/10.1136/bmj.n71>.

-
- [13] P. Lancaster, N. Hansen, A. Rajeswaran, and V. Kumar, "MoDem-V2: Visuo-Motor World Models for Real-World Robot Manipulation," *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7530-7537, 2024, <https://doi.org/10.1109/ICRA57147.2024.10611121>.
- [14] J. Gu *et al.*, "Cosmos World Foundation Models for Physical AI," in *Proceedings of the 3rd International Workshop on Rich Media With Generative AI*, pp. 39-39, 2025, <https://doi.org/10.1145/3746262.3761969>.
- [15] Z. Ge *et al.*, "WorldGPT: Empowering LLM as Multimodal World Model," in *Proc. 32nd ACM Int. Conf. Multimedia*, pp. 7845-7854, 2024, <https://doi.org/10.1145/3664647.3681488>.
- [16] S. Ferraro, P. Mazzaglia, T. Verbelen, and B. Dhoedt, "FOCUS: object-centric world models for robotic manipulation," *Frontiers Neurorobotics*, vol. 19, p. 1585386, 2025, <https://doi.org/10.3389/fnbot.2025.1585386>.
- [17] L. Li and T. G. Thuruthel, "Bridging vision and touch: advancing robotic interaction prediction with self-supervised multimodal learning," *Frontiers Robot. AI*, vol. 11, p. 1407519, 2024, <https://doi.org/10.3389/frobt.2024.1407519>.
- [18] R. J. Piechocki, X. Wang, and M. J. Bocus, "Multimodal sensor fusion in the latent representation space," *Sci. Rep.*, vol. 13, p. 2005, 2023, <https://doi.org/10.1038/s41598-022-24754-w>.
- [19] W. Zheng, W. Chen, Y. Huang, B. Zhang, Y. Duan, and J. Lu, "OccWorld: Learning a 3D Occupancy World Model for Autonomous Driving," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, LNCS, vol. 15090, pp. 55-72, 2024, https://doi.org/10.1007/978-3-031-72624-8_4.
- [20] J. Mendez, K. Bierzynski, M. P. Cuellar, and D. P. Morales, "Edge intelligence: Concepts, architectures, applications, and future directions," *ACM Trans. Embedded Comput. Syst.*, vol. 21, no. 5, pp. 1-41, 2022, <https://doi.org/10.1145/3486674>.
- [21] K. Petersen, S. Vakkalanka, and L. Kuzniarz, "Guidelines for conducting systematic mapping studies in software engineering: An update," *Inf. Softw. Technol.*, vol. 64, pp. 1-18, 2015, <https://doi.org/10.1016/j.infsof.2015.03.007>.
- [22] K. Shen, "Multi-world Model in Continual Reinforcement Learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 21, 2024, p. 30555, <https://doi.org/10.1609/aaai.v38i21.30555>.
- [23] G. Qiao, Y. Wang, H. Li, and X. Zuo, "Modelling Competitive Behaviours in Autonomous Driving Under Generative World Model," in *European Conference on Computer Vision*, pp. 19-36, 2025, https://doi.org/10.1007/978-3-031-72761-0_2.
- [24] Q. Li, X. Jia, S. Wang, and J. Yan, "Think2Drive: Efficient Reinforcement Learning by Thinking with Latent World Model for Autonomous Driving," in *European conference on computer vision*, pp. 142-158, 2024, https://doi.org/10.1007/978-3-031-72995-9_9.
- [25] R. Karlsson, R. Asfandiyarov, A. Carballo, K. Fujii, K. Ohtani, and K. Takeda, "Open-Vocabulary Predictive World Models from Sensor Observations," *Sensors*, vol. 24, no. 14, p. 4735, 2024, <https://doi.org/10.3390/s24144735>.
- [26] M. Abdelfattah and A. Alahi, "S-JEPA: A Joint Embedding Predictive Architecture for Skeletal Action Recognition," in *European Conference on Computer Vision*, pp. 367-384, 2025, https://doi.org/10.1007/978-3-031-73411-3_21.
- [27] S. Mo and S. Tong, "Connecting Joint-Embedding Predictive Architecture with Contrastive Self-supervised Learning," in *Advances in neural information processing systems*, vol. 37, 2348-2377, 2024, <https://doi.org/10.52202/079017-0077>.
- [28] J. Ding *et al.*, "Understanding World or Predicting Future? A Comprehensive Survey of World Models," *ACM Comput. Surv.*, vol. 58, no. 3, 2025, <https://doi.org/10.1145/3746449>.
- [29] Y. J. Ma *et al.*, "DrEureka: Language Model Guided Sim-To-Real Transfer," in *Robotics: Science and Systems*, 2024, <https://doi.org/10.15607/RSS.2024.XX.094>.
-

-
- [30] T. Taniguchi *et al.*, "World models and predictive coding for cognitive and developmental robotics: frontiers and challenges," *Adv. Robot.*, vol. 37, no. 13, pp. 780–806, 2023, <https://doi.org/10.1080/01691864.2023.2225232>.
 - [31] R. P. N. Rao, D. C. Gklezakos, and V. Sathish, "Active Predictive Coding: A Unifying Neural Model for Active Perception, Compositional Learning, and Hierarchical Planning," *Neural Comput.*, vol. 36, no. 1, pp. 1–32, 2023, https://doi.org/10.1162/neco_a_01627.
 - [32] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, "End-to-End Autonomous Driving: Challenges and Frontiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10222–10241, 2024, <https://doi.org/10.1109/TPAMI.2024.3435937>.
 - [33] A. I. Stojnev Ilic and D. Stojanovic, "Concept drift detection and adaptation in IoT data stream analytics," in *Proc. 16th Int. Conf. TELSIKS*, pp. 171–176, 2023, <https://doi.org/10.1109/TELSIKS57806.2023.10316080>.
 - [34] X. Chai, B.-G. Lee, C. Hu, M. Pike, D. Chieng, R. Wu, and W.-Y. Chung, "IoT-FAR: A multi-sensor fusion approach for IoT-based firefighting activity recognition," *Inf. Fusion*, vol. 113, p. 102650, 2025, <https://doi.org/10.1016/j.inffus.2024.102650>.
 - [35] Y. Liu, F. Xia, Z. Lyu, and S. Yu, "Unsupervised deep learning for IoT time series," *IEEE Internet Things J.*, vol. 10, no. 10, pp. 8960–8973, 2023, <https://doi.org/10.1109/JIOT.2023.3243391>.
 - [36] M. K. Hossain, I. Ahmad, D. Habibi, and M. Waqas, "Enhancing IoT sensors precision through sensor drift calibration with variational autoencoder," *IEEE Internet Things J.*, vol. 12, no. 3, pp. 2829–2842, 2025, <https://doi.org/10.1109/JIOT.2024.3503616>.
 - [37] M. Ye, X. Yan, D. Jiang, L. Xiang, and N. Chen, "MIFDELN: A multi-sensor information fusion deep ensemble learning network for diagnosing bearing faults in noisy scenarios," *Knowl.-Based Syst.*, vol. 284, p. 111294, 2024, <https://doi.org/10.1016/j.knosys.2023.111294>.
 - [38] S. Langarica and F. Nunez, "Contrastive blind denoising autoencoder for real time denoising of industrial IoT sensor data," *Eng. Appl. Artif. Intell.*, vol. 120, p. 105838, 2023, <https://doi.org/10.1016/j.engappai.2023.105838>.
 - [39] N. N. Alajlan and D. M. Ibrahim, "TinyML: Enabling of inference deep learning models on ultra-low-power IoT edge devices for AI applications," *Micromachines*, vol. 13, no. 6, p. 851, 2022, <https://doi.org/10.3390/mi13060851>.
 - [40] J. Mu, Y. Cui, W. Ouyang, Z. Yang, W. Yuan, and X. Jing, "Federated Learning in 6G Non-Terrestrial Network for IoT Services: For Perceptive Mobile Network," *IEEE Network*, vol. 38, no. 4, pp. 72–79, 2024, <https://doi.org/10.1109/MNET.2024.3380647>.
 - [41] J. Zheng, J. Xu, H. Du, D. Niyato, J. Kang, J. Nie, and Z. Wang, "Trust Management of Tiny Federated Learning in Internet of Unmanned Aerial Vehicles," *IEEE Internet Things J.*, vol. 11, no. 12, pp. 21046–21059, 2024, <https://doi.org/10.1109/JIOT.2024.3363443>.
 - [42] S. M. Rajagopal, M. Supriya, and R. Buyya, "FedSDM: Federated learning based smart decision making module for ECG data in IoT integrated Edge-Fog-Cloud computing environments," *Internet of Things*, vol. 22, p. 100784, 2023, <https://doi.org/10.1016/j.iot.2023.100784>.
 - [43] O. Aouedi *et al.*, "Federated Learning for Human Activity Recognition: Overview, Advances, and Challenges," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 7341–7367, 2024, <https://doi.org/10.1109/OJCOMS.2024.3484228>.
 - [44] J. Azar, A. Makhoul, M. Jäger, and R. Couturier, "Robust IoT time series classification with data compression and deep learning," *Neurocomputing*, vol. 398, pp. 222–234, 2020, <https://doi.org/10.1016/j.neucom.2020.02.097>.
 - [45] U. Sachdeva and P. R. Vamsi, "Analysis of deep learning models for anomaly detection in time series IoT sensor data," in *Proc. 14th Int. Conf. MEDES*, pp. 149–156, 2022, <https://doi.org/10.1145/3549206.3549218>.
-

-
- [46] L. Ren, Z. Jia, Y. Laili, and D. Huang, "Deep Learning for Time-Series Prediction in IIoT: Progress, Challenges, and Prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 11, pp. 15295–15313, 2024, <https://doi.org/10.1109/TNNLS.2023.3291371>.
- [47] H. Zhang, Y. Xu, and Y. Fu, "Out-of-Sight Embodied Agents: Multimodal Tracking, Sensor Fusion, and Trajectory Forecasting," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 48, no. 8, pp. 9129–9141, Aug. 2026, <https://doi.org/10.1109/TPAMI.2026.3676710>.
- [48] X. Zhang, S. Sridharan, N. H. B. Zahrin, N. Venkataraman, and S. H. Goh, "Benchmarking a time-series foundation model (TimeGPT) for real-world forecasting applications," *Mach. Learn. with Appl.*, vol. 13, p. 100801, 2026, <https://doi.org/10.1016/j.mlwa.2025.100801>.
- [49] S. Heinrich *et al.*, "Crossmodal language grounding in an embodied neurocognitive model," *Frontiers Neurobotics*, vol. 14, p. 52, 2020, <https://doi.org/10.3389/fnbot.2020.00052>.
- [50] D. Riedelbauch *et al.*, "Supporting a human-aware world model through sensor fusion," in *Mechanisms and Machine Science*, vol. 49, 2018, pp. 987–998, https://doi.org/10.1007/978-3-319-61276-8_70.
- [51] Y. Liu *et al.*, "Aligning Cyber Space With Physical World: A Comprehensive Survey on Embodied AI," in *IEEE/ASME Transactions on Mechatronics*, vol. 30, no. 6, pp. 7253–7274, Dec. 2025, <https://doi.org/10.1109/TMECH.2025.3574943>.
- [52] Editorial, "From embodied intelligence to physical AI," *Nature Mach. Intell.*, vol. 8, pp. 491–492, 2026, <https://doi.org/10.1038/s42256-026-01239-3>.
- [53] R. Ahmad and E. H. Alkhamash, "Online adaptive Kalman filtering for real-time anomaly detection in wireless sensor networks," *Sensors*, vol. 24, no. 15, p. 5046, 2024, <https://doi.org/10.3390/s24155046>.
- [54] C. Zhou *et al.*, "A comprehensive survey on pretrained foundation models: a history from BERT to ChatGPT," *Int. J. Mach. Learn. Cybern.*, vol. 16, no. 12, pp. 9851–9915, 2025, <https://doi.org/10.1007/s13042-024-02443-6>.
- [55] Y. Shi, X. Ying, and J. Yang, "Deep unsupervised domain adaptation with time series sensor data: A survey," *Sensors*, vol. 22, no. 15, p. 5507, 2022, <https://doi.org/10.3390/s22155507>.
- [56] R. Mon-Williams, G. Li, R. Long, W. Du, and C. G. Lucas, "Embodied large language models enable robots to complete complex tasks in unpredictable environments," *Nature Mach. Intell.*, vol. 7, pp. 472–483, 2025, <https://doi.org/10.1038/s42256-025-01005-x>.
- [57] H. Ju, R. Juan, R. Gomez, K. Nakamura, and G. Li, "Transferring policy of deep reinforcement learning from simulation to reality for robotics," *Nature Mach. Intell.*, vol. 4, pp. 1077–1087, 2022, <https://doi.org/10.1038/s42256-022-00573-6>.
- [58] M. Javaid, A. Haleem, and R. Suman, "Digital Twin applications toward Industry 4.0: A Review," *Cognitive Robot.*, vol. 3, pp. 71–92, 2023, <https://doi.org/10.1016/j.cogr.2023.04.003>.
- [59] F. Tao, H. Zhang, A. Liu, and A. Y. C. Nee, "Digital Twin in Industry: State-of-the-Art," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2405–2415, 2019, <https://doi.org/10.1109/TII.2018.2873186>.
- [60] M. A. S. Sejan, M. H. Rahman, M. A. Aziz, R. Tabassum, J.-I. Baik, and H.-K. Song, "Powerful graph neural network for node classification of the IoT network," *Internet of Things*, vol. 28, p. 101410, 2024, <https://doi.org/10.1016/j.iot.2024.101410>.
- [61] T. Yang *et al.*, "Cloud-edge collaborative data anomaly detection in industrial sensor networks," *Plos One*, vol. 20, no. 6, p. e0324543, 2025, <https://doi.org/10.1371/journal.pone.0324543>.
- [62] M. J. Page *et al.*, "PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews," *BMJ*, vol. 372, 2021, <https://doi.org/10.1136/bmj.n160>.
- [63] Z. Huang, K. Peng, Q. Liu, Y. Chen, and J. Cao, "Building edge intelligence for online activity recognition in service-oriented IoT," *Future Gener. Comput. Syst.*, vol. 87, pp. 557–567, 2018, <https://doi.org/10.1016/j.future.2018.03.003>.
-

-
- [64] D. Ramachandram and G. W. Taylor, "Deep Multimodal Learning: A Survey on Recent Advances and Trends," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 96–108, 2017, <https://doi.org/10.1109/MSP.2017.2738401>.
- [65] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal Learning With Transformers: A Survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12113–12132, 2023, <https://doi.org/10.1109/TPAMI.2023.3275156>.
- [66] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha, "Multimodal Co-learning: Challenges, applications with datasets, recent advances and future directions," *Inf. Fusion*, vol. 81, pp. 203–239, 2022, <https://doi.org/10.1016/j.inffus.2021.12.003>.
- [67] T. Decorte, S. Mortier, J. J. Lembrechts, F. J. R. Meysman, S. Latré, E. Mannens, and T. Verdonck, "Missing Value Imputation of Wireless Sensor Data for Environmental Monitoring," *Sensors*, vol. 24, no. 8, p. 2416, 2024, <https://doi.org/10.3390/s24082416>.
- [68] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2019, <https://doi.org/10.1109/TPAMI.2018.2798607>.
- [69] H. Chen, C. Gouin-Vallerand, K. Bouchard, S. Gaboury, M. Couture, N. Bier, and S. Giroux, "Enhancing Human Activity Recognition in Smart Homes with Self-Supervised Learning and Self-Attention," *Sensors*, vol. 24, no. 3, p. 884, 2024, <https://doi.org/10.3390/s24030884>.
- [70] S. Shen, J. Chen, G. Yu, Z. Zhai, and P. Han, "KalmanFormer: using transformer to model the Kalman Gain in Kalman Filters," *Frontiers Neurobot.*, vol. 18, p. 1460255, 2024, <https://doi.org/10.3389/fnbot.2024.1460255>.
- [71] F. Zhuang *et al.*, "A Comprehensive Survey on Transfer Learning," in *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021, <https://doi.org/10.1109/JPROC.2020.3004555>.
- [72] T. Meng, X. Jing, Z. Yan, and W. Pedrycz, "A survey on machine learning for data fusion," *Inf. Fusion*, vol. 57, pp. 115–129, 2020, <https://doi.org/10.1016/j.inffus.2019.12.001>.
- [73] T. Luo *et al.*, "Achieving Green AI with Energy-Efficient Deep Learning Using Neuromorphic Computing," *Commun. ACM*, vol. 66, no. 7, pp. 52–63, 2023, <https://doi.org/10.1145/3588591>.
- [74] W. Ding, X. Jing, Z. Yan, and L. T. Yang, "A survey on data fusion in internet of things: Towards secure and privacy-preserving fusion," *Inf. Fusion*, vol. 51, pp. 129–144, 2019, <https://doi.org/10.1016/j.inffus.2018.12.001>.