Beyond Scaling: From Physics-Grounded Wireless Foundation Models to World Models for Embodied-Intelligent 6G Networks

Qingyu Shi, Zhen Gao, Shijian Gao, Yijie Mao, Jintao Zhang, Berk Canberk, and Kapal Dev

***Abstract*—Wireless foundation models support radio tasks, yet evidence for configuration transfer and system gains varies. This review evaluates eleven models across representation reuse, generalization ability, physical interpretability, adaptation efficiency, and system utility, with evidence labeled demonstrated, partial, or unassessed. It separates physics-informed design from independent consistency tests and examines physically indexed CSI representations, predictive pretraining, digital-twin memory, and action-conditioned world models. Integrated closed-loop validation remains open.**

***Index Terms*—wireless foundation models; AI-native 6G; physics-grounded learning; predictive pretraining; digital twins; wireless world models**

I. Introduction

Artificial intelligence (AI) research for sixth-generation (6G) networks spans diverse methods, technologies, and applications [1]. At the physical layer, studies investigate AI and machine learning (ML) for channel state information (CSI) feedback, beam management, and positioning [2]. Foundation models offer a possible unifying abstraction for these tasks [3], but their value depends on more than scale.

Most physical-layer learning systems remain tied to a specific task or observation format, or to a narrow range of channel and hardware configurations. Consequently, strong in-distribution performance can degrade with changes in carrier frequency, array geometry, bandwidth, mobility, or environment. LLM4WM, a large language model (LLM) framework for wireless multitasking, partly addresses this fragmentation through pretrained backbones and parameter-efficient adaptation [4]. However, wireless observations are physically indexed rather than naturally linguistic. They include CSI, channel impulse response (CIR), raw in-phase/quadrature (IQ) samples, orthogonal frequency division multiplexing (OFDM) resource grids, and beamspace responses.

Wireless foundation models should reuse pretrained representations across tasks or radio settings. Beyond this reuse, physics grounding requires testable preservation of complex-valued structure, physical coordinates, time–frequency–space correlations, and propagation relations. WiFo, a wireless foundation model for channel prediction, illustrates pretraining on heterogeneous channel data [5]. Broader studies emphasize tokenization, normalization, multimodality, and variable system dimensions [6]. Yet scaling models and datasets does not automatically recover electromagnetic transformation laws, motivating explicit, testable physics equivariance [7].

Even physically structured channel representations provide only a partial view of the wireless environment. A measured channel reflects geometry, materials, mobility, blockage, hardware states, and system actions rather than describing the wireless world in full. Digital twin channels provide a mechanism for maintaining an updateable virtual representation of these factors. ChannelLM illustrates how reconstructed environments and propagation-guided features can support environment-generalizable channel prediction [8]. Wireless world models extend this view by learning latent spatiotemporal dynamics from channel, geometry, and trajectory observations [9]. Such predictions also need to account for actions that modify propagation. For example, a reconfigurable intelligent surface (RIS) enables programmable control of electromagnetic propagation [10]. Closed-loop models should therefore predict how beam, pilot, RIS, scheduling, or mobility actions alter future observations.

Moving from channel prediction to radio control requires more than an accurate forecast. The unresolved question is whether shared representations retain useful physical structure beyond training conditions and when predictions guide actions. This review assesses reported transfer, physical tests, and control validation to examine that question. Its equations provide common notation for reviewed concepts and prospective interfaces rather than describe an implemented system. Accordingly, comparisons synthesize capabilities reported under each study's evaluation protocol; they do not establish performance superiority of the conceptual framework.

Related surveys offer complementary perspectives on physical-layer WFM task families and transfer protocols [11] and on architecture and network applications [12]. Analyses of generalization extend this discussion to analytical hardware-feasibility constraints on memory, compute and latency [13].

Qingyu Shi and Zhen Gao are with School of Information and Electronics, Beijing Institute of Technology, Beijing, 100081, China (e-mail: qys45363@gmail.com; gaozhen16@bit.edu.cn).

Shijian Gao is with The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, 511453, China (e-mail: shijiangao@hkust-gz.edu.cn).

Yijie Mao is with School of Information Science and Technology, ShanghaiTech University, Shanghai, 201210, China (e-mail: maoyj@shanghaitech.edu.cn).

Jintao Zhang is with Marinesat Network Technology Co., Ltd, 581 Huoju Avenue, Binjiang District, Hangzhou, 310053, China (e-mail: zhangjintao@marinesat.com).

Berk Canberk is with Edinburgh Napier University, Sighthill Court, Edinburgh, EH11 4BN, United Kingdom (e-mail: b.canberk@napier.ac.uk).

Kapal Dev is with Munster Technological University, Rossa Avenue, Cork, T12 P928, Ireland (e-mail: kapal.dev@ieee.org).

Corresponding author: Zhen Gao.

Broader syntheses connect off-the-shelf, wireless-native and agentic foundation models with network autonomy [14], and organize multimodal wireless intelligence around contextual understanding, prediction and control [15].

Alongside these network perspectives, digital-twin and world-model reviews already address latent dynamics and action-conditioned planning for edge intelligence [16]. A telecom world-model architectural study combines field and control/dynamics models with uncertainty-aware planning and reports a simulated network-slicing proof of concept [17]. Table I compares these complementary perspectives. This review asks which wireless capability claims are supported under the reported protocols. It separates physics incorporation from physical-consistency validation and identifies the additional evidence required for measured radio-control loops.

This review uses literature searches combining "wireless foundation models", "physical layer", "survey", "digital twin", and "world model". Table II emphasizes primary studies with reusable wireless representations, explicit pretraining or adaptation procedures, and identifiable downstream evaluations. This selection spans backbone adaptation, wireless-native representation learning, physics/environment grounding, and system optimization; foundational methods and architectural perspectives provide context.

II. Evolution and defining criteria of wireless foundation models

Wireless physical-layer learning has progressed through four overlapping stages: task-specific neural processing, adaptation of general-purpose pretrained models, wireless-native foundation modeling, and environment-interactive intelligence. The first stage established feasibility. Neural OFDM receivers recovered transmitted symbols directly, with implicit channel estimation, and were evaluated under nonlinear clipping noise [18], while CsiNet compressed and reconstructed massive multiple-input multiple-output (MIMO) CSI through a learned latent code [19]. Model-driven architectures such as DetNet further unfolded iterative detection into trainable layers [20]. These systems were effective, but narrow.

Their representations were usually bound to a fixed task, tensor shape, channel distribution, or transceiver configuration. A model trained for CSI feedback could not directly support localization, and a detector developed for one antenna or user configuration often required redesign. The underlying objective remained a conditional mapping,

$$f_\theta^{(\tau,c)} : \mathcal{X}_{\tau,c} \to \mathcal{Y}_{\tau,c}, \tag{1}$$

where θ denotes model parameters, τ the task, c the radio configuration, and $\mathcal{X}_{\tau,c}$ and $\mathcal{Y}_{\tau,c}$ the corresponding input and output spaces. This specialization conflicts with AI-native 6G, which requires rapid reuse across tasks and changing radio conditions.

The second stage reused pretrained language or vision backbones. LLM4CP showed that a mostly frozen Generative Pretrained Transformer 2 (GPT-2) backbone could be adapted to channel prediction through wireless preprocessing and lightweight trainable modules [21]; LLM4WM subsequently extended this route to multiple channel-associated tasks [4]. The main distinction from wireless-native models lies less in the use of a "language tokenizer" than in how wireless structure enters the backbone. LLM4WM, for example, flattens spatial–frequency channel features for channel reconstruction tasks and projects L input tokens into a GPT-2 feature space of width 768, leaving much of the structural alignment to task-specific preprocessing and adapters [4]. WiFo instead tokenizes the channel itself in space, time, and frequency. For a CSI tensor $\mathbf{H} \in \mathbb{C}^{T\times K\times N}$, T, K, N count time samples, frequency samples, and antennas; t, k, n are the corresponding patch dimensions. Non-overlapping patches give $L = (T/t)(K/k)(N/n)$ tokens when the dimensions are divisible by the patch sizes. As an illustrative example, with the reported $4 \times 4 \times 4$ patches, a $16 \times 64 \times 8$ tensor contains 8192 raw indexed entries but only 128 structured patches. For the random-masking pretraining objective, an 85% masking ratio corresponds to $128 \times 0.15 = 19.2$ visible patches before integer discretization, or approximately 19 patches in practice depending on the masking implementation [5]. This reduction in token count does not by itself establish lower computational cost, latency, or physical interpretability; rather, it illustrates how structured tokenization reduces the encoder input length while preserving explicit space–time–frequency organization.

Ying et al. connect specialist semantic-aware MIMO designs for user activity detection, CSI feedback, and source–channel precoding to a proposed evolution toward shared large models [22]. This perspective motivates distinguishing task-level gains from independently demonstrated backbone reuse and transfer.

In this review, a wireless foundation model (WFM) denotes a model pretrained on diverse data whose shared representations support reuse across wireless tasks or radio settings [3]–[5]. Its defining evidence comprises a documented pretraining corpus, an identifiable shared backbone, and transfer evaluated with stated adaptation procedures. Reuse within a task family and reuse across heterogeneous tasks represent different scopes of capability. Wireless-native foundation models learn from wireless measurements or simulated channels. Physics-grounded wireless foundation models incorporate physical knowledge through identifiable mechanisms. These properties can overlap. WiFo supports time- and frequency-domain channel prediction [5]. The Large Wireless Model (LWM) provides task-agnostic channel embeddings, while WavesFM reuses representations for sensing, communication, and localization [23], [24]. Digital twins and wireless world models add geometry, scene context, or trajectories to represent environmental causes and latent dynamics [8], [9].

Scale requires cautious interpretation. WiFo uses 160K heterogeneous CSI samples, and LWM pretrains on over one million channels from 15 scenarios [5], [23], but large corpora do not establish a wireless scaling law. A controlled study of representative non-terrestrial network (NTN) channels reports rapid gains at small scales and diminishing returns beyond approximately 30 million parameters [25]. Scaling from 12M to 96M yields about 0.8 dB lower downstream normalized mean-square error (NMSE) at a signal-to-noise ratio of 20 dB; scaling from 96M to 150M adds only 0.52 dB in the

TABLE I
SCOPE OF SELECTED RELATED REVIEWS AND ARCHITECTURAL STUDIES. YEARS REFER TO THE CITED PUBLICATIONS OR PREPRINTS; INSPECTED VERSIONS ARE SPECIFIED IN THE REFERENCES.

Study (author, year)	Primary scope	Distinct analytical focus	Complementary lens of this review
Wireless foundation model surveys			
Pacheco Huachaca et al. (2026) [11]	Physical-layer wireless foundation models	PHY task families; model-level adaptation and in-/out-of-distribution transfer protocols.	Joint capability evidence linking transfer, physical tests and control-validation boundaries.
Khan et al. (2026) [12]	WFM design and applications across 6G	Architecture, pretraining and deployment-layer taxonomies; benchmarks and physical interpretability.	Physical indexing, concrete consistency tests and specified radio-configuration shifts.
Cheraghinia et al. (2026) [13]	Physical-layer FMs from learning to deployment	Five-axis design/generalization taxonomy; analytical mapping of memory, compute and latency constraints.	Protocol-aware five-dimensional assessments and separate physical-validation tests.
Network intelligence and control			
Liang et al. (2026) [14]	PHY intelligence and network autonomy	Off-the-shelf, wireless-native and agentic paradigms from PHY processing to orchestration.	Backbone reuse and radio-configuration transfer with explicit adaptation budgets.
Zhang et al. (2026) [15]	Multimodal prediction and control in wireless networks	Multimodal context understanding; prediction/control taxonomy; datasets and wireless-specific training.	Preservation of physical coordinates and propagation relations under configuration shifts.
Digital twins and world models			
Zheng et al. (2026) [16]	Digital twins and world models for edge intelligence	Digital-twin comparison; latent dynamics and action-conditioned planning for edge applications.	Linking pretrained PHY representations, environment memory and radio-action validation.
Zou et al. (2026) [17]	Telecom world-model architecture with a slicing proof of concept	Field and action-conditioned dynamics models; uncertainty-aware planning; simulated slicing validation.	Wireless capability evidence and requirements for measured physical control loops.

reconstruction metric [25]. The study relates saturation to the channels' comparatively low intrinsic dimensionality. Evidence on data scaling, training compute, and energy remains sparse and not yet comparable across studies. Scale thus matters, but alone cannot guarantee physical generalization or deployment utility.

Fig. 1a summarizes this progression from task-specific models through backbone adaptation and wireless-native models to environment-interactive intelligence. Scale, self-supervised pretraining, and multiple task heads alone do not establish foundational capability. Five complementary dimensions guide the assessment (Fig. 1b).

Representation reuse evaluates whether the same pretrained backbone supports multiple downstream tasks or related task variants, with the breadth of reuse (within a task family versus across heterogeneous tasks) and the required adaptation explicitly reported [4], [23], [24], [26]. Generalization ability tests zero-shot transfer to deliberately held-out system settings, such as antenna configurations, bandwidths, carrier frequencies, or mobility regimes [5], [7], [27], [28]. Physical interpretability requires explicit tests of physically meaningful relations, including transformation equivariance, CSI–CIR consistency, geometric consistency, or output feasibility [7], [28], [29]. Adaptation efficiency measures the target data, trainable parameters, and optimization effort required under domain shift [4], [23]–[26]. Finally, system utility evaluates whether representation gains translate into operational benefits such as reduced overhead, higher achievable rate, lower latency, or improved reliability [8], [27], [28].

For channel prediction, generalization ability tests whether prediction remains accurate at a carrier frequency excluded from training [5]. Physical interpretability instead tests whether a frequency-domain phase ramp applied to the input CSI induces the corresponding transformation in the predicted CSI [7]. For the same carrier shift, adaptation efficiency quantifies the target CSI samples, trainable parameters, and optimization effort needed to attain a stated prediction accuracy. A zero-shot route has no parameter updates, but target-data preprocessing costs still require separate reporting. These dimensions assess transfer across configurations, consistency with a specified physical relation, and adaptation cost, respectively.

Representation reuse establishes the shared backbone's transfer capability; the other dimensions assess its breadth, physical consistency, adaptation cost, and deployment value. Transfer claims should specify training–test separation, adaptation budgets, and task-specific baselines under matched evaluation conditions. These dimensions provide complementary evidence rather than universal pass–fail criteria: physical interpretability and system utility need separate validation, and untested dimensions remain unassessed. No common numerical thresholds are imposed. Table II summarizes representative evidence.

Table II codes evidence against the five dimensions. D requires direct testing within the stated scope; P denotes related or incomplete assessment; U denotes no reported assessment, not failure. Adapted or environment-only transfer is P for generalization ability. Design priors alone are U for physical interpretability. Incomplete adaptation-budget assessment is P for adaptation efficiency. Inference timing alone is P for system utility. Simulation-based D does not imply field validation.

Three distinctions emerge. First, LWM tests frozen repre-

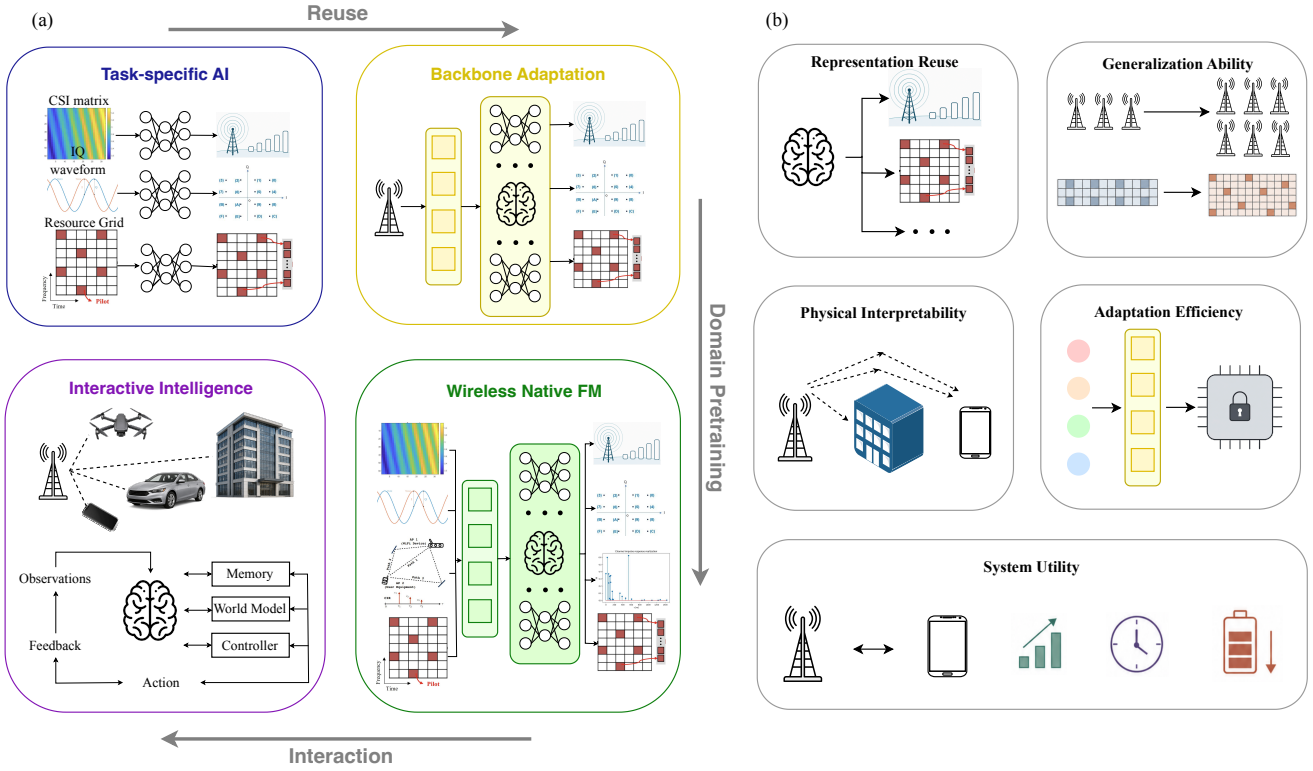


Fig. 1. **Evolution and defining criteria of wireless foundation models. (a)** Conceptual progression from isolated task-specific physical-layer models to general-purpose backbone adaptation, wireless-native foundation models, and environment-interactive intelligence. **(b)** Five complementary evidence dimensions for assessing foundation model capability in wireless systems: representation reuse, generalization ability, physical interpretability, adaptation efficiency, and system utility. AI, artificial intelligence; CSI, channel state information; IQ, in-phase/quadrature; FM, foundation model. Original diagram created by the authors using diagrams.net (draw.io).

sentations with labeled downstream training in cities absent from pretraining, whereas WiFo predicts channels at held-out carrier frequencies without fine-tuning [5], [23]. Different supervision and shifted factors preclude a single generalization ranking. Second, phys-WFM's phase-ramp equivariance and cross-dataset prediction support the tested transformation class without establishing robustness to arbitrary hardware changes or propagation mismatch [7]. Third, WWM's measured CSI prediction prototype and the hierarchical WFM's simulated optimization and runtime gains represent distinct validation stages [9], [28]; neither establishes general action-conditioned control in a physical network. These results support reusable, structured representations. Deployment readiness additionally requires evaluating sensing, action selection, and physical execution under a stated adaptation budget.

The five dimensions assess reported evidence rather than prescribe a model architecture. Table II documents direct backbone-reuse evidence for all eleven models, while physical tests and system gains are more unevenly assessed. This disparity motivates examining which physical coordinates and relations are retained when wireless observations enter a shared representation.

III. Physics-Grounded Wireless Representations

Physics grounding refers here to the identifiable mechanisms through which physical knowledge enters a model, while physical interpretability requires separate evidence that the resulting representations or predictions satisfy specified physical relations. Four complementary, nonhierarchical incorporation pathways are: (i) physics-aware inputs or preprocessing preserving complex measurements, coordinates, calibration variables, or meaningful transforms; (ii) physics-structured architectures encoding symmetries, equivariances, locality, correlation ranges, or geometry; (iii) physics-constrained objectives imposing Fourier, coordinate, propagation, or feasibility relations; and (iv) physics-informed supervision from measurements, ray tracing, electromagnetic simulation, or calibrated digital twins. These pathways are distinguished from explicit physical validation, which tests transformation consistency, output feasibility, geometric or cross-representation consistency, or agreement with measured behavior. No incorporation pathway alone guarantees physical generalization.

In Fig. 2, physical binding maps radio observations, environment and memory, and candidate actions into typed representations sharing a reusable latent space.

Training on simulated or ray-traced channels alone does not establish physical interpretability: models may learn dataset-specific correlations without transformation laws. Likewise, real–imaginary decomposition preserves channel information but does not ensure phase equivariance or robustness to physical phase transformations.

TABLE II

EVIDENCE ACROSS FIVE EVALUATION DIMENSIONS FOR REPRESENTATIVE WIRELESS FOUNDATION MODELS. EACH ASSESSMENT CELL SUMMARIZES REPORTED EVIDENCE AND ENDS WITH [D] FOR DIRECTLY DEMONSTRATED CAPABILITY UNDER THE STATED PROTOCOL, [P] FOR PARTIAL OR INDIRECT EVIDENCE, OR [U] FOR UNASSESSED CAPABILITY. NUMERICAL RESULTS RETAIN THEIR ORIGINAL PROTOCOLS AND ARE NOT DIRECTLY COMPARABLE. SIMULATION-BASED EVIDENCE DOES NOT IMPLY FIELD DEPLOYMENT READINESS.

Model	Representation Reuse (R)	Generalization Ability (G)	Physical Interpretability (I)	Adaptation Efficiency (A)	System Utility (S)
General-purpose Backbone Adaptation					
LLM4WM [4]	Shared GPT-2/MoE-LoRA supports six jointly adapted tasks with task-specific heads. **[D]**	UMa-to-RMa and 1.9-to-2.4-GHz transfer use adaptation; 10% target data stated for RMa. **[P]**	Angle preprocessing provides a design prior; physical-relation tests are unreported. **[U]**	Data/parameter efficiency: 10% RMa data; 1.13M/ 88.71M per-task trainable/ total averages; target optimization budget incomplete. **[D]**	Simulated SE at 10-dB SNR: CSI 6.068–6.081; beam selection 8.557 bit/ s/Hz. **[D]**
Wireless-native Foundation Models for Representation and Optimization					
WiFo [5]	One pretrained encoder-decoder predicts time/frequency channels across 16 training configurations. **[D]**	Three zero-shot holdouts: 3.5/6.7/28 GHz; D19 also introduces 0.25-ms sampling. **[D]**	STF encoding is ablated; independent physical-consistency tests are unreported. **[U]**	Zero parameter updates; dataset-specific CSI standardization uses an unspecified statistics-data budget. **[D]**	WiFo-base: 9.659-ms D1 inference (batch eight; RTX4090 machine), 1.456 GFLOPs; communication benefits untested. **[P]**
LWM [23]	Shared pretrained embeddings support beam prediction and LoS/NLoS classification heads. **[D]**	Unseen-city transfer uses labeled heads or Denver-specific fine-tuning; zero-shot configuration tests are unreported. **[P]**	Attention visualizations provide no independent physical-relation or feasibility test. **[U]**	Label-efficiency curves with frozen embeddings/ 500K beam head; fine-tuning parameter/ optimization budgets incomplete. **[D]**	Beam/LoS classification evaluated; beam gain, rate and operational overhead unassessed. **[U]**
WavesFM [24]	Shared ViT supports sensing, RF classification, positioning and channel estimation with adaptation. **[D]**	Cross-dataset transfer uses labeled adaptation; unseen radio configurations lack controlled zero-shot tests. **[P]**	Wireless grids and SNR-weighted training lack independent physical-relation validation. **[U]**	Parameter efficiency: RF LoRA adds about 1.5M parameters; target-label budgets/splits unspecified. **[D]**	Task metrics reported; operational rate, reliability and communication overhead unassessed. **[U]**
IQFM [26]	Joint-SSL encoder supports four tasks; task-specific SSL maxima use separate encoders. **[D]**	Unseen datasets/tasks require labeled probing or LoRA; zero-shot configuration transfer untested. **[P]**	Phase-aware augmentations are ablated; model equivariance and Fourier consistency untested. **[U]**	OOD adaptation: 1–500 labels/class, 100 epochs, approximately 84K convolutional LoRA parameters, excluding task heads. **[D]**	Over-the-air task accuracy measured; beam gain, rate and operational overhead unassessed. **[U]**
CSI-CLIP [29]	Shared contrastive CSI encoder supports positioning, beam-index prediction and channel identification heads. **[D]**	New-environment Sionna RT transfer retains identical configurations; zero-shot configuration extrapolation unestablished. **[P]**	CSI–CIR contrastive alignment is a training objective; independent Fourier-consistency/ feasibility tests unreported. **[U]**	Labeled fine-tuning uses batch 128/up to 300 epochs; trainable counts/ target-specific stopping budgets unreported. **[P]**	Positioning/beam-index metrics reported; received beam gain and operational overhead unassessed. **[U]**

Wireless foundation models require representations that preserve how observations are generated, indexed, and transformed by the physical channel. A wireless sample should therefore be expressed as

$$\mathbf{x}_i = (v_i, \mathbf{p}_i, \mathbf{c}_i, t_i)\,, \tag{2}$$

where i indexes observations and v_i denotes a real- or complex-valued measurement, $\mathbf{p}_i$ contains its time, frequency, antenna, delay, Doppler, beam, or spatial coordinates, $\mathbf{c}_i$ specifies the transceiver and waveform configuration, and t_i records acquisition time. The coordinate is part of the observation. Flattening such samples into an ordinal sequence may simplify implementation, but it can erase the relationships needed for transfer across bandwidths, arrays, and mobility regimes.

For measured radio channels, physical binding must retain measurement coordinates, radio configuration, and acquisition time so that propagation relations can be tested. Semantic tokens instead serve content transmission and recovery: TokCom exploits cross-modal context, while ToDMA uses token-associated CSI under a slow-fading assumption for stream assignment [31], [32]. ToDMA also uses physical-layer candidates to guide masked-token prediction. Successful semantic reconstruction alone does not establish that a representation preserves antenna geometry, frequency coordinates, or propagation relations.

Signal-level representations should retain the structure of their acquisition domain. CSI is a complex transfer function indexed over antennas, subcarriers, and OFDM symbols, whereas CIR describes the corresponding delay-domain response; their Fourier relation motivates paired views for representation learning. CSI-CLIP aligns their embeddings using a contrastive loss, without explicitly enforcing Fourier consistency between reconstructed outputs [29]. For models reconstructing both CSI and CIR, a consistency test compares independently predicted CSI with the Fourier transform of predicted CIR on matched sampling grids. Phase-ramp equivariance can be quantified by comparing predictions from transformed CSI with correspondingly transformed original predictions [7]. A geometric-consistency test varies the transmitter–receiver distance and checks whether the predicted absolute

TABLE II
EVIDENCE ACROSS FIVE EVALUATION DIMENSIONS FOR REPRESENTATIVE WIRELESS FOUNDATION MODELS (CONTINUED).

Model	Representation Reuse (R)	Generalization Ability (G)	Physical Interpretability (I)	Adaptation Efficiency (A)	System Utility (S)
Wireless-native Foundation Models for Representation and Optimization (continued)					
Hierarchical WFM [28]	Shared frozen encoder supports estimation, beamforming, scheduling and beam selection modules. **[D]**	Zero-shot tests cover 27 unseen user/subcarrier/ antenna combinations across 30 held-out simulated scenes. **[D]**	Simulated per-user rate-feasibility tests report 98.7% in-domain QoS satisfaction. **[D]**	Frozen-backbone adaptation reported; trainable counts, optimization budgets and hardware incomplete. **[P]**	Simulated beamforming reaches 98.5% of WMMSE rate; decoder timing excludes shared encoder. **[D]**
Physics- and Environment-grounded Models					
phys-WFM [7]	Shared MAE covers three acquisition tasks; reported RENEW results cover completion/frequency extrapolation. **[D]**	Unseen 3GPP configurations tested zero-shot; holdout parameter values unspecified. **[D]**	Training-data sweeps measure time/frequency/ space phase-ramp consistency via NMSE-WE. **[D]**	Unseen simulated/ measured datasets tested without fine-tuning or parameter updates. **[D]**	Rate, operational overhead, reliability and end-to-end latency unassessed. **[U]**
ChannelLM [8]	Shared environment encoder supports path-loss maps and sparse-pilot CSI recovery. **[D]**	Scene holdouts retain fixed carrier, antenna count and subcarrier settings. **[P]**	Localization-error sensitivity provides indirect geometry evidence; channel-output physical-relation tests unreported. **[P]**	Target-domain adaptation costs/efficiency unassessed despite specified initial training. **[U]**	Approx. 75-ms sensing-to-PL/CSI prediction latency (RTX4090, batch 1; ray-traced labels); direct communication utility untested. **[P]**
FM on LAE Twin-XL [30]	Shared backbone supports four tasks; seven of ten backbone blocks frozen. **[D]**	Masked-antenna recovery/ city holdouts retain fixed array/carrier; unseen system configurations untested. **[P]**	Single near-field example assesses polar reconstruction/focusing; broader independent physical validation remains absent. **[P]**	Target samples, 300 epochs, three updated blocks reported; trainable counts/optimization costs incomplete. **[P]**	Simulated-task timing reported; communication-rate and operational overhead benefits untested. **[P]**
WWM [9]	Frozen multimodal backbone supports four simulated tasks; measured CSI prediction requires fine-tuning. **[D]**	Simulated zero-shot mobility holdouts: 40/70 km/h, absent from pretraining. **[D]**	Geometry-conditioned JEPA ablated; independent channel-output physical-consistency or feasibility tests unreported. **[U]**	Measured-data adaptation reports 13.2% of simulated pretraining volume; parameter/ optimization budgets incomplete. **[P]**	Simulated cross-band beam selection reports approximately 99% oracle-relative beam gain. **[D]**

Abbreviations: CSI, channel state information; CIR, channel impulse response; CLS, classification token; STF, space–time–frequency; MAE, masked autoencoder; UMa/RMa, urban/rural macro; LoS/NLoS, line-of-sight/non-line-of-sight; AoA, angle of arrival; SSL, self-supervised learning; LoRA, low-rank adaptation; MoE, mixture of experts; ViT, vision transformer; RF, radio frequency; NMSE, normalized mean-square error; JEPA, joint-embedding predictive architecture; SE, spectral efficiency; QoS, quality of service; WMMSE, weighted minimum mean-square error. Adaptation-efficiency D for LLM4WM, LWM and WavesFM is limited to reported data/parameter-efficiency comparisons; complete optimization accounting is not established.

delay of a resolved line-of-sight path changes by the distance increment divided by the speed of light. This assumes free-space propagation, a common calibrated timing reference, and sufficient delay resolution; reflected paths follow their own path lengths. Raw IQ streams expose carrier phase, synchronization errors, modulation, hardware fingerprints, and spatial signatures before receiver processing. IQFM demonstrates that reusable representations can be learned directly from over-the-air multi-antenna IQ data [26]. OFDM resource grids should additionally encode pilots, guard bands, antenna ports, and numerology, while beamspace and delay–Doppler representations should retain angular sparsity and mobility-dependent support. Complex structure matters. A real–imaginary decomposition is information preserving and may be sufficient when the phase reference and calibration state remain consistent. However, ordinary real-valued processing does not guarantee equivariance to physically meaningful phase transformations. Changes in array geometry, coordinate systems, or phase reference may therefore require explicit equivariant structure or transformation-aware augmentation [7]. Complex-valued arithmetic can preserve the algebra of the input, but does not by itself guarantee the required equivariance. Practical impairments further complicate these symmetries: carrier-frequency offset induces time-varying phase rotation, phase noise causes stochastic phase perturbations, and RF calibration errors introduce antenna- or subcarrier-dependent distortions. Such effects should be calibrated, represented as configuration or latent-state variables, or included explicitly in robustness evaluation.

A complete wireless state also extends beyond radio measurements. Base station (BS) and user equipment (UE) locations, building geometry, material properties, red–green–blue images, light detection and ranging (LiDAR), radar, trajectories, moving objects, and blockage states are partial observations of the same propagation environment. Real-world and simulation-oriented multimodal datasets have begun to align these sources with communication measurements [33], [34]. Channel charting further shows that CSI alone can preserve local radio geometry in a learned latent space [35]. Multimodal fusion need not rely on direct concatenation. Co-registered radio and environmental grids can support early feature fusion, whereas heterogeneous observations such as CSI, images, and point clouds are better handled by modality-specific encoders followed by cross-attention. Gated or late fusion is particularly useful when environmental observations are asynchronous, incomplete, or stale. In all cases, environmental

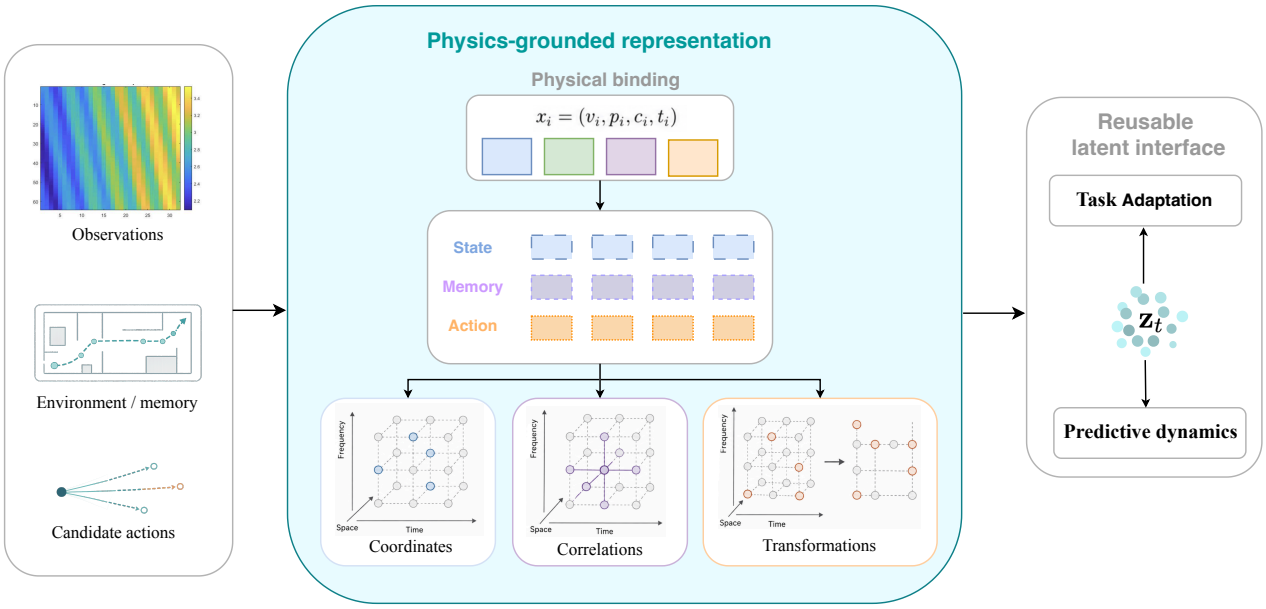


Fig. 2. **Conceptual physics-grounded representation interface for wireless foundation models.** Radio observations, environment and memory, and candidate actions are bound to physical values, coordinates, configurations, and acquisition times. The resulting typed tokens interact under coordinate, correlation, and transformation constraints to form a reusable latent representation for task adaptation and predictive dynamics.

features should provide causal context for channel variation rather than replace the physically indexed radio representation.

Physical coordinates must also constrain token interaction. Real time intervals, subcarrier spacing, carrier frequency, array-element positions, and wavelength provide more transferable information than integer indices. Moreover, dependencies weaken outside the coherence time, coherence bandwidth, spatial correlation range, or physically plausible delay–Doppler support. CSI-native modeling consequently advocates physical coordinate embeddings and correlation-bounded attention [27]. A stronger requirement is equivariance under specified physical transformations: the representation or prediction should follow the corresponding physical transformation law [7].

Finally, environment-interactive models require structured state, memory, and action interfaces. State tokens summarize the current channel, interference, geometry, and hardware condition. Memory tokens retrieve persistent or recently updated knowledge from digital twins, channel maps, and historical trajectories. Action tokens encode beams, powers, pilots, RIS states, schedules, or platform motion. These groups should not be flattened into one undifferentiated sequence. Hierarchical encoders, typed queries, and cross-attention can preserve their distinct physical roles while enabling controlled information exchange, as a prospective extension of the environment-conditioned representations demonstrated by ChannelLM and WWM [8], [9]. These studies do not establish the complete state–memory–action interface proposed here.

Physics grounding depends on how structure enters the model, rather than the backbone alone. Its strongest evidence combines preserved coordinates, physical transformations, and feasible behavior under configuration or environmental shifts, providing an interface for learning reusable predictive dynamics.

IV. Physics-Guided Predictive Pretraining

Pretraining should learn transferable propagation structure rather than merely compress the training distribution. Its objectives must reflect sampling, corruption, physical transformations, and system actions; reconstruction alone is insufficient.

Masked reconstruction should reflect physical partial observability: sparse pilots, unavailable subcarriers, blocked antennas, missing bands, interrupted trajectories, or incomplete sensing. WiFo uses structured space–time–frequency masking for cross-domain channel prediction [5], inferring plausible missing states from valid context. Denoising separates persistent environmental and propagation factors from thermal noise, interference, quantization, synchronization errors, phase noise, and hardware mismatch; corruption models should therefore vary in type and severity beyond additive white noise.

Cross-domain consistency provides label-free supervision. CSI-CLIP contrasts CSI–CIR embeddings [29]; explicit Fourier consistency of reconstructed signals is a further possible constraint. Time-, frequency-, and spatial-domain observations should share multipath structure, while uplink and downlink may share geometry-dependent scattering despite frequency-selective differences. CSI, LiDAR, images, positions, and trajectories should align by shared scene, not superficial feature similarity. Contrastive pairs should respect physical neighborhoods, and physics-equivariant losses can enforce predictable responses to coordinate or array transformations [7].

Reconstruction primarily captures observability. Predictive learning targets dynamics. Transformer-based channel prediction has shown that future CSI can be inferred in parallel

to mitigate mobility-induced channel aging [36]. Channel extrapolation studies relate prediction across time, frequency, and space to partial reciprocity, spatial consistency, and non-stationarity [37]. A latent predictor may be expressed as

$$\hat{\mathbf{z}}_{t+1} = F_\phi\left(\mathbf{z}_{1:t}\right), \tag{3}$$

where $\mathbf{z}_t$ summarizes channel, environment, mobility, and hardware states at time t; $\mathbf{z}_{1:t}$ is its history through t, and F_ϕ, with parameters ϕ, predicts the next latent state $\hat{\mathbf{z}}_{t+1}$. Joint-embedding predictive architecture (JEPA)-style learning is relevant because it predicts target representations rather than every noisy observation coefficient [38]. Wireless latent-dynamics learning applies this principle to measured CSI, predicting channel-chart embeddings conditioned on user-velocity sequences rather than forecasting from CSI history alone [39].

Predictive WFMs forecast future channel states from observations. For control, a wireless world model additionally conditions predictions on candidate actions, allowing alternative decisions to be evaluated before execution. This action-conditioned role is represented by:

$$\hat{\mathbf{z}}_{t+1} = F_\phi\left(\mathbf{z}_t, \mathbf{a}_t\right). \tag{4}$$

Here, $\mathbf{a}_t$ denotes the action at time t and may encode a beam switch, pilot allocation, RIS configuration, transmit power, schedule, or unmanned aerial vehicle (UAV) motion. The model can then estimate how an intervention changes blockage risk, uncertainty, achievable rate, or the future radio field. Digital twins provide a natural environment for generating such counterfactual transitions. The K candidate actions $\{\mathbf{a}_k\}_{k=1}^{K}$ can be simulated to obtain alternative futures $\left\{\hat{\mathbf{z}}_{t+1}^{(\mathbf{a}_k)}\right\}_{k=1}^{K}$, enabling constrained planning before physical execution. This resembles latent imagination in model-based control [40], but wireless actions must additionally satisfy waveform, latency, energy, and protocol constraints. Existing wireless world models provide an early latent-dynamics basis for this direction [9]. Recent low-altitude systems combining vision-language-action (VLA) models with world models have reported action-conditioned UAV control in simulation [41]. General action-conditioned prediction with physical closed-loop validation across wireless tasks and infrastructures nevertheless remains largely open.

As illustrated in Fig. 3, physics-guided predictive pretraining extends state recovery toward latent dynamics modeling and action-conditioned future prediction.

A possible composite pretraining loss $\mathcal{L}$ combines active objectives indexed by $\mathcal{S}$, with weights λ_j:

$$\begin{gathered}\mathcal{L} = \sum_{j\in\mathcal{S}} \lambda_j \mathcal{L}_j, \\ \mathcal{S} \subseteq \{\mathrm{rec}, \mathrm{den}, \mathrm{con}, \mathrm{pred}, \mathrm{act}, \mathrm{phy}\},\end{gathered} \tag{5}$$

where $\mathcal{L}_{\mathrm{rec}}$ recovers physically plausible wireless states from partial observations; $\mathcal{L}_{\mathrm{den}}$ suppresses noise and hardware corruptions; $\mathcal{L}_{\mathrm{con}}$ enforces cross-domain or multimodal consistency, including CSI–CIR and radio–environment relations; $\mathcal{L}_{\mathrm{pred}}$ learns passive latent dynamics; $\mathcal{L}_{\mathrm{act}}$ learns action-conditioned transitions or counterfactual futures; and $\mathcal{L}_{\mathrm{phy}}$ imposes equivariance, Fourier consistency, geometry, or propagation constraints.

The active terms depend on the available observations and supervision. Reconstruction and denoising require partial or corrupted observations; consistency requires paired domains, modalities, or known transformations. Temporal prediction requires sequences, while action conditioning requires observed or simulated action–transition pairs. Controlled evidence on negative transfer remains limited, so joint training is an empirical choice rather than a guaranteed benefit. The coefficients λ_j should reflect supervision availability, data reliability, loss scale, and downstream utility. The combined objective is a prospective template; its benefit requires controlled comparisons and ablations with matched data and training budgets.

V. Digital twins for environment-aware wireless foundation models

A channel observation describes an effect rather than its complete physical cause. The channel H_t at time t is abstracted through a propagation mapping $\mathcal{F}$:

$$H_t = \mathcal{F}(E_t, P_t, M_t, C_t), \tag{6}$$

where E_t represents geometry and material properties, P_t denotes the positions of transceivers and relevant scatterers, M_t captures mobility and dynamic blockage, and C_t specifies the frequency, array, waveform, and hardware configuration. Learning only the channel distribution $p(H)$ may capture recurring statistics, but it does not necessarily explain why the channel changes.

Environment representations range from implicit to explicit. Channel charting extracts local spatial relations directly from CSI [35], whereas channel knowledge maps store or infer location-dependent path loss, shadowing, delay, angle, or beam information [42]. RadioUNet illustrates a more explicit alternative by mapping city geometry and transmitter positions to dense path-loss fields [43]. Hybrid representations combine scene priors with radio measurements, but environmental observations remain uncertain estimates. Detection/localization errors, camera–LiDAR miscalibration, and timestamp offsets can corrupt geometry and blockage states; rain and fog can degrade camera visibility and LiDAR returns [33], [44]. If sensing errors enter twin memory, they may bias subsequent channel predictions, beam choices, and planned actions. Reported sensing-to-channel tests combine measured localization with simulated radio outputs [8]; their robustness evidence is confined to the tested conditions. Fusion alone does not guarantee reliability [44]. Quality weighting in fusion, checks against radio measurements, and conventional fallback should be evaluated under calibration errors, adverse weather, and modality loss using channel error, beam gain, and constraint violations.

These environment representations should not be conflated. Statistical channel simulators generate realizations from predefined models, while radio maps or channel knowledge maps store location-dependent radio information [42], [45], [46]. Generative channel models learn channel distributions and synthesize plausible samples, but generation alone does not

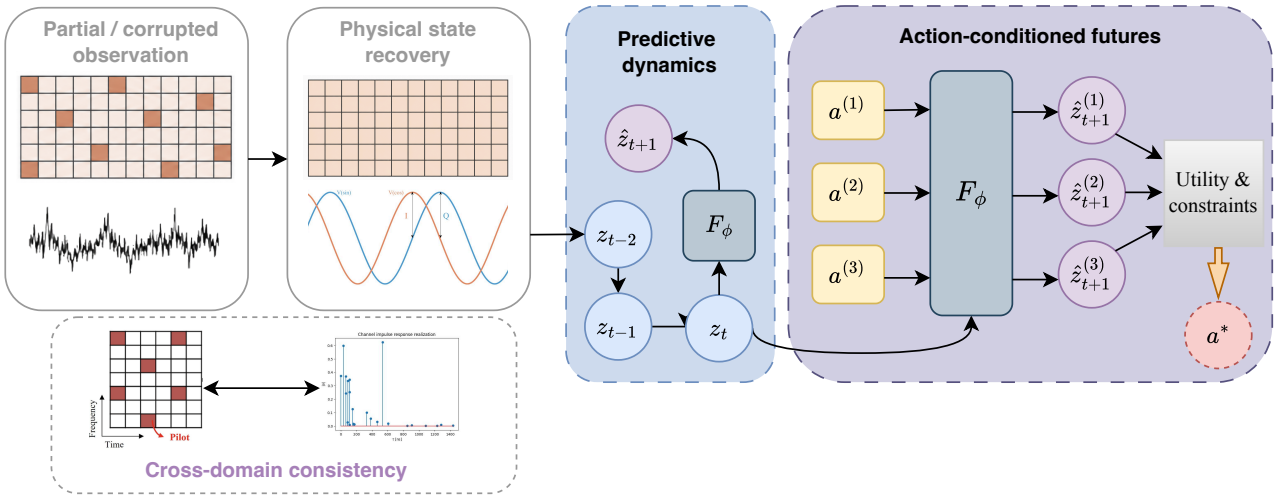


Fig. 3. **Conceptual physics-guided predictive pretraining for wireless foundation models.** Partial or corrupted wireless observations are mapped to a recovered physical state, with cross-domain consistency linking time–frequency measurements and delay-domain responses. The resulting latent-state sequence supports predictive dynamics, while the current state z_t and candidate actions $a^{(k)}$ are jointly processed by the action-conditioned transition model F_ϕ to generate counterfactual futures $\hat{z}_{t+1}^{(k)}$. These predicted futures are evaluated under utility and system constraints to select the action a^*.

constitute a digital twin. In this review, a wireless digital twin denotes a site- or system-specific virtual representation that is continuously aligned with its physical counterpart through measurement, synchronization, and calibration [45], [46]. It should therefore maintain explicit physical–digital state correspondence, an update mechanism, and state freshness or uncertainty. In control-oriented settings, the twin may additionally support counterfactual simulation and decision validation. A world model differs in that it primarily learns compact state dynamics for prediction and planning, as discussed in the section "From wireless world models to embodied wireless intelligence".

A wireless digital twin can comprise five functional layers. The *physical-world layer* contains BSs, terminals, UAVs, RISs, scatterers, and surrounding objects. The *observation and synchronization layer* collects CSI, IQ, images, LiDAR, radar, positions, and device states. The *digital twin state layer* maintains synchronized estimates of geometry, propagation, hardware and network configurations, dynamics, historical measurements, and uncertainty. The *intelligence layer* comprises physics-grounded wireless foundation models and physics-grounded dynamics models. The *service/control layer* supports channel prediction, beam management, positioning, scheduling, and resource control. The state layer holds an updateable digital belief about the physical-world layer's entities. Digital twin channel frameworks emphasize real-time physical-to-digital mapping and progressively autonomous operation [45]; radio-environment knowledge pools formalize reusable environment–channel relations [46].

The twin maintains site-specific state and measurement synchronization, with counterfactual simulation and decision validation in control-oriented settings. The foundation model extracts cross-site propagation regularities for multiple tasks. ChannelLM combines multimodal reconstruction, propagation-guided features, and environment-generalizable prediction [8]. LAETwin-XL supplies a low-altitude XL-MIMO ray-tracing toolchain and dataset for training a generative foundation model from incomplete channels [30]; it demonstrates simulated training and transfer, not a twin continuously synchronized with measurements. Beyond synthetic data, a twin provides evolving context for grounding and updating reusable models.

A prospective interface separates reusable model parameters from external twin memory:

$$z_t = E_\theta\left(o_t,\ \mathrm{Retrieve}\left(\mathcal{M}_t^{\mathrm{DT}}, q_t\right)\right), \tag{7}$$

where E_θ is an encoder with parameters θ, mapping the observation o_t and retrieved content to the latent state z_t. Here, $\mathcal{M}_t^{\mathrm{DT}}$ is external digital twin memory, q_t a task or state query, and $\mathrm{Retrieve}$ retrieves content relevant to q_t. Parameters encode broad propagation priors and reusable task relations; memory holds evolving site geometry, blockage, mobility, calibration, local measurements, and short-term dynamics. Whether memory updates improve safety or efficiency over model adaptation requires evaluation.

The remaining challenge is not a single synthetic-to-real domain shift, but a combination of geometric, electromagnetic, hardware, and temporal mismatches. Measurement-calibrated ray tracing reveals that these discrepancies can be substantial. At 6.75 and 16.95 GHz, site-specific validation of NYU-Ray showed that location calibration reduced the combined matching loss for power delay profiles by 42.28% in LOS and 13.52% in NLOS conditions, while the ray tracer still systematically underestimated multipath richness, including root-mean-square delay spread by approximately 61% and 30% in LOS and NLOS scenarios, respectively [47]. Material uncertainty produces another source of bias. Differentiable ray-tracing calibration against measured indoor channels substantially reduced errors in path gain and delay spread relative to nominal International Telecommunication Union material

parameters, with neural material models providing the most accurate results among the tested calibration approaches [48]. Residual errors nevertheless remained where the geometric scene itself was incomplete, indicating that learned calibration cannot fully compensate for missing physical structure.

Such modeling errors propagate directly to downstream wireless decisions. In a 60-GHz real-world beam-prediction experiment, directly transferring the beam selected from a digital replica achieved only 56.1% top-1 accuracy, although it retained 84.8% of the optimal received power [49]. A model trained on synthetic data achieved 91.4% top-2 accuracy on measured channels when the measured beam codebook was reproduced, but this decreased to 84.8% when an idealized uniform codebook was used, demonstrating the impact of hardware-model mismatch. Fine-tuning with fewer than 20 real-world samples was sufficient to approach the optimal performance, whereas training from real data alone required approximately 100 samples [49]. Complementarily, output-level calibration can reduce the fidelity–complexity trade-off: a lightweight calibration model combined with low-complexity ray tracing required 0.0610 s per sample, compared with 1.2019 s for the high-fidelity digital twin, while approaching its CSI accuracy in the reported simulator comparison [50]. The reported experiment, however, calibrates one simulator against a higher-fidelity simulator rather than against physical measurements. These results suggest that different mismatch timescales should be handled differently. Fast variations such as mobility and blockage are better absorbed through synchronized twin-state updates, whereas persistent geometry, material, or hardware bias may require explicit calibration or restricted model adaptation. Domain adaptation, meta-learning, and test-time adaptation [51]–[53] provide methodological starting points. These cited works evaluate general machine-learning tasks, not wireless digital twin drift. In particular, Tent minimizes classification entropy; CSI regression requires a suitable alternative adaptation objective. Wireless-specific evidence under long-term calibration drift remains limited.

VI. From wireless world models to embodied wireless intelligence

Environment awareness becomes useful when it changes how a wireless system acts. Reactive intelligence uses a policy π to map the current channel H_t to an action $a_t = \pi(H_t)$. Predictive intelligence additionally evaluates how a decision may affect what is observed next. The feedback loop is broader than a change in the propagation environment alone: platform motion or RIS control may modify the physical scene, while beam, power, pilot, or scheduling decisions alter the radio configuration. This interaction is expressed as

$$a_t \rightarrow (E_{t+1}, C_{t+1}) \rightarrow o_{t+1}, \tag{8}$$

where E_{t+1} denotes the physical environment state and C_{t+1} the wireless configuration. Physical-AI proposes a related architecture that maps distributed radio observations into latent environmental states for proactive wireless decisions, with simulation evidence for predictive beam control [54].

Li et al. propose large-model-enabled Intelligent Base Station Agents (IBSAs) combining perception, cognition, and execution with cloud–edge–end collaboration and digital-twin support [55]. Their architectural roadmap motivates examining action-conditioned state prediction and its validation against physical radio feedback.

Fig. 4 illustrates the prospective integration of a physics-grounded wireless foundation model, synchronized digital-twin context, and action-conditioned prediction into a closed-loop framework for wireless control.

The world model adds an action-conditioned transition distribution $p_\phi(z_{t+1} \mid z_t, a_t)$ to the twin's synchronized site-specific state. It retains decision-relevant evolution for evaluating candidate actions before execution [16]. Under partial observability, calibrated twin states and counterfactual data provide context for these predictions. Existing wireless world models learn latent spatiotemporal dynamics from CSI, geometry, and trajectories [9]. Their reported evaluations do not establish a general action-conditioned physical control loop.

Table III summarizes their complementary roles in environment representation and predictive decision making.

Wireless actions themselves are heterogeneous. One prospective interface is

$$\begin{aligned} a_t &= \left(a_t^{\mathrm{d}}, a_t^{\mathrm{c}}\right), \\ a_t^{\mathrm{d}} &\in \mathcal{A}_{\mathrm{d}}, \\ a_t^{\mathrm{c}} &\in \mathcal{A}_{\mathrm{c}} \subseteq \mathbb{R}^m, \\ e_t^a &= E_{\mathrm{d}}\left(a_t^{\mathrm{d}}\right) \oplus E_{\mathrm{c}}(a_t^{\mathrm{c}}), \\ \hat{z}_{t+1} &= F_\phi(z_t, e_t^a), \end{aligned} \tag{9}$$

where $\mathcal{A}_{\mathrm{d}}$ and $\mathcal{A}_{\mathrm{c}}$ are discrete and continuous action spaces, and m counts continuous control variables. Here, a_t^{d} covers categorical or combinatorial decisions such as beam indices, scheduled-user sets, pilot assignments, or quantized RIS states, and a_t^{c} contains variables such as transmit power, continuous precoding coefficients, antenna position, or platform motion. The boundary is hardware dependent: an RIS, for example, may expose either a finite codebook or a finely quantized phase-control interface [56]. Here, E_{d} encodes discrete actions using learned embeddings or set encoders; E_{c} encodes continuous controls using normalized coordinates or features. The fusion operator $\oplus$ forms the action representation e_t^a, which may condition predictions through action tokens, cross-attention, or modulation of latent transition features. What matters is that the model retains the distinction between the state being predicted and the intervention applied to it.

The following abstraction is therefore introduced as a *prospective* action-conditioned control loop rather than an experimentally established wireless foundation model pipeline. Given the current state, candidate actions may in principle be evaluated through model-predicted futures before physical execution:

$$\begin{aligned} o_t &\xrightarrow{E_\theta} z_t, \\ \left\{a_t^{(k)}\right\}_{k=1}^{K} &\xrightarrow{E_a} \left\{e_t^{a,(k)}\right\}_{k=1}^{K} \\ \xrightarrow{F_\phi(z_t,\cdot)} \left\{\hat{z}_{t+1}^{(k)}\right\}_{k=1}^{K} &\xrightarrow{\pi_\psi} a_t^* \xrightarrow{\text{Executor}} o_{t+1}. \end{aligned} \tag{10}$$

Here, E_a denotes action encoding, $k = 1, \ldots, K$ indexes candidate actions, and π_ψ is a planner with parameters ψ

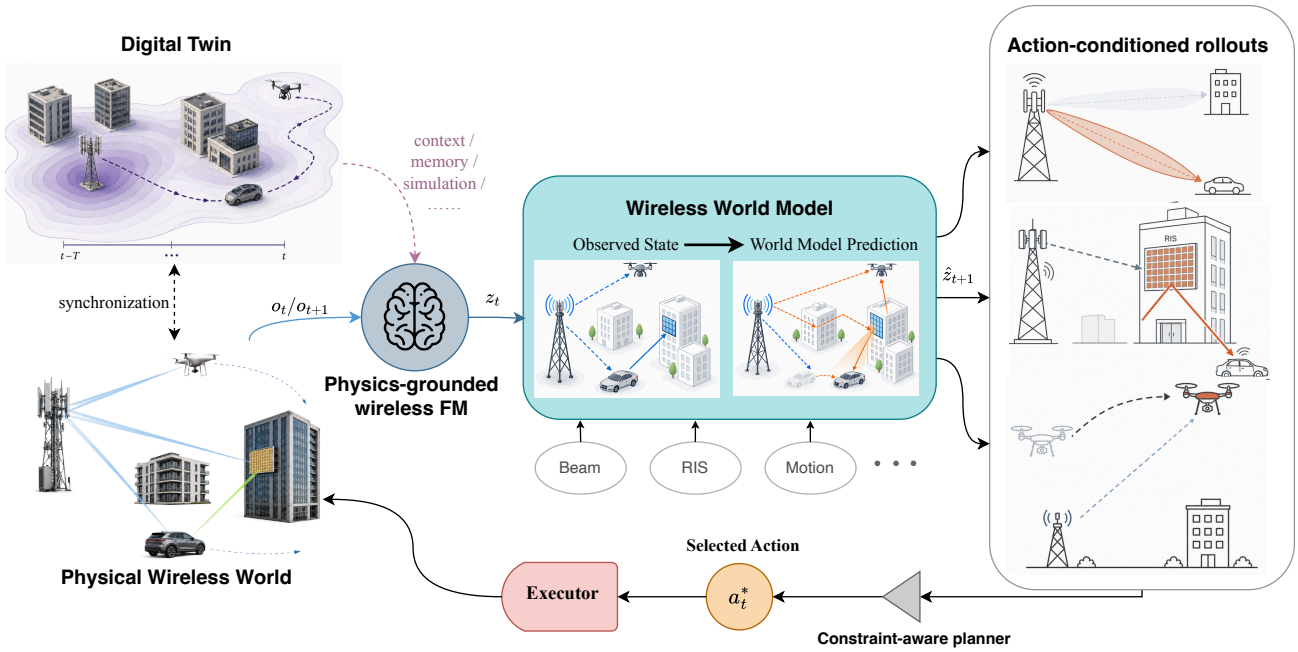


Fig. 4. **Conceptual research framework from environment awareness supported by digital twins to embodied wireless intelligence.** A synchronized digital twin provides environmental context and memory for grounding the current wireless state. The wireless world model predicts counterfactual futures for candidate radio or physical actions, and a constraint-aware planner selects one for execution. Feedback from the physical wireless world updates the state and twin, closing the perception–prediction–action loop. FM, foundation model; RIS, reconfigurable intelligent surface. Original diagram created by the authors using diagrams.net (draw.io).

TABLE III
CONCEPTUAL DISTINCTION BETWEEN WIRELESS DIGITAL TWINS AND WIRELESS WORLD MODELS.

Aspect	Digital twin	Wireless world model
State	Synchronized site- or system-specific physical and radio state.	Compact latent state retaining decision-relevant information.
Dynamics	Physics/simulator-driven evolution with measurement-based updates.	Learned state transitions for prediction and planning.
Actions	Can simulate or validate candidate actions in control-oriented settings.	Actions condition future-state predictions in action-conditioned models.
Update mechanism	Measurement synchronization, calibration, state refresh, and memory updates.	State inference; parameter or adapter updates when needed.
Uncertainty	State freshness, calibration error, and physical/model mismatch.	Predictive uncertainty over future states or rollouts.
Primary objective	Maintain physical correspondence with a specific wireless environment.	Predict decision-relevant evolution for planning and control.
Deployment role	Environment memory and calibration, simulation, and validation space.	Prospective predictive engine for action-conditioned planning and control.

selecting a_t^* from their predicted futures. The executor applies the selected action, after which o_{t+1} is observed. Current evidence supports prediction and task-specific control at different validation levels. The cited wireless world model evaluates temporal channel prediction in simulation and frequency extrapolation on measured CSI [9]. Simulated predictive beam switching uses a standalone predictor without a foundation backbone [54]; UAV navigation and multi-target completion are also evaluated in simulation [41]. These demonstrations do not validate the complete loop in Eq. (10), which combines a reusable backbone, action-conditioned radio dynamics, and measured feedback.

Realizing this loop requires synchronized records of observations, executed actions, and subsequent observations to learn action-conditioned transitions [40], [41]. The twin requires measurement-based calibration and assessment of residual geometry and hardware mismatch [48], [49]. Sensing, inference, rollouts, and actuation must fit the target control timescale [41], [54]. Planning must assess accumulated prediction error and uncertainty, while runtime checks enforce transmit power, spectrum, protocol, mobility, and actuator constraints, with conventional fallback when predictions are unreliable. Physical closed-loop tests should report utility, constraint violations, and control overhead against conventional controllers.

Predictive beam management selects discrete beams using predictions that may depend on continuous mobility. UAV systems couple motion with beam or user selection; energy consumption and sensing coverage are treated as costs or objectives rather than action coordinates. The low-altitude VLA–world-model framework [41] demonstrates simulated UAV navigation; its experiments do not establish joint radio-resource and trajectory optimization. RIS hardware can likewise expose discrete codebooks or finer phase-control spaces [56]. Slower trajectory changes and faster beam, scheduling, or power adaptation favor hierarchical control.

Predictive control depends on physical assumptions and their cost. Priors require testing over their intended transformation classes; incorrect propagation or hardware assumptions may hinder adaptation. Twin-based beam prediction and differentiable ray tracing show both the need for measurement correction and residual scene-model errors after calibration [48], [49]. System benefits must account for sensing, synchronization, state refresh, and rollouts, with controller queueing additionally relevant to software-defined RIS networks [57]. Conventional methods, task-specific models, and compressed foundation models remain necessary baselines for stable environments with few tasks [58]. Richer environment-aware architectures must justify these costs through transfer or decision gains under deployment constraints.

VII. Benchmarks, deployment, trustworthiness, and standardization

Wireless foundation models require evaluation beyond task accuracy. Measured-data studies support representation reuse and channel prediction, including zero-shot transfer and adaptation [7], [9], [26], but do not by themselves establish gains in radio operation. Stronger field validation should hold out sites, acquisition sessions, and hardware. Calibration/adaptation data should be separate from test data. Performance before and after calibration or adaptation should be reported with stated data and compute budgets. On the intended hardware, comparisons with conventional and task-specific baselines under matched budgets should measure achieved rate, reliability, pilot overhead, end-to-end latency, and energy. Twins require measurement-based assessment of residual geometry and material errors [47], [48], channel consistency, synchronization delay, and calibrated uncertainty. Benchmarks should separate configuration, environment, cross-simulator, synthetic-to-real, temporal, and hardware shifts. World-model tests should cover short and long horizons and counterfactual consistency; action-conditioned predictions require comparison with measured outcomes after executed actions.

Deployment should follow radio timescales. Conventional processing and small task heads serve the fastest physical-layer (PHY) functions. Foundation backbones and predictors operate at channel, mobility, or planning timescales. Twins, agents, and cloud training evolve more slowly. Each component need not meet the innermost radio loop's latency.

Radio-side models must become lighter. Task-specific early exits avoid full-depth execution when intermediate representations transfer better; the cited study favors fixed per-task over sample-dependent exits [59]. Tiny-WiFo distills most original capability into a smaller channel predictor [58], while ConsisFormer aggregates correlated tokens before self-attention [60]. Pruning, low-bit quantization, and mixed precision can further reduce memory traffic and arithmetic. Benefits depend on operator fusion, pipelining, and on-chip reuse on neural processing units (NPUs), field-programmable gate arrays (FPGAs), or dedicated accelerators. Microsecond-class deadlines still require measurements on intended hardware; these studies do not establish them for complete backbones or generic task heads.

Trustworthiness spans the model lifecycle. Adversarial channel-estimator inputs degraded predictions on MATLAB-generated fifth-generation (5G) data [61], demonstrating input-space vulnerability rather than an over-the-air attack on a wireless foundation model. Corrupted measurements, maps, or updates could persist in external twin memory and affect later predictions. Wireless federated-learning surveys review related data and model poisoning [62]; extending these threats to twin memory is prospective, not a demonstrated survey result. Representation richness also raises privacy concerns: channel charting exploits CSI's local spatial geometry [35], while CSI amplitude and phase enabled person re-identification in a dedicated experiment [63]. Raw CSI, trajectories, memory, checkpoints, and calibration data therefore require authenticated provenance, access control, integrity checks, and auditable versions.

Reliability should also be measured rather than asserted. For probabilistic classification tasks, calibration can be reported using expected calibration error (ECE), negative log-likelihood (NLL), or the Brier score. Set- or interval-valued prediction should report empirical coverage together with prediction-set or interval size. Conformal prediction provides one wireless-specific route to such coverage guarantees for demodulation, modulation classification, and channel prediction [64]. Standard offline guarantees require assumptions such as exchangeability and concern marginal coverage. The online channel-prediction setting evaluates long-run coverage. Neither provides unconditional per-state guarantees under arbitrary distribution shift. Out-of-distribution detection and drift monitoring should additionally report discrimination metrics such as the area under the receiver operating characteristic curve (AUROC) or the false-positive rate (FPR) at a prescribed true-positive rate, rather than relying on confidence scores alone. At runtime, excessive uncertainty, failed coverage, detected distribution shift, or violated physical constraints should trigger abstention, re-calibration, or fallback to a conventional receiver or controller. A practical lifecycle can assign training-data and checkpoint provenance to cloud or Service Management and Orchestration (SMO) functions. Edge or Radio Access Network (RAN) Intelligent Controller (RIC) functions can monitor drift and adaptation. The radio-side controller can enforce hard constraints and execute fail-safe actions. Direct evidence for such end-to-end safeguards in wireless foundation models remains limited, making trustworthy lifecycle evaluation an important benchmark dimension.

Standardization should be interpreted according to its maturity level and the September 2026 status of the cited records.

In the 3rd Generation Partnership Project (3GPP), Release 20 remains primarily the study phase for 6G. The *FS_6G_Radio* study (work item WI 1080072) is associated with the draft Technical Report (TR) 38.960 listed in the 3GPP portal [65]. Ongoing Release-20 work items *NR_AIML_air_Ph2* (WI 1080085) and *NR_AIML_NGRAN_Ph3* (WI 1110099) extend AI/ML support for the New Radio (NR) air interface and next-generation radio access network (NG-RAN), respectively [66]. For 6G-specific AI protocols, the Core Network and Terminals Working Group 3 (CT3) study *FS_6G_AIP_CT* (WI 1120022) has initiated the draft TR 29.832, *Study on the Protocol for Artificial Intelligence in 6G* [67]. The cited 6G TR records are study documents, not completed normative 6G specifications. The NR AI/ML work items should also be distinguished from these 6G-specific studies.

The O-RAN Alliance has progressed further toward deployable AI/ML lifecycle support. Completed Release 5 includes AI/ML workflow functions spanning the non-real-time (Non-RT) and near-real-time (Near-RT) RICs and was completed with the November 2025 specifications [68]. Subsequent specification updates announced on 27 February 2026 should be distinguished from that release: R1 General Aspects and Principles v13.00 updates model-training capability registration, and R1 Test Specification v5.00 includes model-training, deployment, and discovery tests. A1 Application Protocol v6.00 defines the A1-ML/v1 model-management service, and the Near-RT RIC APIs (application programming interfaces) v3.00 specification introduces AI/ML APIs [69]. These specifications provide lifecycle, interface, and deployment mechanisms for AI-enabled RAN operation, but they do not define a wireless foundation model as a standardized network entity. These specifications can therefore serve as enabling infrastructure for future wireless foundation models rather than evidence that such models themselves have already been standardized.

A compact benchmarking and deployment checklist is: (i) physical consistency tests and their applicable transformations or relations; (ii) held-out configurations and adaptation settings; (iii) sim-to-real performance on measured channels before and after calibration or adaptation; (iv) target-data, trainable-parameter, and optimization budgets; (v) latency, energy, and memory on named hardware with stated batch size and precision; and (vi) system utility and, where applicable, constraint violations relative to conventional and task-specific baselines.

VIII. Conclusion

Wireless foundation models should be judged not by scale alone, but by whether their learned representations remain useful across tasks, configurations, environments, and deployment conditions. Physics-grounded representations, predictive pretraining, and environment-aware modeling provide ways to retain physical structure and learn decision-relevant dynamics. Across the reviewed studies, representation reuse is more extensively documented than independent physical-consistency tests, measured-channel transfer, or physical closed-loop gains. Adaptation gains often concern data or parameter efficiency, with incomplete accounting of optimization costs.

The remaining technical question is whether physical structure retained during pretraining supports reliable predictions under radio-configuration shifts and candidate actions. Digital twins can supply calibrated context, while action-conditioned world models predict the consequences of interventions. The integrated loop remains prospective: its value depends on predictions checked against measured outcomes and control gains that survive model mismatch.

Beyond scaling does not discount the value of larger models or datasets. It asks whether their gains persist when physical predictions guide radio decisions under actual deployment constraints.

Data Availability

No, I do not have any research data outside the submitted manuscript file.

Code Availability

No new software or computational code was developed as part of this review.

Author Contributions

Q.S. and Z.G. conceived and designed the review. Q.S. conducted the literature survey, synthesized the findings, and prepared the original manuscript. S.G., Y.M., J.Z., B.C., and K.D. contributed to the critical discussion, review, and revision of the manuscript. Z.G. supervised the work. All authors reviewed and approved the final manuscript.

Acknowledgements

Not applicable.

Funding

This work was supported in part by the Beijing Natural Science Foundation under Grant L242011; in part by the National Natural Science Foundation of China (NSFC) under Grant 62471036; in part by the Natural Science Foundation of Shandong Province under Grant ZR2025QA30; and in part by the Fundamental Research Funds for the Central Universities.

The funders provided financial support only and had no role in the design of this review, literature selection, evidence analysis and interpretation, or preparation of the manuscript.

Competing Interests

All authors declare that they have no financial or non-financial competing interests related to this manuscript, including relevant patent interests or editorial roles.

ABBREVIATIONS

3GPP — 3rd Generation Partnership Project.
5G — fifth generation.
6G — sixth generation.
AI — artificial intelligence.
AoA — angle of arrival.
API — application programming interface.
AUROC — area under the receiver operating characteristic curve.
BS — base station.
CIR — channel impulse response.
CLS — classification token.
CSI — channel state information.
CT3 — Core Network and Terminals Working Group 3.
DT — digital twin.
ECE — expected calibration error.
FM — foundation model.
FPGA — field-programmable gate array.
FPR — false-positive rate.
GFLOPs — giga floating-point operations.
GPT-2 — Generative Pretrained Transformer 2.
IBSA — Intelligent Base Station Agent.
IQ — in-phase/quadrature.
JEPA — joint-embedding predictive architecture.
LiDAR — light detection and ranging.
LLM — large language model.
LoRA — low-rank adaptation.
LoS / LOS — line-of-sight.
LWM — Large Wireless Model.
MAE — masked autoencoder.
MIMO — multiple-input multiple-output.
ML — machine learning.
MoE — mixture of experts.
Near-RT — near-real-time.
NG-RAN — next-generation radio access network.
NLL — negative log-likelihood.
NLoS / NLOS — non-line-of-sight.
NMSE — normalized mean-square error.
NMSE-WE — normalized mean-square error for wave equivariance.
Non-RT — non-real-time.
NPU — neural processing unit.
NR — New Radio.
NSFC — National Natural Science Foundation of China.
NTN — non-terrestrial network.
OFDM — orthogonal frequency division multiplexing.
OOD — out-of-distribution.
O-RAN — open radio access network.
PHY — physical layer.
PL — path loss.
QoS — quality of service.
RAN — radio access network.
RF — radio frequency.
RIC — RAN Intelligent Controller.
RIS — reconfigurable intelligent surface.
RMa — rural macro.
RT — ray tracing.
SE — spectral efficiency.
SMO — Service Management and Orchestration.
SNR — signal-to-noise ratio.
SSL — self-supervised learning.
STF — space–time–frequency.
TR — Technical Report.
UAV — unmanned aerial vehicle.
UE — user equipment.
UMa — urban macro.
ViT — vision transformer.
VLA — vision-language-action.
WFM — wireless foundation model.
WI — work item.
WMMSE — weighted minimum mean-square error.
WWM — wireless world model.
XL-MIMO — extremely large-scale multiple-input multiple-output.

REFERENCES

[1] B. R. Das, S. R. Hasan, S. R. Sabuj, M. A. Hossain, and S. K. Ray, "A comprehensive survey on emerging AI technologies for 6G communications: Research direction, trends, challenges, and opportunities," *International Journal of Intelligent Networks*, vol. 6, pp. 113–150, 2025.

[2] 3GPP, "Study on artificial intelligence (AI)/machine learning (ML) for NR air interface," 2023, 3GPP TR 38.843, version 18.0.0, Release 18, December 2023.

[3] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," 2021, preprint at https://arxiv.org/abs/2108.07258.

[4] X. Liu, S. Gao, B. Liu, X. Cheng, and L. Yang, "LLM4WM: Adapting LLM for wireless multi-tasking," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 3, pp. 835–847, 2025.

[5] B. Liu, S. Gao, X. Liu, X. Cheng, and L. Yang, "WiFo: Wireless foundation model for channel prediction," *Sci. China Inf. Sci.*, vol. 68, p. 162302, 2025.

[6] D. Buffelli, S. Das, Y.-W. Lin, S. Vakili, C.-Y. Wang, M. Attarifar, P. Nath, and D.-S. Shiu, "Towards a foundation model for communication systems," 2025, preprint at https://arxiv.org/abs/2505.14603.

[7] H. Wang, X. Gao, Z. Sun, and Z. Wang, "Physics equivariance for robust generalization in wireless foundation model," 2026, preprint at https://arxiv.org/abs/2606.28847.

[8] Y. Cai, Y. Qiu, J. Zhang, L. Yu, Y. Zhang, Z. Zhang, and G. Liu, "Paradigm shift from statistical channel modeling to digital twin prediction: an environment-generalizable ChannelLM for 6G AI-enabled air interface," *npj Wireless Technology*, vol. 2, p. 78, 2026.

[9] Z. Chen, Y. Ren, Y. Huang, Q. Sun, N. Li, Y. Huang, C.-L. I, Y. Li, and L. Xia, "A wireless world model for AI-native 6G networks," 2026, preprint at https://arxiv.org/abs/2603.25216.

[10] S. R. Hasan, S. R. Sabuj, M. Hamamura, and M. A. Hossain, "A comprehensive review on reconfigurable intelligent surface for 6G communications: Overview, deployment, control mechanism, application, challenges, and opportunities," *Wireless Personal Communications*, vol. 139, no. 1, pp. 375–429, 2024.

[11] A. M. Pacheco Huachaca, J. J. Rodriguez Rodriguez, A. Aboulfotouh, N. L. S. da Fonseca, C. A. Astudillo, and H. Abou-Zeid, "Wireless foundation models: State-of-the-art and open challenges," 2026, preprint at https://arxiv.org/abs/2609.04707v1.

[12] N. Khan, B. Al Sbeihi, M. Alshehhi, and N. Saeed, "A comprehensive survey of wireless foundation models for AI-native 6G networks," 2026, preprint at https://arxiv.org/abs/2608.14694v1.

[13] M. Cheraghinia, D. Buffelli, L. Li, M. Alhellani, J. Fontaine, S. Vakili, M. Guenach, E. De Poorter, and A. Shahid, "Wireless physical-layer foundation models: Architectures, learning paradigms, applications, and deployment," 2026, arXiv preprint at https://arxiv.org/abs/2608.20486v1.

[14] L. Liang, J. Guo, J. Zhang, C.-B. Chae, L. Lu, S. Xu, O. A. Dobre, S. Jin, and G. Y. Li, "Foundation models for wireless communications: From PHY intelligence to network autonomy," 2026, arXiv preprint at https://arxiv.org/abs/2606.06239v2.

[15] H. Zhang, M. Farzanullah, M. Ghassemi, A. Bin Sediq, A. Afana, and M. Erol-Kantarci, "Multi-modal data-enhanced foundation models for prediction and control in wireless networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 28, pp. 4359–4393, 2026, author preprint inspected at https://arxiv.org/abs/2601.03181v1.

[16] J. Zheng, D. Niyato, C. Zhao, J. Kang, and J. Wang, "From digital twins to world models: Opportunities, challenges, and applications for mobile edge general intelligence," 2026, preprint at https://arxiv.org/abs/2603.17420v1.

[17] H. Zou, Y. Yang, L. Bariah, Y. Tian, Y. Lu, B. Wang, A. Bara, B. Mefgouda, H. Liu, Y. Tao, S. Petrov, S. Cheour, N. Sehad, S. Samarakoon, C. Huang, S. Lasaulce, M. Bennis, and M. Debbah, "Telecom world models: Unifying digital twins, foundation models, and predictive planning for 6G," 2026, arXiv preprint at https://arxiv.org/abs/2604.06882v1.

[18] H. Ye, G. Y. Li, and B.-H. Juang, "Power of deep learning for channel estimation and signal detection in OFDM systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 114–117, 2018.

[19] C.-K. Wen, W.-T. Shih, and S. Jin, "Deep learning for massive MIMO CSI feedback," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 748–751, 2018.

[20] N. Samuel, T. Diskin, and A. Wiesel, "Learning to detect," *IEEE Trans. Signal Process.*, vol. 67, no. 10, pp. 2554–2564, 2019.

[21] B. Liu, X. Liu, S. Gao, X. Cheng, and L. Yang, "LLM4CP: Adapting large language models for channel prediction," *J. Commun. Inf. Netw.*, vol. 9, no. 2, pp. 113–125, 2024.

[22] K. Ying, Z. Gao, T. Yang, J. Zhang, X. Cheng, T. Q. S. Quek, and H. V. Poor, "From specialist to large models: A paradigm evolution towards semantic-aware MIMO," *IEEE Communications Magazine*, vol. 64, no. 5, pp. 124–131, 2026.

[23] S. Alikhani, G. Charan, and A. Alkhateeb, "Large wireless model (LWM): A foundation model for wireless channels," 2024, preprint at https://arxiv.org/abs/2411.08872.

[24] A. Aboulfotouh, E. Mohammed, and H. Abou-Zeid, "6G WavesFM: A foundation model for sensing, communication, and localization," *IEEE Open J. Commun. Soc.*, vol. 6, pp. 6792–6807, 2025.

[25] W.-L. Cheng and W. Liao, "How big should a wireless foundation model be?" 2026, preprint at https://arxiv.org/abs/2605.07266.

[26] O. Mashaal and H. Abou-Zeid, "IQFM: A wireless foundation model for I/Q streams in AI-native 6G," *IEEE Open J. Commun. Soc.*, vol. 7, pp. 1426–1441, 2026.

[27] C. Zhang, X. Lyu, C. Ren, S. Liu, and Q. Cui, "Towards CSI-native foundation models: A channel-adaptive roadmap for 6G," 2026, preprint at https://arxiv.org/abs/2606.20670.

[28] Y. Wang, O. Wang, S. Zhou, and G. Y. Li, "Hierarchical wireless foundation model for multi-task optimization," 2026, preprint at https://arxiv.org/abs/2607.16877.

[29] J. Jiang, W. Yu, Y. Li, Y. Gao, and S. Xu, "A MIMO wireless channel foundation model via CIR–CSI consistency," in *Proc. IEEE Int. Conf. Mach. Learn. Commun. Netw. (ICMLCN)*, 2025.

[30] M. Li, Y. Han, J. Tian, C.-K. Wen, and S. Jin, "Digital twin-based channel generation toolchain and foundation model for low-altitude XL-MIMO," 2026, preprint at https://arxiv.org/abs/2606.14114.

[31] L. Qiao, M. Boloursaz Mashhadi, Z. Gao, R. Tafazolli, M. Bennis, and D. Niyato, "Token communications: A large model-driven framework for cross-modal context-aware semantic communications," *IEEE Wireless Communications*, vol. 32, no. 5, pp. 80–88, 2025.

[32] L. Qiao, M. Boloursaz Mashhadi, Z. Gao, R. Schober, and D. Gündüz, "ToDMA: Large model-driven massive token communications for semantic multiple access," 2026, arXiv preprint, version 3 (9 July 2026), at https://arxiv.org/abs/2505.10946v3.

[33] A. Alkhateeb, G. Charan, T. Osman, A. Hredzak, J. Morais, U. Demirhan, and N. Srinivas, "DeepSense 6G: A large-scale real-world multi-modal sensing and communication dataset," *IEEE Commun. Mag.*, vol. 61, no. 9, pp. 122–128, 2023.

[34] X. Cheng, Z. Huang, L. Bai, H. Zhang, M. Sun, B. Liu, S. Li, J. Zhang, and M. Lee, "M3SC: A generic dataset for mixed multi-modal sensing and communication integration," *China Commun.*, vol. 20, no. 11, pp. 13–29, 2023.

[35] C. Studer, S. Medjkouh, E. Gönültaş, T. Goldstein, and O. Tirkkonen, "Channel charting: Locating users within the radio environment using channel state information," *IEEE Access*, vol. 6, pp. 47 682–47 698, 2018.

[36] H. Jiang, M. Cui, D. W. K. Ng, and L. Dai, "Accurate channel prediction based on transformer: Making mobility negligible," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2717–2732, 2022.

[37] Z. Zhang, J. Zhang, Y. Zhang, L. Yu, and G. Liu, "AI-based time-, frequency-, and space-domain channel extrapolation for 6G: Opportunities and challenges," *IEEE Veh. Technol. Mag.*, vol. 18, no. 1, pp. 29–39, 2023.

[38] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 15 619–15 629.

[39] C. Bou Chaaya, A. M. Girgis, and M. Bennis, "Learning latent wireless dynamics from channel state information," *IEEE Wireless Commun. Lett.*, vol. 14, no. 2, pp. 489–493, 2025.

[40] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.

[41] F. Jiang, L. Dong, L. Mao, K. Wang, C. Pan, D. I. Kim, and N. Al-Dhahir, "Vision-language-action models meet world models: Embodied agentic AI for low-altitude wireless networks," 2026, preprint at https://arxiv.org/abs/2606.11618.

[42] Y. Zeng, J. Chen, J. Xu, D. Wu, X. Xu, S. Jin, X. Gao, D. Gesbert, S. Cui, and R. Zhang, "A tutorial on environment-aware communications via channel knowledge map for 6G," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 3, pp. 1478–1519, 2024.

[43] R. Levie, C. Yapar, G. Kutyniok, and G. Caire, "RadioUNet: Fast radio map estimation with convolutional neural networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 6, pp. 4001–4015, 2021.

[44] T. Beemelmanns, Q. Zhang, C. Geller, and L. Eckstein, "MultiCorrupt: A multi-modal robustness dataset and benchmark of LiDAR-camera fusion for 3D object detection," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, 2024, pp. 3255–3261.

[45] H. Wang, J. Zhang, G. Nie, L. Yu, Z. Yuan, T. Li, J. Wang, and G. Liu, "Digital twin channel for 6G: Concepts, architectures and potential applications," *IEEE Commun. Mag.*, vol. 63, no. 3, pp. 24–30, 2025.

[46] J. Wang, J. Zhang, Y. Zhang, Y. Sun, G. Nie, L. Shi, P. Zhang, and G. Liu, "Radio environment knowledge pool for 6G digital twin channel," *IEEE Commun. Mag.*, vol. 63, no. 5, pp. 158–164, 2025.

[47] M. Ying, D. Shakya, P. Ma, G. Qian, and T. S. Rappaport, "Site-specific location calibration and validation of ray-tracing simulator NYURay at upper mid-band frequencies," *npj Wireless Technology*, vol. 2, p. 8, 2026.

[48] J. Hoydis, F. Ait Aoudia, S. Cammerer, F. Euchner, M. Nimier-David, S. ten Brink, and A. Keller, "Learning radio environments by differentiable ray tracing," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 2, pp. 1527–1539, 2024.

[49] S. Jiang and A. Alkhateeb, "Digital twin based beam prediction: Can we train in the digital world and deploy in reality?" in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2023, pp. 36–41.

[50] H. Luo, S. R. Khosravirad, and A. Alkhateeb, "Wireless digital twin calibration: Refining DFT-domain channel information," 2026, preprint at https://arxiv.org/abs/2603.16126.

[51] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 59, pp. 1–35, 2016.

[52] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 70, 2017, pp. 1126–1135.

[53] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[54] F. H. Kumbhar, K. Dev, S. A. Khowaja, A.-A. A. Boulogeorgos, M. Bennis, and Y. Liu, "Physical-AI: From channel awareness to environmental intelligence in 6G wireless networks," 2026, preprint at https://arxiv.org/abs/2606.20583.

[55] Z. Li, Z. Gao, X. Liu, Z. Wang, X. Zhou, L. Liu, Y. Wu, W. Feng, and Y. Huang, "Large model enabled embodied intelligence for 6G integrated perception, communication, and computation network," *Chinese Journal of Electronics*, vol. 35, no. 5, pp. 1–25, 2026, online First (uncorrected proof).

[56] Y. Liu, X. Liu, X. Mu, T. Hou, J. Xu, M. Di Renzo, and N. Al-Dhahir, "Reconfigurable intelligent surfaces: Principles and opportunities," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1546–1577, 2021.

[57] S. R. Hasan, M. Hamamura, and S. R. Sabuj, "Modeling of a software-defined reflective intelligent surface network using queueing theory," *IEEE Access*, vol. 14, pp. 7543–7556, 2026.

[58] H. Zhang, S. Gao, and X. Cheng, "Tiny-WiFo: A lightweight wireless foundation model for channel prediction via multi-component adaptive knowledge distillation," *IEEE Wireless Communications Letters*, vol. 15, pp. 1846–1850, 2026.

[59] O. Mashaal and H. Abou-Zeid, "Fast wireless foundation models with early exits," 2026, preprint at https://arxiv.org/abs/2606.29640.

[60] Y. Wang, L. Sun, T. Yang, L. Jing, Y. Shi, M. Elkashlan, and M. Debbah, "ConsisFormer: Compute-efficient transformer for wireless foundation models based on channel consistency," 2026, preprint at https://arxiv.org/abs/2606.19953.

[61] F. O. Catak, M. Kuzlu, E. Catak, U. Cali, and O. Guler, "Defensive distillation-based adversarial attack mitigation method for channel estimation using deep learning models in next-generation wireless networks," *IEEE Access*, vol. 10, pp. 98 191–98 203, 2022.

[62] Y. Wan, Y. Qu, W. Ni, Y. Xiang, L. Gao, and E. Hossain, "Data and model poisoning backdoor attacks on wireless federated learning, and the

defense mechanisms: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 3, pp. 1861–1897, 2024.

[63] D. Avola, M. Cascio, L. Cinque, A. Fagioli, and C. Petrioli, "Person re-identification through Wi-Fi extracted radio biometric signatures," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 1145–1158, 2022.

[64] K. M. Cohen, S. Park, O. Simeone, and S. Shamai, "Calibrating AI models for wireless communications via conformal prediction," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 1, pp. 296–312, 2023.

[65] 3GPP, "Study on 6G radio," 3rd Generation Partnership Project (3GPP), Tech. Rep. TR 38.960, 2026, release 20, draft specification record; related Work Item FS_6G_Radio (WI 1080072). Accessed 12 September 2026. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=4472

[66] ——, "Release-20 AI/ML work items for the NR air interface and NG-RAN," https://portal.3gpp.org/desktopmodules/WorkItem/WorkItemDetails.aspx?workitemId=1080085 and https://portal.3gpp.org/desktopmodules/WorkItem/WorkItemDetails.aspx?workitemId=1110099, 2026, nR_AIML_air_Ph2 (WI 1080085) and NR_AIML_NGRAN_Ph3 (WI 1110099), PCG-approved Release-20 work-item records. Accessed 12 September 2026.

[67] ——, "Study on the protocol for artificial intelligence in 6G," 3rd Generation Partnership Project (3GPP), Tech. Rep. TR 29.832, 2026, release 20, draft specification record; related Work Item FS_6G_AIP_CT (WI 1120022). Accessed 12 September 2026. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=5548

[68] O-RAN Alliance, "O-RAN alliance completed its specification release 5 (O-RAN-R005)," Jun. 2026, accessed September 2026. [Online]. Available: https://www.o-ran.org/blog/o-ran-alliance-completed-its-specification-release-5-o-ran-r005

[69] ——, "71 new or updated O-RAN technical documents released since november 2025," https://www.o-ran.org/blog/71-new-or-updated-o-ran-technical-documents-released-since-november-2025, 2026, specification updates announced 27 February 2026; R1 General Aspects and Principles v13.00, R1 Test Specification v5.00, A1 Application Protocol v6.00, and Near-RT RIC APIs v3.00. Accessed 12 September 2026.