ConTwin: Contrastive Learning for Robust Digital Twin CSI Prediction

Sagnik Bhattacharya*
Department of Electrical Engineering
Stanford University
Stanford, CA 94305
sagnikb@stanford.edu

Abhiram Gorle*
Department of Electrical Engineering
Stanford University
Stanford, CA 94305
abhiramg@stanford.edu

John M. Cioffi
Department of Electrical Engineering
Stanford University
Stanford, CA 94305
cioffi@stanford.edu

Abstract

Despite recent progress in deep learning for wireless systems, current *digital twin–based* CSI prediction models often struggle with two core challenges: poor generalization under network distribution shifts and high retraining costs. These issues stem from heavy reliance on synthetic data, overparameterized model designs, and rigid update mechanisms. We introduce **ConTwin**, a contrastive learning framework tailored for *6G digital twin* environments, that enables efficient and generalizable CSI prediction. ConTwin leverages digital twin simulations to construct domain-aware positive and hard negative pairs, enabling the model to learn representations that are robust across varying user mobility, LOS/NLOS scenarios, and antenna configurations. Experiments on 3GPP-aligned benchmarks demonstrate that ConTwin improves CSI prediction NMSE by up to 5.4 dB under distribution shift and by up to 1.2 dB in-distribution, outperforming leading baselines such as SwinCFNet and CNN-based models. These results highlight ConTwin's potential as a foundational component for robust, data-driven 6G digital twin networks.

1 Introduction

In next-generation 6G systems, accurately *predicting* future channel state information (CSI) is vital for enabling sub-ms latency, cm-level positioning, & high reliability in applications like XR, AR, and autonomous vehicles (1). While traditional pilot-aided methods (Kalman Filtering, MMSE) are effective for estimating instantaneous CSI (2), they lack predictive capability, especially under deep fades or NLOS conditions. Accurate **ahead-of-time** CSI is crucial in applications like indoor positioning, AR/VR gesture recognition, and vehicular beam selection (3). For example, when a user enters a parking structure, loses line-of-sight (LOS) for hundreds of milliseconds, the absence of reliable CSI prediction could prevent the transceiver from accurately tracking position.

To address the limitations of pilot-based methods, deep learning has advanced CSI prediction in wireless systems (4; 5). Multi-frame temporal modeling methods and digital twin frameworks enable zone-specific subspace calibration with 40% lower feedback vs. MMSE (6). In high-Doppler

*Equal contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI and ML for Next-Generation Wireless Communications and Networking.

scenarios, CNN-LSTM hybrids outperform Kalman filters by 5.8 dB NMSE, while contrastive learning aids RIS beamforming under partial CSI (7). However, UE-side ML remains a challenge: BS-based prediction (CSILaBS) cuts UE power use by 63% with 92% accuracy (8). Despite this progress, three core barriers still hinder deployment in commercial-grade UE hardware as follows:

1. **Train–test mismatch:** Synthetic engines (Quadriga, Sionna, MATLAB 3GPP toolbox) assume idealized settings (fixed downtilt, urban macro, static users), but real-world deviations (e.g: 15° downtilt shift) degrade accuracy by 7-10% (9; 10; 11), Fig. **??** shows CSI power-delay profiles in such cases.

2. **Heavy inference cost:** Many models exceed 10 GFLOPs/frame and tens of millions of parameters, straining mobile SoCs with limited power and memory.

3. **Retraining overhead:** Even BS-pretrained models need periodic updates, requiring uplink transmissions of raw CSI or gradients—incurring high overhead.

Contrastive learning (CL), a self-supervised paradigm, derives supervision directly from data unlike supervised deep CSI models that depend on costly, non-generalizable ray-tracing labels. By treating *adjacent* CSI samples as positives & distant ones as negatives, CL captures temporal dynamics and leverages abundant unlabeled field logs for accurate 6G CSI prediction–vital since field logs are plentiful yet unlabeled and synthetic labels seldom generalize.

Related Work: Contrastive learning (CL) has shown promise in wireless tasks such as hybrid beamforming at THz (improving rate and robustness) (12), RIS-assisted MU-MIMO (28% spectral efficiency gain) (7), and Wi-Fi-based sensing from unlabeled CSI (13). However, its use for robust, predictive CSI modeling remains underexplored, which this work aims to address.

Contributions of ConTwin. This work proposes *ConTwin*, where digital-twin channel engine spans a *family* of network configurations; hard-negative mining couples CSI samples drawn from disparate domains (e.g. mobile vs. static, LOS vs. NLOS) to learn a domain-robust representation via contrastive loss. Extensive Sionna-raytracing simulations under realistic 3GPP NR settings confirm that ConTwin improves CSI prediction NMSE by up to 5.4 dB under train-test network distribution mismatch and up to 1.2 dB in-distribution, compared to strong baselines like SwinCFNet(14) and CNN-based predictors, demonstrating enhanced robustness to mobility, downtilt, and LOS/NLOS shifts. The remainder of the paper details the system model (2), the channel model (3), a statistical analysis of CSI data (4), the contrastive pre-training pipeline (5), and a thorough performance evaluation (6).

2 System Model

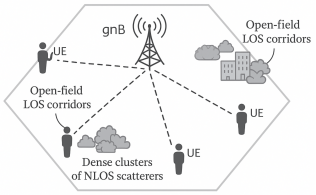


Figure 1: System Model

Fig. 1 depicts a single-cell downlink with a gNB at the center of a hexagonal site serving U UEs distributed across mixed rural–urban LOS corridors and dense NLOS scatterers (e.g., city blocks, vegetation). We assume $f_c = 3.5\,\text{GHz}$, $B = 100\,\text{MHz}$, and OFDM with K equally-spaced subcarriers. Further details regarding the system model can be found in Appendix A.

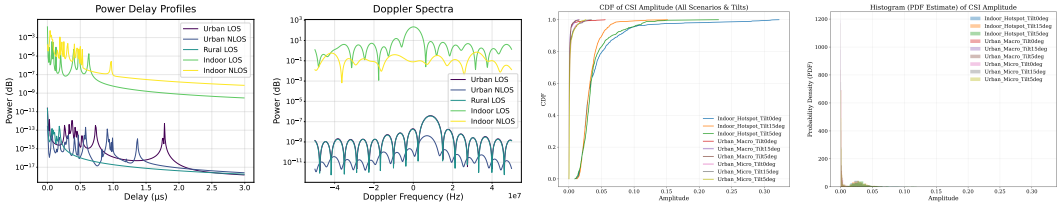
CSI Acquisition Baseline: Downlink CSI is acquired via orthogonal CSI-RS every $T_p = 5\,\text{ms}$, with each UE feeding back a PMI/RI/CQI tuple (≤ 200 bits) under an end-to-end delay $\delta_{\text{fb}} \approx 7\,\text{ms}$ per NR standards (15). These serve as ground-truth labels for our CL-based prediction pipeline (Sections 5–6).

3 Channel Model

The *ConTwin* simulations use a geometry-based stochastic wideband model compatible with `Quadriga 3.6` and `Sionna`. For each user u, the time-variant MIMO IR is

$$\mathbf{H}_u(t,\tau) = \sum_{c=1}^{N_c}\sum_{l=1}^{L_c} \alpha_{ucl}\, e^{j2\pi f_{ucl}(.)}\, \mathbf{a}_t(\boldsymbol{\phi}_{ucl})\mathbf{a}_r^{\mathsf{H}}(\boldsymbol{\theta}_{ucl})\delta(\tau-\tau_{ucl}), \tag{1}$$

where N_c is the number of clusters, L_c the rays per cluster, $\alpha_{ucl} \sim \mathcal{CN}(0,\beta_{ucl})$ the path gain, $f_{ucl} = \frac{v_u f_c \cos\vartheta_{ucl}}{c}$ the Doppler shift at speed v_u, $\mathbf{a}_t, \mathbf{a}_r$ unit-norm TX/RX array responses at azimuth/elevation $\boldsymbol{\phi}_{ucl}, \boldsymbol{\theta}_{ucl}$, and τ_{ucl} ray delay. Now, we look at CSI modeling scenarios in Fig. 2.



(a) Power delay profile (b) Doppler spectra (c) CDF estimate of CSI (d) PDF estimate of CSI

Figure 2: Various Contrastive CSI Modeling Scenarios and CSI Statistics (across all scenarios & LOS/NLOS)

Further details regarding the channel model can be found in Appendix B.

4 Statistical Analysis of CSI Distribution

Prior work in **ergodic spectrum management** (16) establishes foundations for modeling time-varying channel dynamics, an essential precursor to learning-based methods for CSI prediction. This section outlines the synthetic CSI generation framework and analyzes amplitude distributions across downtilt angles, propagation scenarios, and other key parameters.

4.1 Synthetic Data Generation

User Position Sampling: The model uniformly samples N user positions (x_i, y_i) within a circular cell of radius R. Positions are in polar coordinates: $r_i = R\sqrt{u_i}$, $\theta_i = 2\pi v_i$, $(x_i, y_i) = (r_i\cos\theta_i, ; r_i\sin\theta_i)$, where $u_i, v_i \sim \mathcal{U}(0,1)$.

CSI Computation: For each downtilt angle α, the 3D distance from a base station (BS) at height h to user i is $d_i = \sqrt{x_i^2 + y_i^2 + h^2}$. Under narrowband assumptions, the received power follows the Friis transmission equation: $P_{\text{rx},i} = P_{\text{tx}} \cdot G(\theta_{\text{el},i}, \alpha) \cdot \left(\frac{\lambda}{4\pi d_i}\right)^2$, where $\lambda = c/f_c$ is the carrier wavelength, P_{tx} is the BS transmit power, and $G(\cdot)$ is a simplified antenna gain model with $G = \exp\left(-\frac{(\theta_{\text{el},i}-\alpha)^2}{2\sigma_g^2}\right)$, and $\theta_{\text{el},i} = \arctan 2\left(\sqrt{x_i^2+y_i^2}, ; h\right)$, $\sigma_g = 0.1$ rad. Small-scale fading is incorporated by multiplying $P_{\text{rx},i}$ by a complex Gaussian coefficient $h_i \sim \mathcal{CN}(0,1)$, $|h_i| \sim \text{Rayleigh}(1/\sqrt{2})$. The instantaneous amplitude and phase per user link are saved: $r_i = |h_i|\sqrt{P_{\text{rx},i}}$, $\phi_i \sim \mathcal{U}(0, 2\pi)$.

Parameter Configurations: Table 3 (Appendix) shows the simulation configurations (via QuaDRiGa (9)).

Amplitude: For freq.-domain CSI, per-link amplitudes are computed via norm across antennas & subcarriers as $\text{amp}_i = \sqrt{\sum_{r,t,s} |H_{r,t,s,i}|^2}$, where r, t, s index RX, TX antennas and subcarriers.

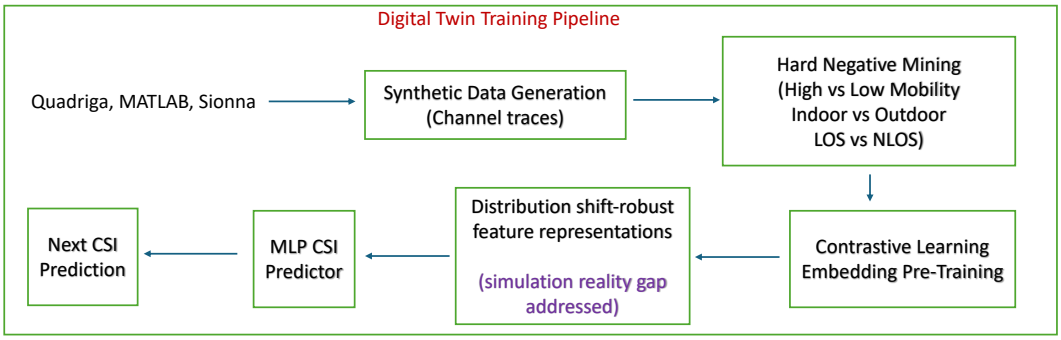


Figure 3: Contrastive pre-training pipeline. Synthetic CSI tensors are domain-labelled in the digital twin, augmented into two views, embedded by f_ϕ, and processed by the projection head g_ψ. The NT-Xent loss aligns positive pairs while repelling *hard negatives* drawn from dissimilar domains.

4.2 Data Augmentation for Contrastive Learning

To simulate realistic channel variability, results generate augmented datasets by applying some physical "transforms" to the QuaDRiGA-generated CSI data like i) adding phase jitter (ϕ_i) and ii) random scaling factors $\sim \mathcal{U}[0.9, 1.1]$. Such augmentations support contrastive learning by creating positive-negative pairs under physically plausible perturbations. Representative CDF and PDF plots for each scenario are presented in Figure 2, along with brief observations on distribution shifts.

4.3 Observations on Empirical Distributions

Fig. 2c shows ordered amplitude CDFs: indoor hotspots rise steeply (narrower dynamic range), followed by urban micro & macro. In macro scenarios, greater downtilt shifts CDFs left: indicating reduced edge coverage but stronger near-BS signals while micro cells show minimal downtilt impact due to shorter links. In Fig. 2d, amplitude histograms exhibit Rayleigh-like behavior at lower values, with right-skew due to LoS components in urban settings. Goodness-of-fit tests (p-values > 0.05) confirm Rayleigh fits for indoor/micro scenarios, while Rice distributions ($b \approx 0.5$) better capture macro LoS. These results validate the diversity of our contrastive-learning dataset and highlight the limitations of parametric models, motivating a learning-based approach for complex environments.

Fig. 4 (Appendix) shows a **7%** average performance drop when models are deployed under mismatched network conditions, especially with spatial distribution shifts. This **train-test mismatch** motivates the use of self-supervised approaches like contrastive pre-training, which outperform traditional supervised methods in such scenarios. The statistical analyses in this section reveal substantial distributional diversity across scenarios, indoor vs. outdoor, LOS vs. NLOS, and varying mobility patterns, validating the richness of our contrastive learning dataset. Moreover, the observed 7% performance degradation under distribution shift (Fig. 4) motivates our approach: rather than training separate models for each deployment scenario, we seek a unified representation that captures domain-invariant channel characteristics. The following section details our contrastive learning pipeline, which explicitly addresses these distribution shifts through hard-negative mining and domain-aware loss formulation.

5 Contrastive–Learning Pipeline for Distribution Shift-Robust CSI Prediction

This section details the proposed ConTwin framework: the contrastive self-supervised pre-training stage that endows the CSI encoder with strong *domain invariance* before CSI prediction. The overall flow is summarised in Fig. 3; the individual blocks are formalised below.

Digital-Twin Data Generation & Domain Labelling: A ray–tracing digital twin (`QuaDRiGa`, `Sionna`) renders channel realisations under a Cartesian product of *domain axes* $\mathcal{A} = \{a^{\mathtt{sc}}, a^{\mathtt{los}}, a^{\mathtt{mob}}, a^{\mathtt{tilt}}, a^{\mathtt{freq}}, \dots\}$, where, e.g., $a^{\mathtt{sc}} \in \{\text{urban-macro}, \text{urban-micro}, \text{indoor}, \text{rural}\}$, $a^{\mathtt{los}} \in \{\text{LOS}, \text{NLOS}\}$, $a^{\mathtt{mob}} \in \{\text{static}, \text{pedestrian}, \text{vehicular}\}$, $a^{\mathtt{tilt}} \in \{0^\circ, 5^\circ, 10^\circ, 15^\circ\}$, $a^{\mathtt{freq}} \in \{3.5,\ 28,\ 60\ \text{GHz}\}$.

A *domain label* is the ordered tuple $d = \big(a^{\mathtt{sc}}, a^{\mathtt{los}}, a^{\mathtt{mob}}, a^{\mathtt{tilt}}, a^{\mathtt{freq}}, \ldots\big) \in \mathcal{D}$. Throughout, this notation abbreviates $\mathrm{sim}(d_i, d_j) = \sum_{a\in\mathcal{A}} w_a[a_i = a_j]$, a weighted Hamming similarity with tunable weights $w_a \geq 0$.

Augmentation & Positive-Pair Definition: Each CSI tensor is transformed into two stochastic *views* $\tilde{\mathbf{H}}^{(1)}$ and $\tilde{\mathbf{H}}^{(2)}$ via a stochastic composition of: 1) complex-valued Gaussian jitter, 2) frequency mask (sub-carrier dropout), 3) time shifting, 4) antenna permutation, and 5) phase rotation. We consider a pair $\big(\tilde{\mathbf{H}}^{(1)}, \tilde{\mathbf{H}}^{(2)}\big)$ *positive* if $\mathrm{sim}(d_i, d_j) \geq \rho$, where $\rho = 5$ in our experiments.

Hard Positive Generation: To construct positive pairs, we sample channel realizations from the same user within a 25 ms window, and occasionally from different users with similar mobility, Rician K factor, and antenna tilt. However, such pairs are often overly similar pre-training, limiting contrastive effectiveness. To mitigate this, we apply CSI-specific augmentation via random phase jitter, injecting offsets $\theta \sim \mathcal{N}(0, \sigma^2)$ to promote more robust and generalizable representations.

Scenario-Driven Negative-Pair Generation: Negative pairs are designed to present *hard distribution shifts*. Let Δ_a be a distance metric on axis a; we mark two domain labels (d_i, d_j) as an *adversarial negative* if

$$\sum_{a\in\mathcal{A}} \lambda_a \, \mathbf{1}\{\Delta_a(a_i, a_j) > \tau_a\} \;\geq\; 1, \tag{2}$$

with tunable $\lambda_a \in [0, 1]$ and thresholds τ_a. Some examples are shown in Table 4, along with LOS vs. NLOS, and scenario mismatch. The union of such adversarial dimensions yields the *candidate pool* for domain label d_i, $\mathcal{C}(i) = \{j \mid (d_i, d_j)$ satisfy (2)$\}$.

Hard-Negative Mining: Within $\mathcal{C}(i)$, we perform **distance-aware mining**. Let $\mathbf{q}_i = g_\psi(\mathbf{z}_i)$ be the ℓ_2-normalised projection. Define the angular distance $\delta_{ij} = 1 - \mathbf{q}_i^\top \mathbf{q}_j$. The top–$H$ samples are chosen as $\mathcal{N}(i) = \operatorname{argmax}_{\substack{\mathcal{S}\subseteq\mathcal{C}(i)\\ |\mathcal{S}|=H}} \sum_{j\in\mathcal{S}} \delta_{ij},$, i.e. the farthest representations in the current batch, following the hard-negative mining principle in (17). Empirically, it is observed that $H = B/8$ yields stable gradients without memory explosion. Negative pair criteria are outlined further in Table 4 (Appendix).

Domain-Aware NT-Xent Loss: To train our contrastive encoder across heterogeneous propagation scenarios, we adopt a multi-positive, hard-negative NT-Xent objective from (18). For any anchor i, let $\rho_{ij} = \exp(\frac{\mathbf{q}_i^\top \mathbf{q}_j}{\tau})$ where $\mathbf{q} = g_\psi(\mathbf{z})$ is the L_2-normalized projection head and $\tau > 0$ is a learnable temperature parameter. We denote by $\mathcal{P}(i)$ the set of all positives for anchor i (including both intra- and inter-domain views) and by $\mathcal{N}(i)$ a curated set of hard negatives mined from other domains. Our *domain-aware NT-Xent* loss over a batch of size B is:

$$\mathcal{L}_{\mathrm{CL}} \;=\; \frac{1}{B}\sum_{i=1}^{B}\Bigg[-\sum_{j\in\mathcal{P}(i)} \frac{1}{|\mathcal{P}(i)|} \log \frac{\varrho_{ij}}{\sum\limits_{k\in\mathcal{P}(i)\cup\mathcal{N}(i)} \varrho_{ik}}\Bigg] \tag{3}$$

Compared to SimCLR (18), our formulation (i) supports multiple positives ($|\mathcal{P}(i)| \geq 2$) with uniform averaging to stabilize training across domains, (ii) incorporates top-K hard negatives ($K = 50$) using cosine similarity, and (iii) learns the temperature τ (initialized at 0.1) to improve early clustering.

Training Procedure: The training pipeline consists of two phases:

1. **Phase I: Contrastive CSI Encoding:** Let f_ϕ be our CNN–Transformer encoder (14). We train f_ϕ for T_1 epochs by minimizing the domain-aware NT-Xent loss in (3) using AdamW with learning rate $\eta = 10^{-3}$ and cosine decay. A large effective batch size ($B = 512$) is achieved via cross-GPU hard-negative mining.
2. **Phase II: Next-CSI Prediction:** We freeze the encoder f_ϕ and append a lightweight prediction head h_ϑ, consisting of four fully-connected layers. We then minimize the mean-squared error $\mathcal{L}_{\mathrm{MSE}} = \big\| h_\vartheta\big(f_\phi(\mathbf{H}_{\text{input}})\big) - \mathbf{H}_{\text{target}}\big\|_2^2$ over a small labeled subset ($< 5\%$ of the data), achieving label efficiency comparable to semi-supervised SimCLR (18).

Theoretical Insight: Under the InfoNCE framework, minimizing (3) provably maximizes a lower bound on the mutual information $I(\mathbf{z}; d)$ while *minimally* encoding domain-specific nuisance factors (19; 20). Recent theoretical work demonstrates that contrastive learning naturally induces domain-invariant features (21), with representation distance between domains bounded by the H-divergence

(22). Our hard-negative mining strategy specifically targets samples near domain boundaries (e.g., transitional LOS/NLOS conditions), which tightens the InfoNCE bound in regions most critical for OOD generalization (23). Furthermore, our multi-positive averaging (|P(i)| 2) provides gradient stabilization and improved robustness, as formally established in supervised contrastive learning (24). Together, these theoretical foundations explain the empirically observed 5.4 dB NMSE improvement under distribution shift: the learned representation captures domain-invariant multipath structure, temporal correlation, and angular characteristics while remaining robust to domain-specific factors like absolute downtilt angle or instantaneous LOS/NLOS state. The hard-negative curriculum further tightens the bound in high-SNR regions, leading to the empirically observed $7{-}10\%$ NMSE reduction under antenna-tilt shifts reported. Having detailed the contrastive pre-training methodology, we now empirically validate ConTwin's robustness to distribution shifts across diverse 6G deployment scenarios.

6 Performance Evaluation

This section quantifies the gains of the proposed contrastive–learning pipeline against four baselines across a wide spectrum of deployment scenarios.

6.1 Simulation Platform and Parameter Suite

Digital–Twin Channel Generation We employ **NVIDIA Sionna 1.1** to synthesise both outdoor 3GPP TR 38.901 (15) and indoor IEEE 802.11ax (25) channels. Table 5 (Appendix) lists the used experimental parameters. Each scenario generates 2×10^5 snapshot triples $\big(\mathbf{H}, \widetilde{\mathbf{H}}, \text{label}\big)$, where $\widetilde{\mathbf{H}}$ is the ground truth next timestep CSI.

Implementation: Model training for both contrastive learning based CSI embedding generator, and next CSI prediction are carried out in *PyTorch 2.1* on four A10 GPUs for 500 epochs. All implementation hyperparameters are outlined in Table 6 (Appendix).

6.2 Baselines and Metrics

- **B1: Swin Transformer** (14) proposes a novel CSI compression and feedback network called SwinCFNet, which leverages Swin Transformer blocks to extract long-range dependencies in the angular-delay domain of the CSI matrix. The core contribution is a two-stage autoencoder architecture with shifted window-based multi-head self-attention, enabling efficient CSI encoding/decoding that significantly outperforms prior CNN or Transformer-based baselines.
- **B2: in-distribution CSI predictors**: A CNN-transformer based architecture is trained for supervised CSI prediction in-distribution, i.e. the training distribution (e.g., indoor, LOS, stationary users) is same as the inference distribution. Since the training and test distributions match, this architecture will give the best performance, and forms the upper bound on the proposed algorithm. Of course, in any practical scenarios, this is unrealistic, since we do not know the exact network distribution beforehand.
- **B3: OOD CSI predictor** where a CNN-transformer based model is trained for CSI prediction on OOD data.

6.3 Robustness to Domain Shifts

Tables 1 and 2 present the NMSE performance (in dB) of various CSI prediction methods across in-distribution and out-of-distribution (OOD) scenarios. In the in-distribution case, where the training and test distributions are aligned (e.g., same downtilt, velocity, and LOS/NLOS conditions), all methods perform well. The proposed algorithm, which achieves indoor and outdoor NMSEs of -29.10 dB and -22.10 dB, respectively, outperforms SwinCFNet, which achieves NMSEs of -28.94 dB (indoor) and -21.68 dB.

However, under distribution shifts—such as a 15° antenna downtilt added during inference, high mobility (30 m/s) inference as opposed to stationary training, or LOS$\leftrightarrow$NLOS changes—baseline performance deteriorates sharply. For instance, SwinCFNet drops to about -17.80 dB and simple CNN+Transformer baselines trained on different distribution degrade in performance. In contrast, our CL+KD approach remains highly robust, achieving NMSEs of about -23.50 dB and better across

all OOD conditions. This consistent generalization demonstrates the strength of our contrastive representation learning, which captures semantically aligned latent features invariant to domain shifts. By combining stability, scalability, and strong generalization to OOD network data during inference, the proposed method establishes a new state-of-the-art in robust CSI prediction.

Table 1: Average NMSE (dB) for CSI Prediction in In-Distribution Settings

Method	**Indoor LOS** Stationary	**Outdoor NLOS** Moving
B1 SwinCFNet (14)	-28.94	-21.68
B2 CNN+Transformer (In-Dist)	**-29.20**	**-22.30**
Ours (In-Dist)	-29.10	-22.10

Table 2: NMSE (dB) for CSI Prediction in Out-of-Distribution (OOD) Settings

Method	**Downtilt** 15°	**High Mobility (30 m/s)**	**LOS/NLOS Shift**
B1	-17.76	-17.30	-16.60
B2	-16.85	-16.38	-15.80
B3	-17.12	-16.50	-15.87
Ours	**-23.51**	**-22.91**	**-22.32**

Results of further ablation studies to show the efficacy of hard-negative mining and longer (temporal) window can be found in Appendix D.

7 Conclusion and Future Work

In contrast to prior supervised or compression-based CSI methods, our proposed ConTwin framework incorporates contrastive learning to learn domain-invariant embeddings that remain robust under distribution shifts such as mobility, downtilt variation, and LOS↔NLOS transitions. Compared to strong baselines like SwinCFNet and CNN-based predictors, ConTwin consistently achieves lower NMSE in both in-distribution and out-of-distribution settings—establishing a new state-of-the-art. These gains are crucial for next-generation wireless networks, where CSI models must generalize reliably across diverse deployment environments without costly retraining or hardware overhead.

References

[1] J. G. Andrews, T. E. Humphreys, and T. Ji, "6g takes shape," *IEEE BITS the Information Theory Magazine*, vol. 4, p. 2–24, Mar. 2024.

[2] L. Tong, B. M. Sadler, and M. Dong, "Pilot-assisted wireless transmissions: General model, design criteria, and signal processing," *IEEE Signal Processing Magazine*, vol. 21, no. 6, pp. 12–25, 2004.

[3] N. Garg, I. Shahid, K. Sankar, M. Dasari, R. K. Sheshadri, K. Sundaresan, and N. Roy, "Bringing ar/vr to everyday life - a wireless localization perspective," in *Proceedings of the 24th International Workshop on Mobile Computing Systems and Applications*, HotMobile '23, (New York, NY, USA), p. 142, Association for Computing Machinery, 2023.

[4] M. Nerini, V. Rizzello, M. Joham, W. Utschick, and B. Clerckx, "Machine learning-based csi feedback with variable length in fdd massive mimo," *IEEE Transactions on Wireless Communications*, vol. 21, 2022.

[5] D. Burghal, Y. Li, P. Madadi, Y. Hu, J. Jeon, J. Cho, A. F. Molisch, and J. Zhang, "Enhanced ai-based csi prediction solutions for massive mimo in 5g and 6g systems," *IEEE Access*, vol. 11, pp. 117810–117825, 2023.

[6] S. Alikhani and A. Alkhateeb, "Digital twin aided channel estimation: Zone-specific subspace prediction and calibration," *arXiv preprint arXiv:2501.02758*, 2025.

[7] Z. He, F. Héliot, and Y. Ma, "Self-supervised contrastive learning for joint active and passive beamforming in ris-assisted mu-mimo systems," *arXiv preprint arXiv:2401.12345*, 2025.

[8] M. K. Shehzad, L. Rose, and M. Assaad, "Massive mimo csi feedback using channel prediction: How to avoid machine learning at ue?," *arXiv preprint arXiv:2403.13363*, 2024.

[9] S. Jaeckel, L. Raschkowski, K. Börner, and L. Thiele, "QuaDRiGa: A 3-D Multi-Cell Channel Model with Time Evolution for Enabling Virtual Field Trials," *IEEE Transactions on Antennas and Propagation*, vol. 62, no. 6, pp. 3242–3256, 2014.

[10] J. Hoydis and et. al, "Sionna RT: Differentiable Ray Tracing for Radio Propagation Modeling," *arXiv preprint arXiv:2303.11103*, 2023.

[11] 3GPP RAN1, "R1-240xxxx Chair notes RAN1#119 (9.4 R19 Ambient IoT) v03," 2024. Qualcomm contribution to 3GPP RAN1 #119.

[12] R. U. Murshed, M. S. Ullah, M. Saquib, and M. Z. Win, "Self-supervised contrastive learning for 6g um-mimo thz communications: Improving robustness under imperfect csi," 2024.

[13] K. Xu, J. Wang, H. Zhu, and D. Zheng, "Self-supervised learning for wifi csi-based human activity recognition: A systematic study," 2023.

[14] J. Cheng, W. Chen, J. Xu, Y. Guo, L. Li, and B. Ai, "Swin transformer-based csi feedback for massive mimo," 2024.

[15] 3rd Generation Partnership Project (3GPP), "Study on channel model for frequencies from 0.5 to 100 GHz (Release 16)," Tech. Rep. TR 38.901, ETSI, 2020. Version 16.1.0.

[16] J. M. Cioffi, C.-S. Hwang, and K. J. Kerpez, "Ergodic spectrum management," *IEEE Transactions on Communications*, vol. 68, no. 3, pp. 1794–1821, 2020.

[17] J. Robinson, C.-Y. Chuang, J. Sohl-Dickstein, T. C. Lin, K. Liu, and Q. V. Le, "Debiased contrastive learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 8765–8775, 2021.

[18] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[19] S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi, "A theoretical analysis of contrastive unsupervised representation learning," in *ICML*, 2019.

[20] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning?," in *NeurIPS*, 2020.

[21] C.-H. Yeh, C.-Y. Hong, Y.-C. Hsu, T.-L. Liu, Y. Chen, and Y. LeCun, "Decoupled contrastive learning," in *ECCV*, 2022.

[22] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, pp. 151–175, 2010.

[23] J. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka, "Contrastive learning with hard negative samples," in *ICLR*, 2021.

[24] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *NeurIPS*, 2020.

[25] "IEEE Standard for Information technology—Telecommunications and information exchange between systems Local and metropolitan area networks—Specific requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications - Amendment 6: Enhancements for High Efficiency WLAN," Tech. Rep. IEEE Std 802.11ax-2021, May 2021.

[26] S. Sun, G. R. M. Jr., and T. S. Rappaport, "A novel millimeter-wave channel simulator and applications for 5g wireless communications," *IEEE International Conference on Communications (ICC)*, pp. 1–7, 2017.

[27] P. Kyösti, J. Meinilä, L. Hentilä, X. Zhao, T. Jämsä, A. Alatossava, R. Bultitude, Y. Jiang, W. Fan, and T. Rautiainen, "WINNER II Channel Models," Tech. Rep. D1.1.2 V1.2, IST-4-027756 WINNER II, September 2007.

Appendix

A. More on System Model

We assume $f_c = 3.5\,\text{GHz}$, $B = 100\,\text{MHz}$, and OFDM with K equally-spaced subcarriers.

Base-station array, radio frame and notation: The gNB employs a uniform planar array of $N_t = N_x N_y$ isotropic elements with half-wavelength spacing (default $N_x = N_y = 8$, $N_t = 64$). A radio frame of duration $T_\text{f} = 10\,\text{ms}$ contains N_s OFDM symbols; subcarrier k in symbol ℓ experiences channel vector $\mathbf{h}_{k,\ell} \in \mathbb{C}^{N_t}$. For simplicity, we omit ℓ and consider one coherence block with time-invariant small-scale fading. The received signal at UE u on subcarrier k is

$$y_{u,k} = \mathbf{h}_{u,k}^{\mathsf{H}} \mathbf{f}_{u,k}\, s_{u,k} \;+\; \sum_{j \neq u} \mathbf{h}_{u,k}^{\mathsf{H}} \mathbf{f}_{j,k}\, s_{j,k} \;+\; n_{u,k}, \tag{4}$$

where $\mathbf{f}_{u,k} \in \mathbb{C}^{N_t}$ is precoding vector, $s_{u,k}$ a unit-power QAM symbol $\big(\mathbb{E}\{|s_{u,k}|^2\} = 1\big)$, and $n_{u,k} \sim \mathcal{CN}(0, \sigma_n^2)$ is AWGN.

Geometric Wideband Channel: Following the 3GPP/Quadriga model, the frequency-domain channel at subcarrier k is

$$\mathbf{h}_{u,k} = \sum_{\ell=1}^{L} \alpha_{u,\ell}\, e^{-j2\pi k \Delta f \tau_{u,\ell}}\, \mathbf{a}(\phi_{u,\ell}^{\text{az}}, \phi_{u,\ell}^{\text{el}}), \tag{5}$$

where $\alpha_{u,\ell} \sim \mathcal{CN}(0, \beta_{u,\ell})$ is the complex gain, $\tau_{u,\ell}$ the delay, and $\mathbf{a}(\cdot)$ the UPA steering vector at azimuth/elevation $(\phi_{u,\ell}^{\text{az}}, \phi_{u,\ell}^{\text{el}})$. The term $\beta_{u,\ell}$ captures path-loss, shadowing; LOS dominates rural links, while urban canyons exhibit richer NLOS components with larger delay and angular spreads.

Mobility and Channel Dynamics: UEs follow a random waypoint model with speeds uniformly drawn from 0–120 km/h, covering static IoT, pedestrians, and vehicles. Mobility induces Doppler spreads up to $f_\text{D}^{\max} = v_{\max} f_c / c \approx 390\,\text{Hz}$, giving a coherence time $T_\text{c} \approx 1/(2 f_\text{D}^{\max}) \simeq 1.3\,\text{ms}$, so several OFDM symbols experience quasi-static fading, justifying (4)–(5). Blockage events (e.g., trucks, foliage) are modeled as temporary LOS-to-NLOS transitions via path-gain drops, matching the digital-twin traces used later.

B. More on Channel Model

Further details regarding the channel model are as follows: **Outdoor Macro/Micro (3GPP TR 38.901):** For rural macro (RMa) and urban macro (UMa) scenarios, large-scale parameters (LSPs): including τ_rms, ASD, ASA, DSD, DSA, etc. are drawn from the (15) distributions and clustered via the NYU "cluster-of-clusters" method (26), yielding N_c = 12 clusters with Laplacian angular spreads and an exponential delay-power profile $P_c \propto e^{-\tau_c/\tau_\text{rms}}$. Cross-polarization is modelled via the XPD-matrix from (15) Table 7.5-6, and path-loss for LOS/NLOS links is adopted from (15).

Outdoor-to-Indoor (O2I): For deep-indoor users, the UMa channel is cascaded with the O2I penetration model from (15) §7.8.3. This includes a frequency-selective wall loss $L_\text{O2I}(f_c)$ along with a clustered indoor delay spread of 15–25 ns, appended to the channel impulse response in (1).

Indoor office/hot-spot (WINNER II A1): Dense–urban offices adopt the WINNER II A1 scenario with $N_c = 6$, $L_c = 20$, RMS delay spread $\tau_\text{rms} = 20$ ns and 3-D angular spreads as specified in Table A1-3 of the WINNER report (27); implemented via the native `Quadriga` class `WINNER_2_A1`.

Stationary vs. Mobile Users: Stationary IoT nodes are modeled with zero velocity ($v_u = 0$) and shadowing standard deviation $\sigma_\text{SF} = 4\,\text{dB}$. Pedestrians ($v_u$ = 3–6 km/h) and vehicles (v_u = 30–120 km/h) follow the Doppler statistics inherent in (1). The mobility state modulates cluster Doppler shifts $f_{u,c,\ell}$ and the temporal channel correlation $R_{hh}(\Delta t) = \mathbb{E}\left\{\mathbf{H}_u(t)\mathbf{H}_u^{\mathsf{H}}(t+\Delta t)\right\}$, which is leveraged by the prediction network described in Section 5. The hybrid specification above enables seamless data generation in `Quadriga` and on-device augmentation in `Sionna` while remaining numerically consistent with 3GPP standards.

C. Additional Tables/Figures

1. Simulation Parameters for CSI generation:

Table 3: Simulation Parameters for CSI Generation

Parameter	Value
Downtilt Angles	$0°, 15°, 30°$
Environments	Urban Macro, Urban Micro, Indoor
Carrier Frequency	3.5GHz
Bandwidth	100MHz
Cell Radius	200m (Macro), 100m (Micro), 30m (Indoor)
BS Height	25m (Macro/Micro), 3m (Indoor)
Users per Scenario	$N = 2000$
Transmit Power	43dBm

2. Train-Test Mismatch: The figure below illustrates the train-test mismatch phenomenon outlined earlier:

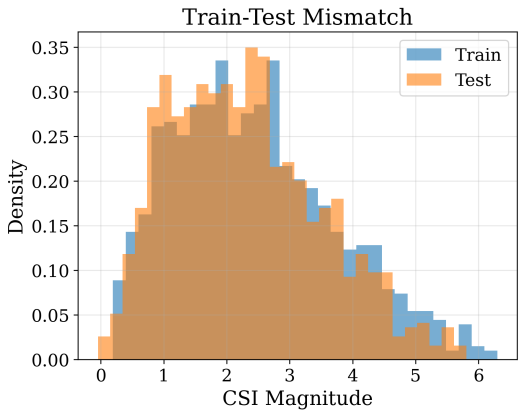


Figure 4: Train-test mismatch

3. Negative-Pair criteria

Domain Label	"Negative" Criterion
Mobility Split	$\Delta_{\texttt{mob}} = \lvert v_i - v_j \rvert > 30$ km/h
Antenna Downtilt	$\Delta_{\texttt{tilt}} = \lvert \theta_i - \theta_j \rvert > 5°$
Carrier Frequency	$\Delta_{\texttt{freq}} = \lvert f_i - f_j \rvert \geq 25$ GHz

Table 4: Examples of adversarial split metrics $\Delta_a(d_i, d_j)$ and their negative-pair criteria.

4. Digital-Twin Channel Generation parameters: All the parameters can be found in Table 5.

Table 5: Key simulation parameters.

Parameter	Outdoor U-Ma	Indoor OFDMA
Carrier freq. f_c	3.5 GHz	5.2 GHz
System bandwidth	100 MHz	80 MHz
Antenna array (BS)	4×4 UPA	4×4 UPA
UE speed	0–120 km/h	0–15 km/h
Antenna downtilt	$[0°, 90°]$	$[0°, 90°]$
Path loss + shadowing	3GPP U-Ma	IEEE TGax D
Timesteps	$5s$ dB	$5s$ dB
CSI Sampling frequency	200 Hz	200 Hz

5. Experimental Hyperparameters: All the hyperparameters are outlined below in Table 6.

D. Further ablation studies

Results of all the classical baselines and further ablations outlined below are shown in Table 7. This table also helps us understand how hard-negative mining and longer window can be affective, especially in the case of OOD CSI prediction.

Table 6: Experimental Hyperparameters for Contrastive Learning and CSI Prediction

Category	**Parameter**	**Value**
Encoder (Contrastive)	Embedding dimension	256
	Patch size	$(4, 5)$
	transformer layers	4
	attention heads	4
	Dropout	0.1
Conv Feature Extractor	Number of Conv2D layers	4
	Conv2D kernel size	3×3 (overlapping)
	Activation	ReLU
	Normalization	BatchNorm2d
Contrastive Loss (NT-Xent)	Temperature (τ)	0.5
	Pos/neg pair formation	Even/odd paired
	Loss function	CrossEntropy
Training Setup	Batch size	32
	Learning rate	10^{-4}
	Optimizer	AdamW
	Epochs	100
MLP Predictor	Architecture	2-layer MLP
	Hidden dimension	512
	Activation	ReLU + Tanh (final)
CSI Domain	Channel domain	Time domain

Table 7: Next-step CSI prediction. IID averages Indoor LOS/Outdoor NLOS; OOD averages Downtilt (15°), High-Mobility (30 m/s), and LOS↔NLOS. Lower NMSE is better.

	NMSE (dB) ↓	
Method	IID	OOD
AR(1) predictor	-19.0	-12.5
Kalman (AR(1) state) filter	-20.5	-14.0
Ablations of ConTwin:		
A1: No hard-negative mining	-25.55	-19.75
A2: Short Δt_{pos}	-25.57	-20.60
SwinCFNet (B1)	-25.44	-17.20
CNN+Transformer (In-Dist, B2)	-25.75	-16.33
CNN+Transformer (OOD-trained, B3)	—	-16.50
ConTwin (full)	**-25.60**	**-22.90**

A1: No hard-negative mining: Here, we replace contrastive hard-negative mining with uniform in-batch negatives (keep batch size/temperature τ fixed). All other training settings remain identical. This tests whether improvements come from the mining heuristic vs. the representation itself.

A2: Positive window sensitivity: We vary the temporal proximity constraint for positive pairs: **Short** window $\Delta t_{\mathrm{pos}}^{\mathrm{short}}$ (e.g., 5 ms) vs. **Long** window $\Delta t_{\mathrm{pos}}^{\mathrm{long}}$ (e.g., 25 ms). Keep negatives unchanged. This probes how temporal invariance width affects the learned representation under different Doppler spreads.

Classical baselines:

We also include the two classical baselines outlined below for comparison:

AR(1) predictor We model each complex CSI component (per subcarrier/antenna) as

$$h_{t+1} = \rho\, h_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{CN}\big(0, \sigma^2 \mathbf{I}\big)\,.$$

Estimate ρ from the training set by least squares:

$$\rho = \Re\!\left(\frac{\sum_t \langle h_t, h_{t+1}\rangle}{\sum_t \|h_t\|_2^2}\right), \text{ clipped to } [0, 1].$$

Apply per subcarrier; predict $\hat{h}_{t+1} = \rho\, h_t$.

Kalman filter (AR(1) state) Use a complex Kalman filter with state $x_t = h_t$, transition

$$x_{t+1} = Fx_t + w_t, \quad F = \rho\,\mathbf{I},$$

and observation $y_t = h_t + v_t$. Set $Q = (1-\rho^2)\sigma_h^2\mathbf{I}$ and $R = \sigma_v^2\mathbf{I}$ from training statistics; run one-step prediction to obtain $\hat{h}_{t+1|t}$. Apply per subcarrier (vectorized across antennas) for efficiency.

Implementation: we operate on stacked real/imaginary parts; vectorize across subcarriers; initialize $\hat{h}_{0|0} = 0,\ P_{0|0} = \sigma_h^2\mathbf{I}$.

Additional experiments:

In future work, we will also evaluate downstream link adaptation via beam choice using a DFT codebook matched to the array (e.g., $|\mathcal{B}|$=64 for an 8×8 UPA), reporting Top-1/Top-k accuracy and rate loss relative to the oracle beam. We will also study sensitivity to codebook size and OOD shifts (mobility, downtilt, LOS$\leftrightarrow$NLOS) and include a small field sanity check to gauge sim-to-real transfer.