CALIBRATION OR LOCATION? PREDICTIVE UNCERTAINTY UNDER GAUGE CONFOUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Representation learning often aims to remove sensor or acquisition variation while preserving task information. These goals conflict when nuisance and task transformations induce the same observation: nuisance invariance then collapses the corresponding target orbit. For coherent arrays, we prove that an unknown sensor phase affine in position is observationally equivalent to a change in incident wavevector. Neither raw I/Q nor conjugate-phase features can then identify absolute direction from one snapshot. Lower bounds for point risk and prediction-set width require valid predictions to represent this ambiguity unless additional information breaks the gauge.

We test these implications through controlled interventions on public 77 GHz raw ADC data. A prespecified label-only pipeline retains 281 automotive frames. Across 843 injected phase ramps, every retained beam peak translates as predicted: mean circular spatial-frequency error rises from 0.042 to 0.144 and returns to 0.042 with oracle correction. An independent audit confirms complete-spectrum translation and noncoherent range–Doppler invariance. In learned benchmarks, three five-model ensembles achieve 88.6–90.2% coverage at the training drift bound but only 6.1–7.1% when the test bound is sixteen times larger, while widths remain nearly constant. Under gauge-paired worlds, byte-identical inputs leave every prediction unchanged and shifted coverage falls to zero. Given a specified nuisance law, an analytic posterior and a learned orbit head recover about 90% coverage by widening along the unidentified orbit; a noisy known-direction anchor restores 90.4%. The theorem covers arbitrary geometries under narrowband far-field propagation; the learned study uses a single-source ULA without multipath.

1 INTRODUCTION

Invariance is a natural goal when learning from physical sensors. A model should not change its prediction because a device was replaced or recalibrated, provided that the quantity of interest did not change. The difficulty is that the sensor and the target may leave the same signature in the data. In that case, suppressing the sensor effect also suppresses part of the target.

Coherent arrays give a concrete example. Direction appears as a phase slope across sensors, but a linear calibration ramp produces the same slope. Angle and calibration are then two explanations for one observation. This ambiguity is known in array self calibration (Wylie et al., 1994; Ramamohan et al., 2021; Liu et al., 2025); here we ask what it means for learned representations and predictive uncertainty. In particular, a measurement collected under drift can be identical to a training measurement from another direction. No uncertainty score computed from that measurement alone can tell which explanation is correct.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

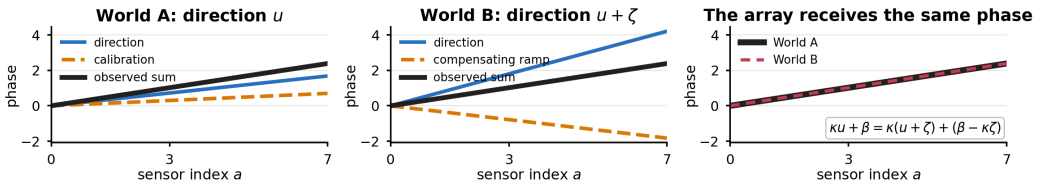


Figure 1: Two parameter settings can produce the same array input. The direction and calibration components differ (left and center), but their observed phase sum is identical (right). A predictor that receives only this input cannot choose between u and $u + \zeta$.

The issue is relevant to radar systems that learn from raw or lightly processed signals (Rebut et al., 2022; Liu et al., 2023a; Giroux et al., 2023). Existing work has studied spatial ambiguity and the deterioration of regression uncertainty under shift (Dong et al., 2020; Wang et al., 2026; Ovadia et al., 2019; Gustafsson et al., 2023; Dheur & Ben Taieb, 2023). The case considered here is different: calibration can move the target while the observed input remains on the source measurement distribution.

The paper develops this argument in four steps:

- Building on the theory of incorrect invariance and a concurrent preprint on the general limit of coverage audits without labels (Wang et al., 2023; Moussa, 2026), we derive the relevant symmetry from a forward operator for the target and nuisance. We then construct deployment shifts with identical input distributions, under which alarms based only on the input have no statistical power beyond their false alarm rate.
- We instantiate the obstruction for coherent arrays and prove that affine sensor phase and incident wavevector are confounded by the gauge. Local conjugate products remove global phase but do not resolve a linear ramp.
- We apply standard two-point arguments to obtain lower bounds for point risk and prediction set width. We then define prediction over the orbit by marginalizing over or bounding the confounded nuisance. A matched learned head predicts the identifiable phase slope and propagates a specified nuisance law; a fixed-drift audit separates average from uniform protection. An anchor with a known direction breaks the shared action.
- We provide reproducible analytic and learned evaluations. Gaussian raw I/Q, Gaussian relational, and conformalized all pair quantile ensembles remain narrow and become severely overconfident outside their training calibration envelope. An exact paired audit holds their noisy input tensors fixed while changing the target and calibration to equivalent values under the gauge. Every output from the saved models remains identical while coverage falls to zero. Prespecified interventions associated with labels on physical automotive ADC verify the predicted localization displacement, with an independent invariance audit on raw ADC data; neither is presented as naturally occurring drift.

2 Shared Nuisance and Target Actions

2.1 Problem Formulation

Let $x \in \mathcal{X}$ denote an observation and $y \in \mathcal{Y}$ its target. A task transformation $g \in G$ acts through τ_g on observations and ρ_g on targets. A nuisance $h \in H$ acts through ν_h on the observation but leaves the target unchanged. A representation $\Phi : \mathcal{X} \to \mathcal{Z}$ is nuisance invariant when

$$\Phi(\nu_h x) = \Phi(x). \tag{1}$$

The representation must still retain enough information for a decoder $d : \mathcal{Z} \to \mathcal{Y}$ to respond to the task action.

The conflict occurs where the two actions overlap on $\mathcal{X}$. More precisely, let K index a *shared action*: there are maps $g_\star : K \to G$ and $h_\star : K \to H$ such that

$$\tau_{g_\star(\xi)} x = \nu_{h_\star(\xi)} x \quad \text{for all relevant } x. \tag{2}$$

Theorem 1 (Impossibility under shared actions). *If Φ satisfies Eq. 1 and Eq. 2 holds, then*

$$\Phi(\tau_{g_\star(\xi)}x) = \Phi(x)$$

for every $\xi \in K$. Hence a deterministic decoder $d \circ \Phi$ cannot be equivariant to a nontrivial target action $\rho_{g_\star(\xi)}$ on examples for which $\rho_{g_\star(\xi)}y \neq y$ while also predicting both targets exactly.

The statement is simple, but the restriction is structural. More data or model capacity cannot recover information that the representation has removed. The full argument appears in Appendix C.

2.2 Observational equivalence

We can state the same problem before choosing a representation. Let $F(y, \eta)$ be a forward operator with nuisance parameter $\eta \in \mathcal{N}$. Define

$$y_0 \sim_F y_1 \quad \Longleftrightarrow \quad \exists \eta_0, \eta_1 : \; F(y_0, \eta_0) = F(y_1, \eta_1).$$

The two parameter pairs give a predictor exactly the same input. To form a conditional distribution, one must specify how probability is assigned inside this equivalence class. Without such a prior, the natural object is a set of compatible targets. We call the class the *target orbit* and use *gauge* for a change of parameters that leaves the observation fixed. This differs from gauge equivariant CNNs, where gauge refers to a local choice of reference frame (Cohen et al., 2019).

Theorem 2 (Invisibility under the input distribution). *Let P be a distribution on pairs of targets and nuisance parameters and let $T : \mathcal{Y} \times \mathcal{N} \to \mathcal{Y} \times \mathcal{N}$ satisfy $F \circ T = F$. If $Q = T_\# P$ is the pushforward deployment distribution, then the observation laws are identical: $F_\# Q = F_\# P$. Consequently, every randomized alarm or uncertainty score computed only from the observation has the same distribution under P and Q, even when T changes the target almost surely. This remains true for any finite i.i.d. target batch and any fixed source dataset available to the procedure.*

For this constructed shift, an OOD detector cannot have more power than its false alarm rate. The result concerns the existence of invisible shifts; it does not say that every calibration change is invisible.

3 Coherent localization under phase calibration

3.1 Array model

Consider sensors at positions $r_a \in \mathbb{R}^p$ and a narrowband source in the far field with complex amplitude s. Ignoring noise for the moment,

$$x_a(k, \delta) = s \exp\big(i(r_a^\top k + \delta_a)\big), \qquad a = 1, \dots, A, \tag{3}$$

Here k is the incident wavevector coordinate, and δ_a is the phase calibration of sensor a. Carrier frequency and field of view determine the admissible values of k.

Theorem 3 (Calibration and wavevector gauge). *For any $\eta \in \mathbb{R}^p$ such that $k + \eta$ is admissible,*

$$x(k, \delta) = x(k + \eta, \delta - R\eta), \tag{4}$$

where row a of R is $r_a^\top$. Therefore k is not identifiable whenever the calibration class is closed under spatially affine phase fields $R\eta$.

ULA corollary. For a ULA indexed by $a = 0, \dots, A-1$, let $\kappa = 2\pi d/\lambda$, $u = \sin\theta$, and $\delta_a = \phi_0 + \beta a + \gamma a^2$. Then

$$x_a(u, \phi_0, \beta, \gamma) = x_a(u + \zeta, \phi_0, \beta - \kappa\zeta, \gamma) \tag{5}$$

for every admissible ζ. The observation depends on u and β only through the combined slope $\kappa u + \beta$.

3.2 Why relative phase does not solve the problem

One might try to remove calibration through local conjugate products:

$$q_{a,\ell} = x_a \overline{x_{a+\ell}}. \tag{6}$$

This feature is invariant to the global phase action $x \mapsto e^{i\phi_0} x$. Under the ULA model,

$$\arg q_{a,\ell} = -\ell(\kappa u + \beta) - \gamma(2a\ell + \ell^2) \pmod{2\pi}. \tag{7}$$

When $\gamma = 0$, the remaining term is still $\kappa u + \beta$. Removing global phase is therefore safe, but removing every linear phase ramp is not: the latter also removes direction. Relative phase remains useful for fault diagnosis across radar captures (Mallick et al., 2026); it simply cannot identify absolute direction under this calibration model.

4 Consequences for predictive uncertainty

4.1 Point and set lower bounds

Proposition 1 (Approximate bounds for two points and alarms)**.** *Let P_0, P_1 be the observation laws at two pairs of target and nuisance parameters, let $D = \|y_0 - y_1\|$, and let $\varepsilon = \mathrm{TV}(P_0, P_1)$. Every possibly randomized point predictor obeys*

$$\max_{j \in \{0,1\}} \mathbb{E}_{P_j} \|\widehat{y}(X) - y_j\| \geq \frac{D}{2}(1 - \varepsilon).$$

For any randomized alarm $\psi(X) \in [0,1]$, $|\mathbb{E}_{P_1}\psi - \mathbb{E}_{P_0}\psi| \leq \varepsilon$. If a possibly randomized prediction set satisfies $P_j\{y_j \in C(X)\} \geq 1 - \alpha$ for $j = 0, 1$, then

$$\min_j P_j\{\operatorname{diam} C(X) \geq D\} \geq [1 - 2\alpha - \varepsilon]_+,$$

where $[z]_+ = \max\{z, 0\}$.

The first inequality is Le Cam's argument for two points (Tsybakov, 2009). The alarm bound follows from the variational definition of total variation. For prediction sets, probability is transferred from one world to the other and a union bound forces both targets into the same set. These are limits of the observation model, not uncertainty about fitted model parameters. If the observation is the same in two worlds, an ensemble has no reason to disagree. The result also covers approximate gauges. With Gaussian noise of covariance $\sigma^2 I$ and forward means separated by Δ, Pinsker's inequality gives $\varepsilon \leq \min\{1, \Delta/(2\sigma)\}$.

4.2 Prediction over the orbit

There are two natural predictions, depending on what is known about the nuisance. If a joint prior is justified, it can be marginalized without assuming that target and nuisance are independent:

$$p(y \mid x) \propto \int p(x \mid y, \eta) p(y, \eta)\, d\eta. \tag{8}$$

The common factorization $p(y, \eta) = p(y)p(\eta)$ is valid only when prior independence is an explicit modeling assumption. If only a certified nuisance set $\mathcal{E}$ is available, prediction can instead return the identified set

$$\mathcal{I}(x) = \{y : \exists \eta \in \mathcal{E},\ F(y, \eta) = x\}.$$

For the noiseless ULA statistic $t = \kappa u + \beta$ and $|\beta| \leq B$,

$$\mathcal{I}(t) = [(t - B)/\kappa, (t + B)/\kappa] \cap [-1, 1]. \tag{9}$$

Known, estimated, and unknown calibration envelopes are different deployment settings. Supplying the test bound gives a method information that an input alone does not contain.

5 EXPERIMENTAL EVALUATION

5.1 QUESTIONS AND PROTOCOL

The experiments address four questions:

1. Does the exact gauge identity hold numerically for the implemented complex forward model and a controlled intervention on physical ADC?
2. Do standard Gaussian regressors and deep ensembles widen under an expanded calibration envelope when the change is confounded by measurement?
3. Are predictions and disagreement exactly unchanged when a target shift is paired with its compensating calibration ramp at fixed noisy input?
4. Can a matched learned head recover calibrated widening by predicting the identifiable slope and propagating a declared nuisance envelope?

The code records configurations, seeds, training histories, checkpoints, and metrics for each case. Each analytic condition contains 10,000 independent samples. The learned experiments use 20,000 training examples, 4,000 validation examples, and 5,000 test examples for each drift condition. We train five independent models. Test directions are restricted to $[-0.4, 0.4]$, inside the training support $[-0.7, 0.7]$, to avoid effects from the support boundary. Uncertainty is reported according to the independent experimental unit; the rules and sample sizes are stated in Appendix D.

5.2 EXACT IDENTITY AUDIT

We first check whether the algebra survives numerical implementation. Across 1,000 ULA parameter triples and their gauge shifts, the mean relative defect in the complex measurement is 1.42×10^{-15}; the maximum is 6.49×10^{-15}. A second audit uses 24 random planar sensor positions, wavevectors in two dimensions, and independent sensor phases. Its maximum defect is 3.19×10^{-16}. Unit tests also confirm that conjugate features are constant along each orbit and that Eq. 9 covers the true direction. The identity is exact for planar waves. With spherical waves, Appendix M finds a p95 defect below 0.0017 at sixteen Fraunhofer distances, while curvature is often visible at one quarter of that distance.

5.3 DIGITAL INTERVENTIONS ON PHYSICAL RAW ADC DATA

We next intervene on public 77 GHz automotive ADC data with metric object labels (Gao et al., 2022a). A label-only rule selects 360 frames from 15 sequences on four dates; a fixed criterion retains 281 frames in 14 sequences. Across three ramps, all 843 peaks move by $-\zeta$ on the 4096-point grid. Mean circular spatial-frequency error rises from 0.042 to 0.144 and returns to 0.042 with the injected ramp supplied as an oracle.

For virtual index $a = 4t + r$, the ramp factors into TX/RX offsets. A frozen follow-up gives exact translation for this mode; TX-only, RX-only, single-RX, and random separable faults have median spectral defects 0.298, 0.191, 0.169, and 0.813. A 65-geometry audit, including 32 dense 12-by-16 planar arrays, retains exact coordinated translation (maximum defect 1.77×10^{-15}). A 150-case RaDICaL audit confirms translation and range–Doppler invariance to 4.66×10^{-16} (Lim et al., 2021). These are controlled interventions or simulations, not natural drift; Appendix F gives the boundaries.

5.4 ANALYTIC PROBABILISTIC BENCHMARK

To isolate the uncertainty calculation, consider the scalar statistic

$$t = \kappa u + \beta + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, 0.03^2),$$

with $u \sim \mathrm{Unif}[-0.7, 0.7]$. We compare four predictors. The fixed Gaussian is calibrated at $|\beta| \leq 0.05$. The gauge posterior uses the test envelope in Eq. 8, while the oracle observes β. The fourth method uses an anchor with known direction. It observes $t_a = \kappa u_a + \beta + \epsilon_a$ at $u_a = 0$, estimates $\widehat{\beta} = t_a$, and uses the resulting Gaussian predictive standard deviation $\sqrt{2}\sigma/\kappa$.

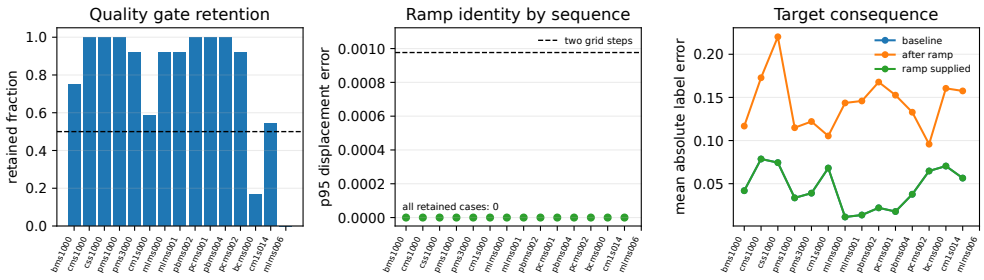


Figure 2: Expanded physical intervention. The fixed criterion retains 281 of 360 frames, spanning 14 of the 15 prespecified sequences (left); retained sequence p95 errors are zero (center), and supplying the ramp restores baseline error (right). Frames within sequences are serially related.

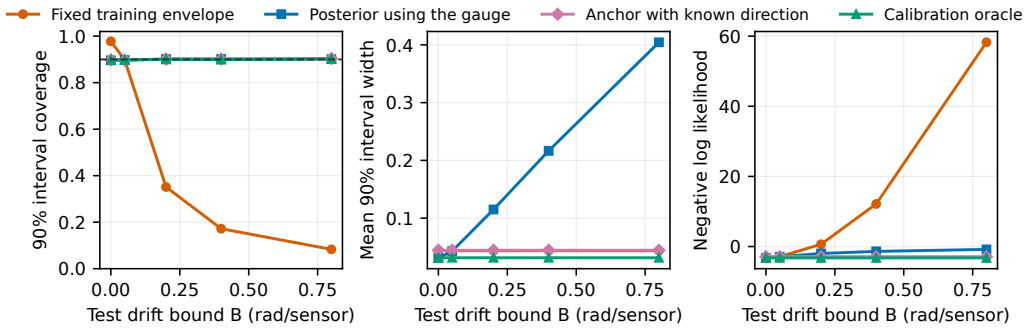


Figure 3: Shift in the calibration envelope in the analytic benchmark. A predictor fixed to the training envelope becomes overconfident. Marginalizing the correct nuisance envelope maintains coverage by increasing width. The oracle and anchor with known direction show the value of calibration information that breaks the symmetry.

At $B = 0.4$, coverage of the fixed Gaussian falls to 17.2% and its NLL rises to 12.15. The gauge posterior reaches 90.2% coverage by increasing mean width from 0.044 to 0.217. Its coverage remains 90.1% at $B = 0.8$, while the oracle keeps width 0.031. At the same bound, the noisy anchor obtains 90.4% coverage, RMSE 0.0133, and width 0.044. The anchor has measured the missing information, so it need not widen along the orbit. Appendix E studies anchor mismatch, and Appendix L checks the circular likelihood.

5.5 Learned probabilistic benchmark

The learned benchmark uses two heteroscedastic Gaussian MLPs. One receives 32 Cartesian I/Q features; the other receives 46 magnitude and adjacent conjugate features. Their widths give 6,402 and 6,433 parameters. A third MLP has 6,535 parameters and predicts the 5th, 50th, and 95th percentiles from magnitudes and all 120 conjugate pairs. Its intervals use split conformal correction on the validation set. Every model is trained with $|\beta| \leq 0.05$.

Relational inputs give narrower intervals at the training bound, but this makes their later overconfidence more severe. From $B = 0.05$ to $B = 0.8$, ensemble width changes by less than 3% while RMSE grows by more than a factor of ten. The quantile model shows that the effect is not specific to Gaussian heads or adjacent relations. Appendix I repeats the comparison at four coverage levels. At 90%, the Wilson intervals for the three models are [0.887, 0.904], [0.877, 0.895], and [0.893, 0.910] at $B = 0.05$; at $B = 0.8$, they are [0.064, 0.079], [0.054, 0.068], and [0.055, 0.068].

The aggregate shift above changes the input distribution slightly, so we also test Theorem 2 exactly. For 5,000 source examples, set $u' = u + \eta$ and $\beta' = \beta - \pi\eta$ for $\eta \in \{-0.2, 0.2\}$, reusing the same noise and input tensor. All three observation hashes match. Centers, bounds, widths, and disagreement from the saved ensembles are also identical at the bit level. Coverage in the source

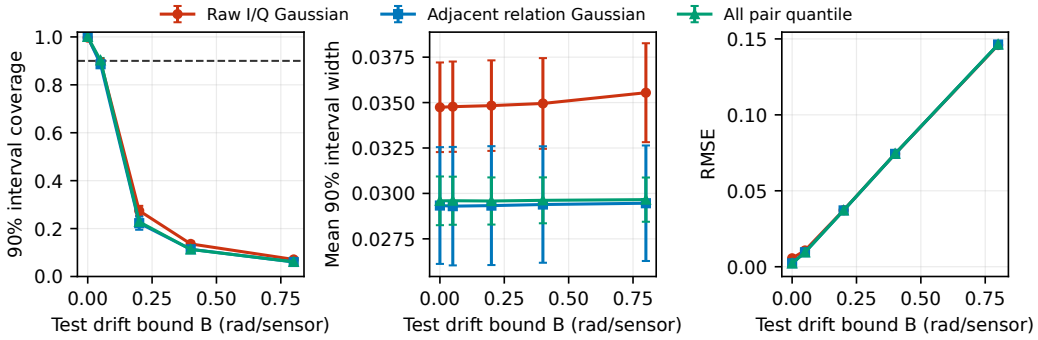


Figure 4: Learned ensembles under calibration shift. Error bars show standard deviation across five seeds. Coverage collapses while width changes little. Neither conjugate inputs nor conformalized quantiles resolve the linear phase ambiguity.

world ranges from 89.3 to 90.7%; after the target shift, it is zero. RMSE rises from about 0.01 to 0.20. The construction rules out a changed input or a weak detector as explanations, but it does not claim that natural drift is always invisible. Appendix K contains the full audit.

5.6 Learned prediction from observables to an orbit

Can a learned model express the missing width when it receives the nuisance law? To test this, we keep the architecture with 6,535 parameters, inputs, and data, but change the prediction target from u to the identifiable slope $q = \pi u + \beta$. A separate split of 4,000 examples estimates the residual distribution in eight fixed bins of predicted q. At deployment, we convolve this residual with the supplied law $\beta \sim \mathrm{Unif}[-B, B]$. The network is not retrained for each bound and does not receive test labels.

Across five values of B, coverage ranges from 89.1 to 90.0%. At $B = 0.8$, the orbit interval widens to 0.459 and covers 90.0%. Direct target CQR covers 5.5% at width 0.029, and an orbit head frozen at the training bound covers 5.6%. Since the architecture is fixed, the difference comes from propagating the nuisance law rather than model capacity. Appendix J also records an earlier protocol that failed its prespecified criteria.

This guarantee is marginal. At 33 fixed values of $\beta_0 \in [-0.8, 0.8]$, Figure 5 shows 90.63% integrated coverage but a zero-to-one conditional range, including zero at the endpoints. The wider certified set (0.514 versus 0.459) has minimum coverage 95.32%. Appendix J.1 gives the frozen protocol, oracle, and simultaneous intervals.

Table 1: Learned ensemble results on 5,000 independent test cases per row. Validation recalibration uses only the training drift envelope $B = 0.05$; Figure 4 reports variation across five training seeds. NLL is "n.d." for the quantile head, which specifies no predictive density.

B	input	NLL	RMSE	cov. 90%	width
0.05	I/Q	-3.111	0.0107	0.896	0.0348
0.05	relational	-3.227	0.0095	0.886	0.0293
0.05	all pair quantile	n.d.	0.0095	0.902	0.0296
0.20	I/Q	2.758	0.0372	0.275	0.0348
0.20	relational	4.944	0.0371	0.223	0.0293
0.20	all pair quantile	n.d.	0.0371	0.228	0.0296
0.40	I/Q	21.732	0.0743	0.136	0.0350
0.40	relational	31.046	0.0743	0.114	0.0294
0.40	all pair quantile	n.d.	0.0744	0.113	0.0296
0.80	I/Q	90.511	0.1460	0.071	0.0355
0.80	relational	129.701	0.1463	0.061	0.0295
0.80	all pair quantile	n.d.	0.1463	0.061	0.0297

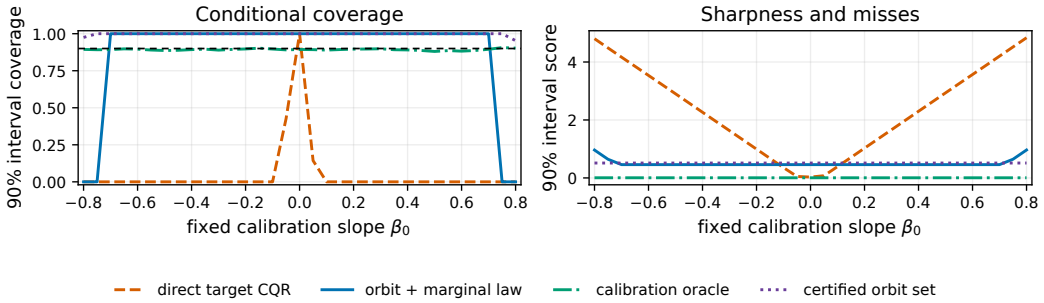


Figure 5: Marginal versus fixed-drift coverage for frozen heads. Each point pools five 1,000-case panels paired across β_0. The orbit interval is nominal on average but not conditional on persistent drift; certification protects the grid at greater width, while observed calibration gives a sharp oracle.

Appendix O replaces the supplied law with an exchangeable deployment panel of known directions. With 4,000 panel examples, both heads cover about 90% under uniform, triangular, and bimodal drift. A stale panel leaves both near 5.8%. Current calibration evidence, rather than the representation by itself, supplies the information needed at deployment.

6 Breaking the Ambiguity

Equation 4 identifies four possible interventions:

1. **Anchor with known direction.** A calibration return with known wavevector identifies the shared affine phase component. This intervention is evaluated in Figure 3.
2. **Calibrated sensor subset.** Fixing enough phase responses removes the corresponding degrees of freedom.
3. **Frequency diversity.** Multiple carriers identify direction only if the frequency law of hardware phase differs from the steering law and is sufficiently restricted.
4. **Restricted calibration subspace.** If admissible calibration modes exclude the steering subspace, the intersection of the shared actions is trivial; Proposition 3 in Appendix C gives the exact rank criterion.

Multiple unknown sources alone identify relative directions but retain a common shift unless scene structure or an anchor fixes it. The following statement makes both qualifications precise.

Proposition 2 (When diversity breaks the gauge)**.** *For J sources with snapshot*

$$x_a = \sum_j s_j \exp i(r_a^\top k_j + \delta_a),$$

the transformation $k_j \mapsto k_j + \eta$ for every j and $\delta \mapsto \delta - R\eta$ leaves the snapshot unchanged. Thus source diversity can reveal relative wavevectors but not their common component. For measurements at distinct angular frequencies $\omega_1 \neq \omega_2$, with $k_\ell = (\omega_\ell/c)q$ and an arbitrary complex amplitude at each frequency, let row $a-1$ of D be $(r_a - r_1)^\top$ for $a = 2, \ldots, A$. A phase field constrained to be the same at both frequencies can admit a common direction shift $q \mapsto q + \zeta$ only if

$$D((\omega_1 - \omega_2)\zeta/c) \in 2\pi\mathbb{Z}^{A-1}. \tag{10}$$

Thus frequency diversity breaks the gauge on an admissible shift domain with no nonzero solution to Eq. 10. Full column rank removes the local unwrapped gauge but, by itself, does not exclude discrete aliases. If an independent affine phase field is allowed at each frequency, the gauge persists.

The proof is in Appendix C. Comparisons of these interventions must give all methods equal access to the same side information. Classical joint calibration needs references, restrictions, or priors (Wylie et al., 1994; Ramamohan et al., 2021; Liu et al., 2025); with the same unrestricted snapshot, it retains the gauge. Our posterior, oracle, and anchor are matched nonlearned references; a solver with stronger calibration information would not be.

7 Related work

Array self calibration. Classical methods jointly estimate DOA and sensor gain or phase and establish that the solution can be nonunique (Wylie et al., 1994; Viberg et al., 2009). Convex formulations provide identifiability conditions (Ramamohan et al., 2021), and sparse Bayesian methods use priors to restrict ambiguity between phase and DOA (Liu et al., 2025). The physical ambiguity is known. Our work connects it to invariant representation design, conditional uncertainty, and learned shift detection.

Raw radar learning. Raw radar models learn angular processing from low level signals (Rebut et al., 2022; Liu et al., 2023a; Giroux et al., 2023). A concurrent preprint uses relative phase for hardware fault diagnosis (Mallick et al., 2026); we study absolute localization under a confounded phase mode.

Invariant representations and invisible shifts. Orbit representations remove nuisance variation (Raj et al., 2017), but a noninvertible invariant map can discard labels (Johansson et al., 2019). Incorrect equivariance theory bounds the resulting errors (Wang et al., 2023) and calibration mismatch (Berman et al., 2026); Theorem 1 derives this conflict from the sensor operator. A concurrent preprint gives a general boundary for coverage audits without labels (Moussa, 2026). Unlabeled conformal adaptation can help when inputs reveal the shift (Kasa et al., 2025; Angelman et al., 2025), but indistinguishable inputs impose limits (Ben-David et al., 2010; Koch et al., 2022). We add the physical gauge, its decision consequences, and tested interventions.

Uncertainty under shift. Common UQ methods, including deep ensembles, can become overconfident under shift (Lakshminarayanan et al., 2017; Ovadia et al., 2019; Gustafsson et al., 2023); regression also requires sharpness (Dheur & Ben Taieb, 2023). Equivariant bootstrapping uses symmetry for UQ (Pereyra & Tachella, 2024). DOA-specific work conformalizes Monte Carlo dropout, ensembles, and quantile networks under exchangeability (Khurjekar & Gerstoft, 2023); a 2026 automotive-radar preprint compares a von Mises ensemble with an evidential head on synthetic range–Doppler–angle regions (Kulkarni & Reddy, 2026). These methods quantify uncertainty when observations vary; we study the complementary case in which a gauge changes the target while preserving the observation law.

8 Limitations and responsible scope

Learned results use one synthetic source and compact MLPs; the identical-input audit is constructed. Physical audits inject ramps into label-associated serial frames, with low retention in two sequences. A frozen 160-frame, 12-TX/16-RX audit confirms complex ADC but lacks release-side channel ordering; no physical planar result is reported. Dense geometry evidence is synthetic and narrowband. None validates natural drift, a named automotive 4D radar, device transfer, physical learned UQ, or fault prevalence.

Table 3 retains failures in near-field sharpness, anchor transfer, panel lag, dense-MIMO degeneracy, and channel ordering.

The posterior and orbit benchmark receive the test envelope or nuisance law; estimation requires an exchangeable panel with known directions. We do not cover unknown envelopes, arbitrary shifts, unknown range, multipath, or general channel errors. The fixed-drift audit separates marginal from conditional coverage: certification protects the bounded grid but needs a valid bound. Appendix E uses contemporaneous references for blockwise mismatch estimation and abstention when recency or support checks fail.

9 Conclusion

When calibration and target transformations share an observation action, nuisance invariance removes target information and disagreement cannot expose it. Valid prediction must cover the orbit, model the nuisance, or break the symmetry. A nuisance law gives marginal coverage; persistent-drift protection requires the certified set or observed calibration.

REFERENCES

Shachar Angelman, Rotem Nizhar, and Jacob Goldberger. Calibrating without labels: Source-free conformal prediction using pseudo-labels. In *Proceedings of the Fourteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 266 of *Proceedings of Machine Learning Research*, pp. 63–81, 2025. URL https://proceedings.mlr.press/v266/angelman25a.html.

Amir Asiaee, Kaveh Aryan, and James P. Long. Partial causal structure learning for valid selective conformal inference under interventions. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 161–184, 2026. URL https://proceedings.mlr.press/v337/asiaee26b.html.

Shai Ben-David, Tyler Lu, Teresa Luu, and David Pal. Impossibility theorems for domain adaptation. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 129–136, 2010.

Edward Berman, Jacob Ginesin, Marco Pacini, and Robin Walters. On uncertainty calibration for equivariant functions. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=rxLUTPLBT3.

Emil Björnson, Özlem Tugfe Demir, and Luca Sanguinetti. A primer on near-field beamforming for arrays and reconfigurable intelligent surfaces. In *2021 55th Asilomar Conference on Signals, Systems, and Computers*, pp. 105–112, 2021. doi: 10.1109/IEEECONF53345.2021.9723331.

Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral CNN. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 1321–1330, 2019.

Victor Dheur and Souhaib Ben Taieb. A large-scale study of probabilistic calibration in neural network regression. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 7813–7836, 2023.

Xu Dong, Pengluo Wang, Pengyue Zhang, and Langechuan Liu. Probabilistic oriented object detection in automotive radar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

Xiangyu Gao, Youchen Luo, Guanbin Xing, Sumit Roy, and Hui Liu. Raw ADC data of 77GHz MMWave radar for automotive object detection. IEEE DataPort, 2022a. URL https://doi.org/10.21227/xm40-jx59.

Xiangyu Gao, Sumit Roy, Hui Liu, Youchen Luo, and Guanbin Xing. Raw ADC data of 2d-MIMO MMWave radar for carry object detection. IEEE DataPort, 2022b. URL https://doi.org/10.21227/begn-ye78.

James Giroux, Martin Bouchard, and Robert Laganiere. T-FFTRadNet: Object detection with swin vision transformers from raw ADC radar signals. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 4030–4039, 2023. URL https://arxiv.org/abs/2303.16940.

Fredrik K. Gustafsson, Martin Danelljan, and Thomas B. Schön. How reliable is your regression model's uncertainty under real-world distribution shifts? *Transactions on Machine Learning Research*, 2023. URL https://arxiv.org/abs/2302.03679.

Fredrik D. Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 527–536, 2019.

Kevin Kasa, Zhiyu Zhang, Heng Yang, and Graham W. Taylor. Adapting prediction sets to distribution shifts without labels. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 1990–2010, 2025. URL https://proceedings.mlr.press/v286/kasa25a.html.

Ishan D. Khurjekar and Peter Gerstoft. Uncertainty quantification for direction-of-arrival estimation with conformal prediction. *The Journal of the Acoustical Society of America*, 154(2):979–990, 2023. doi: 10.1121/10.0020655.

Lisa M. Koch, Christian M. Schürch, Arthur Gretton, and Philipp Berens. Hidden in plain sight: Subgroup shifts escape OOD detection. In *Proceedings of the 5th International Conference on Medical Imaging with Deep Learning*, volume 172 of *Proceedings of Machine Learning Research*, pp. 726–740, 2022.

Vinay Kulkarni and V. V. Reddy. Von Mises-based uncertainty quantification for closely spaced automotive radar targets. *arXiv preprint arXiv:2606.31473*, 2026. URL https://arxiv.org/abs/2606.31473. Version 1, preprint.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Teck-Yian Lim, Spencer A. Markowitz, and Minh N. Do. RaDICaL: A synchronized FMCW radar, depth, IMU and RGB camera data dataset with low-level FMCW radar signals. *IEEE Journal of Selected Topics in Signal Processing*, 15(4):941–953, 2021. doi: 10.1109/JSTSP.2021.3061270.

Yang Liu, Feng Wang, Naiyan Wang, and Zhao-Xiang Zhang. Echoes beyond points: Unleashing the power of raw radar data in multi-modality fusion. In *Advances in Neural Information Processing Systems*, volume 36, 2023a.

Yuanwei Liu, Zhaolin Wang, Jiaqi Xu, Chongjun Ouyang, Xidong Mu, and Robert Schober. Near-field communications: A tutorial review. *IEEE Open Journal of the Communications Society*, 4: 1999–2049, 2023b. doi: 10.1109/OJCOMS.2023.3305583.

Zhexian Liu, Shuai Shao, Hongwei Liu, and Tao Su. Ambiguity-free DOA estimation for linear arrays via sparse bayesian phase error self-calibration. In *9th International Conference on Vision, Image and Signal Processing*, pp. 1–5, 2025. doi: 10.1109/ICVISP68610.2025.11451714.

Mainak Mallick, Junghwan Yim, Alankrit Gupta, and Seung-Kyum Choi. Rad-R: A raw-ADC radar dataset and capture-invariant SSM for hardware-fault diagnosis. *arXiv preprint arXiv:2608.30896*, 2026. URL https://arxiv.org/abs/2608.30896. Version 2, preprint.

Maha Moussa. Auditing conformal prediction under distribution shift: A detectability boundary, exact label-budget design, and repair. *Research Square*, 2026. doi: 10.21203/rs.3.rs-10396495/v1. Preprint.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Marcelo Pereyra and Julián Tachella. Equivariant bootstrapping for uncertainty quantification in imaging inverse problems. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, pp. 4141–4149, 2024.

Anant Raj, Abhishek Kumar, Youssef Mroueh, Tom Fletcher, and Bernhard Schölkopf. Local group invariant representations via orbit embeddings. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1225–1235, 2017.

Krishnaprasad Nambur Ramamohan, Sundeep Prabhakar Chepuri, Daniel Fernandez Comesana, and Geert Leus. Self calibration of scalar and vector sensor arrays. *arXiv preprint arXiv:2104.02561*, 2021. URL https://arxiv.org/abs/2104.02561.

Julien Rebut, Arthur Ouaknine, Waqas Malik, and Patrick Pérez. Raw high-definition radar for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17021–17030, 2022.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1905.03222.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009. doi: 10.1007/b13794.

Mats Viberg, Maria Lanne, and Astrid Lundgren. Calibration in array processing. In *Classical and Modern Direction-of-Arrival Estimation*, pp. 93–124. Academic Press, 2009. doi: 10.1016/B978-0-12-374524-8.00003-9.

Dian Wang, Xupeng Zhu, Jung Yeon Park, Mingxi Jia, Guanang Su, Robert Platt, and Robin Walters. A general theory of correct, incorrect, and extrinsic equivariance. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1740.

Shengpeng Wang, Kuangyu Wang, and Wei Wang. RaUF: Learning the spatial uncertainty field of radar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Marilynn P. Wylie, Sumit Roy, and Hagit Messer. Joint DOA estimation and phase calibration of linear equispaced arrays. *IEEE Transactions on Signal Processing*, 42(12):3449–3459, 1994. doi: 10.1109/78.340779.

APPENDIX CONTENTS

A Notation **14**

B Retained limitations and negative results **14**

C Proofs **15**

D Reproducibility details **18**

E Anchor with a known direction **19**

F Physical ADC intervention **20**

G Dense planar MIMO geometry audit **22**

H Dense physical MIMO phase-content audit **22**

I Calibration at multiple levels **23**

J Learned prediction from observables to an orbit **24**

K Exact learned audit with identical inputs **25**

L Sensitivity to a wrapped likelihood **25**

M Gauge boundary in the near field **26**

N Learned audit in the near field: retained negative result **27**

O Estimating the deployment law from a calibration panel **28**

P Physical anchor audit across sequences **29**

Q Comparison with the closest work **29**

R Lag in a nonstationary calibration panel **30**

S Claim and evidence ledger **31**

T Ethics statement **32**

U Use of Large Language Models (LLMs) **33**

Guide to the supplement. The appendices are grouped by purpose: notation; retained limitations and negative results; proofs; reproducibility and data separation; mitigation studies using anchors, orbit prediction, and calibration panels; boundary studies for wrapped phase, near-field propagation, physical transfer, and panel lag; and finally the comparison and claim ledger. The table of contents above gives the exact section locations.

A NOTATION

Table 2 collects the symbols used across the theoretical and experimental sections. Quantities specific to one audit are defined where they first appear.

Table 2: Core notation.

symbol	meaning
General formulation	
$x \in \mathcal{X},\ y \in \mathcal{Y}$	observation and prediction target
$\eta \in \mathcal{N}$	nuisance parameter
$F(y, \eta)$	forward operator that produces an observation
G, H	task-transformation and nuisance-transformation sets
τ_g, ρ_g, ν_h	action of g on observations, action of g on targets, and action of h on observations
$K, g_\star, h_\star$	shared-action index set and the corresponding task and nuisance maps
Φ, d	learned representation and its decoder
$T, T_\#, P, Q$	gauge transformation, pushforward operation, source law, and pushforward deployment law
Array model	
A, p, r_a, R	number of sensors, spatial dimension, position of sensor a, and geometry matrix with row $r_a^\top$
s, k, δ_a	complex source amplitude, incident wavevector coordinate, and phase calibration of sensor a
d, λ, κ	ULA spacing, wavelength, and normalized spacing $\kappa = 2\pi d/\lambda$
θ, u	physical direction and spatial frequency $u = \sin\theta$
ϕ_0, β, γ	global, linear, and quadratic coefficients of the ULA phase field
ζ	gauge displacement in spatial frequency
$q_{a,\ell}$	conjugate product $x_a \overline{x_{a+\ell}}$ at lag ℓ
Prediction and uncertainty	
P_0, P_1, D, ε	two observation laws, target separation, and their total-variation distance
$\widehat{y}, C(X), \alpha$	point prediction, prediction set, and miscoverage level
$t, B, \mathcal{E}, \mathcal{I}$	identifiable slope, drift bound, certified nuisance set, and identified target set

B RETAINED LIMITATIONS AND NEGATIVE RESULTS

The following cases remain unresolved or fall outside the validated scope.

Table 3: Prespecified cases that did not support the intended extension.

case	retained finding and boundary
spherical wave learner	coverage held, but three of four sharpness gates failed; no learned near-field transition (Appendix N)
physical anchor transfer	retention failed in two splits; largest residual shift 0.092; transfer unresolved (Appendix P)
lagged calibration panel	74.4 to 75.6% coverage versus 89.5 to 90.6% with a current panel; stale data are not exchangeable (Appendix R)
dense-MIMO degeneracy endpoint	affine residual 0.124 clears its 0.02 gate, but spectrum defect 0.013 misses its 0.02 gate; the partial failure is retained (Appendix G)
dense physical channel ordering	160 of 160 frozen frames retain complex $256 \times 61 \times 16 \times 12$ ADC, but no release-side file maps the MAT TX/RX axes to physical antenna IDs; the steering audit stops at its integrity gate (Appendix H)

C PROOFS

Proof of Theorem 1. Fix $\xi \in K$ and an observation x for which the shared action is defined. Equation 2 states that the task transformation and the nuisance transformation produce the same observation:

$$\tau_{g_\star(\xi)} x = \nu_{h_\star(\xi)} x.$$

Applying Φ to both sides and then using nuisance invariance gives

$$\Phi(\tau_{g_\star(\xi)} x) = \Phi(\nu_{h_\star(\xi)} x) = \Phi(x).$$

This proves the first claim. Now write $f = d \circ \Phi$ for the complete predictor. Because the two representations are equal and d is deterministic,

$$f(\tau_{g_\star(\xi)} x) = d(\Phi(\tau_{g_\star(\xi)} x)) = d(\Phi(x)) = f(x).$$

Equivariance to the task action would instead require

$$f(\tau_{g_\star(\xi)} x) = \rho_{g_\star(\xi)} f(x).$$

Combining these two equalities yields $f(x) = \rho_{g_\star(\xi)} f(x)$. Thus the decoded output must be a fixed point of $\rho_{g_\star(\xi)}$. To see the contradiction with exact prediction, suppose that x has target y. Exact prediction at x gives $f(x) = y$, whereas exact prediction after the task transformation gives $f(\tau_{g_\star(\xi)} x) = \rho_{g_\star(\xi)} y$. The equality of the representations also gives $f(\tau_{g_\star(\xi)} x) = y$. Therefore both predictions can be correct only if $y = \rho_{g_\star(\xi)} y$. When $\rho_{g_\star(\xi)} y \neq y$, no deterministic decoder can predict both targets exactly. □

Proof of Theorem 3. Fix an admissible displacement η. The calibration vector in the second parameterization is $\delta - R\eta$, whose entry for sensor a is $\delta_a - r_a^\top \eta$. Substituting $k + \eta$ and this entry into the forward model gives

$$\begin{aligned} x_a(k + \eta, \delta - R\eta) &= s \exp\big(i \left[r_a^\top (k + \eta) + \delta_a - r_a^\top \eta\right]\big) \\ &= s \exp\big(i \left[r_a^\top k + r_a^\top \eta + \delta_a - r_a^\top \eta\right]\big) \\ &= s \exp\big(i (r_a^\top k + \delta_a)\big) \\ &= x_a(k, \delta). \end{aligned}$$

The terms $r_a^\top \eta$ cancel for every sensor a, so the two complete observation vectors are equal. If the calibration class is closed under the affine field $R\eta$, then $\delta - R\eta$ is an allowed calibration whenever δ is allowed. Hence the two admissible parameter pairs (k, δ) and $(k + \eta, \delta - R\eta)$ produce the same observation. For any nonzero admissible η, these pairs contain distinct wavevectors. The observation therefore cannot determine which wavevector generated it, which is precisely the lack of identifiability of k. □

Derivation of the ULA corollary. For a ULA, the phase that depends on direction at sensor a is $\kappa a u$. After applying the shift $u \mapsto u + \zeta$ and the compensating calibration change $\beta \mapsto \beta - \kappa\zeta$, the total phase becomes

$$\begin{aligned} \kappa a(u + \zeta) + \phi_0 + (\beta - \kappa\zeta) a + \gamma a^2 &= \kappa a u + \kappa a \zeta + \phi_0 + \beta a - \kappa a \zeta + \gamma a^2 \\ &= \kappa a u + \phi_0 + \beta a + \gamma a^2. \end{aligned}$$

Thus the shifted parameters give the same phase, modulo 2π, at every sensor. The quadratic term does not change, while the direction term and the linear calibration term enter only through their sum $\kappa u + \beta$.

Proof of Theorem 2. Let $A \subseteq \mathcal{X}$ be any measurable set of observations. By the definitions of a pushforward measure and $Q = T_\# P$,

$$\begin{aligned} (F_\# Q)(A) &= Q(F^{-1}(A)) \\ &= P\big(T^{-1}(F^{-1}(A))\big) \\ &= P\big((F \circ T)^{-1}(A)\big) \\ &= P(F^{-1}(A)) \\ &= (F_\# P)(A). \end{aligned}$$

The fourth equality uses $F \circ T = F$. Because the equality holds for every measurable A, the observation distributions are identical.

For completeness, let a randomized alarm or uncertainty statistic be represented by a Markov kernel $K(x, \cdot)$. For any measurable output set B, its probability under an observation law μ is $\int K(x, B)\, d\mu(x)$. Taking $\mu = F_{\#}P = F_{\#}Q$ gives the same probability in both worlds. This remains true even when T changes the target, because the statistic receives only the observation. For a batch of n i.i.d. observations, the two laws are $(F_{\#}P)^{\otimes n}$ and $(F_{\#}Q)^{\otimes n}$, which are equal. Adding a fixed source dataset to the procedure adds the same constant argument in both worlds and therefore cannot break this equality. □

Proof of Proposition 1. We first include any internal randomization of the predictor. Let U be an auxiliary random variable, independent of the observation and with the same law in both worlds, and write the predictor as $\widehat{y}(X, U)$. Appending the common variable U does not change total variation, so $\mathrm{TV}(P_0 \otimes P_U, P_1 \otimes P_U) = \varepsilon$.

Take a maximal coupling (Z_0, Z_1) of these augmented laws. It satisfies $\mathbb{P}(Z_0 = Z_1) = 1 - \varepsilon$. On the event $Z_0 = Z_1 = z$, both worlds produce the same estimate $\widehat{y}(z)$. The triangle inequality then gives

$$\|y_0 - y_1\| \le \|\widehat{y}(z) - y_0\| + \|\widehat{y}(z) - y_1\|.$$

The left side is D. Multiplying by the indicator of $\{Z_0 = Z_1\}$ and taking expectations yields

$$\mathbb{E}_{P_0}\|\widehat{y} - y_0\| + \mathbb{E}_{P_1}\|\widehat{y} - y_1\| \ge D(1 - \varepsilon).$$

At least one of these two risks must therefore be at least $D(1 - \varepsilon)/2$, which proves the bound for point risk.

For an alarm $\psi \in [0, 1]$, the variational characterization of total variation gives

$$|\mathbb{E}_{P_1}\psi - \mathbb{E}_{P_0}\psi| \le \mathrm{TV}(P_0, P_1) = \varepsilon.$$

The same argument applies to a randomized alarm after adjoining U.

It remains to prove the result for prediction sets. Include the set procedure's randomness in U, and define $A_j = \{y_j \in C(X, U)\}$. Its guarantee in world 1 is $P_1(A_1) \ge 1 - \alpha$. Probability transfer under total variation gives

$$P_0(A_1) \ge P_1(A_1) - \varepsilon \ge 1 - \alpha - \varepsilon.$$

Together with $P_0(A_0) \ge 1 - \alpha$, the union bound gives

$$P_0(A_0 \cap A_1) \ge 1 - 2\alpha - \varepsilon.$$

On $A_0 \cap A_1$, the set contains both y_0 and y_1, so its diameter is at least $\|y_0 - y_1\| = D$. Reversing the roles of the two worlds gives the same bound under P_1. Finally, a probability cannot be negative, which replaces the right side by $[1 - 2\alpha - \varepsilon]_+$.

For the Gaussian statement, let the two means differ by a vector with norm Δ, and let both covariance matrices equal $\sigma^2 I$. Then

$$\mathrm{KL}(P_0 \| P_1) = \frac{\Delta^2}{2\sigma^2}.$$

Pinsker's inequality gives

$$\mathrm{TV}(P_0, P_1) \le \sqrt{\frac{1}{2}\mathrm{KL}(P_0 \| P_1)} = \frac{\Delta}{2\sigma}.$$

Combining this inequality with the trivial bound $\mathrm{TV}(P_0, P_1) \le 1$ gives the stated corollary. □

Proof of Proposition 2. Consider first the snapshot with multiple sources. Under the common shift $k_j \mapsto k_j + \eta$ and calibration change $\delta \mapsto \delta - R\eta$, the exponent of source j at sensor a becomes

$$\begin{aligned} r_a^\top (k_j + \eta) + (\delta_a - r_a^\top \eta) &= r_a^\top k_j + r_a^\top \eta + \delta_a - r_a^\top \eta \\ &= r_a^\top k_j + \delta_a. \end{aligned}$$

Each term in the sum is unchanged, so the complete snapshot is unchanged. Because the same η is added to every k_j, differences such as $k_j - k_{j'}$ are preserved. Multiple sources may therefore identify relative wavevectors, but the observation remains invariant to their common shift.

Now consider two frequencies. Suppose that a direction change $q \mapsto q + \zeta$ can be compensated by one phase calibration that is shared across both frequencies. At frequency ω_ℓ, compare sensor a with sensor 1. This removes the arbitrary complex source amplitude and any phase that is constant across sensors. Equality of the original and transformed relative phases requires

$$(\delta'_a - \delta'_1) - (\delta_a - \delta_1) \equiv -(r_a - r_1)^\top \frac{\omega_\ell}{c}\zeta \pmod{2\pi}.$$

Stacking these relations for $a = 2, \dots, A$ gives

$$\delta'_{\mathrm{rel}} - \delta_{\mathrm{rel}} \equiv -D\frac{\omega_\ell}{c}\zeta \pmod{2\pi}.$$

The left side is the same for $\ell = 1$ and $\ell = 2$ because the hardware phase field is shared. Subtracting the two congruences therefore yields

$$D\frac{\omega_1 - \omega_2}{c}\zeta \in 2\pi\mathbb{Z}^{A-1},$$

which is Eq. 10. If the admissible shift domain contains no nonzero ζ satisfying this condition, no nonzero common direction shift can be hidden by the shared calibration.

In a sufficiently small unwrapped neighborhood, the integer vector must be zero. Because $\omega_1 \neq \omega_2$, the condition then reduces to $D\zeta = 0$. Full column rank of D implies $\zeta = 0$, which proves local identifiability. Globally, however, a nonzero ζ may map to a vector of integer 2π wraps, so full column rank alone does not exclude discrete aliases. If each frequency instead has an independent phase field, choose $\delta'_\ell = \delta_\ell - R(\omega_\ell\zeta/c)$. The cancellation in Theorem 3 then holds separately at each frequency, and the gauge persists. □

Proposition 3 (Criterion for a restricted subspace)**.** *Consider the locally unwrapped phase model $z = Rk + B\theta + \epsilon$, where the columns of B contain every admissible calibration mode, including the mode that is constant across sensors when its phase is unknown. Let $P_B^\perp = I - BB^\dagger$. A direction displacement η is observationally removable by calibration exactly when $P_B^\perp R\eta = 0$. Hence $k \in \mathbb{R}^p$ is locally identifiable iff*

$$\mathrm{rank}(P_B^\perp R) = p, \quad \textit{equivalently} \quad \mathrm{rank}([B\ R]) = \mathrm{rank}(B) + p.$$

For $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, the Fisher information for k after profiling θ is $\sigma^{-2}R^\top P_B^\perp R$, which is nonsingular under the same condition.

Proof. The mean of the observation at parameters (k, θ) is $Rk + B\theta$. Consider a second parameter pair $(k + \eta, \theta + \xi)$. The difference between the two means is

$$R(k + \eta) + B(\theta + \xi) - (Rk + B\theta) = R\eta + B\xi.$$

The displacement η is therefore hidden by a calibration change exactly when there exists a ξ such that

$$R\eta + B\xi = 0.$$

Such a ξ exists if and only if $R\eta$ belongs to the column space of B. Because $BB^\dagger$ is the orthogonal projector onto $\mathrm{range}(B)$, a vector v belongs to this space if and only if $(I - BB^\dagger)v = P_B^\perp v = 0$. Applying this fact to $v = R\eta$ gives the criterion

$$P_B^\perp R\eta = 0.$$

Local identifiability of all p coordinates of k requires the only removable displacement to be $\eta = 0$. Equivalently, $\ker(P_B^\perp R) = \{0\}$, which holds exactly when $\mathrm{rank}(P_B^\perp R) = p$.

To obtain the form using an augmented matrix, decompose each column of R as

$$R = BB^\dagger R + P_B^\perp R.$$

The columns of $BB^\dagger R$ already lie in $\mathrm{range}(B)$, whereas the columns of $P_B^\perp R$ are orthogonal to that space. Adding R to B therefore increases the rank by exactly $\mathrm{rank}(P_B^\perp R)$:

$$\mathrm{rank}([B\ R]) = \mathrm{rank}(B) + \mathrm{rank}(P_B^\perp R).$$

The equality $\mathrm{rank}(P_B^\perp R) = p$ is consequently equivalent to the rank condition in the proposition.

Finally, under Gaussian noise the negative logarithm of the likelihood, up to a constant, is

$$\frac{1}{2\sigma^2}\|z - Rk - B\theta\|_2^2.$$

For a fixed k, minimizing over θ removes the component of $z - Rk$ in $\mathrm{range}(B)$. The remaining residual is $P_B^{\perp}(z - Rk)$, so the profiled negative logarithm of the likelihood is

$$\frac{1}{2\sigma^2}\|P_B^{\perp}(z - Rk)\|_2^2.$$

Because $P_B^{\perp}$ is symmetric and idempotent, the Hessian of this profiled negative log-likelihood with respect to k is

$$\mathcal{I}_k = \sigma^{-2} R^{\top} P_B^{\perp} R.$$

For every vector v,

$$v^{\top}\mathcal{I}_k v = \sigma^{-2}\|P_B^{\perp} R v\|_2^2.$$

Thus $\mathcal{I}_k$ is nonsingular exactly when $P_B^{\perp} R$ has full column rank, which is the same identifiability condition derived above. □

D Reproducibility details

The simulated ULA contains 16 sensors spaced by half a wavelength, which gives $\kappa = \pi$. Each observation has a global phase drawn uniformly from $[-\pi, \pi]$ and circular complex noise with standard deviation 0.05 in each real component. The models use two hidden SiLU layers and AdamW. Hidden widths of 64 and 59 keep their parameter counts close. Training uses minibatches of 512, at most 40 epochs, and early stopping on validation loss.

The analytic posterior integrates a uniform phase slope nuisance convolved with Gaussian measurement error on a grid of 1,201 points. Reported values are read from JSON records rather than entered by hand. These records contain the seeds, data sizes, training histories, validation factors, metrics, and checkpoints. For independent test cases, coverage uses 95% Wilson intervals; at $n = 5{,}000$, their half width is at most 0.014. Repeated seeds or panels use standard deviations or 95% t intervals. Deterministic identity checks and serially related physical frames receive no inferential interval.

For the clean environment audit, we created a disposable Python 3.12 environment from a SHA-256 dependency lock and built the package from source. All 39 tests passed in 7.48 seconds on CPU. The artifact includes the lock, command, and machine-readable result. This check reconstructs the environment on the same machine; it is not an independent replication. That recorded audit predates 21 tests for the fixed-drift, component-level, dense-MIMO, and related follow-ups; a fresh extraction of the current archive passes all 60 tests in the documented CPU environment.

Table 4: Implementation details for the learned models. All models use AdamW with learning rate 10^{-3}, weight decay 10^{-5}, batch size 512, at most 40 epochs, and validation patience 7.

model	input dim.	width	output	parameters	objective/calibration
I/Q Gaussian	32	64	2	6,402	NLL / scale
adjacent Gaussian	46	59	2	6,433	NLL / scale
all pair quantile	256	23	3	6,535	pinball / split conformal
observable slope quantile	256	23	3	6,535	pinball / orbit convolution

D.1 Data separation and executable audit trail

The learned studies use 20,000 training examples, 4,000 validation examples, and 5,000 newly generated test examples per drift condition. Training, validation, and the five test conditions use disjoint simulator seeds (101, 102, and 1000 through 1004); network initializations use seeds 0 through 4. Interval scales and conformal corrections are selected on validation data only. The test target range is narrowed from $[-0.7, 0.7]$ to $[-0.4, 0.4]$, so the reported collapse is not caused by clipping against the edge of the target domain. The orbit head study retains seeds 101 and 102 for training and early stopping, adds a disjoint calibration split of 4,000 examples at seed 103, and uses test seeds 2000

through 2004. Its direct and orbit heads therefore receive the same independent examples for interval calibration.

The fixed-drift audit performs no training or tuning. It verifies every direct and orbit checkpoint hash against the successful orbit-head record, then tests 33 equally spaced values of $\beta_0 \in [-0.8, 0.8]$. Five test panels (seeds 4100 through 4104) contribute 1,000 examples each at every drift. Within a panel, target, global phase, and noise draws are held fixed across drift values. Coverage is reported with pointwise Wilson intervals and Bonferroni–Wilson simultaneous intervals over the 33 conditions; examples estimate coverage of the frozen ensemble and are not treated as independent training runs.

The artifact can replay the deterministic and analytic checks as well as the evaluations of saved models. A separate mode retrains the models before running later evaluations. Experiments on physical data run only when the external data are available. Each run writes a machine-readable record, with reproduced outputs kept separate from the frozen results. The artifact documentation gives the commands, dependencies, checksums, and file locations.

E ANCHOR WITH A KNOWN DIRECTION

For target and anchor statistics

$$t = \kappa u + \beta + \epsilon, \qquad t_a = \kappa u_a + \beta + \epsilon_a,$$

the estimator

$$\widehat{u} = \frac{t - (t_a - \kappa u_a)}{\kappa}$$

cancels β exactly and has Gaussian error $(\epsilon - \epsilon_a)/\kappa$. Under independent noise with equal variance its standard deviation is $\sqrt{2}\sigma/\kappa$, independent of the drift envelope. In 10,000 trials per condition, coverage ranges from 89.7% to 90.4% over all tested bounds, and width is fixed at 0.0444.

We evaluate the intervention at target drift bound $B = 0.8$. Anchor noise ranges from 0.5 to 4 times the target noise, and the difference between target and anchor ramps is drawn uniformly from $[-M, M]$. With equal noise and $M = 0.2$, the naive anchor covers only 34.0%; a posterior that marginalizes the declared mismatch bound achieves 90.0% coverage by widening from 0.044 to 0.120. With $M = 0$, both methods remain near nominal coverage across all noise levels.

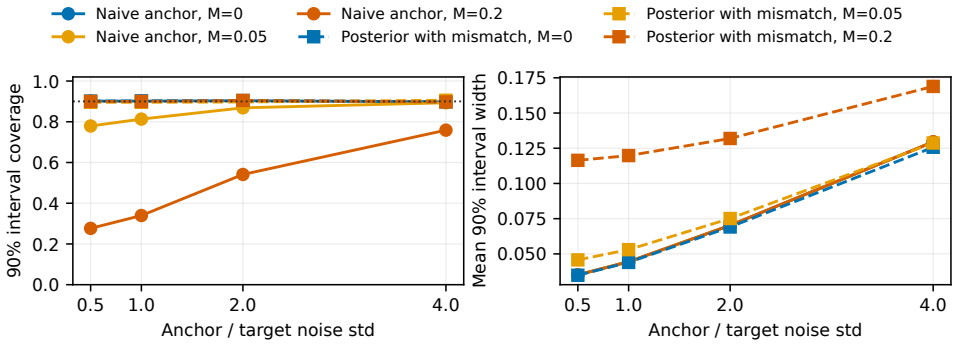


Figure 6: Stress test for an anchor with known direction (10,000 cases per condition). Shared drift cancels regardless of magnitude. An unmodeled difference between the target and anchor ramps causes undercoverage. Integrating over a valid mismatch bound restores coverage by increasing width.

Operational estimation of the mismatch bound. The preceding stress test treats M as supplied. A practitioner can instead estimate a horizon-specific bound using target-like, known-direction witnesses. At commissioning and scheduled recalibration, collect independent blocks that span the intended direction, range, temperature, frequency, and device-state envelope. For anchor a and witness j in block i, estimate the path-specific ramp $b_{q,i}$ from replicate captures as

$$\widehat{b}_{q,i} = \mathrm{loc}_{1 \le r \le R_i}\{t_{q,i}^{(r)} - \kappa u_q\}, \qquad \Delta_i = \max_{j \in \mathcal{R}_i} d(b_{j,i}, b_{a,i}), \qquad U_i = \max_{j \in \mathcal{R}_i} \left\{ d(\widehat{b}_{j,i}, \widehat{b}_{a,i}) + c_{j,i} \right\}.$$

The location estimator and replicate count R_i are fixed before collection. Here $d(x, y) = |x - y|$ for an unwrapped statistic and $d(x, y) = \min_{m\in\mathbb{Z}} |x - y + mP|$ for period P. The radius $c_{j,i}$ must be chosen from the replicate noise calibration so that, with simultaneous failure probability at most α_b, it upper-bounds the estimation error in $b_{j,i} - b_{a,i}$ for every retained pair. Thus $U_i \geq \Delta_i$ on that event. Using a block maximum, rather than treating serial frames as independent, makes the unit of calibration a deployment episode. For n exchangeable blocks, choose a second error budget α_M, set

$$k = \lceil (n + 1)(1 - \alpha_M) \rceil , \qquad \widehat{M} = U_{(k)},$$

and require $k \leq n$; otherwise the data cannot support a finite bound at that level. Because the k-th order statistic of the U_i's upper-bounds that of the latent Δ_i's on the simultaneous confidence event, the exchangeable-rank argument gives $\Pr\{\Delta_{n+1} > \widehat{M}\} \leq \alpha_b + \alpha_M$ (Romano et al., 2019). The witnesses must use the same target/anchor signal paths and the same time separation as the claimed deployment horizon; pooling across an unvalidated operating regime does not preserve this statement.

At run time, acquire the anchor contemporaneously and apply frozen SNR and repeatability gates. For claimed horizon h, also require $G_t = d(\widehat{b}_{a,t}, \widehat{b}_{a,t-h}) \leq G_{\max}(h)$, where $G_{\max}(h)$ is frozen from same-horizon commissioning blocks; this is a recency diagnostic, not proof that target/anchor mismatch is small. Recompute $\widehat{M}$ only at a prespecified cadence, and abstain or recalibrate when the operating stratum or any gate is violated, the elapsed horizon is longer than validated, or the interval exceeds an application width limit. The rank guarantee is pointwise for each scheduled update; repeated monitoring requires a sequential error budget or a time-uniform bound. Estimating a bound alone does not justify the uniform mismatch law used in the stress test. For a distribution-free bound claim, enlarge a $(1-\alpha_\epsilon)$ interval for $(\epsilon-\epsilon_a)/\kappa$ by $\widehat{M}/|\kappa|$ on each side and allocate $\alpha_\epsilon + \alpha_b + \alpha_M \leq \alpha$. A narrower marginal posterior requires estimating the full mismatch law on separate exchangeable blocks and calibrating the complete pipeline end to end. This protocol is a proposed deployment procedure, not a result validated by our three-sequence physical audit; temporal drift statistics and denser physical reference data remain necessary to test its exchangeability, width, and abstention rate.

F PHYSICAL ADC INTERVENTION

Automotive ADC with associated labels. The source release contains synchronized images, metric labels for object centers, and raw frames with shape 128 samples by 255 chirps by 4 receivers by 2 TDM transmitters (Gao et al., 2022a). We pin the public mirror revision and hash the label plan, raw path manifest, and downloaded contents. The expanded plan scans 22 available sequences using labels alone. Eligibility requires exactly one object, range 2 to 20 m, and label spatial frequency $u = p_x/\sqrt{p_x^2 + p_y^2} \in [-0.75, 0.75]$. Every sequence with at least 24 eligible labels enters the audit. We divide its eligible timeline into 24 equal strata and take each midpoint. This selects 360 frames from 15 sequences on four dates before any newly downloaded ADC array is examined.

Before the intervention, we recorded the selection, processing, metrics, thresholds, exclusions, and stopping rule in a repository contract. The plan was fixed in advance but was not publicly preregistered. The source revision is `27fe4a0bb0bfd80614ab0e64b862934fcf7ed321`; the SHA-256 digests of the label plan and raw path manifest are, respectively, `6a11614d2941d7f691e26956b0ba0cf170f8c80fc4f0864e1298df163e822717` and `75ef5bf801e0f24459b8c95aa2424a1011a15eac233a0bc146396c861e2b8069`.

We apply range and Doppler FFTs after Hann windowing and concatenate the documented 4×2 channels into an eight-element virtual ULA. The channel order is fixed from the initial single sequence audit and is not selected again for each sequence. Within two bins of the labeled range, the label selects the range–Doppler cell whose peak best matches the steering direction. A baseline peak is associated within circular radius 0.25. It must be a strict local maximum and at least 10 dB above the median. The selected cell remains fixed during every intervention on that frame.

The fixed criterion retains 281 of 360 frames from 14 sequences. Thirteen of 15 sequences meet the prespecified minimum of 12 retained frames; one retains four and one retains none. The 79 exclusions comprise 62 labels outside the positive half of the processed range FFT, 10 peaks below

10 dB, and seven peaks that are not strict local maxima. The first execution stopped when it reached the range mismatch: label eligibility extended to 20 m, whereas the processed half spectrum ends near 14.3 m. No complete result was written. Version 1.1 keeps the fixed selection and processing but records these cases as explicit range exclusions.

Table 5 reports the three grid-aligned ramps. For each ramp, the median, p95, and maximum displacement errors are zero. Two global phase controls on every retained frame also give zero displacement. The correction uses the known injected ramp and is therefore an oracle, not an estimated anchor. Frames within a sequence are serially related, so the recording sequence is the replication unit and no frame level significance test is reported.

Table 5: Expanded automotive ADC intervention. Errors are circular spatial frequency units over 281 retained frames from 14 sequences per ramp.

ζ	cases	MAE with unchanged label	MAE with supplied ramp
0.0625	281	0.0664	0.0423
0.1250	281	0.1211	0.0423
0.2500	281	0.2448	0.0423

Physical TX/RX factorization and its boundary. For receiver $r \in \{0, \ldots, 3\}$, transmitter $t \in \{0, 1\}$, and transmitter-major virtual index $a = 4t + r$, the virtual gauge ramp factors as

$$\exp(-i\pi a\zeta) = \exp(-i\pi(4t)\zeta)\exp(-i\pi r\zeta).$$

It is therefore realizable by coordinated phase offsets on the physical TX and RX channels. This algebra does not imply that a TX-only, RX-only, or isolated component fault is gauge-confounded.

We froze a second contract before testing that distinction. It inherits the 281 retained cells, three shifts, channel order, beam grid, and sequence as the replication unit. Besides the coordinated factorization, it applies TX-only, RX-only, single-RX, and fixed RMS-matched random separable fields. For each field, least squares projects the specified unwrapped component phase onto an affine virtual profile; the primary metric is the relative L_2 defect from the baseline spectrum translated by that fitted affine component. The first retained frame in each sequence is also processed from raw ADC after every fault at $\zeta = 0.125$ and both global controls.

All prespecified gates pass. The maximum raw-domain commutation defect over 98 probes is 2.21×10^{-16}; the coordinated factor has zero complex-factor defect, zero peak-displacement error, and a median sequence-level spectral defect of 2.89×10^{-16}. Across the 14 sequences, the corresponding median defects are 0.298 (TX-only), 0.191 (RX-only), 0.169 (single RX), and 0.813 (random separable); every sequence-level median exceeds the frozen 0.05 boundary for every non-affine family. Global phase and oracle reversal are exact to numerical precision. These are controlled injections into measured ADC channels, not observations of natural hardware degradation.

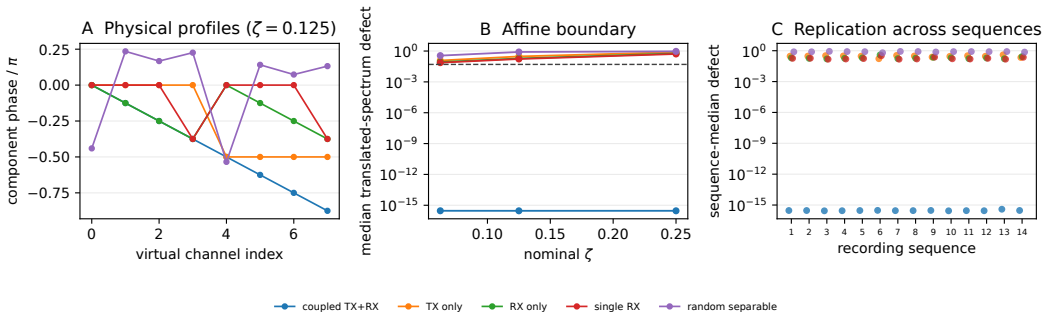


Figure 7: Component-level boundary on the retained automotive ADC cells. The coordinated TX/RX field is affine in virtual-array order (A) and translates the complete spectrum without deformation (B). TX-only, RX-only, single-RX, and random separable faults retain non-affine spectral defects in every recording sequence (C). The dashed line is the frozen defect threshold 0.05.

G DENSE PLANAR MIMO GEOMETRY AUDIT

Array density does not by itself remove the gauge in the narrowband far field. For transmitter locations p_t, receiver locations q_r, and virtual locations $v_{tr} = p_t + q_r$, every wavevector shift η factors as

$$\exp(-i\pi v_{tr}^{\mathsf{T}}\eta) = \exp(-i\pi p_t^{\mathsf{T}}\eta)\exp(-i\pi q_r^{\mathsf{T}}\eta).$$

We froze a geometry-only audit before execution: one 2-by-4 ULA, 32 random full-rank 4-by-4 planar arrays, and 32 jittered 12-by-16 planar MIMO arrays with 192 virtual channels. Each geometry receives 32 target/shift probes, for 2,080 probes total. The six component fields are coordinated TX/RX affine, TX-affine only, RX-affine only, single TX, single RX, and RMS-matched random separable phase. Geometry is the replication unit; probe cases are repeated deterministic checks.

Eleven of twelve frozen gates pass. Across all geometries, maximum coordinated factor, steering, and translated-spectrum defects are 9.64×10^{-16}, 3.16×10^{-15}, and 1.77×10^{-15}, respectively. The coordinated field has the smallest median spectrum defect in every geometry family. All five non-affine families clear both planar gates: their median normalized affine residuals range from 0.394 to 0.923 in the sparse arrays and from 0.293 to 0.986 in the dense arrays; median spectrum defects range from 0.0227 to 0.0584 and from 0.0518 to 0.386, respectively. Thus, more virtual channels expose generic non-affine component faults but cannot remove the coordinated affine gauge.

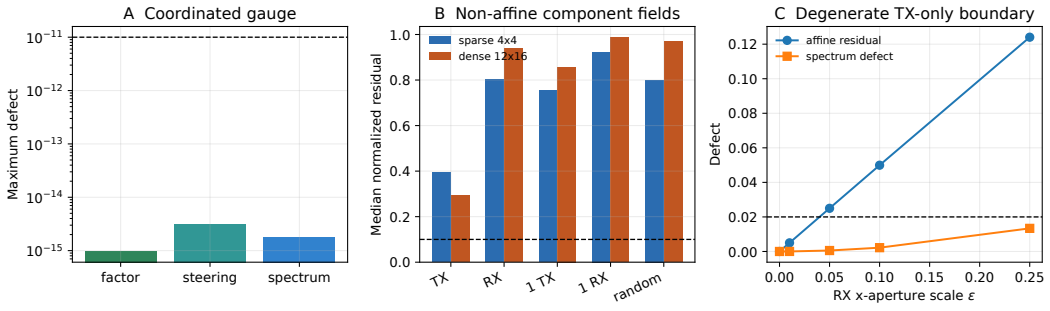


Figure 8: Frozen dense planar MIMO audit. Coordinated TX/RX phase remains an exact gauge at numerical precision (A). All five non-affine component families clear the 0.10 residual gate in both planar families (B). In the prespecified degenerate TX-only construction, RX aperture breaks exact affinity continuously, but the spectrum defect at $\epsilon = 0.25$ is 0.0134 and misses the frozen 0.02 endpoint gate (C).

The failed endpoint is retained. At zero RX aperture in the shift direction, the TX-only field is exactly gauge-confounded: factor and spectrum defects are below 5×10^{-16}. At $\epsilon = 0.25$, normalized affine residual is 0.124, but spectrum defect is only 0.0134. Therefore aperture makes this fault non-affine, yet the chosen Bartlett distortion threshold is not reached at the frozen endpoint. The audit is synthetic and does not model a named radar, mutual coupling, wideband effects, near-field curvature, manufacturing tolerances, fault prevalence, or learned UQ.

H DENSE PHYSICAL MIMO PHASE-CONTENT AUDIT

We next applied a frozen admissibility protocol to the public *Raw ADC Data of 2D-MIMO MMWave Radar for Carry Object Detection* release (Gao et al., 2022b). Its 65.1 GB archive contains 2,917 MAT frames from a 12-TX by 16-RX cascaded radar. Before opening any selected array, we ranked paths by SHA-256 within each sequence and selected eight per sequence. The resulting manifest contains 160 frames from all 20 sequences and has SHA-256 `5a2e9707bf0bb9628997140811c55b830c3b7d43e5242f63912286c47340ea6e`.

All 160 frames pass the phase-content gate. Each contains one native complex `adcData` tensor that permutes uniquely to 256 samples by 61 chirps by 16 receivers by 12 transmitters. No frame is rejected; real and imaginary energy fractions range from 0.49917 to 0.50083. Thus this dense physical release retains the coherent per-channel phase required to test the gauge.

The physical steering audit nevertheless stops at its third frozen integrity gate. Neither the official repository nor the downloaded archive includes a capture, conversion, chirp-order, or antenna-order record mapping the MAT transmitter and receiver axes to the physical antenna IDs. Hardware geometry alone is insufficient because the TDM chirp sequence determines transmitter order. Assigning a convenient permutation after inspecting the data would make the translated-spectrum result untraceable and effectively construct the steering convention being tested. We therefore retain the positive phase-content result but make no physical planar steering claim from this release. This is a data-documentation boundary, not evidence that dense hardware removes the gauge.

Independent audit of the complete spectrum. Figure 9 visualizes one of the 150 RaDICaL interventions. Multiplication by a phase ramp across channels shifts the spatial spectrum but cannot change a noncoherent range–Doppler statistic. The checksum, source configuration, every case, and plotting code are included in the artifact.

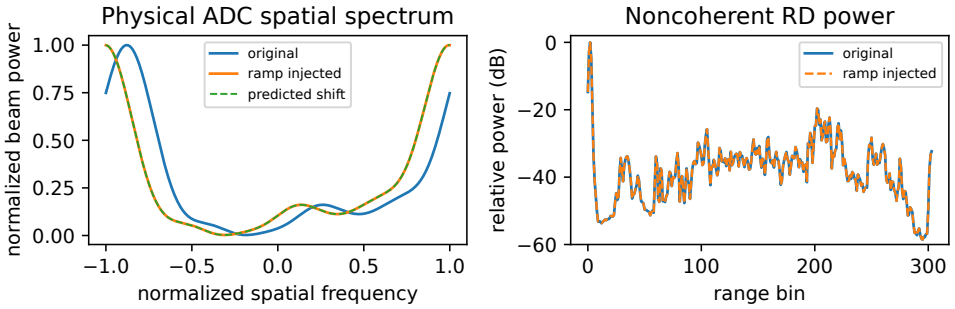


Figure 9: Controlled ramp $\zeta = 0.125$ on one physical RaDICaL ADC frame. The injected spectrum overlaps its analytically translated prediction (left), whereas original and injected noncoherent range–Doppler profiles overlap (right).

I Calibration at multiple levels

We recalibrate each saved ensemble on the same 4,000 validation cases for nominal coverage levels of 50, 80, 90, and 95%. No test examples enter this step. At the training envelope $B = 0.05$, the largest difference from nominal coverage is 3.6 percentage points. At $B = 0.8$, even the nominal 95% intervals cover only 8.2% for I/Q, 6.5% for adjacent relations, and 6.6% for quantiles based on all pairs.

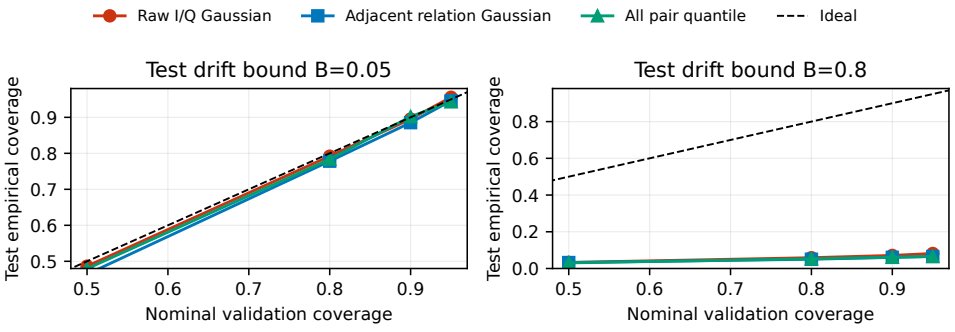


Figure 10: Calibration curves for saved ensembles trained with five seeds and evaluated on 5,000 independent cases per condition. Validation calibration transfers within the training drift envelope (left) but not under the expansion of the envelope that is confounded by gauge (right).

J LEARNED PREDICTION FROM OBSERVABLES TO AN ORBIT

This experiment separates information in the observation from information supplied at deployment. The network estimates $q = \pi u + \beta$ from relational features using all pairs. If $e = \widehat{q} - q$ is its calibration residual and the declared nuisance law is $\beta \sim \mathrm{Unif}[-B, B]$, then

$$u = \widehat{q}/\pi - (e + \beta)/\pi.$$

Without using test labels, we estimate the 5th and 95th percentiles of $e + \beta$ by combining validation residuals with a fixed nuisance grid of 1,001 points. The residuals are grouped into eight prespecified bins of predicted q, with 355 to 554 examples in each bin. This allows residual scale to vary with the network output available at deployment.

Training, early stopping, interval calibration, and testing use disjoint seeds and 20,000, 4,000, 4,000, and 5,000 examples per test condition, respectively. The direct head reuses its five checkpoints, discards the earlier interval correction, and is recalibrated on the same split as the orbit head. Both heads have 6,535 parameters and identical inputs; only the target changes from u to q. We also form a conservative set using the 90th percentile r of the absolute residual: $[(\widehat{q}-B-r)/\pi, (\widehat{q}+B+r)/\pi]$. We do not interpret this set as a sharp probabilistic interval.

Table 6: Results for matched learned heads (5,000 independent test cases per bound; 95% Wilson half width at most 0.014). "Frozen" uses $B = 0.05$ at every condition; "cert." is the conservative orbit set.

B	direct cov.	orbit cov.	orbit width	frozen cov.	oracle cov.	cert. cov.
0.00	1.000	0.891	0.0036	1.000	0.891	0.940
0.05	0.898	0.900	0.0288	0.900	0.903	0.999
0.20	0.232	0.899	0.1147	0.227	0.892	1.000
0.40	0.105	0.900	0.2294	0.105	0.897	1.000
0.80	0.055	0.900	0.4588	0.056	0.895	1.000

At $B = 0.8$, the orbit interval score is 0.483, compared with 2.291 for direct target CQR. RMSE for q stays between 0.0037 and 0.0038 across bounds. All ten prespecified checks pass, including exact simulator replay, independent calibration, monotone widening, the frozen bound ablation, and stable oracle coverage. At $B = 0.8$, the Wilson interval for orbit coverage is [0.891, 0.908].

We retain the first prespecified version as a failed run. Although its coverage was 90.04 to 90.16% for every nonzero bound, it overcovered the case with no drift and the oracle cases at about 94 to 95%. A change in floating point operation order also produced a replay defect of 7.65×10^{-6}. We wrote the conditional protocol and its thresholds only after recording this failure. Both contracts and result records are kept in the artifact.

This experiment supports the learned method only when the nuisance law is supplied in the controlled ULA model. It does not estimate B. The direct head for an unknown envelope receives less information and is therefore not a matched comparison.

J.1 FIXED-DRIFT CONDITIONAL COVERAGE

Before evaluating persistent drift, we froze a separate protocol, retained the successful five-member checkpoints and calibration objects above, and changed only the test data-generating process. For every $\beta_0 \in \{-0.8, -0.75, \ldots, 0.8\}$, five independently seeded panels draw $u \sim \mathrm{Unif}[-0.4, 0.4]$, global phase, and complex noise. The same exogenous draws are reused across drift values within each panel. The estimand is

$$c_m(\beta_0) = \Pr\{u \in C_m(X) \mid \beta = \beta_0\},$$

not variation across new training runs. Checkpoint hashes, exact simulator replay, and common-random-number checks all pass.

The nuisance-marginalized orbit interval has trapezoidal integrated coverage 0.9063, consistent with the earlier i.i.d. result, but conditional coverage is zero at $\beta_0 \in \{-0.8, -0.75, 0.75, 0.8\}$ and one

over the central grid. At a zero-coverage condition, the simultaneous 95% Wilson upper endpoint is 0.002. The direct target interval integrates to 0.0491. In contrast, oracle coverage stays between 0.8802 and 0.9062 at mean width 0.0037, and the certified set has minimum coverage 0.9532 at constant width 0.5139. All eight frozen checks pass. This controlled result distinguishes a valid average over a declared population from protection for an individual sensor with persistent drift; it does not estimate how frequently physical devices occupy each drift.

K EXACT LEARNED AUDIT WITH IDENTICAL INPUTS

Before running this audit, we recorded its paired design in the repository. We draw 5,000 source examples according to

$$u \sim \mathrm{Unif}[-0.4, 0.4], \qquad \beta \sim \mathrm{Unif}[-0.05, 0.05].$$

Global phase is random, and circular complex noise has standard deviation 0.05 in each real component. For each $\eta \in \{-0.2, 0.2\}$, the deployment world sets $u_\eta = u + \eta$ and $\beta_\eta = \beta - \pi\eta$, while reusing the same noise and stored tensor. The clean forward computations have maximum absolute defect 7.21×10^{-15}. All three tensors have the same SHA-256 digest:

`f417ec1f131975b4f8a722e63cb729b1038c9c54b66fac96735dadbff817a2f1.`

All 15 checkpoints are loaded with their original normalization, calibration, and ensemble aggregation. There is no retraining, recalibration, bound input, or shift metadata. For every method, the predictive center, both bounds, width, and ensemble disagreement change by exactly zero. The target dependent results appear in Table 7. The largest numerical violation of the triangle bound for point error is 4.48×10^{-8}.

Table 7: Saved ensembles on noisy inputs that are identical at the byte level, with targets paired by the gauge. Width and disagreement are unchanged in every shifted world.

method	world	RMSE	cov. 90%	width
I/Q	source	0.0108	0.893	0.0347
I/Q	$\eta = \pm 0.2$	0.199 to 0.201	0.000	0.0347
adjacent relational	source	0.0095	0.894	0.0294
adjacent relational	$\eta = \pm 0.2$	0.199 to 0.201	0.000	0.0294
all pair quantile	source	0.0094	0.907	0.0296
all pair quantile	$\eta = \pm 0.2$	0.199 to 0.201	0.000	0.0296

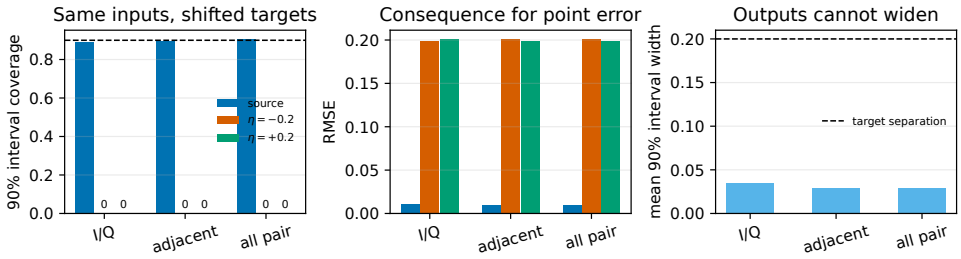


Figure 11: Exact paired consequence for three saved ensembles of five models. Inputs that are identical at the byte level force identical predictions and disagreement. The target shift changes coverage and RMSE, but cannot alter an uncertainty score based only on input. This is a deterministic constructed audit, not a sample of natural shift.

L SENSITIVITY TO A WRAPPED LIKELIHOOD

The main scalar experiment uses $u \in [-0.7, 0.7]$. At its largest bound, $B = 0.8$, the noiseless phase mean still satisfies $|\kappa u + \beta| \le 0.7\pi + 0.8 = 2.999 < \pi$. None of the 10,000 samples crosses

the phase cut. Consequently, a wrapped normal likelihood formed from five aliases gives the same reported metrics as the unwrapped likelihood (Table 8).

As a stress test, we widen the field to $[-0.95, 0.95]$, where 4.29% of observations wrap at $B = 0.8$. Treating phase as linear gives 86.1% coverage and NLL 29.70. The circular likelihood restores 90.0% coverage by widening the equal tail interval from 0.423 to 0.669. Its larger RMSE comes from posterior mass on aliases, not optimization failure.

Table 8: Sensitivity to a circular likelihood (10,000 cases per row). "Naive" treats wrapped phase as unbounded; "circular" sums wrapped normal aliases.

u range	B	wrapped	method	NLL	cov. 90%	width
$[-0.7, 0.7]$	0.05	0.00%	circular	-2.916	0.897	0.0429
$[-0.7, 0.7]$	0.40	0.00%	circular	-1.398	0.900	0.2158
$[-0.7, 0.7]$	0.80	0.00%	circular	-0.820	0.897	0.4035
$[-0.95, 0.95]$	0.80	4.29%	naive	29.703	0.861	0.4229
$[-0.95, 0.95]$	0.80	4.29%	circular	-0.686	0.900	0.6691

M GAUGE BOUNDARY IN THE NEAR FIELD

Theorem 3 assumes a planar wavefront. At shorter ranges, a spherical model may reveal curvature that breaks the gauge (Liu et al., 2023b). We therefore replace the ULA phase with

$$\phi_a(R, u) = -2\pi \left[\sqrt{R^2 - 2Rr_a u + r_a^2} - R \right]$$

in wavelength units. The centered sensors have half wavelength spacing and aperture D. We use the Fraunhofer distance $R_F = 2D^2$ to normalize range, not as a universal operating boundary (Björnson et al., 2021). For each pair $u' = u + \eta$, least squares fits the phase difference with an admissible affine calibration $c + br_a$. This fit establishes that a nuisance value exists; it does not estimate calibration from the snapshot.

The audit fixes $u \sim \mathrm{Unif}[-0.4, 0.4]$, $\eta \in \{-0.2, 0.2\}$, 10,000 paired cases, and circular complex Gaussian noise with standard deviation 0.05 per Cartesian component. If the two compensated complex means are separated by Δ, their Gaussian laws with equal covariance have exact total variation

$$\mathrm{TV} = 2\Phi\left(\frac{\|\Delta\|_2}{2\sigma}\right) - 1.$$

The corresponding lower bound for absolute error is $0.1(1 - \mathrm{TV})$. This conditional quantity is not the TV of a mixture over directions.

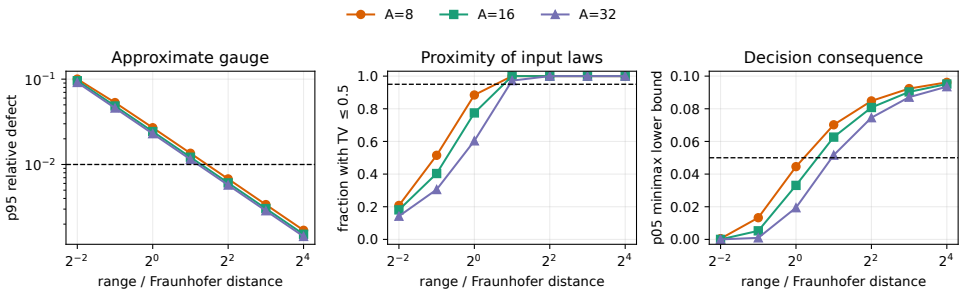


Figure 12: Prespecified boundary audit with spherical waves after the best affine phase compensation. Dashed lines are the locked gates. Curvature distinguishes many pairs at close range, but the residual decreases with range and restores a strong decision lower bound for two points.

Table 9: Near-field boundary for the array with 16 sensors. Each row has 10,000 paired cases; the fifth column is the fraction with TV at most 0.5.

R/R_F	defect p95	TV median	TV p95	TV $\leq .5$	lower p05
0.25	0.0958	0.904	1.000	0.181	0.0000
0.50	0.0485	0.599	0.947	0.404	0.0053
1.00	0.0243	0.326	0.669	0.775	0.0331
2.00	0.0122	0.167	0.373	1.000	0.0627
4.00	0.0061	0.084	0.192	1.000	0.0808
8.00	0.0030	0.042	0.097	1.000	0.0903
16.00	0.0015	0.021	0.049	1.000	0.0951

All eight prespecified checks pass. At $16R_F$, p95 defect ranges from 0.0014 to 0.0017 across $A \in \{8, 16, 32\}$, and every pair has TV ≤ 0.5. For target separation 0.2, the p05 minimax bound is 0.0935 to 0.0962. At $0.25R_F$, median TV rises to 0.904 for $A = 16$ and 0.975 for $A = 32$, showing that curvature can reveal the ambiguity at close range. The planar wave control has maximum defect 2.17×10^{-15}. This audit measures the limit of the far field approximation; it is not a learned experiment and does not include multipath or unknown range.

N LEARNED AUDIT IN THE NEAR FIELD: RETAINED NEGATIVE RESULT

The analytic result leaves a practical question: can a learned predictor use the remaining curvature at the chosen noise level? We fixed a separate contract before training. The simulation uses 16 sensors with half wavelength spacing, exact spherical propagation, known $R/R_F \in \{0.25, 0.5, 1, 2, 4, 8, 16\}$, $u \sim \mathrm{Unif}[-0.4, 0.4]$, independent affine phase nuisance, and complex Gaussian noise of standard deviation 0.05 per Cartesian component. At each range, we generate 4,000 training, 1,000 validation, 1,000 calibration, and 5,000 test examples. Five seeds train quantile MLPs with matched parameter counts. One receives raw I/Q and range; the other receives range and the affine closure

$$c_a(x) = x_{a+1}x_{a-1}\overline{x_a}^2/\big(|x_{a+1}x_{a-1}x_a^2| \vee 10^{-8}\big).$$

The closure is invariant to phase that is constant or linear across sensors: the affine phase terms cancel because $(a + 1) + (a - 1) - 2a = 0$. Split conformal calibration is performed separately at each range, and the resulting 90% intervals are clipped to the known support.

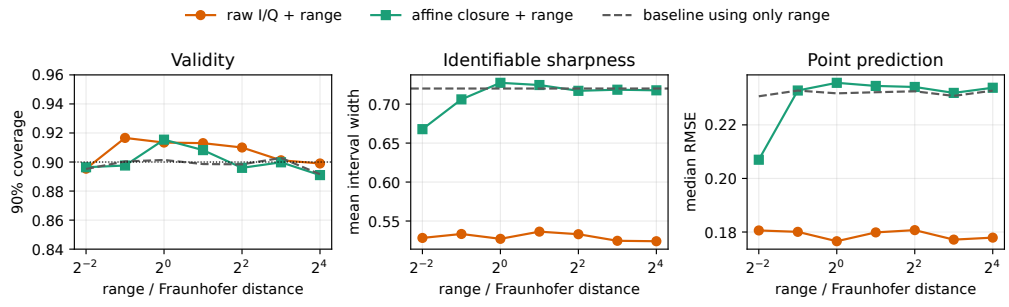


Figure 13: Prespecified learned audit in the near field. Coverage is valid after CQR for each range, but the head based on affine closure shows only a small and nonmonotone sharpness gain at short range. The raw head is not affine invariant and can exploit the supplied nuisance distribution; its narrow intervals therefore do not establish identifiability based on curvature.

The four integrity checks pass. The closure defect under random affine fields is at most 4.77×10^{-7}, the model sizes differ by 1.09%, all values are finite, and ensemble coverage lies in $[0.891, 0.917]$. The proposed transition, however, is not supported. At $0.25R_F$, closure width is 0.668, or 92.8% of the 0.72 range baseline, missing the 90% threshold. Width rises by only 0.050 at $16R_F$, rather than the required 0.10, and its Spearman association with log range is 0.393 rather than 0.75. We

retain this failed result without changing its thresholds. It does not rule out other architectures, but it provides no evidence for the prespecified learned transition. Unknown range, multipath, extended targets, motion, and physical near-field data remain open.

O ESTIMATING THE DEPLOYMENT LAW FROM A CALIBRATION PANEL

Supplying the nuisance law favors the orbit construction. For a fairer comparison, we keep all ten checkpoints and give both heads the same deployment panel. The panel contains raw snapshots with known fixture directions and is exchangeable with the test deployment. It provides measured calibration information rather than inferring drift from a confounded, unlabeled snapshot.

For an orbit prediction $\widehat{q}$, the first panel half estimates the effective error

$$z = \widehat{q}/\pi - u,$$

which includes nuisance variation and model residual. Its 5th and 95th percentiles define an asymmetric base interval. The untouched second half provides the conformal quantile of the CQR score $\max\{\ell - u, u - r\}$ (Romano et al., 2019). The direct head receives the same correction data, so both methods use equal deployment information.

We prespecify uniform and triangular laws on $[-0.8, 0.8]$, plus a symmetric bimodal law supported on $[-0.8, -0.48] \cup [0.48, 0.8]$. For each law and panel size $n \in \{500, 1000, 4000\}$, ten independently seeded panels calibrate a fixed test set of 20,000 examples. Confidence intervals below quantify variation across panels, not across test examples. A negative control calibrates on the original uniform $B = 0.05$ process before testing on uniform $B = 0.8$.

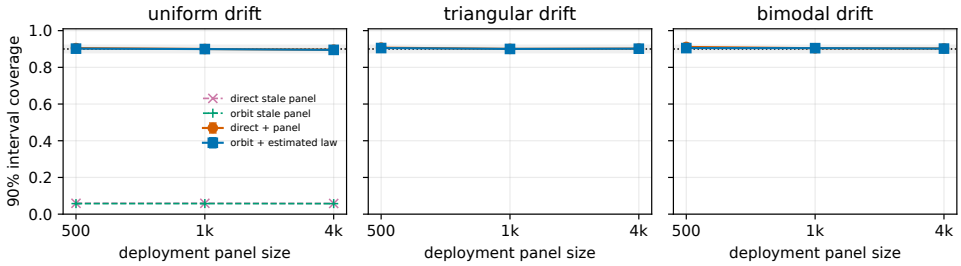


Figure 14: Coverage after estimating deployment uncertainty from a panel with known directions. Error bars are 95% t intervals over ten independent panels; the gray band is the locked 88 to 92% criterion. Both equally informed methods calibrate, whereas the controls using a stale panel remain near 6%.

Table 10: Results with 4,000 panel examples. Coverage, width, and interval score are means over ten independent panels.

law	direct cov.	orbit cov.	direct width	orbit width	direct score	orbit score
uniform	0.8963	0.8952	0.4554	0.4562	0.4835	0.4844
triangular	0.9020	0.9017	0.3516	0.3521	0.4044	0.4053
bimodal	0.9027	0.9022	0.4105	0.4109	0.4282	0.4286

All ten prespecified checks pass. With $n = 4000$, mean coverage lies between 0.895 and 0.903, and every 95% interval across panels intersects $[0.89, 0.91]$. Orbit widths and scores remain within 0.5% of the equally informed direct baseline. The triangular law gives an orbit interval 22.8% narrower than the uniform law, so the estimate reflects distributional shape as well as support. With a stale panel, coverage falls to 0.0582 and 0.0573. The result supports finite panel adaptation under exchangeability, not law estimation from unlabeled inputs or adaptation to nonstationary drift.

P PHYSICAL ANCHOR AUDIT ACROSS SEQUENCES

Does an early physical anchor remain valid in a later sequence? Before reading new ADC values, we selected three single car sequences from the pinned Gao et al. dataset (Gao et al., 2022a), recorded on April 9, April 30, and May 29. Within each sequence, the first 32 label selected frames form a calibration panel and the last 32 form the evaluation set. After label-guided range–Doppler association, an eight channel Bartlett peak gives the residual $r = \mathrm{wrap}_2(\widehat{u} - u_{\mathrm{label}})$, where $\mathrm{wrap}_2(z) = ((z+1) \bmod 2) - 1 \in [-1, 1)$. We subtract the panel circular mean and compare the current panel, both panels from other sequences, no correction, and a control with reversed sign. A frame is retained only when its local peak is at least 10 dB above median power. Each calibration and evaluation half must retain at least 24 of 32 frames.

The first contract is retained as a source integrity failure. Its first eligible April 9 label had no raw file with the same index, so the audit stopped before ADC analysis. Version 2 added raw file existence to the eligibility rule without changing the signal processing or scientific thresholds. Figure 15 and Table 11 report that locked follow up.

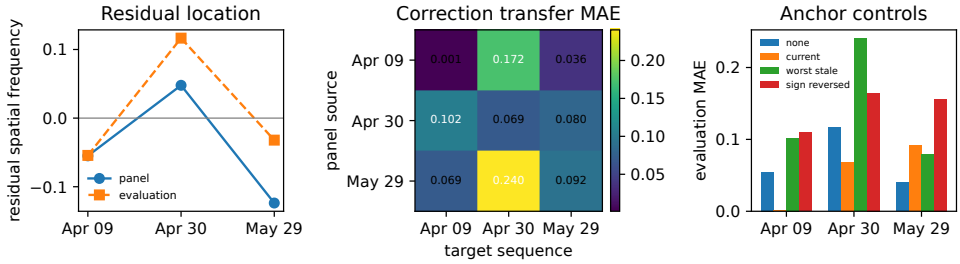


Figure 15: Prespecified physical audit of anchor transfer. Left: residual locations from the early panel and later evaluation frames. Center: evaluation MAE for each pair of panel source and target sequence. Right: no correction, current panel, worst stale panel, and control with reversed sign. The result is classified *unresolved*, not as hardware drift.

Table 11: Audit across sequences associated with a target. "Retained" gives calibration/evaluation counts after the frozen quality rule.

sequence	retained	panel loc.	eval. loc.	raw MAE	current MAE	worst stale MAE
Apr 09	32/32	-0.0550	-0.0543	0.0543	0.0012	0.1021
Apr 30	32/13	0.0478	0.1166	0.1166	0.0687	0.2403
May 29	16/32	-0.1238	-0.0321	0.0403	0.0917	0.0799

The source and numerical integrity checks pass, but minimum retention fails for the April 30 evaluation set and the May 29 calibration set. Changes in residual location are 0.0006, 0.0688, and 0.0917, so only April 9 satisfies the 0.02 criterion. On May 29, the current correction increases MAE from 0.0403 to 0.0917, while the stale April 9 offset gives 0.0363. These three label guided sequences support neither stable transfer nor a claim of hardware drift. Chronology alone does not justify exchangeability; a physical system needs a contemporaneous, validated reference.

Q COMPARISON WITH THE CLOSEST WORK

Table 12 compares this work with its closest theoretical and application precedents. It distinguishes standard components from our combination and evaluation based on the operator. Khurjekar & Gerstoft (2023) conformalize Monte Carlo dropout, ensembles, and quantile regression for acoustic DOA. Their guarantee assumes exchangeability, and their experiments use condition-matched training, calibration, and test data, with a separate calibration set for each environmental condition. In automotive radar, the preprint of Kulkarni & Reddy (2026) compares a three-member von Mises

ensemble with a normal-inverse-gamma evidential head on synthetic 16-channel MIMO range–Doppler–angle regions; its OOD tests alter SNR, phase, spatial leakage, or available channels. These are valuable domain-specific UQ settings in which the observed input changes. Our boundary is different: because Theorem 2 quantifies over every randomized input-only score, neither conformalization nor an evidential or circular head can distinguish an exact gauge-paired equal-law construction without side information. This does not imply failure under the observable shifts studied in those papers.

Table 12: Comparison with the closest work. Scope statements concern the cited work; they are not claims about the entire literature.

prior work	direct overlap	boundary relative to this paper
Wang et al. (2023)	Incorrect model invariance forces classification and regression error across orbits.	Our qualitative Theorem 1 is an instance derived from the operator. Their setup does not derive a gauge between sensor nuisance and target, construct deployment shifts with equal input laws, or study prediction set width and physical interventions.
Berman et al. (2026)	Bounds ECE/ENCE and aleatoric bleed for incorrectly equivariant models.	Their learning setup assumes no distribution shift at test time and studies metrics of calibration error. We study observationally identical shifted worlds, alarm power based only on input, metric risk, prediction set diameter, and side information that breaks the gauge.
Moussa (2026) (concurrent preprint)	Proves exact and approximate limits for coverage audits without labels under unrestricted conditional shift.	The general detectability boundary is prior work. We derive an admissible sensor transformation, derive point risk and set width consequences that depend on target separation, and test the construction on learned models and physical ADC.
Kasa et al. (2025) and Angelman et al. (2025)	Adapt conformal prediction from unlabeled target inputs using model uncertainty or artificial labels.	Such methods can help under structured empirical shifts. No unlabeled batch can distinguish the worlds with equal input laws in Theorem 2; we do not claim failure outside that construction.
Array self calibration (Wylie et al., 1994; Viberg et al., 2009)	Establishes joint lack of uniqueness in direction and phase and the need for constraints or references.	The phase/DOA ambiguity is not new. We connect it to learned invariance and UQ, state exact and approximate decision consequences, and audit learned and physical interventions.
Koch et al. (2022)	Shows that subgroup shifts can escape OOD detection.	Our construction with equal input laws is tied to a gauge of the forward operator and is paired with risk/set bounds and a localization mechanism.
Pereyra & Tachella (2024)	Uses equivariance in inverse problems to bootstrap uncertainty regions.	Their symmetry is exploited for UQ; ours identifies when nuisance invariance erases the target action and what side information restores it.

R LAG IN A NONSTATIONARY CALIBRATION PANEL

An exchangeable panel cannot justify calibration from stale data. This distinction also appears in selective conformal inference under interventions (Asiaee et al., 2026). We fixed a path with six time points, $\beta_t \sim \mathrm{Unif}(c_t - 0.25, c_t + 0.25)$, with $c_t = 0, 0.1, \ldots, 0.5$. Each time has a fixed test set of 20,000 examples and ten independent panels of 4,000 known directions. We compare the current panel, the corresponding panel from time zero, and a panel lagged by one step. The direct and orbit heads receive data that are identical at the byte level.

The first protocol was not a fair width comparison: it used asymmetric residual quantiles for the orbit head but only symmetric expansion for the direct head. At $c = 0.5$, direct width reached 0.447 while orbit width remained 0.144. We kept this result and defined a second protocol. Both heads now estimate the 5th and 95th percentiles of prediction error on half of the panel, then apply the same finite sample CQR correction on the other half.

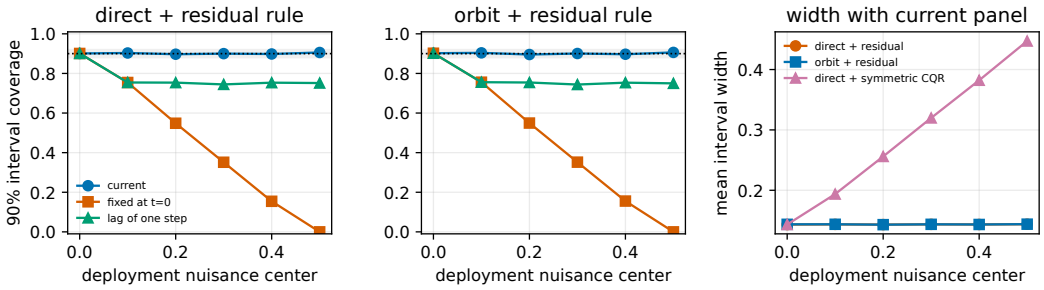


Figure 16: Prespecified fairness audit with a nonstationary panel. Current panels maintain coverage for both heads (left, center). A lag of one step with size 0.1 produces undercoverage, and the panel fixed at time zero decays to zero coverage. Right: the fair residual quantile rule removes the apparent width gap between direct and orbit heads; direct CQR using only a symmetric correction widens instead of recentering.

Table 13: Mean 90% coverage across ten independent panel repeats; 95% t intervals are shown in Figure 16. D is the direct residual head and O is the orbit head; both use the same residual quantile update.

	current		lag of one step		fixed at (t=0)	
c_t	D	O	D	O	D	O
0.0	.901	.902	.901	.902	.901	.902
0.1	.903	.904	.754	.756	.754	.756
0.2	.897	.895	.754	.754	.548	.549
0.3	.900	.901	.745	.744	.351	.352
0.4	.898	.897	.753	.753	.154	.155
0.5	.906	.906	.751	.750	.000	.000

All eleven follow up checks pass. For both methods, current panel coverage is 0.895 to 0.906 and width is 0.1430 to 0.1439. Coverage with one step of lag is 0.744 to 0.756 after time zero, while coverage from the fixed panel eventually reaches zero. On this path, validity depends on panel recency. The experiment does not establish distribution free coverage under nonstationarity or specify a physical update rate.

S Claim and Evidence Ledger

Table 14 separates mathematical conclusions, controlled evidence, and unresolved questions about external validity. It is intended to prevent a reproducible diagnostic result from being read as a broader deployment claim.

Table 14: Alignment between claims and evidence.

claim	evidence	boundary
Invariance to a shared action loses a target orbit	Theorem 1 and proof	Instance derived from the operator of prior theory on incorrect invariance (Wang et al., 2023); exact under Eq. 2, and not every nuisance shares a task action.
Alarms based only on input can be powerless	Theorem 2, proof, and exact paired learned audit	Existence result for gauge pushforwards; not universal failure on all shifts. The general boundary for audits without labels on finite batches is prior work (Moussa, 2026); our construction is derived from sensors.
Affine phase confounds wavevector	Theorem 3, ULA/planar audits, 843 automotive ADC ramp interventions, a component-factorization audit with 4,215 fault cases and 98 raw probes over 281 frames from 14 sequences, the dense-MIMO geometry audit with 65 arrays and 2,080 probes, and a RaDICaL audit with 150 cases	Exact coherent model for the far field; boundary for spherical waves quantified in Appendix M. Association guided by labels and injected faults are not a blind benchmark or an estimate of natural drift prevalence; the exact gauge requires the coordinated component mode tested in Appendices F and G. The dense audit is synthetic; one prespecified degeneracy spectrum gate fails.
Spherical curvature yields learned sharpness	Prespecified quantile audit with known range, affine closure features, five seeds, and CQR for each range	Not supported: three of four transition gates fail (Appendix N). The failure is retained; it does not exclude other architectures or establish behavior with unknown range or physical data in the near field.
Ambiguity forces risk or width	Proposition 1, executable tests, wrapped sensitivity	Exact pairs and approximate pairs based on total variation; the bound does not choose a nuisance prior.
Learned intervals can stay narrow	Three matched ensembles trained with five seeds, checkpoints, Wilson intervals, four nominal levels, and exact pairs with identical inputs	Controlled ULA shifts; not a universal statement about ensembles, conformal prediction, or natural drift.
A learned head over an orbit can widen correctly	Matched heads with 6,535 parameters, five seeds, disjoint tuning and calibration splits, failed first contract, conditional follow up with ten criteria, deployment panel audit with ten repeats, fixed-drift audit over 33 conditions, and fairness follow up for nonstationary lag	Requires either a supplied law or an exchangeable panel with known directions in the controlled ULA model; the equally informed direct head also calibrates. Marginal coverage is not uniform protection for a persistent drift; the certified set is wider. A panel lag of 0.1 already produces undercoverage on the frozen path. Does not infer the law from unlabeled inputs or validate physical UQ.
An anchor restores sharpness	Algebra, benchmark with 10,000 cases, stress test with 12 mismatch conditions, and prespecified physical transfer audit across three sequences	Analytic result for a shared ramp. Physical transfer audit is unresolved and warns against assuming stationarity; it does not identify hardware drift. The blockwise mismatch-bound protocol is proposed, not empirically validated.

T ETHICS STATEMENT

Authors have no conflicts of interest to disclose.

U USE OF LARGE LANGUAGE MODELS (LLMS)

Some of the initial notes and summaries for this work were written in French. We used large language model (LLM) tools to translate these materials into English, assist with summarization, and help with LaTeX formatting. These tools were used to reduce language related inconsistencies and improve the presentation of our own work, not to replace the authors' judgment or contributions. We checked and revised all text produced with LLM assistance for fidelity, clarity, and unintended overlap with existing sources. The authors take full responsibility for the final content.