TokaReconBench: A Benchmark for Transferable AI in Tokamak Equilibrium Reconstruction

Anonymous authors
Paper under double-blind review

Abstract

Developing reliable and transferable AI models for plasma equilibrium reconstruction is essential yet challenging for tokamak fusion reactors. Existing AI-based equilibrium reconstruction methods have shown promising performance as fast surrogates for traditional solvers. However, their predominantly in-distribution evaluation provides limited evidence of reliability under domain and device shifts. To address this limitation, we introduce **TokaReconBench**, a benchmark for transferable tokamak equilibrium reconstruction. It integrates heterogeneous experimental and synthetic datasets, including public MAST measurements, ITER-like simulations, and our newly contributed, configuration-aligned EXL-50U simulation and experimental data. We define unified task formulations and evaluation protocols covering three levels of generalization: simulation-to-simulation (*sim2sim*), simulation-to-real (*sim2real*) transfer, and *cross-device* generalization. We benchmark representative neural surrogate models, operator learning approaches, and physics-based baselines under diverse training loss settings. Metrics assess field accuracy, geometry, and physics consistency against reference source terms. Experiments reveal significant bottlenecks in cross-domain and cross-device transfer for existing learning-based reconstruction models. We hope **TokaReconBench** provides a unified platform for developing transferable AI methods towards future multi-device fusion applications. Our benchmark, datasets, and evaluation protocols are available upon acceptance.

1 Introduction

Global decarbonization efforts and the growing electricity demand of artificial intelligence (AI) strengthen the need for reliable low-carbon energy (International Energy Agency, 2025). Fusion offers a potential long-term solution. Tokamaks, a leading approach to magnetic-confinement fusion, use magnetic fields to confine extremely hot plasma inside a toroidal chamber (Pironti & Walker, 2005) (Figure 6). Developing these devices requires substantial investment: large projects such as ITER involve nearly EUR 30 billion commitments and three decades of development (European Court of Auditors, 2025; ITER Organization, n.d.). Yet experimental plasma data from a new device are unavailable before operation and cover only a limited range of conditions during early campaigns. Models intended to support its commissioning and operation must therefore draw on simulations and experience from existing devices. This combination of complex physics, costly experimental validation, and limited target-device data makes tokamak development a demanding application domain for AI in science and engineering. The challenge is to build physically reliable models that remain useful beyond the devices and conditions on which they were trained.

Equilibrium reconstruction is a central component of tokamak data analysis, providing the magnetic geometry needed for diagnostic interpretation, stability analysis, and plasma control (Pironti & Walker, 2005). It infers the poloidal magnetic flux distribution and associated geometric features, including the plasma boundary, magnetic axis, and X-points, from external diagnostic measurements. However, this is a severely ill-posed inverse problem: sparse, noisy measurements may be consistent with multiple equilibria (Braams, 1991). Under the axisymmetric equilibrium assumption, physically admissible solutions must also satisfy the nonlinear Grad–Shafranov (GS) equation and appropriate boundary conditions (Grad & Rubin, 1958).

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.

Table 1: Comparison of representative tokamak equilibrium-learning benchmarks.

Benchmark	Reconstruction MainInputs	Same-device Sim2Real	Cross-device evaluation	Target-device supervision
TokaMark (Rousseau et al., 2026)	Magnetics	×	×	×
FusionEQ (Michoski et al., 2026)	Profiles	×	✓	×
CrossGS (Yoo et al., 2026)	Profiles	×	✓	✓
TokaReconBench	Magnetics	✓	✓	✓

Existing reconstruction methods include iterative GS numerical solvers such as EFIT and its real-time variants (Lao et al., 1985; 1990; Ferron et al., 1998), as well as learning-based methods that map diagnostic measurements to equilibrium fields and incorporate physical constraints (Joung et al., 2020; Lao et al., 2022). These approaches offer different balances of speed, accuracy, and physical consistency (Zheng et al., 2026; Nakkina et al., 2026). Learning-based methods are particularly attractive for their low-latency inference and potential to learn equilibrium representations that can be reused across devices and operating conditions.

However, these learning-based studies remain focused on in-distribution evaluation on individual devices for experiments. Few studies systematically assess in-distribution and out-of-distribution generalization under a unified protocol, especially simulation-to-real, and cross-device transfer for adapting models to new tokamaks with limited experimental labels. We still lack a clear understanding of how changes in plasma conditions and diagnostic observation processes affect reconstruction accuracy and physical consistency. High simulation accuracy on one device therefore provides limited evidence of reliability in real experiments or unseen devices. As tokamaks progress toward fusion reactor scale, standardized cross-domain and cross-device evaluation with clearly specified target supervision is needed to assess whether these models can reliably support new devices.

To address this gap, we introduce **TokaReconBench**, a benchmark for reconstructing static tokamak magnetic equilibria governed by the GS equation. It integrates multi-device datasets, transfer tasks, complementary metrics, reconstruction baselines, and a modular codebase for reproducible evaluation and method development. Table 1 compares **TokaReconBench** with existing benchmarks and more related work are discussed in Appendix A. Our contributions are as follows.

1. **A benchmark for target-device adaptation.** We define five tasks covering in-distribution sim2sim reconstruction, zero-shot same-device *sim2real*, and *cross-device* adaptation with limited target labels. The evaluation jointly measures field accuracy, geometric reconstruction, and physics consistency against reference source terms.
2. **Simulation and experimental equilibrium data from EXL-50U.** We release EXL-50U simulated and experimental data under a common reconstruction interface, enabling same-device *sim2real* evaluation. Together with public MAST experimental data and ITER-like simulations, these data support transfer studies across data sources and device configurations.
3. **Empirical findings on transfer gaps and metric trade-offs.** Our comparison of architectures and training objectives reveals persistent transfer gaps and metric-dependent rankings. Lower flux error does not consistently imply better geometry or physics consistency, and auxiliary constraints provide uneven benefits. These findings highlight cross-domain generalization and physical reliability as priorities for adapting reconstruction models to new tokamaks.

2 Benchmark Framework

This section details **TokaReconBench**. We introduce the reconstruction problem, transfer scenarios, training loss, datasets and evaluation metrics in sequence. Figure 1 summarizes this framework.

2.1 The Equilibrium Reconstruction Problem

Many scientific problems can be formulated as reconstructing a continuous physical field from sparse and indirect observations. Tokamak magnetic *equilibrium reconstruction* is one such inverse problem with explicit physical constraints. Given magnetic diagnostic measurements m, coil currents I, and optional device geometry information g, the objective is to reconstruct the poloidal

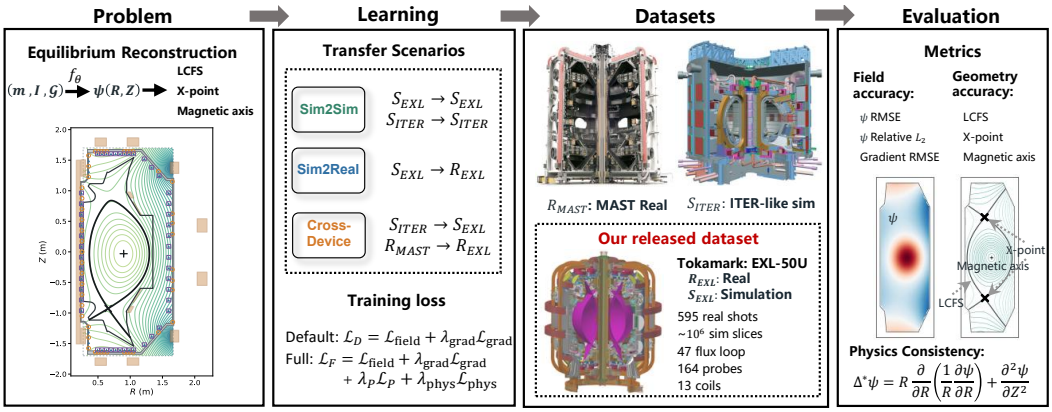

Figure 1: Overview of **TokaReconBench**.

flux field ψ on the device-native grid $\mathcal{G}$. In an axisymmetric tokamak, the poloidal flux ψ determines the magnetic topology and equilibrium geometry. We therefore treat ψ as the sole prediction target, while geometric structures such as the last closed flux surface (LCFS), magnetic axis, and X-points are derived from the predicted field during evaluation. This avoids inconsistencies that may arise from predicting these quantities independently and allows each device to retain its native sensor layout, geometry, and output grid. Formally, given a training set of N equilibrium samples, the task can be formulated as a field-reconstruction problem constrained by the Grad–Shafranov (GS) equation:

$$\begin{aligned} \widehat{\psi}_i &= f_\theta(m_i, I_i, g_i; \mathcal{G}_i), \\ \theta^\star &= \arg\min_\theta \frac{1}{N}\sum_{i=1}^{N} \left\| \widehat{\psi}_i - \psi_i \right\|_{\Omega_i}^2, \\ \text{s.t.} \quad \mathcal{R}_{\text{GS}}(\widehat{\psi}_i) &\equiv \Delta^*\widehat{\psi}_i + \mu_0 R^2 p_i'(\widehat{\psi}_i) + F_i(\widehat{\psi}_i)F_i'(\widehat{\psi}_i) = 0, \end{aligned} \tag{1}$$

where θ denotes the model parameters, $\widehat{\psi}_i$ and ψ_i are the predicted and reference poloidal flux fields, respectively, and Ω_i is the valid evaluation region on the native grid, R and Z denote the radial and vertical coordinates of the cylindrical coordinate system. Here, $\Delta^* = R\partial_R(R^{-1}\partial_R) + \partial_Z^2$, $p(\psi)$ is the plasma pressure, and $F(\psi)$ is the poloidal current function.

The reconstruction task is severely ill-posed and constrained by nonlinear governing equations. Magnetic diagnostics are typically sparse, noisy, and potentially incomplete, and a limited set of observations may be compatible with multiple candidate equilibria. Diagnostic calibration errors, unmodeled currents in conducting structures, and uncertainty in boundary information introduce additional measurement and model errors. For deployment beyond the training conditions, including on new devices, source-domain accuracy alone is insufficient; transfer across data sources, diagnostic systems, and device configurations must also be evaluated. Detailed background on tokamak equilibrium physics and the GS formulation is provided in Appendix B.1.

2.2 Transfer Scenarios

To evaluate transferability, we organize the benchmark along two dimensions. The **device dimension** indicates whether the device identity changes, whereas the **data-domain dimension** indicates whether the data are obtained from simulation or experiment. Based on these dimensions, **TokaReconBench** includes three complementary evaluation regimes: in-domain reconstruction, same-device *sim2real* transfer, and *cross-device* transfer. These regimes respectively evaluate basic reconstruction capability, transfer from simulation-based design to experimental operation, and the reuse of knowledge from existing devices on a target device.

We instantiate these regimes through five tasks, as summarized in Table 2. $\mathbb{T}_1$ and $\mathbb{T}_2$ perform in-domain training and evaluation on ITER-like and EXL-50U simulation data, respectively, providing reference performance without a deliberately introduced domain shift. $\mathbb{T}_3$ trains on EXL-50U simulation data and evaluates on experimental data from the same device. By keeping the nominal

Table 2: Benchmark tasks organized by device and data-domain variation.

Task	Training domain	Test domain	Evaluation regime	Fine-tuning data size
$\mathbb{T}_1$	ITER-like simulation	ITER-like simulation	In-domain simulation	0
$\mathbb{T}_2$	EXL-50U simulation	EXL-50U simulation	In-domain simulation	0
$\mathbb{T}_3$	EXL-50U simulation	EXL-50U experiment	Same-device Sim2Real	0
$\mathbb{T}_4$	ITER-like simulation	EXL-50U simulation	Cross-device Sim2Sim	5% of training
$\mathbb{T}_5$	MAST experiment	EXL-50U experiment	Cross-device Real2Real	5% of training

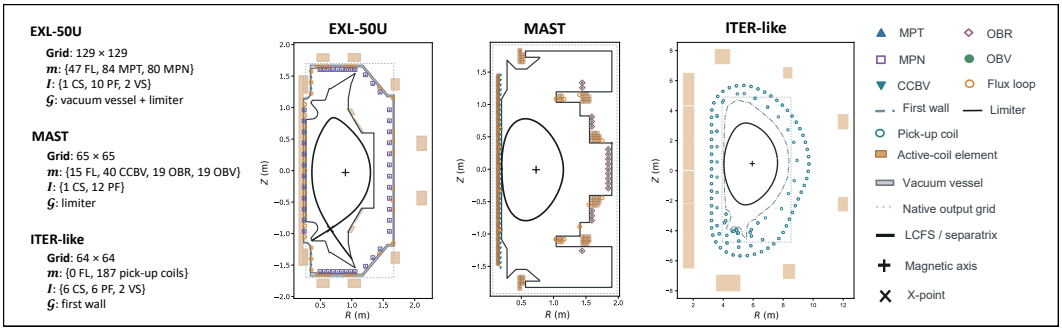

Figure 2: **Heterogeneity across tokamak domains.** Representative device-native poloidal cross-sections, preserving the original sensor layouts, active coil geometry, equilibrium boundaries, and available device geometry. EXL-50U includes vacuum vessel and limiter geometry, MAST includes only limiter geometry, while ITER-like provides the first wall. The side annotation summarizes the native flux-grid resolution and sensing/coil composition.

device geometry and diagnostic layout fixed, $\mathbb{T}_3$ assesses the *sim2real* gap transfer without a change of device. $\mathbb{T}_4$ transfers from ITER-like simulation data to EXL-50U simulation data and evaluates *cross-device* generalization between substantially different scales and configurations. $\mathbb{T}_5$ transfers from MAST experimental data to EXL-50U experimental data and evaluates transfer between two spherical tokamaks with different diagnostic systems and operating conditions. $\mathbb{T}_1$ and $\mathbb{T}_2$ use disjoint training, validation, and test splits. For $\mathbb{T}_3$ to $\mathbb{T}_5$, models are first trained using source-domain data only and evaluated on a fixed target-domain test set. $\mathbb{T}_3$ evaluates zero-shot transfer, while $\mathbb{T}_4$ and $\mathbb{T}_5$ use fine-tuning on limited target-domain data, motivated by our observation that zero-shot evaluation performs poorly in this challenging scenario.

2.3 Datasets

We release simulation and experimental data from EXL-50U, the largest spherical tokamak in China (Shi et al., 2025a;b;c). Both share the device geometry, diagnostic definitions, coil configuration, output grid, and reconstruction procedure: simulated measurements are generated by the FGE (Heiß et al., 2026) forward model and, like experimental measurements, are processed by PT-EFIT (Zheng et al., 2026) to obtain the reference flux field. The two releases are aligned in configuration and processing but are not paired sample by sample, which avoids extra discrepancies from inconsistent sensor layouts, target definitions, or label generation. Table 3 summarizes the aligned configuration and current scale, namely 595 carefully selected experimental shots and a simulated ensemble from ten validated configurations with 300 runs each. Provenance, generation, channel configuration, cleaning rules, and statistics are given in Appendix C.

To establish a consistent interface without removing device differences, we unify the three datasets at the level of physical semantics. As shown in Figure 2, the domains differ greatly in diagnostic layout, coil configuration, operating range, and grid. Magnetic diagnostics are standardized by measurement type and unit, coil currents are converted to consistent polarity and ampere-turn conventions, coordinates use (R, ϕ, Z), the poloidal flux gets an explicit unit and gauge, and the LCFS, magnetic axis, and X-points are treated as quantities derived from the flux. Each sample is a common container of measurements m, coil currents I, geometry g, the output grid $\mathcal{G}$, the reference flux ψ, and validity masks. Device-specific adapters expose these records without padding or in-

Table 3: Current EXL-50U experimental and simulated releases. The pair shares the same diagnostic layout, coil configuration, and PTEFIT label source, while release statistics differ.

	#shots	I_p range (kA)	Data size	#Sample	Measure	Coil	Label source
Experiment	595	$[100, 678]$	33.05GB	250658	47 loops $84 + 80$ probes	1 CS 10 PF 2 VS	PTEFIT
Simulation	2877	$[100, 688]$	64.73GB	608283			

dex alignment. This preserves genuine heterogeneity while removing avoidable discrepancies from units, indexing, and target definitions.

EXL-50U is the central device connecting transfer scenarios. Its data play two roles. First, they form the configuration-aligned same-device pair for *sim2real* evaluation ($\mathbb{T}_3$). Second, they support two *cross-device* evaluations: $\mathbb{T}_4$ transfers between ITER-like(Bonotto et al., 2021; Abate & Bettini, 2019) and EXL-50U simulations, which differ greatly in scale, configuration, diagnostics, and grid, while $\mathbb{T}_5$ performs real-device transfer between EXL-50U and MAST experimental data(Jackson et al., 2025; 2024), which differ in sensor layouts, coils, regimes, and reconstruction pipelines. Beyond these tasks, aligned EXL-50U data support parameter extrapolation from low to high current, noise robustness under controlled perturbations, and diagnostic design through probe masking.

2.4 Training loss

We adopt four loss terms to train neural models. The field loss $\mathcal{L}_{\mathrm{field}}$ measures pointwise flux error and the gradient loss $\mathcal{L}_{\mathrm{grad}}$ matches first spatial derivatives. The probe-consistency loss $\mathcal{L}_{\mathrm{probe}}$ uses the instrument model and the measurement vector. The Grad–Shafranov residual defines the physical loss $\mathcal{L}_{\mathrm{phys}}$, which uses the toroidal current density. Field plus gradient is the default: both are computed from the flux label alone, and the gradient is the quantity the geometry and topology metrics depend on. The probe measurement is available at deployment, but the term also needs the instrument model, so it is not folded into that label-only default. The physics term is optional for a different reason: j_ϕ is a known source when the profile formulas are available, and is otherwise taken from the supplied equilibrium record.

$$\begin{aligned} \mathcal{L}_D &= \mathcal{L}_{\mathrm{field}} + \lambda_{\mathrm{grad}}\mathcal{L}_{\mathrm{grad}}, && (\underline{D}\textit{efault}), \\ \mathcal{L}_F &= \mathcal{L}_D + \lambda_{probe}\mathcal{L}_{\mathrm{probe}} + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}}, && (\underline{F}\textit{ull}). \end{aligned} \tag{2}$$

The **field** term is available across datasets but alone does not capture all benchmark metrics. The **grad** term constrains spatial derivatives relevant to geometry and topology, requiring grid coordinates but no additional labels; together, these terms form the default objective. The **phys** term constrains the second-order GS operator through the reference toroidal current density j_ϕ. When profile formulas are available, j_ϕ can be derived from them; otherwise, it is taken from the supplied equilibrium record. For fixed source terms, the residual is unchanged by adding solutions of the homogeneous GS operator. Without sufficient boundary or diagnostic constraints, reducing it therefore does not ensure accurate separatrix or X-point locations. Because it targets the same GS residual as E_{phys}, the +phys objective tests whether improving this criterion also benefits reconstruction. The **probe** term compares the predicted exterior magnetic field with measurements outside the plasma, encouraging diagnostic consistency without uniquely determining the internal equilibrium. These measurements are available at deployment, unlike reference j_ϕ when neither source profiles nor equilibrium records are available. All combinations use a pseudo-Huber penalty, with $\lambda_{\mathrm{grad}} = \lambda_{\mathrm{probe}} = \lambda_{\mathrm{phys}} = 0.05$. Appendix F.7 provides definitions, supports, and normalization conventions.

2.5 Evaluation Metrics

Equilibrium reconstruction requires more than pointwise agreement of the predicted flux field. We therefore evaluate complementary properties of the solution at the field, geometry, topology, and physics levels. Detailed formulation of these metrics are provided in Appendix D.

Field reconstruction accuracy. We report the flux RMSE (Rousseau et al., 2026), E_ψ^{RMSE}, and the relative field error (Kovachki et al., 2023), E_ψ^{rel}. The former measures absolute reconstruction error of the flux field, while the latter normalizes the error by the reference field scale and is therefore more

Table 4: **Main results across eight models.** Within each scenario and metric, darker colors indicate lower errors; boldface marks minima. An asterisk denotes a mean conditional on successful geometric extraction[2]. Flux RMSE is in mWb/rad; relative errors are dimensionless. These architecture-study runs are separate from the controlled loss study in Tables 5 and 12.

Transfer Scenarios	Model	Params (M)	Field Reconstruction ↓			Geometry ↓			GS Diagnostic ↓
			E_{ψ}^{RMSE}(mWb/rad)	E_{ψ}^{rel}	$E_{\nabla\psi}^{\mathrm{rel}}$	E_{axis}(cm)	E_{LCFS}^{95}(cm)	$E_{\mathrm{X-pt}}$(cm)	E_{phys}
$\mathbb{T}_1$ $S_{\mathrm{ITER}} \to S_{\mathrm{ITER}}$ *sim2sim*	U-Net	1.95	39.38	0.009894	0.01913	1.009	**0.890**	4.303	0.1139
	Transformer	4.47	39.03	0.009546	**0.01685**	**0.7768**	1.127	4.646	**0.06105**
	PlaNet	1.56	56.24	0.01322	0.02446	1.434	1.088	4.856	0.1624
	DeepONet	0.596	77.38	0.01842	0.03037	2.028	1.748	6.344	0.09505
	FNO	2.06	**33.70**	**0.007956**	0.01706	4.893	1.095	**4.042**	0.1188
	SetONet	0.103	106.5	0.02764	0.04117	4.874	2.478	7.742	0.3571
	GINO	0.184	61.44	0.01527	0.03035	13.19	1.831	8.130	0.4518
	GNOT	0.629	52.81	0.01282	0.02153	1.221	1.146	5.959	0.07000
$\mathbb{T}_2$ $S_{\mathrm{EXL}} \to S_{\mathrm{EXL}}$ *sim2sim*	U-Net	1.95	0.2766	**0.008816**	0.01114	0.3647	**0.288**	2.043	0.2455
	Transformer	4.47	**0.2699**	0.009344	0.01283	0.3506	0.318	0.8846	0.1817
	PlaNet	1.56	0.5247	0.01728	0.02144	1.328	0.472	2.740	0.8043
	DeepONet	0.596	0.5615	0.01998	0.03609	1.151	0.586	5.775	0.3034
	FNO	2.06	0.2795	0.009273	**0.01096**	0.5156	0.294	**0.7504**	0.1943
	SetONet	0.103	0.5451	0.01963	0.03259	0.9990	0.604	2.836	0.9428
	GINO	0.184	0.3157	0.01149	0.02772	0.8990	0.385	1.514	1.242
	GNOT	0.629	0.2827	0.009872	0.01262	**0.2208**	0.327	0.8469	**0.1677**
$\mathbb{T}_3$ $S_{\mathrm{EXL}} \to R_{\mathrm{EXL}}$ *sim2real*	U-Net	1.95	**2.199**	**0.1029**	**0.1182**	**2.224**	2.336	1.662	0.5363
	Transformer	4.47	5.270	0.2004	0.2071	5.730	3.360	3.641	0.4134
	PlaNet	1.56	3.041	0.1392	0.1489	4.745	3.154	2.548	0.8836
	DeepONet	0.596	10.21	0.3227	0.3389	8.514	5.107	3.333	0.7476
	FNO	2.06	3.078	0.1325	0.1703	4.573	3.177	**1.539**	1.057
	SetONet	0.103	4.081	0.1606	0.1783	4.404	3.574	3.907	1.108
	GINO	0.184	3.375	0.1312	0.1746	13.95	1.973	5.138	6.316
	GNOT	0.629	4.340	0.1785	0.1900	5.003	**1.840**	2.191	**0.4030**
$\mathbb{T}_4$ $S_{\mathrm{ITER}} \to S_{\mathrm{EXL}}$ *cross-device* 5% target fine-tuning	U-Net	1.95	1.549	0.07547	**0.08150**	2.17	5.26	2.57	1.836
	Transformer	4.47	**1.208**	**0.05932**	0.1004	**1.75**	**4.62**	2.50	0.3671
	PlaNet	1.56	2.024	0.09828	0.1098	4.33	5.57	**2.44**	1.245
	DeepONet	0.596	2.924	0.1405	0.2026	2.66	6.34	3.26	0.6250
	FNO	2.06	3.821	0.1843	0.1509	3.20	8.37*	4.20	1.072
	SetONet	0.103	3.832	0.1844	0.1967	4.06	8.74*	3.95*	1.901
	GINO	0.184	4.284	0.2086	0.3545	6.31	28.60*	5.57*	15.26
	GNOT	0.629	2.286	0.1127	0.1387	2.73	6.00	2.68	**0.3094**
$\mathbb{T}_5$ $R_{\mathrm{MAST}} \to R_{\mathrm{EXL}}$ *cross-device* (target adaptation)	U-Net	1.95	3.350	0.2278	0.317	12.62	19.05	4.13	1.058
	Transformer	4.47	**1.750**	**0.1204**	**0.184**	**7.41**	**10.98**	**0.90**	**0.476**
	PlaNet	1.56	5.510	0.3517	0.442	14.45	25.19	7.26	1.553
	DeepONet	0.596	3.890	0.2592	0.419	8.97	15.12	6.75	1.173
	FNO	2.06	3.480	0.2336	0.312	13.93	21.89	3.81	0.954
	SetONet	0.103	4.800	0.3022	0.429	9.42	17.13	7.85	1.356
	GINO	0.184	4.370	0.3082	0.522	83.49	55.96	19.34	18.53
	GNOT	0.629	3.140	0.2121	0.301	10.08	15.36	3.86	0.560

suitable for comparison across equilibria with different amplitudes. To further resolve local spatial structure, we evaluate the gradient RMSE, $E_{\nabla\psi}^{\mathrm{RMSE}}$, and its normalized counterpart, $E_{\nabla\psi}^{\mathrm{rel}}$. These gradient metrics capture local distortions that can remain weak in field-value errors but directly affect magnetic geometry.

Geometric fidelity. We evaluate the reconstructed plasma boundary using LCFS HD95 (Nakkina et al., 2026; Michoski et al., 2026), E_{LCFS}^{95}, and measure the localization errors of the magnetic axis and X-points (Rousseau et al., 2026; Nakkina et al., 2026) by E_{axis} and $E_{\mathrm{X-pt}}$. These quantities probe distinct geometric structures derived from flux field: the plasma boundary, equilibrium center, and separatrix-defining saddle points. Explicit evaluation is necessary because similar field errors can still lead to substantially different equilibrium geometries, which are directly relevant to downstream plasms analysis and control.

Physics consistency. Physical consistency is measured by the Grad–Shafranov plasma residual (Joung et al., 2023; Bonotto et al., 2024; Zhou & Zhu, 2025), E_{phys}, which quantifies the extent to which the predicted field satisfies the equilibrium equation. It complements supervised field and geometry metrics by testing the governing physical constraint directly.

Table 5: **Training-loss comparison on selected reconstruction metrics.** Within each scenario and metric, darker colors indicate lower errors; boldface marks minima. An asterisk (*) denotes means over successful geometric extractions. More results are in Appendix G.

Transfer Scenario	Loss	Field E_{ψ}^{RMSE} (mWb/rad)			LCFS E_{LCFS}^{95} (cm)			X-point $E_{\mathrm{X-pt}}$ (cm)		
		U-Net	FNO	GNOT	U-Net	FNO	GNOT	U-Net	FNO	GNOT
$\mathbb{T}_1$	$\mathcal{L}_D$	44.23	42.00	58.29	**2.878***	3.085*	4.279*	4.377*	4.737*	10.05*
	$\mathcal{L}_D + \lambda_{\mathrm{probe}}\mathcal{L}_{\mathrm{probe}}$	46.40	42.62	55.11	3.184*	3.174*	3.548*	4.273*	5.568*	5.264*
	$\mathcal{L}_D + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}}$	49.06	41.30	65.89	3.232*	2.966*	3.963*	4.114	5.635*	10.17*
	$\mathcal{L}_F$	51.22	**41.04**	56.88	3.190*	3.090*	3.436*	**3.619***	5.102*	5.879*
$\mathbb{T}_2$	$\mathcal{L}_D$	0.2212	0.2225	0.2660	0.9589*	0.9165*	1.064*	0.2380	**0.2299**	0.3309
	$\mathcal{L}_D + \lambda_{\mathrm{probe}}\mathcal{L}_{\mathrm{probe}}$	0.2244	**0.2210**	0.2766	1.004*	**0.9161***	1.123*	0.2339	0.2627	0.3140
	$\mathcal{L}_D + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}}$	0.2230	0.2317	0.2877	0.9694*	0.9794*	1.165*	0.2407	0.2359	0.3725
	$\mathcal{L}_F$	0.2245	0.2365	0.2509	0.9652*	0.9972*	0.9689*	0.2308	0.2458	0.2687
$\mathbb{T}_3$	$\mathcal{L}_D$	1.897	2.425	3.204	6.616*	7.376*	8.621*	2.410	2.352*	**1.754**
	$\mathcal{L}_D + \lambda_{\mathrm{probe}}\mathcal{L}_{\mathrm{probe}}$	1.908	2.404	3.022	6.551*	7.365*	8.894*	2.564	2.357	1.938
	$\mathcal{L}_D + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}}$	**1.857**	2.366	3.119	**6.498***	7.248*	8.770*	2.261	2.210	1.799
	$\mathcal{L}_F$	1.879	2.334	3.074	6.534*	7.393*	8.503*	2.311	2.132*	1.854
$\mathbb{T}_4$	$\mathcal{L}_D$	1.914	3.907	**1.553**	5.979*	8.171*	5.233*	2.750	4.771*	1.804
	$\mathcal{L}_D + \lambda_{\mathrm{probe}}\mathcal{L}_{\mathrm{probe}}$	1.908	4.033	1.602	5.656*	8.538*	5.142	3.180*	5.067*	1.963*
	$\mathcal{L}_D + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}}$	1.923	4.250	1.658	5.813*	7.971*	**5.110**	2.657	4.800*	**1.739**
	$\mathcal{L}_F$	1.695	3.996	1.591	5.756*	7.827*	5.146*	2.413	4.383*	1.880
$\mathbb{T}_5$	$\mathcal{L}_D$	**0.8519**	1.250	1.249	**3.764***	5.342*	4.853*	1.029	1.596*	1.693
	$\mathcal{L}_D + \lambda_{\mathrm{probe}}\mathcal{L}_{\mathrm{probe}}$	0.8722	1.299	1.237	3.911*	5.338*	4.193*	**0.9770**	1.760*	1.390
	$\mathcal{L}_D + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}}$	0.8976	1.252	1.169	3.911*	5.332*	4.230*	1.003	1.800*	1.480*
	$\mathcal{L}_F$	0.9352	1.246	1.282	4.035*	5.222*	4.640*	1.078*	1.769*	1.479

3 Experiments

3.1 Experimental Setup

To comprehensively evaluate the performance of existing methods, we design comparative experiments from two dimensions: model architecture and training loss. We consider two architecture categories. The first category contains three neural surrogate models: the general U-Net (Ronneberger et al., 2015) and Transformer (Qin et al., 2023), as well as PlaNet (Bonotto et al., 2024), a dedicated model recently developed for equilibrium reconstruction. The second category includes five neural operator models, namely FNO (Li et al., 2020), DeepONet (Lu et al., 2021), SetONet (Tretiakov et al., 2025), GINO (Li et al., 2023), and GNOT (Hao et al., 2023), which are widely adopted for physical system prediction. Except for PlaNet, the other seven architectures have not been formally adopted or systematically compared for equilibrium reconstruction in existing studies. We tailor the structure and parameter size of each model to adapt to our data format. Additionally, due to the unique structural constraints of different models, we design separate geometric embedding strategies to integrate probe position and angle information. In terms of training objectives, we employ four loss combinations as elaborated in Section 2.4. Detailed model architectures, hyperparameters and training procedures are documented in Appendix E and Appendix F.

3.2 Analysis of Main Results

Table 4 summarizes eight architectures trained with the default loss. Three key observations emerge.

In-domain accuracy does not establish transferability. Table 4 shows low relative flux errors on the in-domain simulated tasks, but larger errors under transfer. For U-Net, the error is 0.8816% on $\mathbb{T}_2$ and 10.29% on $\mathbb{T}_3$. Under the 5% target-data adaptation protocol reported in this table, the lowest errors are 5.932% on $\mathbb{T}_4$ and 12.04% on $\mathbb{T}_5$, both obtained by Transformer. FNO leads on $\mathbb{T}_1$, whereas U-Net leads on $\mathbb{T}_2$ and $\mathbb{T}_3$. These comparisons concern different test distributions; they neither isolate the effect of domain shift nor establish a universal architecture ranking. They show why new-device adaptation requires evaluation beyond simulated in-domain reconstruction.

[2]Asterisked geometry metrics are computed only on samples with successful structure extraction; failed extractions, typically due to topological inconsistencies in the predicted flux field, are reported as unavailable.

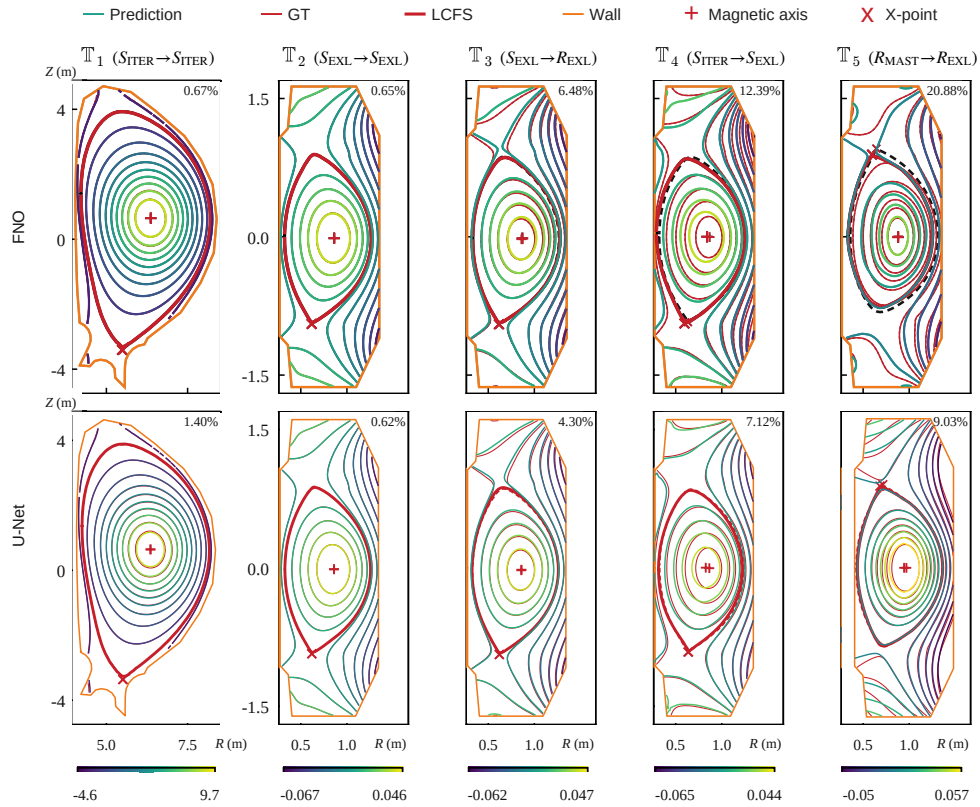

Figure 3: **Representative profiles across tasks and architectures.** E_{ψ}^{rel} is shown in each profile. More results are shown in Appendix G.

Field, geometry, and physics consistency are complementary. In $\mathbb{T}_3$, U-Net has lower relative flux error than GNOT (10.29% versus 17.85%), but higher LCFS error (2.336 versus 1.840 cm; Table 4). Boundary and critical-point errors concern geometry used in equilibrium analysis and plasma control (Pironti & Walker, 2005), which field error alone does not capture. Figure 3 illustrates deviations near outer flux surfaces and the separatrix. Geometric errors must be interpreted alongside extraction coverage and topology recovery (Appendix D.2–D.3). The physics consistency instead tests agreement with a supplied reference source (Appendix D.4 and F.7); it does not independently validate the experimental equilibrium.

Auxiliary constraints have model-dependent benefits. In $\mathbb{T}_1$, probe supervision reduces GNOT's flux RMSE from 58.29 to 55.11 mWb/rad but increases U-Net's from 44.23 to 46.40 mWb/rad (Table 5). For U-Net in the same loss study, Appendix G, Table 12 reports that adding the physical term lowers the physical score from 0.1345 to 0.08392 while increasing flux RMSE from 44.23 to 49.06 mWb/rad. Thus, improving a constraint-related score need not improve reference-field accuracy. The Full loss does not consistently outperform simpler objectives. Under the tested weights, these tradeoffs do not imply an inherent physics–data conflict.

3.3 Diagnostic sensitivity analysis and target-device adaptation

***sim2real*:** Figure 4 tests measurement sensitivity in $\mathbb{T}_3$. Noise/gain augmentation and retrained diagnostic ablations have model-dependent effects. With the same number of retained probes, FNO's errors differ between dispersed and localized masks, indicating that spatial coverage matters as well as sensor count. Inference-only masking and retrained ablations test different conditions and do not identify the dominant cause of the *sim2real* gap.

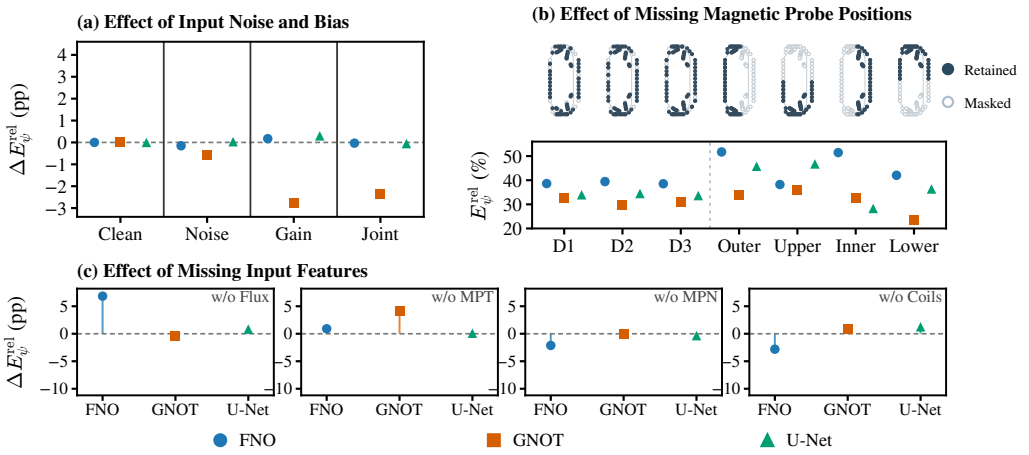

Figure 4: **Measurement robustness and diagnostic dependence under *sim2real* transfer.** (a) Field-error changes after noise and gain bias augmentation to input; (b) absolute field error when 82 of 164 probes are masked at inference; (c) four changes after retraining without flux loops, MPT, MPN, or coil currents. Panels (a) and (c) show percentage-point changes from their respective controls; (b) shows absolute error.

***cross-device*:** Figure 5 varies labeled EXL experimental data for fine-tuning MAST-pretrained models from zero to 20%. At 0.001%, U-Net's field error falls from roughly 350% to 50%, and Transformer's from roughly 110% to 35%. Gains then diminish and fluctuate: at 20%, Transformer is near 12% error, while several other models remain around 20–30%. These are approximate values read from the supplied plot. Thus, a small amount of target data helps substantially, but more data does not consistently close the *cross-device* gap. The remaining error motivates more robust diagnostic handling and structured adaptation; this curve alone doesn't identify the mechanism.

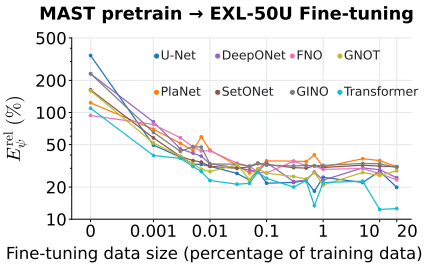

Figure 5: Effect of fine-tuning data size on the *cross-device* task.

4 Conclusion

In this work, we introduce TokaReconBench, a benchmark for transferable tokamak equilibrium reconstruction across simulation, experiment, and device domains. By unifying multi-device datasets, transfer protocols, and multi-level evaluation, TokaReconBench enables systematic study of same-device Sim2Real transfer and cross-device adaptation with limited target labels. Our results show that strong in-domain accuracy does not ensure comparable performance under transfer, and lower flux error does not necessarily imply better geometry or reference-source GS consistency. The benefits of auxiliary constraints depend on the architecture, task, and evaluation metric. These findings highlight the need to assess field accuracy, geometry, and physics consistency jointly. TokaReconBench provides a common testbed for developing reconstruction methods that use simulations and existing-device data to support adaptation to new tokamaks.

AI use statement

Generative AI tools were used in this work for two purposes. First, they were used to assist with language editing and polishing, including improving grammar, clarity, concision, and presentation of author-written text. Second, they were used for literature retrieval and discovery, such as identifying potentially relevant prior work and related references. All AI-assisted text was reviewed and revised

by the authors, and all retrieved references and associated claims were manually checked against the original sources. The authors independently verified the final text, references, and scientific claims and take full responsibility for the content of the paper.

ACKNOWLEDGMENTS

REFERENCES

Domenico Abate and Paolo Bettini. An inverse equilibrium tool to define axisymmetric plasma equilibria. *Plasma Physics and Controlled Fusion*, 61(10):105016, 2019.

R Aymar, P Barabaschi, and Y Shimomura. The iter design. *Plasma physics and controlled fusion*, 44(5):519–565, 2002.

Jacques Blum, Cedric Boulbe, and Blaise Faugeras. Reconstruction of the equilibrium of the plasma in a tokamak and identification of the current density profile in real time. *Journal of computational physics*, 231(3):960–980, 2012.

M Bonotto, D Abate, P Bettini, and F Villone. A coupled fem-bem approach for the solution of the free-boundary axi-symmetric plasma equilibrium problem. *Communications in Computational Physics*, 31(1):27–59, 2021.

Matteo Bonotto, Domenico Abate, and Leonardo Pigatto. Reconstruction of plasma equilibrium and separatrix using convolutional physics-informed neural operator. *Fusion Engineering and Design*, 200:114193, 2024.

Bastiaan J Braams. The interpretation of tokamak magnetic diagnostics. *Plasma Physics and Controlled Fusion*, 33(7):715–748, 1991.

European Court of Auditors. Annual report on EU joint undertakings for the 2024 financial year. Technical report, European Court of Auditors, 2025. URL https://www.eca.europa.eu/ECAPublications/SAR-JUS-2024/SAR-JUS-2024_EN.pdf.

John R Ferron, ML Walker, LL Lao, HE St John, DA Humphreys, and JA Leuer. Real time equilibrium reconstruction for tokamak discharge control. *Nuclear fusion*, 38(7):1055–1066, 1998.

Harold Grad and Hanan Rubin. Hydromagnetic equilibria and force-free fields. In *Proceedings of the second United Nations international conference on the peaceful uses of atomic energy. V. 31. Theoretical and experimental aspects of controlled nuclear fusion*, number INIS-XU–021, pp. 190–197, 1958.

Zhongkai Hao, Zhengyi Wang, Hang Su, Chengyang Ying, Yinpeng Dong, Songming Liu, Ze Cheng, Jian Song, and Jun Zhu. Gnot: A general neural operator transformer for operator learning. In *International conference on machine learning*, pp. 12556–12569. PMLR, 2023.

Cosmas Heiß, Antoine Merle, Francesco Carpanese, Federico Felici, Craig Donner, Stefano Marchioni, Alessandro Mari, and Olivier Sauter. Fge: a fast free-boundary grad–shafranov evolutive solver. *Plasma Physics and Controlled Fusion*, 68(4):045031, 2026.

Y Huang, ZP Luo, BJ Xiao, LL Lao, Adriano Mele, A Pironti, M Mattei, G Ambrosino, QP Yuan, YH Wang, et al. Gpu-optimized fast plasma equilibrium reconstruction in fine grids for real-time control and data analysis. *Nuclear Fusion*, 60(7):076023, 2020.

International Energy Agency. Energy and AI. Technical report, International Energy Agency, Paris, April 2025. URL https://www.iea.org/reports/energy-and-ai. Published 10 April 2025. Accessed 26 September 2026.

ITER Organization. FAQs. Official website, n.d. URL https://www.iter.org/faqs. Accessed 26 September 2026.

Samuel Jackson, Saiful Khan, Nathan Cummings, James Hodson, Shaun De Witt, Stanislas Pamela, Rob Akers, Jeyan Thiyagalingam, MAST Team, et al. Fair-mast: A fusion device data management system. *SoftwareX*, 27:101869, 2024.

Samuel Jackson, Saiful Khan, Nathan Cummings, James Hodson, Shaun De Witt, Stanislas Pamela, Rob Akers, Jeyan Thiyagalingam, MAST Team, et al. An open data service for supporting research in machine learning on tokamak data. *IEEE Transactions on Plasma Science*, 53(9):2440–2449, 2025.

Semin Joung, Jaewook Kim, Sehyun Kwak, JG Bak, SG Lee, HS Han, Hyun-Seok Kim, Geunho Lee, Daeho Kwon, and Y-C Ghim. Deep neural network grad–shafranov solver constrained with measured magnetic signals. *Nuclear Fusion*, 60(1):016034, 2020.

Semin Joung, Y-C Ghim, Jaewook Kim, Sehyun Kwak, Daeho Kwon, C Sung, D Kim, Hyun-Seok Kim, JG Bak, and SW Yoon. Gs-deepnet: mastering tokamak plasma equilibria with deep neural networks and the grad–shafranov equation. *Scientific Reports*, 13(1):15799, 2023.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

Lang L Lao, H St. John, RD Stambaugh, AG Kellman, and W Pfeiffer. Reconstruction of current profile parameters and plasma shapes in tokamaks. *Nuclear fusion*, 25(11):1611–1622, 1985.

Lang L Lao, Scott Kruger, Cihan Akçay, Prasanna Balaprakash, Torrin A Bechtel, Eric Howell, Jaehoon Koo, Jarrod Leddy, Matthew Leinhauser, YQ Liu, et al. Application of machine learning and artificial intelligence to extend efit equilibrium reconstruction. *Plasma Physics and Controlled Fusion*, 64(7):074001, 2022.

LL Lao, JR Ferron, RJ Groebner, W Howl, H St. John, EJ Strait, and TS Taylor. Equilibrium analysis of current profiles in tokamaks. *Nuclear Fusion*, 30(6):1035–1049, 1990.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

Zongyi Li, Nikola Kovachki, Chris Choy, Boyi Li, Jean Kossaifi, Shourya Otta, Mohammad Amin Nabian, Maximilian Stadler, Christian Hundt, Kamyar Azizzadenesheli, et al. Geometry-informed neural operator for large-scale 3d pdes. *Advances in Neural Information Processing Systems*, 36:35836–35854, 2023.

Yunfeng Liang, Huasheng Xie, Yuejiang Shi, Xiang Gu, Xinchen Jiang, Lili Dong, Xueyun Wang, Danke Yang, Wenjun Liu, Tiantian Sun, et al. Overview of the physics design of the ehl-2 spherical torus. *Plasma Science and Technology*, 27(2):024001, 2025.

Jingjing Lu, Youjun Hu, Nong Xiang, and Youwen Sun. Fast equilibrium reconstruction by deep learning on east tokamak. *AIP Advances*, 13(7), 2023.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature machine intelligence*, 3(3):218–229, 2021.

Sandeep Madireddy, Cihan Akçay, Scott E Kruger, T Bechtel Amara, Xuan Sun, J McClenaghan, Jaehoon Koo, Anirban Samaddar, Y Liu, Prasanna Balaprakash, et al. Efit-prime: Probabilistic and physics-constrained reduced-order neural network model for equilibrium reconstruction in diii-d. *Physics of Plasmas*, 31(9), 2024.

J McClenaghan, C Akçay, TB Amara, X Sun, S Madireddy, LL Lao, SE Kruger, and OM Meneghini. Augmenting machine learning of grad–shafranov equilibrium reconstruction with green's functions. *Physics of Plasmas*, 31(8), 2024.

Craig Michoski, Matthew Waller, Brian Sammuli, William Boyes, Mitchell Clark, Sterling Smith, Tapan Ganatma Nakkina, David R. Hatch, and Raffi Nazikian. The fusion equilibrium challenge: Predicting plasma shape from control inputs. NeurIPS 2026 Competition Track, 2026. Fusion Equilibrium Challenge.

J-M Moret, BP Duval, HB Le, S Coda, F Felici, and H Reimerdes. Tokamak equilibrium reconstruction code liuqe and its real time implementation. *Fusion Engineering and Design*, 91:1–15, 2015.

AW Morris. Mast: results and upgrade activities. *IEEE Transactions on plasma science*, 40(3): 682–691, 2012.

Tapan Ganatma Nakkina, Matthew Waller, Craig Michoski, Brian Sammuli, David R Hatch, William Boyes, Mitchell Clark, Raffi Nazikian, and Sterling Smith. The fusion equilibrium challenge: Inferring magnetic geometry without magnetic diagnostics. *arXiv preprint arXiv:2609.01750*, 2026.

Alfredo Pironti and Michael Walker. Fusion, tokamaks, and plasma control: an introduction and tutorial. *IEEE Control Systems Magazine*, 25(5):30–43, 2005.

Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, Slobodan Ilic, Dewen Hu, and Kai Xu. Geotransformer: Fast and robust point cloud registration with geometric transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9806–9821, 2023.

Marku Rampp, Rolan Preuss, Raine Fischer, and ASDEX Upgrad Team. Gpec: A real-time–capable tokamak equilibrium code. *Fusion Science and Technology*, 70(1):1–13, 2016.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Cécile Rousseau, Samuel Jackson, Rodrigo H Ordonez-Hurtado, Nicola C Amorisco, Tobia Boschi, George K Holt, Andrea Loreti, Eszter Székely, Alexander Whittle, Adriano Agnello, et al. Tokamark: A comprehensive benchmark for mast tokamak plasma models. *arXiv preprint arXiv:2602.10132*, 2026.

Yuejiang Shi, Xianming Song, Dong Guo, Xinchen Jiang, Xiang Gu, Lili Dong, Xueyun Wang, Tiantian Sun, Muzhi Tan, Zhengyuan Chen, et al. Strategy and experimental progress of the exl-50u spherical torus in support of the ehl-2 project. *Plasma Science and Technology*, 27(2): 024003, 2025a.

Yuejiang Shi, Xianming Song, YK Martin PENG, and the EXL-50U Team. Achievement of 1 ma discharges in hydrogen–boron plasmas in the exl-50u. *Plasma Science and Technology*, 27(9): 092002, 2025b.

Yuejiang Shi, Yumin Wang, Bing Liu, Xianming Song, Shaodong Song, Xinchen Jiang, Dong Guo, Di Luo, Xiang Gu, Tiantian Sun, et al. Overview of exl-50 research progress. *Nuclear Fusion*, 65 (9):092004, 2025c.

M Shimada, DJ Campbell, V Mukhovatov, M Fujiwara, N Kirneva, K Lackner, MPVD Nagami, VD Pustovitov, N Uckan, J Wesley, et al. Overview and summary. *Nuclear Fusion*, 47(6):S1, 2007.

A Sykes, RJ Akers, LC Appel, ER Arends, PG Carolan, NJ Conway, GF Counsell, G Cunningham, A Dnestrovskij, Yu N Dnestrovskij, et al. First results from mast. *Nuclear Fusion*, 41(10):1423–1433, 2001.

Stepan Tretiakov, Xingjian Li, and Krishna Kumar. Setonet: A deep set-based operator network for solving pdes with permutation invariant variable input sampling. *arXiv e-prints*, pp. arXiv–2505, 2025.

U.S. Department of Energy, Office of Science. DOE explains...Tokamaks. `https://www.energy.gov/science/doe-explainstokamaks`, n.d. Tokamak diagram courtesy of EUROfusion.

JT Wai, MD Boyer, and E Kolemen. Neural net modeling of equilibria in nstx-u. *Nuclear Fusion*, 62(8):086042, 2022.

Chenguang Wan, Zhi Yu, Alessandro Pau, Olivier Sauter, Xiaojuan Liu, Qiping Yuan, and Jiangang Li. A machine-learning-based tool for last closed-flux surface reconstruction on tokamaks. *Nuclear Fusion*, 63(5):056019, 2023.

Huasheng Xie, Ruohan Zhang, Xingyu Li, Feng Zhang, and Zhengxiong Wang. VEQDB: A compact and reconstructible multi-device tokamak equilibrium database, 2026.

Jay Phil Yoo, William Howes, Yashika Ghai, Kazuma Kobayashi, Souvik Chakraborty, and Syed Bahauddin Alam. Towards data-efficient cross-device generalization of grad-shafranov equilibria via transfer learning neural operator. *arXiv preprint arXiv:2606.15512*, 2026.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

GH Zheng, ZY Yang, SF Liu, Rui Ma, XW Gong, Ao Wang, Shuo Wang, and WL Zhong. Real-time equilibrium reconstruction by multi-task learning neural network based on hl-3 tokamak. *Nuclear Fusion*, 64(12):126041, 2024.

GH Zheng, SF Liu, HS Xie, X Gu, ZY Chen, XC Lun, Y Liu, J Li, D Guo, RY Tao, et al. Efit-mini: an embedded, multi-task neural network-driven equilibrium inversion algorithm. *Nuclear Fusion*, 65(10):106008, 2025.

GH Zheng, SF Liu, X Gu, YP Zhang, J Li, Y Liu, XC Lun, L Xing, JG Chen, ZY Chen, et al. A novel numerical algorithms optimization method with machine learning frameworks: application on real-time plasmas equilibrium reconstruction in exl-50u spherical torus. *arXiv preprint arXiv:2601.12378*, 2026.

Cuizhi Zhou and Kaien Zhu. Physics-informed neural networks for high-precision grad-shafranov equilibrium reconstruction. *arXiv preprint arXiv:2507.16636*, 2025.

A Related Work

Numerical equilibrium reconstruction. Classical methods such as EFIT reconstruct plasma equilibria by fitting diagnostic measurements subject to the Grad–Shafranov equation and are widely used for offline analysis (Lao et al., 1985; 1990) However, these solvers are often too slow for feedback control, which typically needs update rates of 1 kHz or higher (Ferron et al., 1998). Acceleration therefore uses GPU parallelization, model simplification, reduced-order modeling, or approximate solvers, including rt-EFIT (Ferron et al., 1998), GPU-parallel P-EFIT (Huang et al., 2020), LIUQE (Moret et al., 2015), real-time current-profile inversion (Blum et al., 2012), GPEC (Rampp et al., 2016), and PTEFIT (Zheng et al., 2026), which reaches sub-millisecond reconstruction on EXL-50U by optimizing the numerical stack. These solvers remain interpretable, but cost grows with grid resolution and model complexity, and each device hosts its own probe layout, vessel model, and regularization, so equilibria are not immediately comparable across devices.

Learning-based equilibrium reconstruction. Neural surrogates learn mappings from diagnostic measurements to equilibrium quantities, enabling rapid inference after training. (Joung et al., 2020; Lu et al., 2023; Wai et al., 2022; Wan et al., 2023; Zheng et al., 2024). Later work followed three routes: physics-constrained GS solvers, from a supervised network to GS-DeepNet on KSTAR (Joung et al., 2020; 2023); the EFIT-AI stack of portable solvers, a DIII-D database, and model-order-reduction surrogates (Lao et al., 2022); and multi-task maps such as EFITNN on HL-3 (Zheng et al., 2024). Hybrid and physics-informed designs inject governing equations, boundary conditions, or other constraints into training or into the solver, including EFIT-mini (Zheng et al., 2025), EFIT-Prime (Madireddy et al., 2024), Green's-function surrogates (McClenaghan et al., 2024), and PINN reconstructions (Zhou & Zhu, 2025). Although inference is fast, the accuracy depends strongly on the training distribution and may violate physical consistency (Zheng et al., 2026), while cross-domain transferability has yet to be systematically established (Nakkina et al., 2026). Their progress motivates a controlled comparison of architectures and training objectives: whether improvements persist after transfer, and whether better flux predictions also yield more accurate geometry and physical consistency.

Datasets and transfer benchmarks. Recent resources for tokamak plasma modeling and equilibrium-related prediction have been released: VEQDB primarily provides parameter-sampled, fixed-boundary synthetic equilibria (Xie et al., 2026); TokaMark provides MAST experimental data for magnetic inverse reconstruction (Rousseau et al., 2026); The recent Fusion Equilibrium Challenge includes a cross-device track, while the target is shape rather than magnetic inverse(Nakkina et al., 2026); CrossGS studies cross-device GS generalization with neural operators (Yoo et al., 2026).

TokaReconBench complements these resources by combining simulation data and experimental data from three devices. It evaluates same-device Sim2Real and cross-device reconstruction under explicit target-supervision settings, with separate metrics for field accuracy, geometry, and physical consistency. (Table 1.). **This joint evaluation tests how effectively simulations and existing-device data support reconstruction on a new device when experimental labels are limited.**

B Fusion Background

B.1 Tokamak Equilibrium and Reconstruction

Magnetic confinement and coordinates. A tokamak confines a fully ionized plasma with a strong magnetic field inside an axisymmetric vacuum vessel. Throughout this paper we use cylindrical coordinates (R, ϕ, Z), where R is the distance from the machine symmetry axis, Z the vertical position, and ϕ the toroidal angle. The confining field has two components: a toroidal field produced by external toroidal-field coils, and a poloidal field produced by the plasma current together with the central solenoid (CS) and the poloidal-field (PF) coils. Their superposition yields helical field lines that trace out nested toroidal surfaces, on which charged particles are confined. Figure 6 shows this coil and field layout, together with photographs and CAD models of the three devices in this benchmark, EXL-50U, MAST, and ITER.

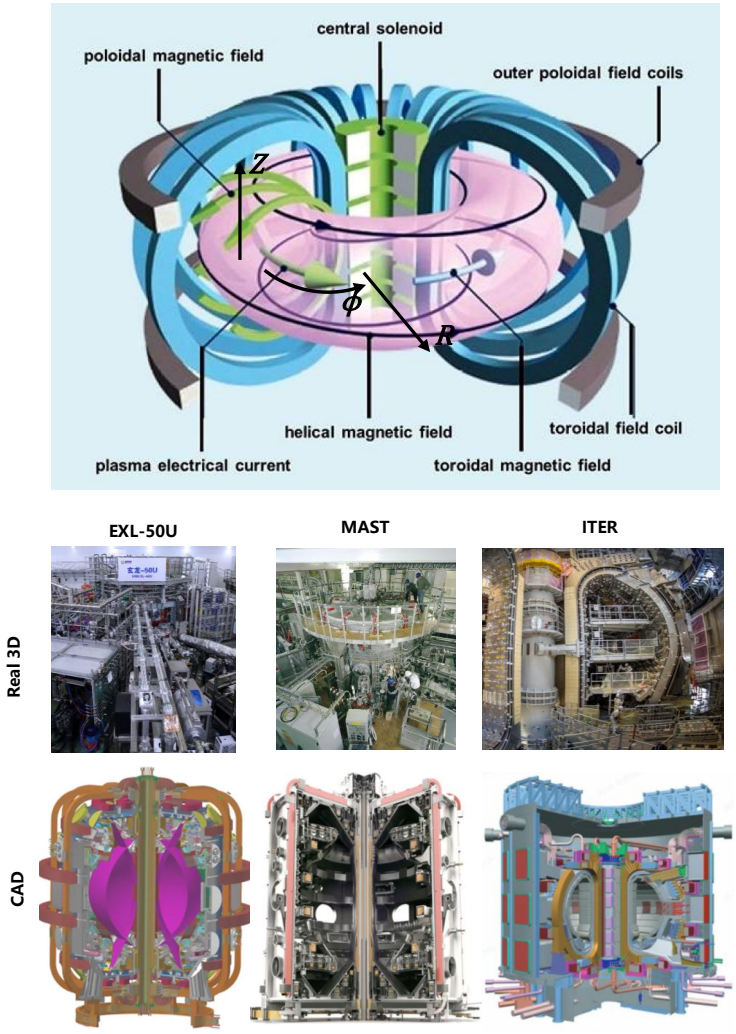

Figure 6: A typical tokamak and the three devices in this benchmark. Top: coil and field layout (image courtesy of EUROfusion, via (U.S. Department of Energy, Office of Science, n.d.)). Bottom: photographs and CAD models of EXL-50U, MAST, and ITER.

Poloidal flux and magnetic topology. Axisymmetry allows the poloidal field to be written in terms of a single scalar, the poloidal flux function $\psi(R, Z)$, defined as the magnetic flux through the horizontal disc of radius R centred on the symmetry axis at height Z (per radian in our convention, in $\mathrm{Wb/rad}$):

$$B_R = -\frac{1}{R}\frac{\partial \psi}{\partial Z}, \qquad B_Z = \frac{1}{R}\frac{\partial \psi}{\partial R}.$$

Field lines lie on level sets of ψ, called flux surfaces. The innermost surface degenerates to a point, the *magnetic axis* (R_a, Z_a), where ψ attains its extremum ψ_a. Moving outwards, the last surface that closes inside the vessel is the *last closed flux surface* (LCFS), with flux value ψ_b; everything inside it is confined plasma. In a *diverted* configuration the LCFS passes through one or more *X-points* (R_x, Z_x), saddle points of ψ where the poloidal field vanishes; in a *limited* configuration the LCFS instead touches the first wall. The normalized flux $\psi_N = (\psi - \psi_a)/(\psi_b - \psi_a)$ labels surfaces from the axis (ψ_N=0) to the boundary (ψ_N=1). All geometric targets used in our evaluation—LCFS, magnetic axis, and X-point—are therefore functionals of the single field $\psi(R, Z)$, which is why ψ is the only prediction target of the benchmark task.

Grad–Shafranov equilibrium. In ideal magnetohydrodynamic equilibrium the pressure gradient balances the Lorentz force, $\nabla p = \mathbf{j} \times \mathbf{B}$. Under axisymmetry this reduces to the Grad–Shafranov (GS) equation

$$\Delta^* \psi \;\equiv\; R\frac{\partial}{\partial R}\Big(\frac{1}{R}\frac{\partial \psi}{\partial R}\Big) + \frac{\partial^2 \psi}{\partial Z^2} \;=\; -\mu_0 R^2\, p'(\psi) \;-\; F(\psi)F'(\psi),$$

with the toroidal current density

$$j_\phi \;=\; R\,p'(\psi) \;+\; \frac{F(\psi)F'(\psi)}{\mu_0 R},$$

where $p(\psi)$ is the plasma pressure and $F(\psi) = RB_\phi$ the poloidal current function; both are free functions of ψ alone. In the *free-boundary* setting relevant here, ψ is generated jointly by the plasma current density and the known external CS/PF coil currents, and the plasma boundary is not prescribed but emerges from the solution. *Equilibrium reconstruction* is the corresponding inverse problem: given external magnetic measurements—flux loops sampling ψ at fixed wall positions, pick-up probes sampling local poloidal-field components, and the coil currents—infer parameterizations of p' and FF' (and hence ψ, j_ϕ, and the derived geometry) consistent with the GS equation. This is exactly the mapping $f_\theta : (m, I, g) \mapsto \psi(R, Z)$ that our benchmark task emulates, and the GS residual used in our physics-level evaluation measures how well a predicted ψ satisfies the equation above.

The standard solution is the equilibrium-fitting code EFITLao et al. (1985), which parametrizes $\mathrm{d}p/\mathrm{d}\psi$ and $F\,\mathrm{d}F/\mathrm{d}\psi$ by a small set of physical quantities—typically the total plasma current I_p, the poloidal beta β_p, the internal inductance l_i, and the edge (or axial) safety factor—and reconstructs the equilibrium by least-squares fitting. The synthetic magnetic signals computed from the reconstructed ψ are matched to the measurements, with the ill-conditioned fit regularized by singular-value decomposition. Because the source terms depend nonlinearly on ψ, the reconstruction alternates between solving the GS equation and re-fitting the profile parameters (a Picard iteration) until convergence. EFIT additionally accommodates both limiter and diverted configurations, where the plasma boundary is set by wall contact or by a magnetic X-point, respectively.

Equilibrium reconstruction is therefore a nonlinear inverse problem that requires many GS solves per discharge, which is computationally demanding for real-time control and for processing large experimental or synthetic datasets. This motivates learning-based surrogates that map available diagnostic signals directly to the equilibrium field $\psi(R, Z)$ and to derived quantities (plasma boundary, magnetic axis, shape parameters), without explicitly solving the GS equation—precisely the task formalized by our benchmark.

Geometric and global parameters. Figure 7 illustrates the main geometric quantities on a poloidal cross-section. With $R_{\min}$ and $R_{\max}$ the innermost and outermost radii of the LCFS midplane, the geometric centre and minor radius are $R_0 = (R_{\max}+R_{\min})/2$ and $a = (R_{\max}-R_{\min})/2$; the aspect ratio is $A = R_0/a$. Spherical tokamaks such as EXL-50U and MAST operate at low aspect ratio ($A \lesssim 2$), while ITER is a conventional large-aspect device ($A \approx 3.1$). The elongation

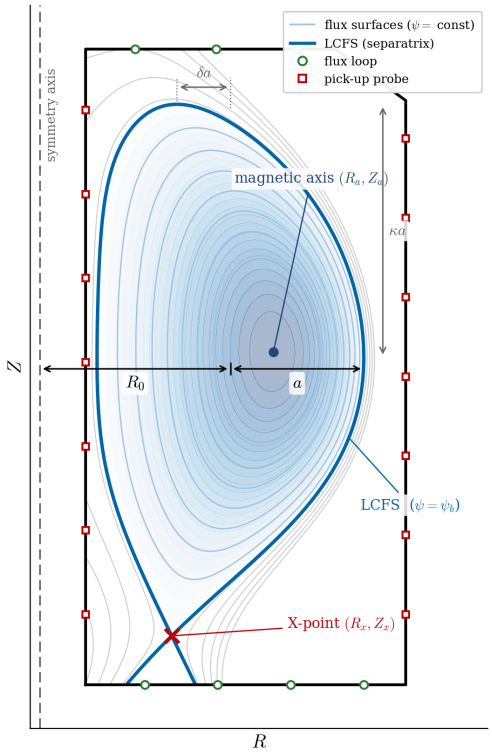

Figure 7: Poloidal cross-section of a diverted tokamak equilibrium. Nested contours of the poloidal flux $\psi(R, Z)$ are shown inside the first wall, together with the magnetic axis (R_a, Z_a), the last closed flux surface (LCFS), the X-point (R_x, Z_x), the geometric centre R_0, the minor radius a, and the elongation κ and triangularity δ of the boundary. Flux loops and magnetic pick-up probes located outside the plasma provide the measurements m used for reconstruction.

$\kappa = (Z_{\max} - Z_{\min})/(2a)$ and the upper/lower triangularities δ_u, δ_l describe the vertical stretching and the inward shift of the top/bottom of the LCFS. Global scalars characterizing the equilibrium include the plasma current $I_p = \int j_\phi \, dA$, the internal inductance ℓ_i, and the poloidal and toroidal betas β_p, β_t (ratios of kinetic to magnetic pressure). These are the shape scalars stored with each sample in Appendix C; together with ψ_a, ψ_b, (R_a, Z_a), and (R_x, Z_x) they summarize the equilibrium geometry that our geometry-level metrics evaluate.

B.2 Fusion equipments

Our benchmark spans three tokamak devices — EXL-50U, MAST, and ITER — that together cover distinct magnetic configurations and a wide range of plasma parameters relevant to equilibrium reconstruction. EXL-50U and MAST are low-aspect-ratio spherical devices, whereas ITER is a large conventional tokamak. Tabel 6 summarizes their main parameters. B_t is the toroidal magnetic field strength, which directly determines plasma confinement performance; R_0 is the major radius, denoting the distance from the device center to the center of the plasma cross-section; a is the minor radius, representing the characteristic size of the plasma cross-section; κ is the elongation, defined as the ratio of the vertical height to the horizontal width of the plasma cross-section, reflecting its ellipticity; and Ip is the plasma current, a critical parameter for maintaining plasma equilibrium and stability.

EXL-50U Shi et al. (2025a) is China's first large spherical torus device with a toroidal field reaching 1 T, upgraded from EXL-50, the first spherical torus constructed by ENN. As an intermediate device between EXL-50 and the next-generation EHL-2Liang et al. (2025), its experiments are designed to support the physical design and operational scenarios of EHL-2 in the areas of current drive, high-ion-temperature exploration, energy transport and confinement, and hydrogen–boron physics. The

Table 6: Main parameters of the three tokamak devices.

Configuration	B_t (T)	R_0 (m)	a (m)	κ	I_p (MA)
EXL-50U	1.0–1.2 (at $R = 0.6$ m)	0.6–0.8	0.32–0.57	1.4–2	0.5–1
MAST	0.51	0.85	0.65	2.2	1.05
ITER	5.3	6.2	2.0	1.7	15

device operates at major radii of 0.6–0.8 m with an aspect ratio of 1.4–1.85, and its toroidal field reaches 1.0 T (1.2 T in the second phase) at R = 0.6 m. In the latest experiments, MA-level discharges were achieved for the first time in hydrogen–boron plasmasShi et al. (2025b): with a boron-rich fueling strategy reaching 10% boron content and 800 kW of ECRH power, the plasma current was ramped to 1 MA with core electron temperatures up to 3 keV, demonstrating the feasibility of high-performance hydrogen–boron plasmas in a spherical torus.

MAST (the Mega Amp Spherical TokamakSykes et al. (2001)), operated at the UK Atomic Energy Authority, is a scaled-up version of the earlier START experiment. It was designed to investigate the properties of the spherical tokamak in large, collisionless plasmas and to provide new perspectives on tokamak physics for ITER and beyond. With a major radius of 0.85 m and a minor radius of 0.65 m (aspect ratio 1.3), MAST has achieved plasma currents above 1 MA, elongated plasmas with $\kappa \approx$ 2.2, and a toroidal field of about 0.5 TMorris (2012). Since beginning operation in 2000, MAST has produced extensive results on edge localized mode mitigation, divertor physics, and confinement, and has been upgraded to MAST-U, which adds a super-X divertor.

ITER is designed as the world's first reactor-scale magnetic fusion experimentShimada et al. (2007); Aymar et al. (2002), with the goal of demonstrating the scientific and technological feasibility of fusion energy for peaceful purposes. Its reference design targets controlled ignition and extended burn of D–T plasmas, combining superconducting magnet systems with a single-null divertor. The ITER design specifies a major radius of 6.2 m, a minor radius of 2.0 m, a toroidal field of 5.3 T, a nominal plasma current of 15 MA, and an elongation of about 1.7.

Taken together, the three devices span the low-aspect-ratio spherical regime (EXL-50U, MAST) and the conventional high-aspect-ratio regime (ITER), across more than an order of magnitude in major radius and plasma current. This diversity in configuration and scale directly underpins the generalization levels defined in SectionC, where experimental and synthetic equilibria are drawn from EXL-50U, experimental equilibria from MAST, and synthetic equilibria from ITER.

C Dataset Details

TokaReconBench is assembled from three sources that already exist as public or gated releases: PTEFIT reconstructions of the EXL-50U spherical tokamak, obtained with a numerical Grad–Shafranov solver rather than a learned model, FAIR-MAST Level 2 experimental shots, and the ITER-like synthetic free-boundary catalogue. Table 7 outlines the five releases. Channel counts are native to each device and are never aligned by array index across machines. The EXL-50U simulation trajectories are generated by FGE and inverted by the same PTEFIT operator as the experiment; the shared layout and scale of that pair are in Table 3, and the generation and pairing process is in Section C.1.

C.1 EXL-50U simulation provenance

The EXL-50U *sim2real* pair uses one device configuration and three codes with distinct roles. FGE is the *forward* model: given the voltages applied to the poloidal-field coils, it advances the free-boundary Grad–Shafranov equilibrium together with the circuit dynamics of the active coils and passive conductors, and synthesizes the magnetic diagnostics those equilibria would produce (Heiß et al., 2026). PTEFIT is the *inverse* model that supplies learning labels: an EFIT-class numerical Grad–Shafranov reconstruction, not a learned surrogate, applied to both experimental and synthetic magnetics (Zheng et al., 2026). LIUQE is used only to validate FGE against experiment, not to label the benchmark (Moret et al., 2015). All three share the EXL-50U device description: one central solenoid, fourteen PF circuits, 47 flux loops, the 164-probe magnetic set (84 tangential and

Table 7: Dataset details across the five releases used in **TokaReconBench**.

Source	Magnetics	Coil currents	Grid / time	Access
EXL-50U exp. 100-probe	52 tan. + 48 nor. B 47 loops	CS and 12 PF	129×129 shot, ms	ModelScope gated
EXL-50U exp. 164-probe	84 tan. + 80 nor. B 47 loops	CS and 12 PF	129×129 shot, ms	ModelScope gated
EXL-50U sim	same as 164-probe	CS and 12 PF	129×129 shot, ms	ModelScope SIM gated
MAST	78 B (40/19/19) 15 loops	solenoid and 12 PF	65×65 3 clocks, s	FAIR-MAST S3 public
ITER-like	187 pick-up coils no loops	6 CS, 6 PF, 2 VS	64×64 none	HuggingFace public

80 normal), the vacuum Green's table on a 129×129 mesh, and the operational constraint weights and disabled-channel list used for device reconstruction.

Validation of the forward model. The EXL-50U implementation of FGE was constructed from device engineering data and calibrated with dedicated experiments. Ten simulation cases, each corresponding to a distinct discharge, were retained only when FGE, driven by the recorded coil-voltage commands, reproduced the evolution of plasma current, position, and shape obtained from LIUQE on the same shots. Figure 8 shows a representative comparison using I_p, $R_{\max}$, $R_{\min}$, and κ.

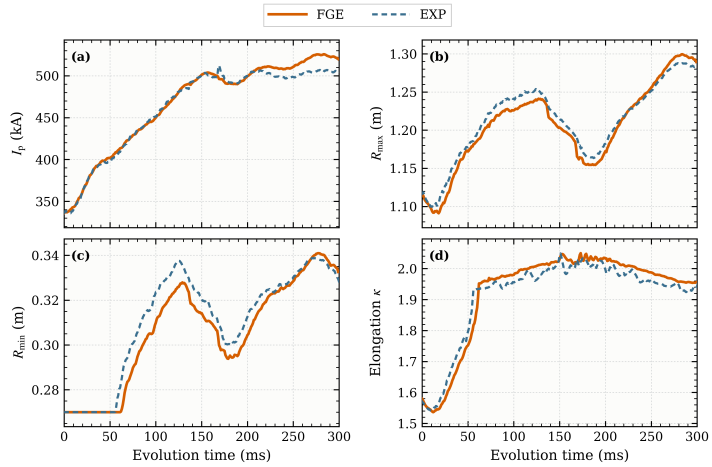

Figure 8: FGE validation for a representative EXL-50U discharge used to build the simulated ensemble. The FGE simulation (solid) and the experimental LIUQE reconstruction of the same shot (dashed; labelled EXP in the panel) are compared on I_p, $R_{\max}$, $R_{\min}$, and κ, which characterize the equilibrium current, radial extent, and elongation.

Generation of the simulated ensemble. The ten validated cases were re-run under PID and reinforcement-learning (RL) control so that both the internal plasma state and the closed-loop current trajectories vary. In FGE the toroidal current density is the source term of the Grad–Shafranov equation and is parameterized by I_p, β_p, and q_{axis}: I_p sets the magnitude, β_p the pressure-driven contribution, and q_{axis} the peaking of the current profile. PID runs hold the nominal $500\,\text{kA}$ flat-top while sampling β_p and q_{axis}; RL runs close the loop with a plasma-current controller that was trained in FGE and subsequently demonstrated on the device. Each validated discharge contributes 200 PID and 100 RL runs of 300–500 time steps. The resulting ensemble size and sampled ranges are those of Table 3.

Pairing with the experimental configuration. FGE supplies the physical trajectory and the synthetic magnetics; it does not supply the learning target. The process that turns these trajectories into the benchmark pair is as follows.

1. **Forward evolution.** FGE is integrated from the experimental coil-voltage protocol of a validated discharge, under PID or RL control as above.
2. **Diagnostic synthesis.** At each time step the same 47 loops and 164 probes used on the device are evaluated on the FGE equilibrium, producing a synthetic measurement vector in the device channel order and units.
3. **Shared inversion.** That vector, together with the 15-dimensional CS/PF current, is inverted by the same real-time PTEFIT operator that labels the experimental archive, carrying solver state across the time slices of each shot as in online operation. The learning target is the reconstructed ψ map, not the FGE equilibrium. Slices with $I_p < 100\,\mathrm{kA}$ (breakdown and termination) and slices that fail to invert are discarded.
4. **Matched release.** Experiment and simulation are stored with the same HDF5 groups, coil and loop identities, grid, and millisecond clock (Table 8). Matched channels and a shared reconstruction operator remove those two discrepancies from the comparison. They do not isolate which physical source dominates the *sim2real* gap.

C.2 EXL-50U Magnetic Diagnostic Uncertainty

Although the simulated and experimental EXL-50U data are aligned in sensor layout and physical variables, their magnetic measurements remain subject to different sources of error. Synthetic signals are generated by an idealized magnetic forward model, while experimental signals also reflect diagnostic calibration errors and device-dependent electromagnetic effects. Quantifying this discrepancy is important for interpreting the *sim2real* gap and for designing realistic perturbations of the simulated inputs. To isolate the magnetic forward model discrepancy from plasma-connected effects, we analyze several plasma-free EXL-50U measurement operated with the same PF coil program. Under this controlled setting, the vacuum forward prediction $\mathbf{y} = \mathbf{GI}$ can be directly compared channel-wise with the three types of measured magnetic signals: MPT, MPN, and flux loops.

Systematic discrepancy. The calibration results show that systematic effects dominate the magnetic measurement discrepancy. Across the three diagnostic types, the median relative RMS error is approximately 1%, while the mean reaches 2–3% because of a small number of abnormal channels. About 80% of the channels remain within 3% error and about 90% remain within 5%. The channel-wise error pattern is highly reproducible across calibration discharges, showing that most of the discrepancy is persistent across shots.

Random fluctuations. High-frequency fluctuations are more than one order of magnitude smaller than the systematic error and contribute only weakly under quasi-static conditions. Their dominant component is also common across channels. Independent Gaussian noise therefore captures only a limited part of the experimental measurement uncertainty.

Implications for the *sim2real* task. The discrepancy is mostly a reproducible channel-wise bias, with a median relative RMS near 1%. Gain and additive perturbations therefore cover only part of the measurement mismatch, and high-frequency noise is a weak model of it. The sim-to-real checks use these perturbations as sensitivity tests. They do not identify the dominant cause of the *sim2real* gap.

C.3 EXL-50U

The first source of the benchmark is the PTEFIT numerical reconstruction archive of EXL-50U. One sample is one time slice of a shot at which the inversion succeeded. Diagnostics and inversion outputs share the same clock, in milliseconds, so no extra time alignment is required.

The two experimental batches differ only in the magnetic-probe set. The first uses a 100-probe layout with 52 tangential and 48 normal probes. The second uses a 164-probe layout with 84 tangential and 80 normal probes, covering 595 shots. The simulation release follows the 164-probe layout. Combined with the 47 flux loops, the magnetic measurement vector therefore has 147 entries on the 100-probe batch and 211 entries on the 164-probe batch and the simulation.

Table 8: Stored fields in each EXL-50U shot file.

Quantity	Field	Shape or count	Notes
Flux loops	`chns/flux`	$(T, 47)$	Wb/rad; default input
Tangential B probes	`chns/mpt`	$(T, 52)$ or $(T, 84)$	tesla; 52 on the 100-probe batch
Normal B probes	`chns/mpn`	$(T, 48)$ or $(T, 80)$	tesla; 48 on the 100-probe batch
Coil currents	`chns/CS`, `PF`, `TF`	1 CS, 12 PF, 1 TF	amperes; TF is auxiliary
Plasma current	`chns/IP`	scalar per slice	measured I_p
Loop voltage	`chns/vloops`	$(T, 47)$	stored but unused (weight 0)
Eddy current	`chns/eddy_current`	scalar	placeholder, filled with zeros; not a measured eddy current
Solver weights	`weights/*`	no time axis	circuit and channel flags
Poloidal flux	`inv_result/psi`	$(T, 129, 129)$	float32
Toroidal current density	`inv_result/jphi`	$(T, 129, 128)$	
LCFS	`inv_result/surfs`	90 vertices	$(T, 2, 90, 2)$ on 100-probe; $(T, 2, 90)$ otherwise
Axis and X-points	`ra`, `za`, `rx`, `zx`	2 or 4 X-points	two candidates on 100-probe; four otherwise
Shape scalars	a, κ, $\delta_{l,u}$, ℓ_i, β_p, β_t, ψ_a, ψ_b	per slice	

All three releases share the same coil set: one central solenoid, fourteen poloidal-field circuits, and an auxiliary toroidal-field current. The reconstruction current used by the benchmark is

$$I = [I_{\mathrm{CS}}, I_{\mathrm{PF},1}, \ldots, I_{\mathrm{PF},12}] \in \mathbb{R}^{13}.$$

Solver weights on these circuits are flags, not least-squares measurement weights. Measured I_p is stored with the diagnostics and enters the inversion as a weak least-squares constraint, with weight about 10^{-6} , so it does not set the equilibrium. The reconstructed plasma current is the integral of J_ϕ, not a copy of the Rogowski value. The solution is set by the flux loops, the magnetic probes, and the coil currents.

Native records are HDF5, one file per shot, indexed by shot number and time. The time axis lists only slices that inverted. Full-layout files record physical units. Experiment and simulation files use the same groups. The stored fields are listed in Table 8. Access is through ModelScope and is gated.

For a learning model the default input is that magnetic vector together with the 15-dimensional coil current, and the target is the PTEFIT ψ map on the 129×129 grid. Experiment and simulation therefore share a numerical label source and differ as in-domain versus *sim2real*; the 100-probe and 164-probe batches are a sensor-layout shift, not a change of physics. Loop voltage and the eddy-current channel sit in the same file but are excluded from the default input and must not be concatenated just because they are stored.

C.4 MAST

The first public transfer source is the experimental Level 2 shots from FAIR-MAST. One sample is one Level 2 time slice.

Equilibrium, magnetics, and poloidal-field currents are archived on three different clocks, in seconds. The 15 flux loops and 78 probes share the magnetics time base, not the equilibrium clock. Time alignment is therefore part of preprocessing rather than a property of the raw record.

Coil currents form a 13-dimensional vector: the central solenoid together with P2 (inner and outer) and P3 to P6, each split into upper and lower. The layout is the same across Level 2 campaigns. Circuit names follow IMAS and must not be matched by index to the EXL-50U poloidal-field circuits.

Native records are Level 2 Zarr on the public STFC S3 bucket, indexed by shot number and IMAS group. The catalogue is the public FAIR-MAST release. The stored fields are listed in Table 9. Access is public.

Table 9: Stored fields used from FAIR-MAST Level 2 shots.

Quantity	Field	Shape or count	Notes
Flux loops	`magnetics/flux_loop_flux`	15	Wb; default input
Centre-column B probes	`ccbv`	40	tesla
Outboard radial B probes	`obr`	19	tesla
Outboard vertical B probes	`obv`	19	tesla
Plasma current	`summary/ip`	scalar per slice	also available under magnetics
Coil currents	`pf_active/coil_current`	13 circuits	solenoid and 12 PF
Poloidal flux	`equilibrium/psi`	$(65, 65, T_{\mathrm{eq}})$	float64
Toroidal current density	`equilibrium/j_phi`	same grid	
LCFS	`equilibrium/lcfs_r`, `lcfs_z`	contour	
Shape scalars	elongation, ℓ_i, β	per slice	equilibrium scalars
Kinetics	Thomson scattering, CX	raw	privileged; not a default input

The 15 loops and 78 probes give 93 magnetic inputs. These counts hold for the FAIR-MAST Level 2 campaigns used here, not for Level 1 or other exports.

A model therefore sees those 93 channels plus the 13-dimensional coil vector and predicts ψ on a 65×65 grid. Pairing magnetics with the equilibrium label is a preprocessing choice, not a property of the raw record; an undeclared nearest-neighbour alignment can look like leakage. The 13 circuits are not the EXL-50U set padded with zeros, and Thomson scattering and charge exchange are privileged kinetics, not default inputs.

C.5 ITER-LIKE

The ITER-like source is a table of static synthetic free-boundary equilibria computed with the IET/FRIDA solvers. One sample is one table row. There is no shot number and no time base.

The magnetic input models the pick-up coil set of the ITER magnetic diagnostic system and contains no flux loops. Sensor positions are not shipped with the HuggingFace table or the wrapper package. The package adds only the shared computational grid and the first wall. Geometry of the 187 probes, including R, Z, inclination angle, and ITER-convention channel names, comes from the FRIDA device description. The same 187-channel table is used for every row, and the magnetic measurements are read in that native order.

Each row stores a 12-dimensional coil vector: six central-solenoid circuits, six poloidal-field circuits, and two vertical-stability circuits, packed as one vector with no split among circuit families. The last two entries are often zero. Currents are never concatenated across devices without a circuit identifier and a geometry table.

The stored fields are listed in Table 10. Access is through HuggingFace and is public.

The learning task is a static map from 187 pick-up measurements and the 12-dimensional coil vector to a 64×64 ψ target, with no time axis and no flux loops. The HuggingFace field named `flux` is that target, not a loop measurement, and must not be read as EXL-50U `chns/flux`. This is the same reconstruction task on another device, not a downsampled EXL-50U grid.

C.6 PREPROCESSING PIPELINE

Native records and model views are kept distinct. Each model view holds an identifier, a time (if any), measurements, currents, the grid, the target, masks, and recorded transforms. It is built in six steps.

1. **Identity.** Source, release, shot or row, and a content hash define an object. A time slice (EXL-50U, MAST) or a table row (ITER-like) is one sample.
2. **Channel map.** Channels are selected by a channel identifier. The two EXL-50U experimental batches differ only in the number of magnetic probes. When a task must mix the two batches, a 147-channel common view keeps all 47 loops plus the 100-probe subset. MAST

Table 10: Stored fields in each ITER-like table row.

Quantity	Field	Shape or count	Notes
Magnetic probes	`mag_measures`	187	pick-up coils only; no flux loops
Coil currents	`coils_current`	14	6 CS, 6 PF, 2 VS
Poloidal flux	`flux`	64×64	ψ target, not a loop measurement; float64
Right-hand side	`rhs`	same grid	$j_\phi = -\mathrm{rhs}/(\mu_0 R)$ is derived, not stored
LCFS	`separatrix`	200 points	
Diverted flag	`is_diverted`	scalar	used with the separatrix in place of X-point labels
Pressure profile	`p_profile`	101	privileged; not a default input

is mapped through its own channel table. ITER-like keeps its native 187-dimensional vector. Neither is padded into the 211-dimensional EXL-50U vector by position.

3. **Time.** EXL-50U time bases are already aligned, as described above. For MAST's three clocks, at each t_k the nearest neighbour across groups is taken and the slice is dropped if $|\Delta t|$ exceeds a tolerance. ITER-like has no time coordinate.
4. **Units and sign.** Loops are in Wb or $\mathrm{Wb/rad}$, B in T, and currents in A. Polarity and ampere-turns are applied per circuit. An additive gauge $\psi \leftarrow \psi - \psi_{\mathrm{ref}}$ is recorded in the transform log and is not written back into the raw file.
5. **Grid.** Native ψ stays on the source grid (129×129 for EXL-50U, 65×65 for MAST, 64×64 for ITER-like). EXL-50U j_ϕ remains 129×128.
6. **Masks and normalization.** Invalid probes, zero solver weights, and non-finite ψ are masked. Normalizers are fit only on the source-training split of each task.

D Evaluation Metrics

Metrics are computed on the device-native (R, Z) grid. Let Ω be the valid reference region. Area weights w_i use tensor-product trapezoidal quadrature and are normalized as

$$\tilde{w}_i = \frac{w_i}{\sum_{j\in\Omega} w_j}. \tag{3}$$

A small constant $\epsilon = 10^{-12}$ is used only to avoid division by zero.

D.1 Field metrics

The poloidal flux is defined up to an additive constant. The predicted field is aligned to the reference by

$$\delta^\star = \sum_{i\in\Omega} \tilde{w}_i \big(\psi_i^{\mathrm{true}} - \psi_i^{\mathrm{pred}}\big). \tag{4}$$

The flux RMSE is the area-weighted residual after this alignment:

$$e_\psi = \left[\sum_{i\in\Omega} \tilde{w}_i \big(\psi_i^{\mathrm{pred}} + \delta^\star - \psi_i^{\mathrm{true}}\big)^2\right]^{1/2}. \tag{5}$$

Let

$$\bar{\psi}_w^{\mathrm{true}} = \sum_{i\in\Omega} \tilde{w}_i\, \psi_i^{\mathrm{true}}, \qquad A_\psi = \left[\sum_{i\in\Omega} \tilde{w}_i \big(\psi_i^{\mathrm{true}} - \bar{\psi}_w^{\mathrm{true}}\big)^2\right]^{1/2} \tag{6}$$

be the area-weighted mean and the RMS spatial variation of the reference. The relative flux error used in the current evaluator is

$$E_\psi^{\mathrm{rel}} = \frac{e_\psi}{A_\psi + \epsilon}. \tag{7}$$

When the axis and boundary fluxes are available, a second relative flux error is

$$E_{\psi,\mathrm{ab}} = \frac{e_\psi}{|\psi_{\mathrm{boundary}} - \psi_{\mathrm{axis}}| + \epsilon}. \tag{8}$$

Coordinates are normalized to $[-1, 1]$ before derivatives,

$$\hat{R} = 2\frac{R - R_{\min}}{R_{\max} - R_{\min}} - 1, \qquad \hat{Z} = 2\frac{Z - Z_{\min}}{Z_{\max} - Z_{\min}} - 1. \tag{9}$$

Spatial derivatives use centered three-point differences on the native grid. Let Ω_∇ be the points whose center and four axial neighbors lie in Ω, with corresponding weights $\tilde{w}_i^\nabla$. The gradient RMSE is

$$e_{\nabla\psi} = \left[\sum_{i\in\Omega_\nabla} \tilde{w}_i^\nabla \|\nabla\psi_i^{\mathrm{pred}} - \nabla\psi_i^{\mathrm{true}}\|_2^2\right]^{1/2}. \tag{10}$$

The relative gradient error is

$$E_{\nabla\psi}^{\mathrm{rel}} = \frac{e_{\nabla\psi}}{\left[\sum_{i\in\Omega_\nabla} \tilde{w}_i^\nabla \|\nabla\psi_i^{\mathrm{true}}\|_2^2\right]^{1/2} + \epsilon}. \tag{11}$$

D.2 Geometrical metrics

Geometry is extracted from the predicted and reference flux fields with the same post-processing. E_{axis} and $E_{\mathrm{X-pt}}$ are in centimeters.

The magnetic-axis error is

$$E_{\mathrm{axis}} = \|\mathbf{x}_{\mathrm{axis}}^{\mathrm{pred}} - \mathbf{x}_{\mathrm{axis}}^{\mathrm{true}}\|_2, \qquad \mathbf{x}_{\mathrm{axis}} = (R_{\mathrm{axis}}, Z_{\mathrm{axis}}). \tag{12}$$

Let $\mathcal{C}_{\mathrm{pred}}$ and $\mathcal{C}_{\mathrm{true}}$ be the predicted and reference LCFS contours, each resampled to 256 points in arc length. With $d(\mathbf{x}, \mathcal{C})$ the Euclidean point-to-polyline distance,

$$D_{95}(\mathcal{C}_1 \to \mathcal{C}_2) = P_{95}\big(\{d(\mathbf{x}, \mathcal{C}_2) : \mathbf{x} \in \mathcal{C}_1\}\big), \tag{13}$$

and the symmetric LCFS HD95 is

$$E_{\mathrm{LCFS}}^{95} = \max\big[D_{95}(\mathcal{C}_{\mathrm{pred}} \to \mathcal{C}_{\mathrm{true}}),\ D_{95}(\mathcal{C}_{\mathrm{true}} \to \mathcal{C}_{\mathrm{pred}})\big]. \tag{14}$$

An additional overlap score, not named in the outline, is also implemented. Let A_{pred} and A_{true} be the regions enclosed by the two LCFS contours. Then

$$\mathrm{IoU}_{\mathrm{LCFS}} = \frac{\sum_i w_i\, \mathbf{1}[i \in A_{\mathrm{pred}} \cap A_{\mathrm{true}} \cap \Omega]}{\sum_i w_i\, \mathbf{1}[i \in (A_{\mathrm{pred}} \cup A_{\mathrm{true}}) \cap \Omega]}. \tag{15}$$

For diverted equilibria, let $\mathcal{X}_{\mathrm{pred}}$ and $\mathcal{X}_{\mathrm{true}}$ be the predicted and reference active X-point sets. When the two sets have the same cardinality, the X-point localization error is the Hausdorff distance

$$E_X = \max\left\{\max_{\mathbf{x}\in\mathcal{X}_{\mathrm{pred}}} \min_{\mathbf{y}\in\mathcal{X}_{\mathrm{true}}} \|\mathbf{x} - \mathbf{y}\|_2,\ \max_{\mathbf{y}\in\mathcal{X}_{\mathrm{true}}} \min_{\mathbf{x}\in\mathcal{X}_{\mathrm{pred}}} \|\mathbf{x} - \mathbf{y}\|_2\right\}. \tag{16}$$

Limiter equilibria are excluded from $E_{\mathrm{X-pt}}$. Let $\mathcal{D}$ be the diverted frames and $\mathcal{D}_{\mathrm{det}}$ the subset with $|\mathcal{X}_{\mathrm{pred}}| = |\mathcal{X}_{\mathrm{true}}|$. The X-point detection rate is

$$R_X = \frac{|\mathcal{D}_{\mathrm{det}}|}{|\mathcal{D}|}. \tag{17}$$

E_X is averaged only on $\mathcal{D}_{\mathrm{det}}$. Frames in $\mathcal{D} \setminus \mathcal{D}_{\mathrm{det}}$ are detection failures, not a finite distance.

D.3 TOPOLOGICAL METRICS

Each frame is assigned one topology class from the same extraction used for the geometrical metrics:

- limiter: no active X-point and the LCFS lies on the first wall;
- single-null: one active X-point;
- double-null: two active X-points.

Let y_k and $\hat{y}_k$ be the reference and predicted classes on frame k, and let N be the number of frames. Topology accuracy is

$$\mathrm{Acc}_{\mathrm{topo}} = \frac{1}{N}\sum_{k=1}^{N} \mathbf{1}[\hat{y}_k = y_k]. \tag{18}$$

Macro-F1 is the unweighted mean of the three per-class F1 scores.

D.4 PHYSICAL CONSISTENCY

The physics-consistency score is the area-weighted residual of $\Delta^*\psi - (-\mu_0 R J_\phi) = 0$ on the predicted flux, using the reference toroidal current and the same centered differences as the gradient metric:

$$e_{\mathrm{phys}} = \left[\sum_{i\in\Omega_\nabla} \tilde{w}_i^{\nabla}\left(\Delta^*\psi_i^{\mathrm{pred}} - (-\mu_0 R_i J_{\phi,i}^{\mathrm{ref}})\right)^2\right]^{1/2}. \tag{19}$$

The relative residual is

$$E_{\mathrm{phys}} = \frac{e_{\mathrm{phys}}}{\left[\sum_{i\in\Omega_\nabla} \tilde{w}_i^{\nabla}\left(\mu_0 R_i J_{\phi,i}^{\mathrm{ref}}\right)^2\right]^{1/2} + \epsilon}. \tag{20}$$

J_ϕ^{ref} is taken from the label record when a current density is stored, otherwise from that label solver's $P'(\psi)$ and $FF'(\psi)$. It is never inferred from the predicted ψ. The label solvers are PTEFIT on EXL-50U, the FAIR-MAST Level 2 equilibrium reconstruction on MAST, and the IET/FRIDA free-boundary solvers on the ITER-like catalogue.

D.5 STATISTICAL REPORTING

Implemented in the current appendix: metrics are computed per frame, averaged within each benchmark group, then macro-averaged over groups. For geometry metrics, a frame with an eligible reference structure but a failed prediction-side extraction is recorded as a failure rather than dropped.

E BASELINE DETAILS

We select baseline models to cover complementary assumptions about tokamak equilibrium reconstruction rather than simply increasing model complexity. The benchmark includes convolutional neural surrogates, neural operators, set-based architectures, and geometry-aware models.

This diversity enables us to investigate how different inductive biases, including dimensionality reduction, spatial locality, operator learning, and geometric encoding, affect equilibrium reconstruction accuracy and transfer performance.

All baselines receive the same canonical benchmark sample

$$\mathcal{S} = \{m, I, g, \mathcal{G}, \psi_{\mathcal{G}}\},$$

through model-specific adapters. The adapters only reshape or encode the available benchmark information into the representation required by each architecture, without introducing additional physical measurements.

Unless otherwise stated, all models predict the same poloidal-flux field ψ and are evaluated on the device-native target grid.

E.1 Fixed-grid Neural Surrogate Models

U-Net. The U-Net baseline formulates equilibrium reconstruction as a fixed-grid image-to-image regression problem.

The model adopts the U-Net encoder–decoder architecture Ronneberger et al. (2015), where convolutional encoders extract hierarchical spatial features and decoders reconstruct the target flux field through progressive upsampling.

Skip connections preserve high-resolution spatial information during decoding.

Non-image diagnostic features are first transformed into spatial feature channels through the benchmark adapter before entering the convolutional network.

FNO. We adopt the Fourier Neural Operator (FNO) Li et al. (2020) as a fixed-grid neural operator baseline.

The model first lifts the input representation into a latent feature space and then applies Fourier spectral convolution layers to learn global interactions between spatial locations.

The final projection layer reconstructs the target poloidal flux field.

E.2 Operator-learning Baselines

PlaNet. We include PlaNet, a convolutional physics-informed neural operator for plasma equilibrium reconstruction proposed by Bonotto et al. Bonotto et al. (2024).

Unlike purely data-driven surrogate models, PlaNet incorporates physics-aware operator learning for reconstructing equilibrium quantities including magnetic flux and separatrix information.

In our benchmark, the original implementation is adapted through a unified input interface while preserving its operator-learning structure.

E.3 Set-based and Geometry-aware Operator Models

SetONet. SetONet treats diagnostic measurements as unordered sets instead of fixed-size vectors.

Following the general principle of set-based neural architectures Zaheer et al. (2017), individual sensor elements are encoded through shared feature mappings and aggregated using permutation-invariant operations.

The resulting global representation is combined with spatial queries to predict the equilibrium field.

DeepONet. DeepONet extends the original DeepONet framework by introducing geometry-aware input representations.

Sensor measurements, coil information, and spatial coordinates are encoded with their associated geometric information. The resulting representations are combined through branch–trunk operator learning to reconstruct the flux field.

E.4 Geometry-aware Neural Operator and Transformer Models

GINO. GINO adopts a geometry-informed neural operator formulation Li et al. (2023).

The model maps irregular physical inputs into a latent geometric representation and applies neural operator layers to learn mappings between input functions and output fields.

This allows the model to handle nonuniform geometries more naturally than standard grid-based neural operators.

GNOT. GNOT is included as a transformer-based neural operator baseline Hao et al. (2023).

Table 11: Training and hyperparameter settings for the benchmark baselines.

Model	Optimizer	LR	Batch	Epochs/Steps	Weight Decay	Loss	Seeds
U-Net	AdamW	3×10^{-4}	32	20k	1×10^{-4}	ψ	1101
Transformer	AdamW	1×10^{-3}	32	20k	1×10^{-4}	ψ	1101
PlaNet	AdamW	1×10^{-3}	32	20k	1×10^{-4}	ψ	1101
DeepONet	AdamW	1×10^{-3}	32	20k	1×10^{-4}	ψ	1101
FNO	AdamW	3×10^{-3}	32	20k	1×10^{-4}	ψ	1101
SetONet	AdamW	3×10^{-3}	32	20k	1×10^{-4}	ψ	1101
GINO	AdamW	3×10^{-3}	32	20k	1×10^{-4}	ψ	1101
GNOT	AdamW	3×10^{-3}	32	20k	1×10^{-4}	ψ	1101

The architecture represents physical measurements and spatial queries as tokens, which are processed through operator transformer blocks to capture long-range dependencies between input measurements and output locations.

Transformer. Transformer introduces geometry-aware attention mechanisms Qin et al. (2023).

In our benchmark, the model represents heterogeneous tokamak information as geometry-associated tokens and applies attention-based interactions before decoding the queried equilibrium quantities.

The architecture provides a comparison between explicit geometric encoding and standard vector-based surrogate models.

F Training and Reproducibility

F.1 Common Training Protocol

To make model comparisons reproducible, all benchmark runs are defined by a task configuration specifying the source and target domains, train/validation/ test splits, input adapter, normalization state, model configuration, optimizer, and random seed.

Whenever shot identities are available, all slices from the same shot are kept within a single split. This prevents temporally adjacent equilibria from appearing simultaneously in training and evaluation sets. For catalogue-style ITER-like data, the split is performed at the equilibrium-row level.

All preprocessing statistics are fitted using the training split only and are then frozen. Validation data may be used for model selection and hyperparameter selection, but the target test split is never used for either purpose.

F.2 Zero-shot Transfer and Target-domain Adaptation

For zero-shot evaluation, a model is trained using only the source-domain training split and evaluated directly on the fixed target test split. Same-device sim-to-real transfer (T3) uses this protocol.

Cross-device tasks (T4 and T5) fine-tune on 5% of the labeled target training split. The target validation and test sets stay fixed. The loss comparison uses training seed 1101.

F.3 Model Selection

Training settings are listed in Table 11. Cross-device tasks fine-tune on 5% of the labeled target training split. The loss comparison uses training seed 1101.

F.4 Random Seeds and Uncertainty

The loss comparison is run at training seed 1101, including in-domain reconstruction, same-device sim-to-real transfer, and cross-device fine-tuning on 5% of the labeled target training data.

F.5 Computational Resources

For every official run, we record the hardware type, number of accelerators, software versions, training time, parameter count, peak memory usage, and single-sample inference latency.

Inference time is measured using **hardware**, with **warm-up iterations** warm-up steps followed by **number** timed repetitions. Synchronization is performed before and after timing when GPU execution is asynchronous.

F.6 Checkpoint and Configuration Release

Each benchmark result is associated with a configuration file containing the dataset release, preprocessing version, split identifier, model hyperparameters, random seed, training objective, and software environment. Official checkpoints and evaluation scripts will be released together with the benchmark code.

F.7 Loss Function

Notation and normalization. For sample i, let ψ_i and $\widehat{\psi}_i$ denote the reference and predicted flux on the native grid $\mathcal{G}_i$, in the canonical flux convention. Write $u_i = T_d(\psi_i)$ and $\widehat{u}_i = T_d(\widehat{\psi}_i)$ for their normalized representations in domain d, and let $e_i = \widehat{u}_i - u_i$. The frozen affine transform is

$$T_d(\psi) = \frac{\psi/c_d - \mu_{\psi,s}}{\sigma_{\psi,s}}, \qquad T_d^{-1}(u) = c_d(\sigma_{\psi,s}u + \mu_{\psi,s}), \tag{21}$$

where $\mu_{\psi,s}$ and $\sigma_{\psi,s}$ are fitted on the source pool prescribed by the task. For the final same-device *sim2real* source model, this pool combines the SIM training and validation splits; the other settings use their source-training split. Within-device tasks use $c_d = 1$, whereas *cross-device* tasks use the fixed physical flux scale $c_d = B_{0,d}a_{0,d}^2$ before statistical normalization, with $B_{0,d}$ and $a_{0,d}$ the registered reference magnetic field and length. The same source statistics are retained during target adaptation. Training uses the prepared canonical gauge; no prediction-dependent offset is fitted inside the loss.

For a nonempty finite support A, define $\langle v \rangle_A = |A|^{-1} \sum_{a \in A} v_a$. All terms use the scalar pseudo-Huber penalty

$$\rho(t) = \sqrt{1 + t^2} - 1, \tag{22}$$

corresponding to a transition parameter of one after normalization.

Field reconstruction. Let Ω_i be the valid native-grid mask. The field loss is

$$\ell_{\mathrm{field},i} = \big\langle \rho(e_i) \big\rangle_{\Omega_i}. \tag{23}$$

First-gradient supervision. Let Ω_i^{∇} contain interior grid nodes whose center and four axial neighbors are all valid. Using the prepared coordinate axes $(\bar{R}, \bar{Z})$, the gradient loss is

$$\ell_{\mathrm{grad},i} = \langle \rho(D_{\bar{R}}e_i) + \rho(D_{\bar{Z}}e_i) \rangle_{\Omega_i^{\nabla}}. \tag{24}$$

Within-device coordinates use the native adapter's fixed affine geometry normalization; cross-device coordinates are $\bar{R} = (R - R_{0,d})/a_{0,d}$ and $\bar{Z} = (Z - Z_{0,d})/a_{0,d}$. Here $(R_{0,d}, Z_{0,d})$ is the registered coordinate origin. For either coordinate x, the nonuniform three-point derivative is

$$D_x f_j = -\frac{h_+}{h_-(h_- + h_+)} f_{j-1} + \frac{h_+ - h_-}{h_- h_+} f_j + \frac{h_-}{h_+(h_- + h_+)} f_{j+1}, \tag{25}$$

where $h_- = x_j - x_{j-1}$ and $h_+ = x_{j+1} - x_j$.

Probe consistency. Let $\mathcal{P}_i$ contain usable magnetic probes with a supported forward operator, and let b_{ij} be the measured magnetic-field component of probe j, in tesla. With device geometry implicit in $\mathcal{H}_{d,i}$,

$$\ell_{\mathrm{probe},i} = \frac{1}{|\mathcal{P}_i|} \sum_{j \in \mathcal{P}_i} \rho\left(\frac{\mathcal{H}_{d,i}(\widehat{\psi}_i, I_i)_j - b_{ij}}{s_{p,d}} \right). \tag{26}$$

Only measured magnetic-field components enter this loss; flux loops are excluded. For EXL-50U and MAST, the operator computes

$$\widehat{B}_R = -\frac{1}{R} D_Z \widehat{\psi}, \qquad \widehat{B}_Z = \frac{1}{R} D_R \widehat{\psi}, \tag{27}$$

on physical native coordinates, bilinearly interpolates these components to the probe location, and projects them onto the recorded probe direction. All four interpolation corners require valid derivative support; the field is not extrapolated outside the grid.

For external ITER-like probes, the operator computes $\widehat{j}_\phi = -\Delta_h^* \widehat{\psi}/(\mu_0 R)$ on supported native stencils and uses axisymmetric Green integration plus the known active-coil fields before directional projection. Reference flux and current density are not substituted into this predicted observation operator.

GS-equation residual. The physical operator is evaluated in meters with flux in Wb/rad:

$$\begin{aligned} \Delta_h^* f &= D_{RR} f - \frac{1}{R} D_R f + D_{ZZ} f, \\ D_{xx} f_j &= \frac{2}{h_- + h_+} \left(\frac{f_{j+1} - f_j}{h_+} - \frac{f_j - f_{j-1}}{h_-} \right). \end{aligned} \tag{28}$$

Let q_i be the supplied physical right-hand side. When a toroidal current-density reference is supplied, the adopted convention gives $q_i(R,Z) = -\mu_0 R j_{\phi,i}^{\mathrm{ref}}(R,Z)$. Here μ_0 is the vacuum permeability. When the profile formulas are available, j_ϕ is a known source. The ITER-like release provides the right-hand side directly; EXL-50U and MAST supply the reconstructed current density after native-grid alignment. The supplied q_i is not inferred from the predicted flux.

Let Ω_i^{GS} contain complete valid five-point stencils with positive radii whose centers lie inside the supplied plasma mask. The equation loss is

$$\ell_{\mathrm{GS},i} = \left\langle \rho\left(\frac{\Delta_h^* \widehat{\psi}_i - q_i}{s_{\mathrm{GS},d}} \right) \right\rangle_{\Omega_i^{\mathrm{GS}}}. \tag{29}$$

The affine offset in Equation 21 vanishes under differentiation; its scale factor is retained. The supplied q_i is fixed, rather than inferred from $\widehat{\psi}_i$. Replacing it by $\Delta_h^* \psi_i$ would define reference-operator matching, a different objective.

Physical-loss scales. In the controlled loss comparison, probe and equation residuals use separate fixed scales. From a common deterministic subset $\mathcal{C}$ of at most 256 source-training records, independent of model, loss combination and training seed, we compute

$$\begin{aligned} s_{p,s}^2 &= \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \frac{1}{|\mathcal{P}_i^{\mathrm{cal}}|} \sum_{j \in \mathcal{P}_i^{\mathrm{cal}}} b_{ij}^2, \\ s_{\mathrm{GS},s}^2 &= \frac{1}{|\mathcal{C}_{\mathrm{GS}}|} \sum_{i \in \mathcal{C}_{\mathrm{GS}}} \langle q_i^2 \rangle_{\Omega_i^{\mathrm{GS}}}, \end{aligned} \tag{30}$$

where $\mathcal{C}_{\mathrm{GS}}$ contains calibration records with available GS references. These are uncentered root-mean-square reference scales with equal weight per record. The calibration subset is selected by a fixed hash ordering of sample identifiers. By default, $\mathcal{P}_i^{\mathrm{cal}} = \mathcal{P}_i$. An explicit measurements-only calibration instead uses all valid source magnetic probes, without enabling an unsupported probe-loss run.

Within-device comparisons retain the source scales. Cross-device comparisons use $s_{k,d} = s_{k,s} B_{0,d}/B_{0,s}$ for $k \in \{p, \mathrm{GS}\}$, with the registered physical reference fields. No target, validation or test observations are used to refit these scales. The component weights therefore act on normalized discrepancies.

Aggregation and missing references. All four per-record losses use uniform averages over their valid nodes or probes. Let $\mathcal{I}_g$ index records in group g and let $n_g = |\mathcal{I}_g|$. Groups are the shot or

state groups defined by the data split. For component k, the training-set objective is

$$\mathcal{L}_k = \frac{1}{G} \sum_{g=1}^{G} \frac{1}{n_g} \sum_{i \in \mathcal{I}_g} \ell_{k,i}. \tag{31}$$

For a batch of size B from N permitted records, we average $w_i \ell_{k,i}$ with $w_i = N/(G n_{g(i)})$ over the full batch. Explicitly unavailable MAST auxiliary GS references remove only the corresponding GS contributions; other losses, weights and the denominator are unchanged, and coverage is reported. Evaluation metrics use their separately defined gauge alignment, area weighting and relative normalization, rather than these training penalties.

G Additional Results

G.1 Main Results

Cross-device tasks T4 and T5 fine-tune on 5% of the labeled target training data. Entries in this table use training seed 1101. Field error, geometry, and physics consistency are reported together.

Table 12: **Complete results for the controlled training-loss comparison.** Within each transfer scenario and metric, darker colors indicate lower prediction errors; boldface marks the lowest available value, not a selected loss recipe. An asterisk (*) marks a conditional mean over successful geometric extractions. Entries use training seed 1101 and the frozen test sets. Cross-device settings (T4 and T5) fine-tune on 5% of the labeled target training data. Flux RMSE is in Wb/rad; the main text reports the same quantity in mWb/rad.

Transfer Scenarios	Model	Loss	Field Reconstruction ↓			Geometry ↓			Reference GS Operator ↓
			E_ψ^{RMSE}(Wb/rad)	E_ψ^{rel}	$E_{\nabla\psi}^{\mathrm{rel}}$	E_{axis}(cm)	E_{LCFS}^{95}(cm)	$E_{\mathrm{X-pt}}$(cm)	E_{op}
$\mathbb{T}_1$ $S_{\mathrm{ITER}} \to S_{\mathrm{ITER}}$ *sim2sim*	U-Net	$\mathcal{L}_D$	0.04423	0.01217	**0.0221**	1.099	**2.878***	4.377*	0.1345
	U-Net	$\mathcal{L}_D + \lambda_{\mathrm{probe}}\mathcal{L}_{\mathrm{probe}}$	0.0464	0.01266	0.02239	1.096	3.184*	4.273*	0.1335
	U-Net	$\mathcal{L}_D + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}}$	0.04906	0.01352	0.02294	1.067	3.232*	4.114	0.08392
	U-Net	$\mathcal{L}_F$	0.05122	0.01402	0.02344	**1.059**	3.190*	**3.619***	0.08362
	FNO	$\mathcal{L}_D$	0.042	0.01212	0.02302	1.213	3.085*	4.737*	0.1477
	FNO	$\mathcal{L}_D + \lambda_{\mathrm{probe}}\mathcal{L}_{\mathrm{probe}}$	0.04262	0.0123	0.02325	1.267	3.174*	5.568*	0.1466
	FNO	$\mathcal{L}_D + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}}$	0.0413	0.0119	0.02259	1.088	2.966*	5.635*	0.1052
	FNO	$\mathcal{L}_F$	**0.04104**	**0.01173**	0.02244	1.143	3.090*	5.102*	0.1025
	GNOT	$\mathcal{L}_D$	0.05829	0.01628	0.02519	1.351	4.279*	10.05*	0.07523
	GNOT	$\mathcal{L}_D + \lambda_{\mathrm{probe}}\mathcal{L}_{\mathrm{probe}}$	0.05511	0.01572	0.0257	1.412	3.548*	5.264*	0.07583
	GNOT	$\mathcal{L}_D + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}}$	0.06589	0.01762	0.02699	1.387	3.963*	10.17*	0.0757
	GNOT	$\mathcal{L}_F$	0.05688	0.01568	0.02537	1.249	3.436*	5.879*	**0.07355**
$\mathbb{T}_2$ $S_{\mathrm{EXL}} \to S_{\mathrm{EXL}}$ *sim2sim*	U-Net	$\mathcal{L}_D$	0.0002212	**0.01073**	**0.01285**	0.2128	0.9589*	0.2380	0.2563
	U-Net	$\mathcal{L}_D + \lambda_{\mathrm{probe}}\mathcal{L}_{\mathrm{probe}}$	0.0002244	0.0109	0.01316	0.2161	1.004*	0.2339	0.2581
	U-Net	$\mathcal{L}_D + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}}$	0.000223	0.01082	0.01309	0.2159	0.9694*	0.2407	0.2302
	U-Net	$\mathcal{L}_F$	0.0002245	0.0109	0.01301	**0.2127**	0.9652*	0.2308	0.2316
	FNO	$\mathcal{L}_D$	0.0002225	0.01086	0.01286	0.2246	0.9165*	**0.2299**	0.2041
	FNO	$\mathcal{L}_D + \lambda_{\mathrm{probe}}\mathcal{L}_{\mathrm{probe}}$	**0.000221**	0.0108	0.01289	0.2229	**0.9161***	0.2627	0.2069
	FNO	$\mathcal{L}_D + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}}$	0.0002317	0.01131	0.01319	0.231	0.9794*	0.2359	0.194
	FNO	$\mathcal{L}_F$	0.0002365	0.01152	0.01349	0.2312	0.9972*	0.2458	0.1974
	GNOT	$\mathcal{L}_D$	0.000266	0.01317	0.01559	0.3231	1.064*	0.3309	0.1736
	GNOT	$\mathcal{L}_D + \lambda_{\mathrm{probe}}\mathcal{L}_{\mathrm{probe}}$	0.0002766	0.01387	0.01571	0.3289	1.123*	0.3140	0.1739
	GNOT	$\mathcal{L}_D + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}}$	0.0002877	0.01419	0.01672	0.3191	1.165*	0.3725	0.1765
	GNOT	$\mathcal{L}_F$	0.0002509	0.01246	0.0147	0.2693	0.9689*	0.2687	**0.1686**
$\mathbb{T}_3$ $S_{\mathrm{EXL}} \to R_{\mathrm{EXL}}$ *sim2real*	U-Net	$\mathcal{L}_D$	0.001897	0.1025	**0.1188**	**2.409**	6.616*	2.410	0.5604
	U-Net	$\mathcal{L}_D + \lambda_{\mathrm{probe}}\mathcal{L}_{\mathrm{probe}}$	0.001908	0.1028	0.1196	2.508	6.551*	2.564	0.5725
	U-Net	$\mathcal{L}_D + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}}$	**0.001857**	**0.1003**	0.1221	2.519	**6.498***	2.261	0.4868
	U-Net	$\mathcal{L}_F$	0.001879	0.1016	0.1193	2.624	6.534*	2.311	0.4823
	FNO	$\mathcal{L}_D$	0.002425	0.1347	0.1716	3.857	7.376*	2.352*	1.069
	FNO	$\mathcal{L}_D + \lambda_{\mathrm{probe}}\mathcal{L}_{\mathrm{probe}}$	0.002404	0.1331	0.1726	4.132	7.365*	2.357	1.059
	FNO	$\mathcal{L}_D + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}}$	0.002366	0.1307	0.1689	4.244	7.248*	2.210	0.8924
	FNO	$\mathcal{L}_F$	0.002334	0.1289	0.1664	3.929	7.393*	2.132*	0.8917
	GNOT	$\mathcal{L}_D$	0.003204	0.181	0.1941	5.038	8.621*	**1.754**	0.4105
	GNOT	$\mathcal{L}_D + \lambda_{\mathrm{probe}}\mathcal{L}_{\mathrm{probe}}$	0.003022	0.171	0.1853	4.791	8.894*	1.938	0.415
	GNOT	$\mathcal{L}_D + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}}$	0.003119	0.1768	0.1924	4.804	8.770*	1.799	0.4121
	GNOT	$\mathcal{L}_F$	0.003074	0.1743	0.1891	4.678	8.503*	1.854	**0.4066**
$\mathbb{T}_4$ $S_{\mathrm{ITER}} \to S_{\mathrm{EXL}}$ *cross-device* (5% target adaptation)	U-Net	$\mathcal{L}_D$	0.001914	0.09323	0.09453	2.357	5.979*	2.750	2.217
	U-Net	$\mathcal{L}_D + \lambda_{\mathrm{probe}}\mathcal{L}_{\mathrm{probe}}$	0.001908	0.09275	0.09322	3.077	5.656*	3.180*	2.458
	U-Net	$\mathcal{L}_D + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}}$	0.001923	0.09302	0.09441	2.303	5.813*	2.657	1.535
	U-Net	$\mathcal{L}_F$	0.001695	0.08204	**0.08695**	2.255	5.756*	2.413	1.463
	FNO	$\mathcal{L}_D$	0.003907	0.1865	0.1659	3.487*	8.171*	4.771*	1.644
	FNO	$\mathcal{L}_D + \lambda_{\mathrm{probe}}\mathcal{L}_{\mathrm{probe}}$	0.004033	0.1927	0.1751	3.674	8.538*	5.067*	1.83
	FNO	$\mathcal{L}_D + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}}$	0.00425	0.2043	0.1776	3.509	7.971*	4.800*	1.029
	FNO	$\mathcal{L}_F$	0.003996	0.1912	0.1662	3.576	7.827*	4.383*	1.242
	GNOT	$\mathcal{L}_D$	**0.001553**	**0.07529**	0.09707	2.225	5.233*	1.804	**0.3575**
	GNOT	$\mathcal{L}_D + \lambda_{\mathrm{probe}}\mathcal{L}_{\mathrm{probe}}$	0.001602	0.0779	0.1025	**2.162**	5.142	1.963*	0.3586
	GNOT	$\mathcal{L}_D + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}}$	0.001658	0.07994	0.1068	2.21	**5.110**	**1.739**	0.3699
	GNOT	$\mathcal{L}_F$	0.001591	0.07661	0.1026	2.236	5.146*	1.880	0.3599
$\mathbb{T}_5$ $R_{\mathrm{MAST}} \to R_{\mathrm{EXL}}$ *cross-device* (5% target adaptation)	U-Net	$\mathcal{L}_D$	**0.0008519**	**0.04499**	**0.06269**	1.431	**3.764***	1.029	0.618
	U-Net	$\mathcal{L}_D + \lambda_{\mathrm{probe}}\mathcal{L}_{\mathrm{probe}}$	0.0008722	0.04594	0.08285	**1.394**	3.911*	**0.9770**	1.759
	U-Net	$\mathcal{L}_D + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}}$	0.0008976	0.0473	0.06456	1.407	3.911*	1.003	0.5094
	U-Net	$\mathcal{L}_F$	0.0009352	0.04914	0.08327	1.486	4.035*	1.078*	1.778
	FNO	$\mathcal{L}_D$	0.00125	0.06704	0.08994	2.202	5.342*	1.596*	0.6968
	FNO	$\mathcal{L}_D + \lambda_{\mathrm{probe}}\mathcal{L}_{\mathrm{probe}}$	0.001299	0.06938	0.1065	2.751	5.338*	1.760*	2.108
	FNO	$\mathcal{L}_D + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}}$	0.001252	0.06707	0.08896	2.02	5.332*	1.800*	0.6084
	FNO	$\mathcal{L}_F$	0.001246	0.06661	0.1037	2.124	5.222*	1.769*	1.971
	GNOT	$\mathcal{L}_D$	0.001249	0.06626	0.08911	2.136	4.853*	1.693	0.3956
	GNOT	$\mathcal{L}_D + \lambda_{\mathrm{probe}}\mathcal{L}_{\mathrm{probe}}$	0.001237	0.06558	0.1132	1.767	4.193*	1.390	1.339
	GNOT	$\mathcal{L}_D + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}}$	0.001169	0.0619	0.08675	1.7	4.230*	1.480*	**0.3821**
	GNOT	$\mathcal{L}_F$	0.001282	0.06833	0.103	2.062	4.640*	1.479	0.4933

G.2 OOD

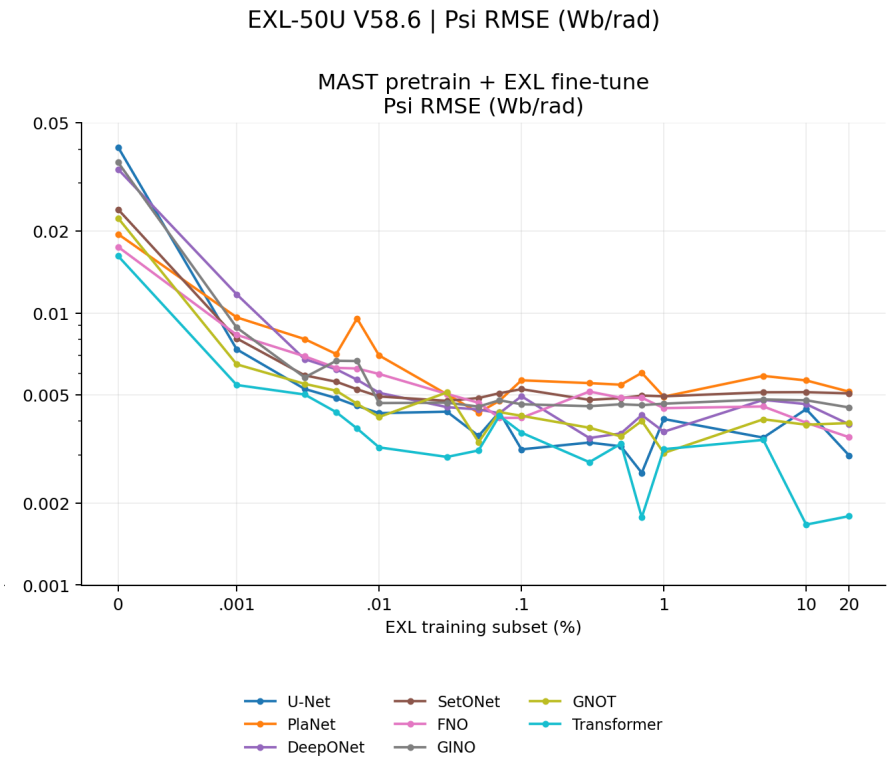

Figure 9: Cross-device fine-tuning on 5% of the labeled target training data.

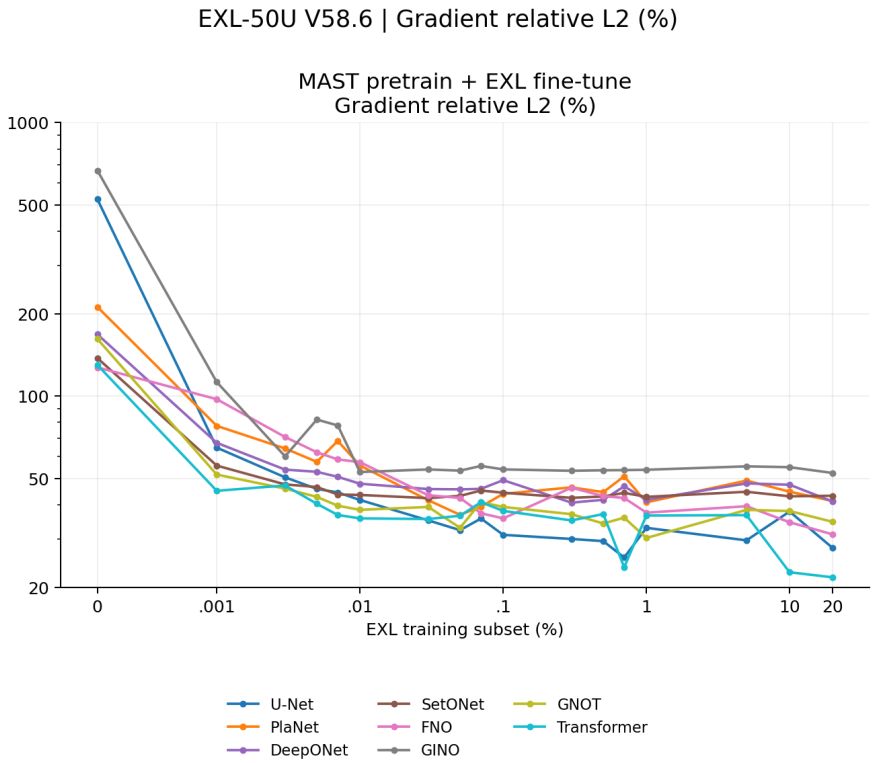

Figure 10: Cross-device fine-tuning on 5% of the labeled target training data. Relative gradient error.

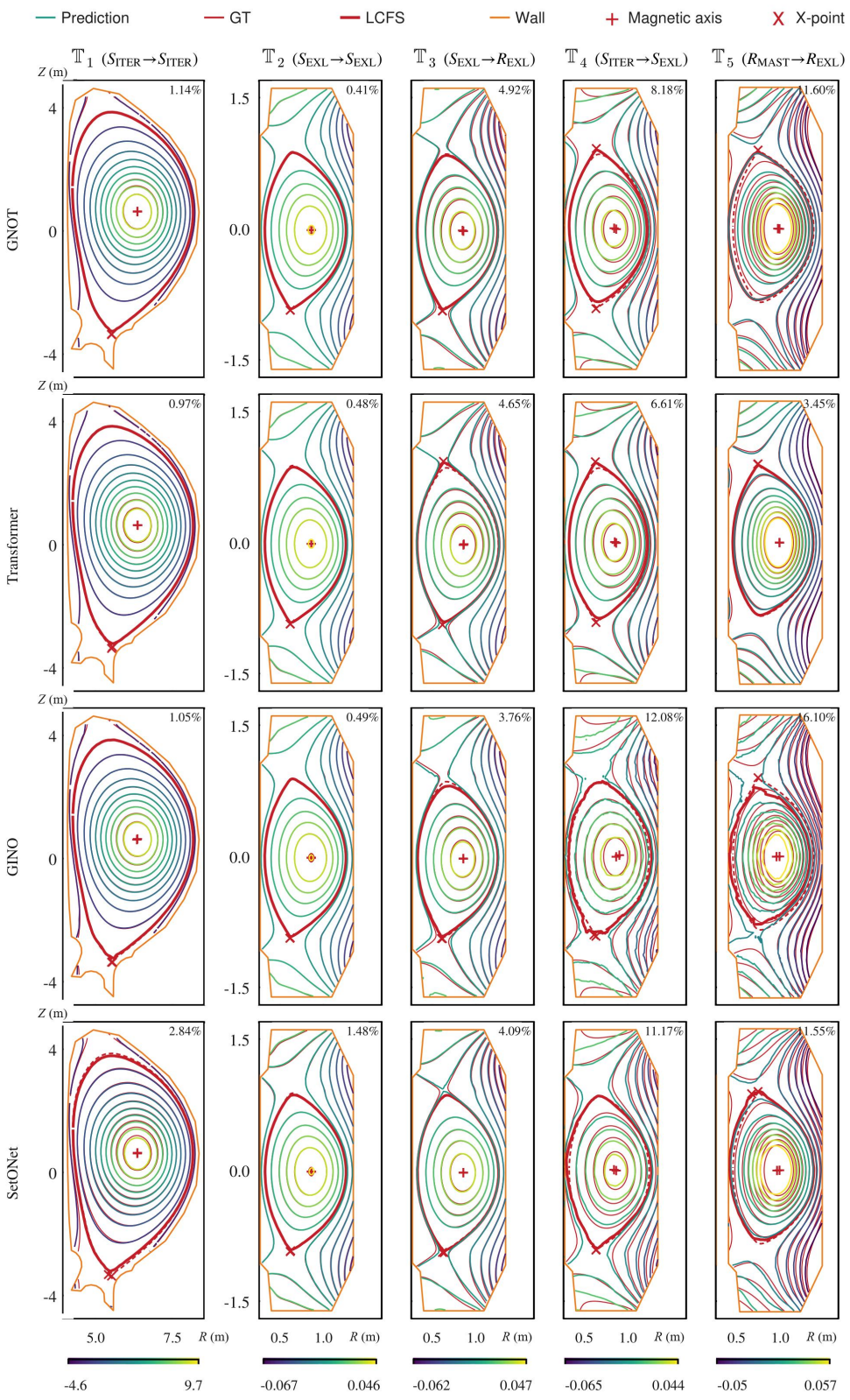

Figure 11: Additional representative profiles across tasks and architectures. Cross-device panels use 5% target-domain fine-tuning.

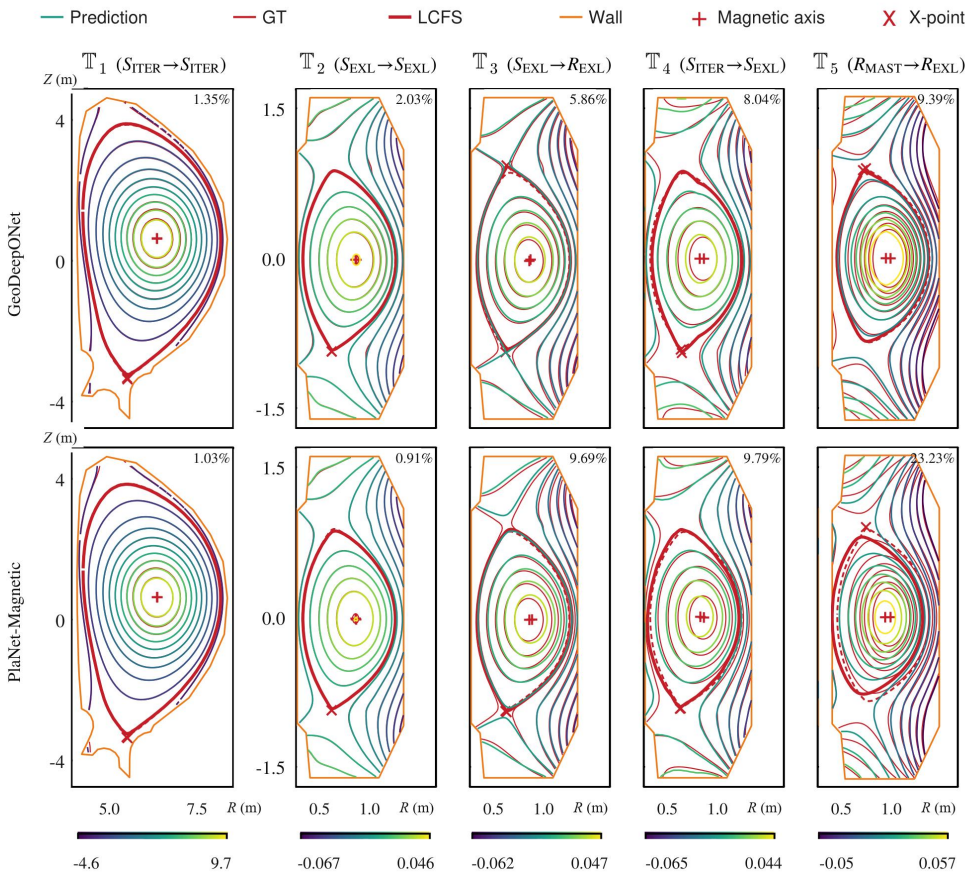

Figure 12: Additional representative equilibrium profiles across tasks and architectures. Cross-device panels use 5% target-domain fine-tuning.