TOKAMACON: AN EXPERIMENT-DERIVED PARALLEL ENVIRONMENT SUITE FOR TOKAMAK PLASMA CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-time tokamak plasma control is a prerequisite for commercial magnetic-confinement fusion. Operating high-performance tokamak plasmas constitutes an extreme control challenge: controllers rely on sparsely indirect diagnostics, face distribution shift both within and across plasma discharges, adapt to multi-time-scale transport coupling and the effects of magnetohydrodynamic (MHD) instabilities, and obey hard safety limits where violations can terminate the discharge or damage hardware. Yet standard reinforcement-learning environments rarely expose this fusion-specific combination as a simultaneous, interacting control problem, while fusion-control studies lack reusable scenarios and metrics for studying it reproducibly. We introduce TokamaCon, which decouples tokamak real-time control into two representative environment slices, both grounded in EAST data and checked against reference simulators, thereby enabling the development and evaluation of reinforcement-learning algorithms for fusion-control problems. `IsoFlux` regulates magnetic equilibrium at 1 kHz on a high-frequency switching plant with input delay; `Profile` regulates kinetic profiles at 100 Hz on a stiff 1D transport PDE with instability-driven events. Both environments provide native PyTorch GPU-vectorised rollouts, reaching roughly 10^3 `IsoFlux` and 10^3 `Profile` environment steps per second at the recommended batch sizes on a single GPU. We release standardised scenarios, per-episode metrics, provenance metadata, and reference controllers for reproducible comparison of control approaches. Code and scenarios are released at `https://anonymous.4open.science/r/TokamaCon_anonymous-CED5`.

1 INTRODUCTION

Controlled nuclear fusion promises dense, carbon-free power, and tokamaks remain one of the leading routes toward that goal (Campbell et al., 2025). A tokamak confines a reactor-relevant plasma with magnetic fields (Degrave et al., 2022; Gribov et al., 2007). Real-time operation involves two coupled control problems on different time scales (Moreau et al., 2008). The two loops drive distinct actuators against a single, partially observed plasma state. This coupling makes them difficult to design or evaluate in isolation, even when they run on separate hardware. This split is illustrated in Figure 1, which shows the two control tasks alongside an illustrative EAST[1] equilibrium (Li et al., 2013).

On the faster time scale, magnetic control regulates plasma current and the boundary geometry of the discharge. On the slower time scale, profile control regulates how heating, current drive, and density evolution reshape the radial profiles that set both performance and stability margin (Moreau et al., 2008; Felici et al., 2011). Both control problems must steer the plasma along an operator-specified trajectory while keeping it inside hard operational limits, and the two loops fail in different ways. A tracking failure or disruption in magnetic control ends the discharge before any profile objective can be delivered, so this loop underwrites every other objective in the shot (De Tommasi, 2019).

[1]EAST (Experimental Advanced Superconducting Tokamak) is a fully superconducting tokamak operated by the Institute of Plasma Physics, Chinese Academy of Sciences. It was designed to investigate advanced steady-state tokamak operation and to provide key physics and engineering experience for future long-pulse fusion devices such as ITER and CFETR (Gong et al., 2024).

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

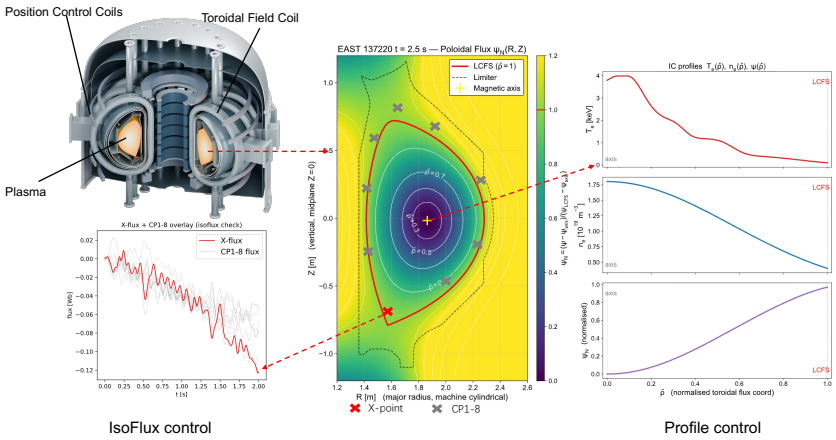


Figure 1: **Overview of TokamaCon.** Two independent, complementary tokamak-control environments with an EAST equilibrium illustration. *Left:* `IsoFlux` magnetic-loop coil geometry and a closed-loop isoflux control-point trajectory. *Centre:* an illustrative EAST shot 137220 poloidal-flux map $\psi_N(R, Z)$ at $t = 2.5$ s, not a shared equilibrium for all experiments. *Right:* `Profile` initial-condition T_e, n_e, ψ_N profiles on the normalised toroidal-flux coordinate $\hat{\rho}$.

Profile control instead determines the achievable performance of the discharge, because the same profiles that support confinement and stored energy also gate large-scale instabilities such as the neoclassical tearing mode (NTM) (La Haye, 2006), and the problem becomes more critical still on next-generation burning-plasma devices where alpha self-heating couples the profile back to itself (Shimada et al., 2007). Both control problems are high-dimensional, actuator-coupled, partially observed, and multi-objective in ways that no single tracking signal captures. These properties make real-time learned feedback control an open research target on both time scales rather than a settled engineering problem.

Conventional tokamak control is often built around local models and reference scenarios (Gribov et al., 2007). Magnetic controllers use reconstructed equilibria and local perturbed equilibrium response model to compute nominal coil trajectories, then use PID feedback to eliminate deviations in plasma current, position, and shape (Degrave et al., 2022). Model-based profile-control studies use transport or profile-response models to plan reference profiles and heating or current-drive schedules, then use feedback to correct tracking errors (Felici et al., 2011; Kim & Lister, 2012). However, this workflow is hard to reuse once the operational target changes. A new shape, profile, or stability margin can demand a fresh reference trajectory and manual re-tuning (Seo et al., 2022). Reference-tracking decomposition also struggles with multi-objective trade-offs across shape, pressure, and stability margin (Seo et al., 2024), since each loop tracks a single deviation signal. And linearised or local component models accumulate error over the longer profile-control horizon (Wang et al., 2025). Reinforcement learning can instead optimize policies directly for these coupled long-horizon objectives.

Recent work argues that learned controllers can enable several tokamak-control capabilities. These include reducing manual controller design effort for new plasma configurations (Degrave et al., 2022; Tracey et al., 2024), learning from historical shots before experimental deployment (Char et al., 2023), designing operation trajectories for multiple plasma targets (Seo et al., 2022), coordinating coupled pressure and safety constraints (Wakatsuki et al., 2023), and guiding operation away from disruptive instability boundaries (Seo et al., 2024). These results show the potential of learned tokamak control, but the studies are difficult to compare directly (Martínez-Plumed et al., 2021). Each uses its own simulator, observation model, actuator interface, reward, and evaluation protocol. As a result, performance can be affected by environment design as much as controller quality (Mnih et al., 2015). A reusable benchmark therefore needs a fixed environment interface before policy comparisons become meaningful (Ott et al., 2022). It must specify scenarios, observations, actions, rewards, baselines, and rollout interfaces (Reuer et al., 2023). The closest existing artefact is Gym-TORAX (Mouchamps et al., 2026), a Gymnasium wrapper around TORAX with a single ITER ramp-up profile-control scenario. It addresses only the slow profile-control loop and ships one scenario per environment, **leaving the fast magnetic-control loop, a multi-scenario reproducible-comparison protocol, and a batched-rollout contract for large-scale RL training unaddressed.**

We introduce TokamaCon, an EAST-grounded environment suite that addresses these limitations. It provides the fast magnetic-equilibrium loop and the slow kinetic-profile loop as two independent, complementary GPU-vectorized PyTorch environments in a multi-scenario benchmark with EAST-derived configurations. Scenario configurations and shot-specific plant bundles are separated from the reusable solvers and rollout interfaces (Appendix C.11). `IsoFlux` covers magnetic shape control using an EAST-grounded linearized plasma-control-system model. Policies regulate plasma current and isoflux boundary targets at kHz control rate. `Profile` covers kinetic profile control using a RAPTOR-style one-dimensional transport model. Policies act on the electron-temperature profile through electron-cyclotron heating and current drive to mitigate seeded NTM dynamics at approximately 100 Hz. Both environments are natively implemented in PyTorch and follow the Isaac Gym vectorised-environment contract for batched GPU rollout. Each environment also ships with standardized scenarios, reward menus, and partial-observation sensor interfaces. This environment suite provides a ready-to-use platform for developing and evaluating learned feedback controllers in fusion-relevant settings.

Concretely, we contribute the following: (1) We formalise tokamak magnetic shape control and kinetic profile control as reinforcement-learning environments with explicit observation, action, reward, and termination interfaces. (2) We release two EAST-grounded PyTorch environments, `IsoFlux` and `Profile`, both with batched GPU rollout and partial-observation sensor interfaces. (3) We provide standardized scenarios for isoflux tracking and NTM mitigation, together with reference baselines, evaluation protocols, and provenance documentation that support reproducible comparison.

2 BACKGROUND

2.1 TOKAMAKS AND THE TWO-TIME-SCALE CONTROL ARCHITECTURE

A tokamak is an axisymmetric magnetic-confinement device whose macroscopic state during a discharge is summarised by the toroidal plasma current I_p, the magnetic equilibrium that fixes the plasma boundary and internal flux-surface geometry, and the kinetic profiles of electron and ion temperature, density, and current density across those surfaces (Wesson & Campbell, 2011). These quantities drift on widely separated timescales: ideal magnetohydrodynamic (MHD) instabilities can grow on Alfvénic timescales of order microseconds, resistive MHD modes grow on intermediate $10\,\mu$s–10 ms timescales, the energy-confinement time on present-day devices is ~ 100 ms, and resistive diffusion of the current profile takes seconds. Throughout a discharge the plasma must also stay inside several operational limits, including the Greenwald density limit (Greenwald, 2002), the Troyon β limit (Troyon et al., 1984), the q-edge kink limit, and the vertical-stability margin, and it crosses those limits at multiple points during the ramp-up, flat-top, and ramp-down phases, so closed-loop real-time feedback is active throughout the shot to keep the plasma on a target trajectory and inside the operational envelope.

Because the relevant control timescales span more than four orders of magnitude, real-time tokamak control is conventionally separated into two cascaded loops (Moreau et al., 2008; De Tommasi, 2019). The fast *magnetic loop* runs at ~ 1 kHz and regulates the plasma current and the equilibrium boundary shape through the voltages applied to the poloidal-field (PF) and central-solenoid (CS) coil sets; its setpoints are typically a target I_p trajectory together with a discrete collection of boundary control points such as poloidal flux at chosen sensor locations, X-point position, and the gap between the last closed flux surface and the wall, plus an inner vertical-stabilisation loop that suppresses the unstable vertical-displacement mode. The slower *profile loop* runs at ~ 100 Hz and regulates kinetic-profile shape and MHD-stability state through auxiliary heating and current-drive actuators such as electron-cyclotron, neutral-beam, and ion-cyclotron systems; its setpoints typically include the electron-temperature profile, the safety-factor q profile, and the active suppression of neoclassical-tearing modes. TokamaCon provides one environment per loop, instantiated on EAST-derived configurations.

2.2 PLASMA MODELS BEHIND THE TWO ENVIRONMENTS

The `IsoFlux` environment uses the linearised free-boundary response of the tokamak around an EFIT-reconstructed equilibrium (Lao et al., 1985), i.e. a Grad–Shafranov solution (Freidberg, 2014)

that fits the available magnetic measurements on a chosen reference shot. The linearised plant takes the form

$$\dot{x} = Ax + Bu, \qquad y = Cx, \tag{1}$$

where x is the state deviation, u the voltage controls, y the sensor outputs, and A, B, C are obtained from the EFIT linearisation.

The `Profile` environment evolves the coupled electron temperature $T_e(\hat{\rho}, t)$ and the poloidal flux $\psi(\hat{\rho}, t)$ on a normalised toroidal-flux coordinate $\hat{\rho} \in [0, 1]$ under a one-dimensional RAPTOR-style transport PDE (Felici et al., 2011). Following standard transport modelling, the temperature equation in compact form reads

$$\frac{3}{2} n_e \frac{\partial T_e}{\partial t} = \frac{\partial}{\partial \hat{\rho}} \left(n_e \chi_e \frac{\partial T_e}{\partial \hat{\rho}} \right) + S_{\mathrm{tot}}, \tag{2}$$

where n_e is the electron density, χ_e is the electron-heat diffusivity provided by a Bohm-gyro-Bohm (Erba et al., 1997) or Felici-style (Felici et al., 2011) turbulent-transport closure, and S_{tot} collects ohmic heating, the electron-cyclotron heating injected by the policy, and additional source and sink terms.

The neoclassical-tearing-mode (NTM) scenario adds a reduced magnetic-island-width equation in the modified Rutherford form (Buttery et al., 2000; Sauter et al., 1997), which models how a $(m, n) = (2, 1)$ magnetic island seeded at the $q = 2$ rational surface evolves under three competing drives:

$$\frac{dw}{dt} \propto \Delta' + \frac{a_1 J_{\mathrm{bs}}}{w^2 + w_d^2} - \frac{a_2 J_{\mathrm{cd}}}{w^2 + w_{\mathrm{marg}}^2}. \tag{3}$$

Here w is the island width and m, n are the poloidal and toroidal mode numbers; Δ' is the classical tearing index set by the equilibrium current profile; the bootstrap-current term $a_1 J_{\mathrm{bs}}/(w^2 + w_d^2)$ destabilises the island once it grows above the threshold w_d; the externally driven current term $a_2 J_{\mathrm{cd}}/(w^2 + w_{\mathrm{marg}}^2)$ is the electron-cyclotron current drive that the policy can deposit at the rational surface to stabilise it. In this reduced model, once the island grows, parallel transport along its perturbed flux surfaces short-circuits the local heat-diffusivity χ_e, which flattens the T_e profile across the island and is therefore visible in the observation channel. This NTM model has not been experimentally validated against island-width dynamics.

3 Environments

TokamaCon contains two independently instantiated environments with EAST-derived configurations, one per control task: `IsoFlux` (Section 3.1) targets the fast magnetic loop and `Profile` (Section 3.2) targets the slow profile loop. The tasks are independent but complementary, without online state exchange or coupled multi-rate closed-loop simulation (Appendix C.2). They share a vectorised PyTorch interface that runs B environments in parallel on the GPU, with all per-step state, observations, actions, and rewards carried as tensors batched along a leading B dimension and no per-step CPU round-trip. `IsoFlux`'s tracking failure is formalised in Section 3.1 and `Profile`'s locked-mode disruption in Section 3.2, both alongside their environment definitions.

Table 1: Scenarios released in TokamaCon.

Scenario id	Environment	Initial condition	Reward kind
`isoflux_eq_track`	`IsoFlux`	EFIT linearisation	flux, B_x, I_p combiner
`ntm_w_energy`	`Profile`	2/1 island $w_0 = 0.15\,\mathrm{m}$	island width with W_{th} boost

3.1 IsoFlux: magnetic-equilibrium control

Physics. `IsoFlux` targets the fast control problem of regulating the toroidal plasma current I_p and the boundary geometry of the equilibrium (De Tommasi, 2019). The plant is the linearised free-boundary plasma response of an EAST equilibrium reconstructed via EFIT (Lao et al., 1985; Albanese & Villone, 1998); one outer step covers 1 ms and a default episode covers 2 s of physical time. Headline results use shot 152444; an additional shot-137220 instance (IsoFlux137220-A) and its held-out-equilibrium evaluation are reported in Appendix E.1. An internal bang-bang vertical-stabilisation loop runs inside the env around the equilibrium's unstable vertical mode (Humphreys et al., 2009) and is invisible to the policy, so the plant exposed to RL is vertically stable.

Action. In RL terms the environment exposes a continuous action $a \in \mathbb{R}^{12}$ of normalised power-supply voltage commands routed to the twelve poloidal-field and central-solenoid circuits, after per-channel saturation and a 5 ms transport delay (Ambrosino & Albanese, 2005). Two profile-perturbation inputs ($\Delta\beta_p$, Δl_i) carried by the underlying simulator are held as fixed disturbances and are not part of the policy interface.

Observation. The privileged plant state $y \in \mathbb{R}^{188}$ available to the critic groups into ten boundary control-point fluxes, the X-point flux and two X-point poloidal-field components, the X-point and current-centroid displacements, twelve coil currents, the fast vertical-stabilisation coil current, 158 vacuum-vessel passive eddy-current modes, and I_p. The actor observes a deployable subset $o \in \mathbb{R}^{12}$ that retains eight isoflux control-point fluxes, the X-point flux and field, and I_p, mirroring the diagnostic signal that an EAST real-time controller would consume (Yuan et al., 2013).[2]

Reward. The reward is a $[0, 1]$-normalised weighted average of I_p, control-point-flux, and X-point-field tracking errors, smoothed across the three components by a tunable mixing coefficient.

Failure mode. The terminal failure mode for `IsoFlux` is a tracking violation, in which I_p or a boundary control point drifts outside the operator-defined tolerance band around its setpoint while no actuator hits a hardware limit. Read in standard reinforcement-learning vocabulary, this is a violation of an ε-tube constraint around a reference trajectory under bounded actions. Tracking failure is operationalised as a per-episode metric in the benchmark protocol of Section 4.1, with scenario-specific thresholds defined inside the environment. Figure 2 illustrates this mode by comparing open-loop baselines against the learned RL controllers of Section 4 on a single `IsoFlux` isoflux-tracking rollout. Tight co-drift between the eight boundary control-point fluxes and the X-point flux is the empirical signature of a closed isoflux loop. The zero-action baseline loses this signature within the first few hundred milliseconds as the control-point fluxes fan out, activating the tracking-failure mode. The Gaussian-random baseline drives the fluxes incoherently and triggers early episode termination, exhibiting the actuator-noise-driven loss of containment that any unmatched open-loop schedule risks. The learned controllers recover tight co-drift along the target trajectory, and Section 4 quantifies the gap between learned baselines and the open-loop floor.

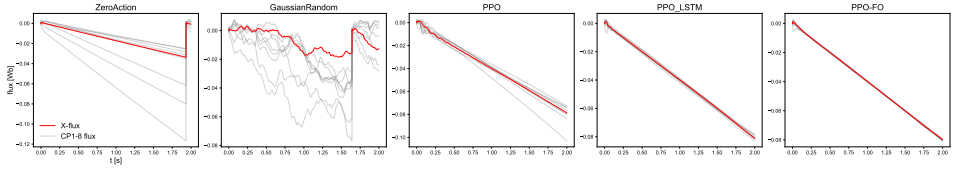


Figure 2: `IsoFlux` isoflux-tracking rollouts on EAST shot 152444 under five controllers.

3.2 PROFILE: KINETIC-PROFILE CONTROL

Physics. `Profile` targets the slower control problem of mitigating the neoclassical-tearing-mode dynamics introduced in Section 2.2 while sustaining the stored thermal energy of the discharge, with electron-cyclotron heating and current drive acting on the electron-temperature profile $T_e(\hat{\rho})$ and the current-density profile that sets the safety-factor profile $q(\hat{\rho})$. The plant is the one-dimensional transport solver of Section 2.2; one outer step covers 10 ms and a default episode covers 4 s of physical time. The initial (T_e, ψ, n_e) profiles are taken from EAST shot 137220, and n_e is held frozen at this initial condition throughout an episode and enters the dynamics only as a coefficient.

Action. The action $a \in \mathbb{R}^{12}$ specifies the deposition geometry and power of $K = 4$ electron-cyclotron beams (Prater, 2004), with three normalised channels per beam, the deposition centre r_{dep} on $\hat{\rho}$, the deposition width w_{dep}, and the injected power P_{in}. Each channel decodes through a hard physical bound that mirrors the EAST gyrotron envelope, with $r_{\mathrm{dep}} \in [0, 0.85]$, $w_{\mathrm{dep}} \in [0.15, 0.25]$, and $P_{\mathrm{in}} \in [0.3, 0.8]$ MW per beam, capping the total injected electron-cyclotron power at 3.2 MW. This 3.2 MW cap is the binding constraint for NTM suppression because it sets the maximum driven current J_{cd} available to overcome the destabilising bootstrap term in Equation 3, so the policy must

[2] An alternative 24-dimensional observation appends the twelve coil-current measurements; we use the 12-d default for headline results.

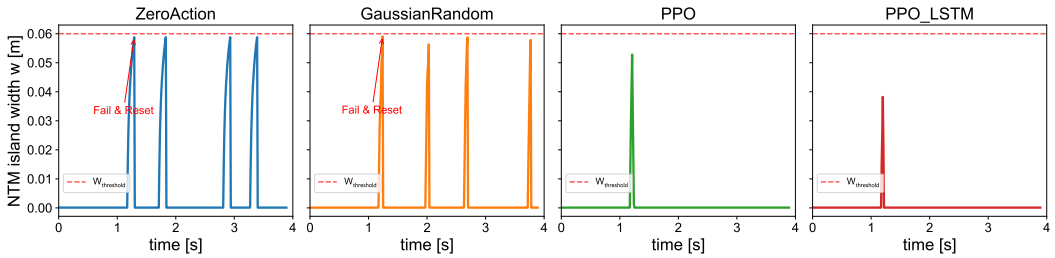


Figure 3: `Profile` NTM-suppression rollouts on the $2/1$ tearing-mode scenario under four controllers, showing the magnetic-island width $w(t)$ over a $4\,\mathrm{s}$ episode against the disruption threshold $W_{\mathrm{threshold}}$ (red dashed).

allocate a finite power budget across the four beams and choose where to deposit each one rather than pumping arbitrary stabilising current at the rational surface. Action clipping, the affine decoding map, and the per-step action-noise model are deferred to Appendix C.6.

Observation. The privileged state $s \in \mathbb{R}^{3N_\rho}$ is the discretised triplet (ψ, T_e, n_e) on a radial grid of $N_\rho = 64$ points. The actor observation $o \in \mathbb{R}^{N_{\mathrm{ECE}}}$ with $N_{\mathrm{ECE}} = 12$ is a synthetic electron-cyclotron-emission (ECE) readout that samples T_e at twelve fixed $\hat{\rho}$ positions calibrated to the EAST ECE diagnostic geometry on shot 137220, reflecting the sparse, partial nature of profile measurements available in real time. No channel sits at $\hat{\rho} = 0$, because the second-harmonic ECE band overlaps third-harmonic emission from the low-field side near the magnetic axis, leaving the on-axis T_e physically unmeasurable and forcing the policy to infer core dynamics from off-axis channels.

Reward. The benchmark reward (La Haye et al., 2006) multiplies an island-width penalty, decreasing in the modified-Rutherford width w of Equation 3, by a stored-thermal-energy boost $(W_{\mathrm{th}}/W_{\mathrm{th,ref}})^b$, where $W_{\mathrm{th}} = \frac{3}{2}\int k_B(n_eT_e + n_iT_i)V'\,d\hat{\rho}$ is the conventional tokamak performance metric and $W_{\mathrm{th,ref}}$ is the nominal stored energy on shot 137220. The multiplicative coupling forces the policy to keep both factors high simultaneously and rules out the degenerate solutions of extinguishing the island while letting T_e collapse, or heating the column without controlling the island. This reward also exposes the central physical tension of the task. Pushing P_{in} up raises T_e and the radial gradient ∇T_e, which inflates the bootstrap-current term $a_1 J_{\mathrm{bs}}/(w^2 + w_d^2)$ in Equation 3 and accelerates island growth, so the policy must instead steer the four beams onto the $q = 2$ rational surface to drive a compensating J_{cd} rather than heating uniformly. Hyperparameter values are listed in Appendix C.7.2.

Failure mode. The disruption modelled by `Profile` is the locked-mode endpoint of the $(2, 1)$ neoclassical tearing mode introduced in Section 2.2, in which the magnetic-island width w grows under Equation 3 until it locks to the vacuum-vessel frame and triggers a thermal quench (Buttery et al., 2000; De Vries et al., 2011). The env condenses this physical chain into a single termination, ending the episode the first time w crosses a configurable disruption threshold (default $0.15\,\mathrm{m}$) and assigning a fixed terminal penalty. In RL vocabulary the post-threshold region is absorbing and no admissible action recovers from it, so the policy must drive w down with electron-cyclotron current drive at the $q = 2$ rational surface before the crossing occurs. Figure 3 contrasts this pathway across four controllers. The ZeroAction baseline issues no actuator command and the GaussianRandom baseline injects power off the rational surface on average, so under both open-loop policies the island width $w(t)$ governed by Equation 3 grows monotonically from its seeded value until the env's locked-mode termination triggers. The PPO and PPO-LSTM controllers instead deposit current near the rational surface and hold $w(t)$ below $W_{\mathrm{threshold}}$ for the full $4\,\mathrm{s}$ episode.

4 Benchmark Tasks, Metrics, and Baselines

4.1 Scenario menu and evaluation protocol

TokamaCon ships two scenarios, one per environment. Table 1 lists each scenario together with its initial condition and reward. The `IsoFlux` scenario covers magnetic-equilibrium tracking around a single EAST EFIT linearisation. The `Profile` scenario covers the neoclassical-tearing-mode failure channel with a synthetic seeded $(2, 1)$ island, using the multiplicative island-width $\times$ stored-energy reward documented in Section 3.2.

Per-episode metrics. Each rollout produces three scenario-level scalars that map onto the canonical failure modes of Section 3.1. The first is a tracking error integrated over the controlled variables of the scenario, normalised so that the scenario reference schedule scores zero. The second is an MHD-violation rate, defined as the fraction of control steps on which the relevant stability proxy crosses its threshold; the proxy is the boundary excursion past the operator deadband for `IsoFlux` and the magnetic-island width past its saturation marker for the NTM scenario. The third is a time-to-failure equal to the step index at which the scenario terminates as a fraction of the maximum episode length, with non-terminating episodes scoring one. A fourth, controller-level scalar (the actuator-saturation fraction) is reserved for the baseline analysis in Section 4.3, since saturation reflects controller behaviour rather than scenario difficulty.

4.2 Provenance, parity, and runtime

Grounding in EAST shot data. Table 5 in Appendix C.3 summarises selected environment inputs, including geometry coefficients, transport-closure parameters, equilibrium references, and initial conditions, together with their documented sources. Experimentally measured quantities include geometry from EFIT g-files, magnetic measurements from the field probes and flux loops, and the ECE-derived T_e traces used for the `Profile` initial conditions. Synthetic quantities include the NTM seed island width and any auxiliary scenario parameters not derived from a measurement, all documented as synthetic and not represented as measured.

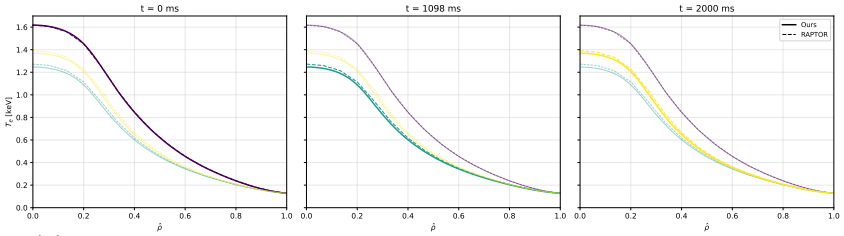


Figure 4: $T_e(\hat{\rho})$ on `Profile` against the RAPTOR launcher for scenario `test_data_3` (EAST shot 137220), at the initial frame, the pre-EC steady state, and the post-EC steady state.

Parity against the reference engine and GPU runtime. Beyond input provenance, we compare GPU outputs with reference engines for matched inputs on prescribed action trajectories (Appendix C.3). The reference engines are the EAST PCSVP plasma-control simulator for `IsoFlux` and the RAPTOR launcher for `Profile`, both single-process float64 codes that accept one trajectory at a time.

The `IsoFlux` backend ports the EAST PCSVP code that is in operational use as part of the EAST plasma-control system. The GPU re-implementation matches the reference to a relative L^2 error of 9.81×10^{-5} on a 100-step rollout at $B = 64$, at the floating-point noise floor of the port.

The `Profile` backend is calibrated against the RAPTOR code. On the scenario of EAST shot 137220 with a 700 kW to 1.35 MW EC step at $t = 1.099$ s, Figure 4 overlays the `Profile` $T_e(\hat{\rho})$ profile against RAPTOR at different timesteps. The two engines agree to a relative L^2 error of 1.94×10^{-2} on the T_e profile across the full 2 s window. The residual is a numerical difference rather than a modelling discrepancy, since the transport equation is stiff and small implementation differences in the implicit-step solver and the radial discretisation accumulate to per-frame deviations at this scale even between float64 implementations of the same closure. Diagnostic-channel residuals are dominated by a small mid-radius offset in χ_e that propagates into $\sigma_{\parallel}$. The supporting diagnostic stack (χ_e, $\sigma_{\parallel}$, q, P_{oh}, P_{ech}, P_{rad}, j_{oh}, j_{bs}, j_{aux}) is deferred to Section E.3. Both TokamaCon environments use a vectorised, GPU-resident contract that runs the environment loop natively on accelerator hardware and amortises the kernel batch across many parallel rollouts. Figure 5 traces per-step wall time as a function of batch size B (`num_envs`) on a single GPU against the CPU reference engines, and Table 2 summarises each environment's parity alongside the resulting throughput at the recommended B. The resulting throughput makes the multi-seed by three-controller evaluation protocol of Section 4.1 fit within a single GPU-week on commodity accelerators rather than being budget-limited.

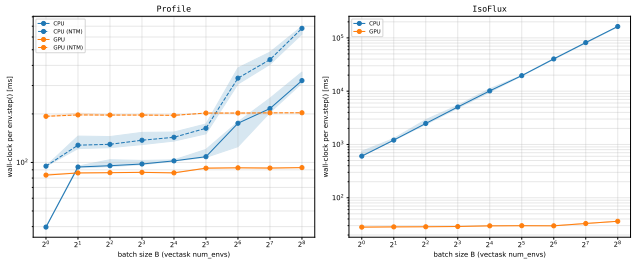


Figure 5: Per-step wall time against batch size B (`num_envs`) for both environments on a single GPU compared against the CPU reference engines.

Table 2: Per-environment parity against the reference engine and runtime comparison. Throughput numbers are environment-level and independent of any specific controller; the relative error L_2 is measured on a previously unseen action trajectory for `IsoFlux` (numerically tight) and `Profile` (residual is a numerical difference under a stiff transport equation, not a modelling gap). Throughput is reported at the recommended `num_envs` $B = 256$ on a single GPU.

Environment	Reference engine	Rel. error	CPU env-steps/s	GPU env-steps/s
`IsoFlux`	EAST PCSVP	9.81×10^{-5}	1.57	7 071
`Profile`	RAPTOR launcher	1.94×10^{-2}	795	2 756

4.3 Reference controllers

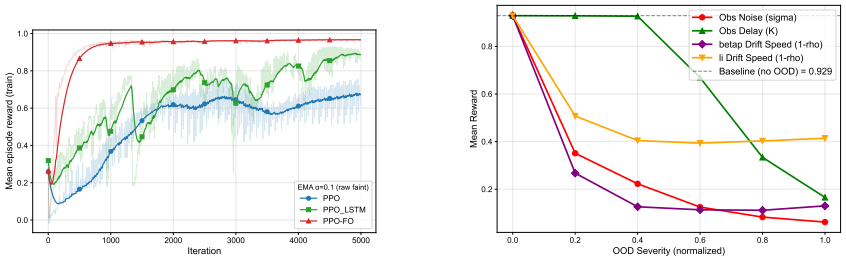
We implement Proximal Policy Optimization (PPO) (Schulman et al., 2017) as the reference reinforcement-learning controller, then keep other training configuration fixed while adding two probes: PPO-LSTM, which adds a long short-term memory (LSTM) recurrent core to study the partially observable Markov decision process (POMDP) regime, and PPO-FO, which gives the actor full privileged plant-state observation to isolate the cost of missing state. The purpose of these baselines is to study the control characteristics and challenges of the fusion-control tasks, *i.e.*, sparse observation limits learning and final tracking, diagnostic noise, delay, and plant drift test robustness, and actuator-rail usage provides a proxy for control-margin consumption rather than a safety guarantee. We also include an engineering PID controller for `IsoFlux` under the same evaluation conditions (Section D.1).

Table 3: Reference controllers on TokamaCon: `IsoFlux` uses EAST shot 152444 and `Profile` uses the shot-137220 case, as independent tasks. The left block reports per-(controller, environment) evaluation reward, and the right block reports actuator-saturation fraction, the share of control steps on which any actuator command lies on its per-channel hardware rail. For `Profile`, – means this engineering baseline was not supplied for the task, not that PID is theoretically inapplicable.

	Evaluation reward		Actuator saturation	
Controller	`IsoFlux`	`Profile`	`IsoFlux`	`Profile`
PID	0.9478	–	0.0000	–
PPO	0.7441	0.6420	0.0048	0.0102
PPO-LSTM	0.9238	0.6412	0.0018	0.0877
PPO-FO	0.9741	0.6329	0.0011	0.2094

Partial observability. On `IsoFlux` the actor sees only 12 of the 188 dimensions available to the critic, so the policy must infer the unobserved coil currents and vacuum-vessel eddy-current modes that drive the dynamics from a sensor channel that does not expose them directly. The PPO-LSTM evaluation reward in Table 3 exceeds PPO by 0.180 and falls 0.050 short of the fully observed PPO-FO, while Figure 6(a) shows the same ordering in convergence speed over the fixed training budget. The resulting 0.230 span between sensor-only PPO and PPO-FO therefore appears in both learning dynamics and final evaluation reward. Recurrent memory closes some of this gap, but the remaining 0.050 residual indicates that observation history mitigates rather than removes the information bottleneck faced by tokamak plasma controllers acting from sparse indirect diagnostics.

Engineering-controller comparison. On the same shot-152444 instance, PID's evaluation reward of 0.9478 exceeds both sensor-only PPO (0.7441) and PPO-LSTM (0.9238). PPO-FO has a higher point-estimate reward (0.9741), but its privileged plant-state access makes this an information-availability comparison, not evidence of a deployable RL advantage over PID.



(a) Training reward **(b) Out-of-distribution evaluation**

Figure 6: Reference baselines on the `IsoFlux isoflux_eq_track` scenario. (a) Training reward of the three baselines over the full training budget. (b) Out-of-distribution evaluation with four perturbation modes.

Distribution shift. Real tokamak operation imposes perturbations within a shot and parameter drift across shots, the second deployment difficulty surfaced by TokamaCon. We capture this with a four-axis severity sweep along observation noise σ, observation delay K, and the β_p and l_i dynamics-drift speeds, all evaluated against the same fixed reference schedule that produced the in-distribution numbers in Table 3. Figure 6 (b) separates the robustness signature by perturbation source: reward decreases under diagnostic noise and delay in the measurement channel, and also under β_p and l_i drift in the plant response. These axes are not arbitrary corruptions, but deployment-time distribution shifts that tokamak real-time controllers must tolerate across and within discharges. The sweeps perturb the shot-152444 instance rather than substitute a newly linearised equilibrium bundle. As a separate supplemental instance, Section E.1 reports newly trained policies on equilibrium A of shot 137220 and evaluates their frozen checkpoints and the same PID on held-out equilibrium B without retraining or retuning; this is not a transfer of the shot-152444 policies.

Hard safety constraints. The third deployment difficulty surfaced by TokamaCon is hard safety, expressed on `IsoFlux` as a magnet-current limit and on `Profile` as a disruption terminal; because trained policies rarely reach the terminal boundary in nominal evaluation, we report actuator-saturation fraction as a control-margin proxy. On `IsoFlux`, saturation follows information availability, with PPO at 0.48%, PPO-LSTM at 0.18%, and PPO-FO at 0.11%, so partial observation consumes roughly four times the rail margin of full observation while recurrence cuts that burden by more than half. These absolute values remain below 1%, but the differential matters because observation error, delay, and plant drift consume actuator margin before they become explicit violations. On `Profile`, the ordering inverts in a way we want to flag as a property the scenario successfully exposes rather than as a baseline shortcoming. Saturation climbs roughly $20\times$ from 1.02% (PPO) through 8.77% (PPO-LSTM) to 20.94% (PPO-FO), while evaluation reward stays in a narrow 0.6329–0.6420 band with PPO-FO at the bottom rather than the top, so the fully observed actor is the most rail-bound and lowest-reward of the three. The mechanism is the 3.2 MW total electron-cyclotron-power cap fixed by the EAST gyrotron envelope (Section 3.2), which bounds the rational-surface current any controller can deliver and therefore caps reward independently of observation quality, while incentivising a confident PPO-FO actor to hold maximum deposition at the rail. `Profile` thereby surfaces an actuator-authority-bound regime that observation-rich benchmarks cannot expose, and the $20\times$ saturation spread under a sub-one-point reward spread shows that in such scenarios, reward alone is insufficient as a deployment-risk signal, a regime tokamak controllers face routinely under hardware caps and operator deadbands.

5 Conclusion

We presented TokamaCon, an EAST-derived environment suite that decouples real-time tokamak control into the magnetic-equilibrium loop (`IsoFlux`) and the kinetic-profile loop (`Profile`), both running natively on GPU through a vectorised PyTorch contract. The released scenarios, per-episode metrics, provenance metadata, and four reference baselines (PPO, PPO-LSTM, PPO-FO, PID) together quantify three of the four challenge axes that fusion deployment imposes on a learned controller, namely partial observation, distribution shift, and hard-safety margin. Code, scenarios, and headline checkpoints are available at the anonymised repository linked from the abstract, and a stable point of contact and versioned issue tracker will be added on de-anonymisation.

AI USE STATEMENT

We used generative AI tools for writing and organising materials, code and scripts, analysis and visualisation, and research and derivations. All AI-assisted work was reviewed by the authors, who take responsibility for the final text, code, analyses, claims, and other artefacts.

ETHICS STATEMENT

TokamaCon supports simulator-based research on tokamak plasma control, not direct deployment on operating devices. Performance in these environments does not establish the safety of real-world deployment, which requires independent validation and device-specific safeguards.

REPRODUCIBILITY STATEMENT

The anonymous materials linked in the abstract provide code, scenarios, and reference checkpoints. Environment definitions and benchmark protocols are described in Sections 3 and 4; implementation and configuration details, reference baselines, and training hyperparameters are provided in Appendices C and D.

REFERENCES

Raffaele Albanese and F Villone. The linearized create-l plasma response model for the control of current, position and shape in tokamaks. *Nuclear Fusion*, 38(5):723–738, 1998.

Giuseppe Ambrosino and Raffaele Albanese. Magnetic control of plasma current, position, and shape in tokamaks: a survey or modeling and control approaches. *IEEE Control Systems Magazine*, 25 (5):76–92, 2005.

Jean-François Artaud, Frédéric Imbeaux, J Garcia, G Giruzzi, T Aniel, V Basiuk, A Bécoulet, C Bourdelle, Y Buravand, J Decker, et al. Metis: a fast integrated tokamak modelling tool for scenario design. *Nuclear Fusion*, 58(10):105001, 2018.

RJ Buttery, S Günter, G Giruzzi, TC Hender, D Howell, G Huysmans, RJ La Haye, M Maraschek, H Reimerdes, O Sauter, et al. Neoclassical tearing modes. *Plasma Physics and Controlled Fusion*, 42(12B):B61–B73, 2000.

DJ Campbell, AJH Donné, I Bandyopadhyay, A Fasoli, ME Fenstermacher, CM Greenfield, RJ Hawryluk, AE Hubbard, SV Konovalov, K Krieger, et al. Introduction and overview: Chapter 1 of the special issue: on the path to tokamak burning plasma operation. *Nuclear Fusion*, 65(9):093002, 2025.

Ian Char, Joseph Abbate, László Bardóczi, Mark Boyer, Youngseog Chung, Rory Conlin, Keith Erickson, Viraj Mehta, Nathan Richner, Egemen Kolemen, et al. Offline model-based reinforcement learning for tokamak control. In *Learning for Dynamics and Control Conference*, pp. 1357–1372. PMLR, 2023.

Jonathan Citrin, Ian Goodfellow, Akhil Raju, Jeremy Chen, Jonas Degrave, Craig Donner, Federico Felici, Philippe Hamel, Andrea Huber, Dmitry Nikulin, et al. Torax: A fast and differentiable tokamak transport simulator in jax. *arXiv preprint arXiv:2406.06718*, 2024.

Gianmaria De Tommasi. Plasma magnetic control in tokamak devices. *Journal of Fusion Energy*, 38 (3):406–436, 2019.

PC De Vries, MF Johnson, B Alper, P Buratti, TC Hender, HR Koslowski, V Riccardo, and Jet-Efda Contributors. Survey of disruption causes at jet. *Nuclear fusion*, 51(5):053018, 2011.

Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, 2022.

M Erba, A Cherubini, VV Parail, E Springmann, and A Taroni. Development of a non-local model for tokamak heat transport in l-mode, h-mode and transient regimes. *Plasma Physics and Controlled Fusion*, 39(2):261–276, 1997.

F Felici, O Sauter, S Coda, BP Duval, TP Goodman, JM Moret, JI Paley, and TCV Team. Real-time physics-model-based simulation of the current density profile in tokamak plasmas. *Nuclear Fusion*, 51(8):083052, 2011.

Jeffrey P Freidberg. *ideal MHD*. Cambridge University Press, 2014.

X Gong, EAST Team, Collaborators, EAST Team:, Yuntao Song, Baonian Wan, Jiangang Li, Yuanxi Wan, Xinchao Wu, Fukun Liu, Junling Chen, et al. Overview of recent experimental results on the east tokamak. *Nuclear Fusion*, 64(11):112013, 2024.

Martin Greenwald. Density limits in toroidal plasmas. *Plasma Physics and Controlled Fusion*, 44(8): R27–R53, 2002.

Yu Gribov, D Humphreys, K Kajiwara, EA Lazarus, JB Lister, Tl Ozeki, A Portone, M Shimada, ACC Sips, and JC Wesley. Chapter 8: Plasma operation and control. *Nuclear Fusion*, 47(6): S385–S403, 2007.

Shengyi Huang, Rousslan Fernand Julien Dossa, Chang Ye, Jeff Braga, Dipam Chakraborty, Kinal Mehta, and João G.M. Araújo. Cleanrl: High-quality single-file implementations of deep reinforcement learning algorithms. *Journal of Machine Learning Research*, 23(274):1–18, 2022. URL http://jmlr.org/papers/v23/21-1342.html.

DA Humphreys, TA Casper, N Eidietis, M Ferrara, DA Gates, IH Hutchinson, GL Jackson, Egemen Kolemen, JA Leuer, J Lister, et al. Experimental vertical stability studies for iter performance and design guidance. *Nuclear Fusion*, 49(11):115003, 2009.

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.

SH Kim and JB Lister. A potentially robust plasma profile control approach for iter using real-time estimation of linearized profile response models. *Nuclear Fusion*, 52(7):074002, 2012.

RJ La Haye. Neoclassical tearing modes and their control. *Physics of Plasmas*, 13(5), 2006.

RJ La Haye, R Prater, RJ Buttery, N Hayashi, A Isayama, ME Maraschek, L Urso, and H Zohm. Cross–machine benchmarking for iter of neoclassical tearing mode stabilization by electron cyclotron current drive. *Nuclear fusion*, 46(4):451–461, 2006.

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.

J Li, HY Guo, BN Wan, XZ Gong, YF Liang, GS Xu, KF Gan, JS Hu, HQ Wang, L Wang, et al. A long-pulse high-confinement plasma regime in the experimental advanced superconducting tokamak. *Nature physics*, 9(12):817–821, 2013.

Spangher Lucas, Matteo Bonotto, William Arnold, Dhruva Chayapathy, Tommaso Gallingani, Alexander Spangher, Francesco Cannarile, Daniele Bigoni, Eliana De Marchi, and Cristina Rea. Disruptionbench: A robust benchmarking framework for machine learning-driven disruption prediction. 2024.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

Fernando Martínez-Plumed, Pablo Barredo, Sean O Heigeartaigh, and Jose Hernandez-Orallo. Research community dynamics behind popular ai benchmarks. *Nature Machine Intelligence*, 3(7): 581–589, 2021.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

D Moreau, D Mazon, M Ariola, Gianmaria De Tommasi, L Laborde, F Piccolo, F Sartori, Tuomas Tala, L Zabeo, A Boboc, et al. A two-time-scale dynamic-model approach for magnetic and kinetic profile control in advanced tokamak scenarios on jet. *Nuclear Fusion*, 48(10):106001, 2008.

Antoine Mouchamps, Arthur Malherbe, Adrien Bolland, and Damien Ernst. Gym-torax: Open-source software for integrating reinforcement learning with plasma control simulators in tokamak research. *Software Impacts*, pp. 100829, 2026.

Simon Ott, Adriano Barbosa-Silva, Kathrin Blagec, Jan Brauner, and Matthias Samwald. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, 13(1):6793, 2022.

Gregorij V Pereverzev and e PN Yushmanov. Astra. automated system for transport analysis in a tokamak. 2002.

Ronald Prater. Heating and current drive by electron cyclotron waves. *Physics of plasmas*, 11(5): 2349–2376, 2004.

Kevin Reuer, Jonas Landgraf, Thomas Fösel, James O'Sullivan, Liberto Beltrán, Abdulkadir Akin, Graham J Norris, Ants Remm, Michael Kerschbaum, Jean-Claude Besse, et al. Realizing a deep reinforcement learning agent for real-time quantum feedback. *Nature Communications*, 14(1): 7138, 2023.

Michele Romanelli, Gerard Corrigan, Vassili Parail, Sven Wiesen, Roberto Ambrosino, Paula Da Silva Aresta Belo, Luca Garzotti, Derek Harting, Florian Köchl, Tuomas Koskela, et al. Jintrac: a system of codes for integrated simulation of tokamak scenarios. *Plasma and Fusion research*, 9: 3403023–3403023, 2014.

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.

Olivier Sauter, Robert J La Haye, Zuoyang Chang, David A Gates, Yutaka Kamada, Hartmut Zohm, Anders Bondeson, Dominique Boucher, James D Callen, Ming S Chu, et al. Beta limits in long-pulse tokamak discharges. *Physics of Plasmas*, 4(5):1654–1664, 1997.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

J Seo, Y-S Na, B Kim, CY Lee, MS Park, SJ Park, and YH Lee. Development of an operation trajectory design algorithm for control of multiple 0d parameters using deep reinforcement learning in kstar. *Nuclear Fusion*, 62(8):086049, 2022.

Jaemin Seo, SangKyeun Kim, Azarakhsh Jalalvand, Rory Conlin, Andrew Rothstein, Joseph Abbate, Keith Erickson, Josiah Wai, Ricardo Shousha, and Egemen Kolemen. Avoiding fusion plasma tearing instability with deep reinforcement learning. *Nature*, 626(8000):746–751, 2024.

M Shimada, DJ Campbell, V Mukhovatov, M Fujiwara, N Kirneva, K Lackner, MPVD Nagami, VD Pustovitov, N Uckan, J Wesley, et al. Chapter 1: Overview and summary. *Nuclear Fusion*, 47 (6):S1–S17, 2007.

Brendan D Tracey, Andrea Michi, Yuri Chervonyi, Ian Davies, Cosmin Paduraru, Nevena Lazic, Federico Felici, Timo Ewalds, Craig Donner, Cristian Galperti, et al. Towards practical reinforcement learning for tokamak magnetic control. *Fusion Engineering and Design*, 200:114161, 2024.

Francis Troyon, R Gruber, H Saurenmann, S Semenzato, and S Succi. Mhd-limits to plasma confinement. *Plasma physics and controlled fusion*, 26(1A):209–215, 1984.

T Wakatsuki, M Yoshida, E Narita, T Suzuki, and N Hayashi. Simultaneous control of safety factor profile and normalized beta for jt-60sa using reinforcement learning. *Nuclear Fusion*, 63(7): 076017, 2023.

Allen M Wang, Alessandro Pau, Cristina Rea, Oswin So, Charles Dawson, Olivier Sauter, Mark D Boyer, Anna Vu, Cristian Galperti, Chuchu Fan, et al. Learning plasma dynamics and robust rampdown trajectories with predict-first experiments at tcv. *Nature Communications*, 16(1):8877, 2025.

John Wesson and David J Campbell. *Tokamaks*, volume 149. Oxford university press, 2011.

QP Yuan, BJ Xiao, ZP Luo, ML Walker, AS Welander, A Hyatt, JP Qian, RR Zhang, DA Humphreys, JA Leuer, et al. Plasma current, position and shape feedback control on east. *Nuclear Fusion*, 53 (4):043009, 2013.

A Related work

A.1 Tokamak prediction benchmarks

Open benchmarks in the tokamak community have so far concentrated on offline supervised forecasting from archival shot data. TokaMark (Rousseau et al., 2026) packages 14 prediction tasks across 11,573 MAST shots into a unified evaluation, with subtasks spanning equilibrium reconstruction, magnetic and profile dynamics, and MHD activity together with disruption precursors; it draws on the FAIR-MAST open data catalog Jackson et al. (2024) for shot-level provenance. DisruptionBench Lucas et al. (2024) formalises cross-machine disruption prediction across Alcator C-Mod, DIII-D, and EAST under three transfer regimes. Together these efforts establish a standard reference pool for fusion-domain machine learning: they define inputs, splits, and metrics, and they ship reproducible baselines. None of them, however, defines an action space, exposes a real-time control interface, or evaluates feedback policies. Closed-loop control therefore lies outside their evaluation protocol, and the prediction-only setting these benchmarks formalise leaves no published numbers a learned controller could be compared against.

A.2 Tokamak control simulators and environments

On the control side, prior work splits into a simulator layer and an environment layer: 1-D transport codes such as TORAX (Citrin et al., 2024) and the RAPTOR family (Felici et al., 2011), alongside broader integrated-modelling tools such as ASTRA (Pereverzev & Yushmanov, 2002), METIS (Artaud et al., 2018), and JINTRAC (Romanelli et al., 2014), supply transport closures and scenario-design machinery but, as released, do not expose a standardised reinforcement-learning interface, while at the environment layer the published artifact most directly aligned with reinforcement learning is Gym-TORAX (Mouchamps et al., 2026), a Gymnasium wrapper around TORAX that demonstrates a single ITER ramp-up profile-control scenario. TokamaCon extends this picture with two environments and a vectorised runtime contract. The profile environment evolves coupled T_e and ψ transport equations and additionally provides a neoclassical-tearing-mode (NTM) (Buttery et al., 2000) scenario in which a $2/1$ island seeded at $w_0 = 0.15\,\mathrm{m}$ evolves under the modified Rutherford equation, with the island width feeding back into local thermal transport. A second environment provides magnetic-equilibrium control via a 172-D EFIT (Lao et al., 1985) linearisation on EAST, driven at $1\,\mathrm{kHz}$ with a $5\,\mathrm{ms}$ PF-coil delay, per-channel voltage saturation, and an internal vertical-stabilisation loop that is not exposed to the policy. Both environments share a vectorised contract built on the Isaac Gym pattern (Makoviychuk et al., 2021), exposing GPU-resident `[num_envs, ...]` PyTorch tensors and a sensor-injection slot for observation ablations. TokamaCon is intended as a reproducible substrate for research on learned tokamak control, of which the magnetic-control demonstration of Degrave et al. (Degrave et al., 2022) and the tearing-instability-avoidance work of Seo et al. (Seo et al., 2024) are representative milestones, rather than as a claim about algorithmic findings.

B Datasheet for the TokamaCon environment suite

B.1 Motivation

TokamaCon is released to give reinforcement-learning research a reproducible substrate for tokamak real-time control, complementing the offline-forecasting benchmarks and the single-scenario Gym-TORAX environment surveyed in Section A. The intended users are reinforcement-learning researchers studying control under the four challenge axes inherited from EAST deployment physics (partial observation, distribution shift, hard safety, and stiff nonlinear dynamics). The two scenarios were chosen to expose distinct failure channels, namely tracking violation on the fast magnetic loop and locked-mode disruption on the slow profile loop, both inherited from EAST deployment physics rather than synthesised for benchmark difficulty.

B.2 Composition

The suite contains two environments (`IsoFlux`, `Profile`; specifications in Section 3), two evaluation scenarios (`isoflux_eq_track` and `ntm_w_energy`, documented in Table 1), and

three reference RL baselines (PPO, PPO-LSTM, PPO-FO; hyperparameters in Table 9). The reported `IsoFlux` evaluations also include an engineering PID reference (Section D.1). Source ships under `plasmabench_src/`, scenario YAMLs alongside, and PyTorch `.pt` checkpoints for the trained baselines, with launchers and evaluation drivers under `scripts/`. The evaluated instances remain limited to EAST: `IsoFlux` includes the original shot 152444 instance and equilibria A and B of shot 137220, while `Profile` uses a shot-137220-derived configuration. This is a limited set of instances, not broad device or equilibrium coverage (Section B.3). The submission is anonymous, the code repository URL embedded in the abstract resolves to an anonymised mirror, and no personally identifying information is shipped with the release.

B.3 Recommended uses and known limitations

Concrete recommended uses include controller-architecture studies that swap the actor backbone while holding the environment, scenario, and reward fixed, and ablations of the partial-observation regime via the sensor-injection slot of the vectorised contract. TokamaCon is not recommended for sim-to-real validation, since the actor observation does not include a calibrated sensor-noise model and transfer claims to deployed EAST controllers cannot be drawn from these environments alone. Cross-machine transfer studies are also out of scope, since both environments are grounded in a single device, and reactor-relevant burning-plasma studies are out of scope, since the `IsoFlux` linearisation is taken at non-burning EAST conditions. Known limitations (limited EAST instances, frozen n_e, synthetic ECE forward model, χ_e parity offset) are detailed in Section F; maintenance is handled through the anonymous code repository linked from the abstract, and a stable point of contact and versioned issue tracker will be added on de-anonymisation.

C Environment implementation details

C.1 Plasma-control primer and notation

A tokamak confines a current-carrying plasma with magnetic fields. Plasma equilibrium describes the balance between magnetic forces and plasma pressure; its boundary and current must be regulated while internal temperature and current profiles evolve. The fast magnetic loop in `IsoFlux` uses poloidal-field (PF) and central-solenoid (CS) coil voltages to control plasma current, boundary flux, and the X-point field. Magnetic diagnostics provide indirect observations rather than the complete plant state. The slower `Profile` loop uses electron-cyclotron heating and current drive to influence temperature and suppress a neoclassical tearing mode (NTM), represented by a magnetic island (Felici et al., 2011; Buttery et al., 2000). Synthetic electron-cyclotron-emission (ECE) channels sample temperature at selected radii. In the RL interface, diagnostics become observations, actuator commands become actions, and tracking or suppression objectives become rewards. Tracking violations and excessive island growth define scenario-specific failure conditions, not comprehensive guarantees of operational safety. The two loops are evaluated separately.

Table 4: Selected plasma-control notation; detailed definitions appear in the environment and reward specifications.

Symbol	Meaning	Unit
I_p	Plasma current	A (reported also in kA)
$B_{x,R}$, $B_{x,Z}$	X-point poloidal-field components	T
$\hat{\rho}$	Normalised toroidal-flux radius	dimensionless
T_e	Electron temperature	eV
n_e	Electron density	m^{-3}
χ_e	Electron thermal diffusivity	m^2/s
w	Magnetic-island width	m
W_{th}	Stored thermal energy	J (reported also in kJ)

C.2 Control-layer decomposition and scope

Magnetic equilibrium and kinetic profiles are physically coupled, but their different response times and actuators motivate distinct control layers. TokamaCon studies these tasks independently rather

Table 5: Source information for selected TokamaCon inputs; synthetic entries are explicitly indicated.

Quantity	Category	Source
`IsoFlux`		
EFIT linearisation A, B, C	measured	EAST shot 152444, EFIT g-file
PF and CS coil topology	measured	EAST PCS hardware
Voltage and rate limits	measured	EAST PF power supply
`Profile`		
Geometry $V', g_{0..3}, F$	measured	EAST shot 137220, EFIT g-file
$\psi(\hat{\rho}, 0)$ initial condition	measured	EAST shot 137220, EFIT
$T_e(\hat{\rho}, 0)$ baseline	measured	EAST shot 137220, ECE
χ_e closure (BgB or Felici)	calibrated form	literature defaults
$2/1$ island seed $w_0 = 0.15\,\mathrm{m}$	synthetic	NTM benchmark scenario

than reproducing their online interaction. IsoFlux treats $\Delta\beta_p$ and Δl_i as configured disturbances, not runtime outputs from Profile; Profile uses its own reference equilibrium geometry. This decomposition supports layer-specific controller development and evaluation, but does not establish performance in coupled whole-discharge operation.

C.3 Experimental provenance and reference validation

Experimental provenance, reference-model calibration, and GPU implementation parity are distinct evidence layers. `IsoFlux` loads an EAST-derived plant bundle containing the equilibrium, linearised response matrices, equilibrium vectors, diagnostic and control-point mappings, and actuator limits. Its upstream PCSVP reference simulator was calibrated with EAST experimental data and validated through closed-loop control in EAST operation; this does not constitute direct experimental validation of the GPU implementation or learned policies. `Profile` uses EAST-derived geometry, initial profiles, diagnostic settings, and actuator parameters. Its RAPTOR comparison uses a shot-137220-calibrated case with a prescribed heating trajectory, rather than a distribution of policy-generated trajectories. The GPU comparisons in Section 4.2 assess trajectory-level agreement under matched inputs, separately from upstream experimental calibration. The NTM component is a reduced, synthetic perturbation model without direct experimental validation. Broader experimental comparisons and redistribution of measurements depend on data authorisation and licensing; additional RAPTOR comparisons likewise require software access and permission.

C.4 Linearised plant derivation and parameter calibration

The original shot-152444 `IsoFlux` instance uses a CREATE-L type linearised free-boundary plasma response model Albanese & Villone (1998); Ambrosino & Albanese (2005) on the EAST equilibrium reconstructed by EFIT Lao et al. (1985), loaded at construction time from the scenario bundle `PCSVP/scenario_files/default.pt`. The bundle-driven implementation can load a new equilibrium (Section C.11); the eigenvalue and numerical parameterisation below describe this original instance used for the main results (Section 3.1, Table 3), not the supplementary shot-137220 equilibrium A (Section E.1). After linearising the coupled circuit, radial Grad–Shafranov, and vertical-motion equations around the equilibrium $\mathbf{x}_{\mathrm{eq}}$, the plant takes the LTI form

$$\dot{\mathbf{x}} = A\mathbf{x} + B\mathbf{u}, \qquad \mathbf{y} = C(\mathbf{x} - \mathbf{x}_{\mathrm{eq}}) + \mathbf{y}_{\mathrm{eq}}, \tag{4}$$

with $\mathbf{x} \in \mathbb{R}^{172}$ stacking twelve active poloidal-field/central-solenoid coil currents, 158 vessel passive-mode currents, the plasma current I_p, and a vertical-motion sub-state; $\mathbf{u} \in \mathbb{R}^{16}$ stacking the twelve coil voltages, a hard-wired vertical-stabilisation (VS) bang-bang voltage, a loop-voltage offset, and two scalar profile disturbances $(\delta\beta_p, \delta l_i)$; and $\mathbf{y} \in \mathbb{R}^{188}$ stacking ten boundary control-point fluxes, the X-point flux and field, displacements, all coil currents, vessel modes, and I_p as detailed in Section 3.1. The single unstable eigenvalue $\lambda \approx +225.85\,\mathrm{rad\,s^{-1}}$ (vertical-displacement mode, growth time $4.43\,\mathrm{ms}$) is closed inside the env by the VS bang-bang loop on a dedicated circuit and is invisible to the policy Humphreys et al. (2009), so the plant exposed to RL is vertically stable.

Equation 4 integrates with a fixed-step Bogacki–Shampine third-order Runge–Kutta scheme at substep $h = 10^{-5}\,\mathrm{s}$, advancing 100 substeps per 1 kHz outer step (this substep keeps the cumulative

parity gap against the reference Simulink engine below 10^{-9} over a 2 s episode in float64). The policy command $a \in [-1, 1]^{12}$ is rescaled to physical units by the per-channel gains in Table 6, then passed through a $\tau_{\mathrm{pf}} = 5$ ms transport delay and per-channel saturation, and finally demultiplexed onto the twelve policy-controlled EAST circuits by the wiring vector `cccirc` (Table 7). For this original instance, the matrices (A, B, C), equilibrium vectors, saturation thresholds, and VS threshold are loaded from the `case` dictionary without further tuning, traceable to the EAST plasma current, shape and vertical position (PCSVP) control bench used on shot 152444 Yuan et al. (2013); full numerical settings are in Table 6.

Table 6: Original shot-152444 `IsoFlux` plant parameters loaded from the `case` dictionary of `scenario_files/default.pt`, together with numerical settings of the discrete integrator.

Parameter	Value
(A, B, C) shape	172×172, 172×16, 188×172
Feedthrough D (`dmat`)	0
Number of vessel modes n_v	158
Number of control points n_{cp} (`n_ctrl_point`)	10 (first 8 used in obs)
Saturation, Ohmic / CS, PS1–PS6 (`Vmax_ohmic`)	$\pm$300 V
Saturation, PF, PS7–PS12 (`Vmax_pf`)	$\pm$900 V
Saturation, VS supply (`Vmax_vs`)	$\pm$2000 V
VS bang-bang threshold on $\lvert Z_p \rvert$ (`dz_bangbang`)	2 cm
Transport delay, PF (`tau_delay_pf`)	5 ms
Transport delay, VS (`tau_delay_vs`)	0.1 ms
Rate limit, $\delta\beta_p$ (`dt_betap`)	200 s^{-1}
Rate limit, δl_i (`dt_li`)	100 s^{-1}
Profile gain, $\delta\beta_p$ (`delta_betap`)	-0.167
Profile gain, δl_i (`delta_li`)	1
ODE substep h	10^{-5} s
Outer-step substep count	100
Episode horizon	2000 outer steps (2 s)

Table 7: Coil-to-circuit wiring vector `cccirc` $= (1, 2, 3, 4, 5, 6, 7, 8, 7, 8, 9, 10, 11, 12, 13, -13)$ for EAST. Sixteen physical conductors are mapped to thirteen independent circuits; the VS pair drives a single circuit with opposite polarity and is hidden from the policy.

Physical coil	$\rightarrow$ Circuit	Action index	Role
PF1–PF6	$\rightarrow$ 1–6	$a_{0:5}$	Ohmic / CS
PF7, PF9 (parallel)	$\rightarrow$ 7	a_6	PF
PF8, PF10 (parallel)	$\rightarrow$ 8	a_7	PF
PF11–PF14	$\rightarrow$ 9–12	$a_{8:11}$	PF
VS+, VS$-$	$\rightarrow$ 13	not exposed	internal VS bang-bang

C.5 PROFILE TRANSPORT CLOSURE

The `Profile` plant is a RAPTOR-style one-dimensional transport solver (Felici et al., 2011) on the normalised toroidal-flux coordinate $\hat{\rho} \in [0, 1]$, evolving the coupled (ψ, T_e) pair on a $N_\rho = 64$ uniform grid with the equilibrium geometry (V', g_{03}, F, q_0) frozen at the EAST shot 137220 EFIT reconstruction and refreshed only through $q(\hat{\rho})$ as ψ evolves. The electron-temperature equation in conservative form is Equation 2; the poloidal-flux equation runs on the same grid and is closed by a Sauter neoclassical conductivity model that yields the parallel conductivity $\sigma_\parallel$ and the bootstrap current density J_{bs} (Sauter et al., 1997). The electron density $n_e(\hat{\rho})$ is held fixed at the shot 137220 initial profile and enters the T_e dynamics only as a coefficient, so the policy never modifies it.

The default heat-diffusivity closure is the Felici four-term ad-hoc form $\chi_e(\hat{\rho}) = \chi_{\mathrm{neo}} + \chi_{\mathrm{anom}}(\hat{\rho}) + \chi_{\mathrm{ITB}}(\hat{\rho}) + \chi_{\mathrm{central}}(\hat{\rho})$, with the anomalous term gated by the magnetic-shear profile and the central term a Gaussian core boost; a Bohm/gyro-Bohm alternative $\chi_e = c_{\mathrm{neo}} + \alpha_B\,\chi_B + \alpha_{\mathrm{gB}}\,\chi_{\mathrm{gB}}$ with $\chi_B \propto (T_e/B)q^2(\nabla T_e/T_e + \nabla n_e/n_e)$ and $\chi_{\mathrm{gB}} \propto \rho^\star(T_e/B)\,\nabla T_e/T_e$ is selectable through `cfg.physics.electron_heat.chi_e_model = "bgb"`.

Each of the $K = 4$ electron-cyclotron beams introduced in Section 3.2 contributes a Gaussian deposition kernel

$$\frac{dP_{\mathrm{ec},k}}{dV}(\hat{\rho}) = \frac{P_{\mathrm{in},k}}{\mathcal{N}_k}\,\exp\!\left(-4\,\frac{(\hat{\rho}-r_{\mathrm{dep},k})^2}{w_{\mathrm{dep},k}^2}\right), \tag{5}$$

with the per-beam normalisation $\mathcal{N}_k$ chosen so that the volume integral $\int dP_{\mathrm{ec},k}/dV \cdot V'\,d\hat{\rho}$ recovers the requested injected power $P_{\mathrm{in},k}$; the same kernel shape with a fixed efficiency η_{eccd} produces the ECCD current density that enters Equation 3.

The island-width state w evolves under the modified-Rutherford balance $\dot{w} = (r_s/\tau_R)\sum_i \Delta_i'$ summed over classical, bootstrap, Glasser–Greene–Johnson curvature, ECCD, and ECRH contributions (Sauter et al., 1997; La Haye et al., 2006), with the resistive time $\tau_R = \mu_0 r_s^2/(1.22\,\eta_{\mathrm{neo}})$ at the rational surface $\hat{\rho}_{m,n}$ and parameter values listed in Table 8. Once $w > 0$, the island flattens T_e across its mouth through a multiplicative thermal short-circuit applied to χ_e,

$$\chi_e(\hat{\rho}) \to \chi_e(\hat{\rho}) \cdot \left[1 + A_{\mathrm{fact}}\,\exp\!\left(-4\,\frac{(\hat{\rho}-\hat{\rho}_{m,n})^2}{(W_{\mathrm{fact}}\,w/a)^2}\right)\right], \tag{6}$$

peaked at the rational surface and with width set by the island half-width normalised to the minor radius w/a. The actor observation samples T_e at the twelve $\hat{\rho}$ channel positions of the shot 137220 ECE geometry by linear interpolation; harmonic selection, R-major frequency cutoff, and optical-depth physics are intentionally left out, so the ECE forward model is a sensor-geometry placeholder rather than a calibrated diagnostic.

Table 8: `Profile` default closure parameters loaded from the shot 137220 scenario yaml. Symbols match Equations 2, 3, 5 and 6.

Group	Parameter	Value
Heat-diffusivity closure		
χ_e model selector	`chi_e_model`	`felici_ff` (default) / `bgb`
BgB Bohm coefficient	α_B	8×10^{-5}
BgB gyro-Bohm coefficient	α_{gB}	3.5×10^{-2}
Neoclassical floor	c_{neo}	$1\,\mathrm{m}^2/\mathrm{s}$
Modified-Rutherford balance (Equation 3)		
Mode numbers	(m, n)	$(2, 1)$
Classical baseline	$\Delta'_{\mathrm{CL},0}$	-0.2
Bootstrap coefficient	$a_{2,\mathrm{bs}}$	1.0
Electron / ion threshold widths	w_{de}, w_{di}	0.01 m
Effective ECCD width	w_{cd}	0.01 m
ECCD geometric efficiency	g	0.6
Resistivity correction	$f_{\eta,\mathrm{corr}}$	1.0
Seed island width	w_{seed}	0.015 m
χ_e *thermal short-circuit (Equation 6)*		
Boost amplitude	A_{fact}	1.0
Boost width factor	W_{fact}	1.0

C.6 PROFILE ACTION INTERFACE

The raw policy command $a \in \mathbb{R}^{12}$ is laid out as four interleaved per-beam triples $(r_{\mathrm{dep},k}, w_{\mathrm{dep},k}, P_{\mathrm{in},k})$ for $k = 1, \ldots, 4$. On every env step the command is first perturbed by an independent Gaussian noise of standard deviation $\sigma_a = 0.1$ on each channel, then clipped to $[-1, +1]^{12}$, and finally decoded through the affine map $a_{\mathrm{phys},c} = a_{\min,c} + \frac{1}{2}(a_c + 1)\,(a_{\max,c} - a_{\min,c})$ for each channel c, with $(a_{\min}, a_{\max})$ given in the main text (Section 3.2). The decoded triple drives a Gaussian electron-cyclotron deposition kernel centred at $r_{\mathrm{dep},k}$ with width $w_{\mathrm{dep},k}$ and total absorbed power $P_{\mathrm{in},k}$, summed over the four beams to form the source term S_{tot} in Equation 2. Internally the env also runs a fixed PI loop on V_{loop} to hold I_p at its scenario target with gains $(K_p, K_i) = (5\times10^{-5}, 10^{-5})$ and $V_{\mathrm{loop}} \in [-20, +20]$ V; this loop is invisible to the policy.

C.7 Reward functions and parameters

C.7.1 IsoFlux isoflux-tracking reward

The isoflux-tracking reward referenced in Section 3.1 is built from three raw error channels evaluated at every control step on the privileged plant state. The plasma-current error is $e_{I_p} = |I_p/1000 - I_p^{\mathrm{ref}}|$ in kA, the control-point isoflux error is $e_\psi = \frac{1}{8}\sum_{i=1}^{8} |\psi_{\mathrm{cp},i} - \psi_X|$ in normalised flux units (the mean absolute mismatch between the eight boundary control-point fluxes and the X-point flux), and the X-point poloidal-field error is $e_{B_x} = \|\mathbf{B}_X\|_2$ taken over the two X-point poloidal-field components. Each raw error is mapped to a per-component reward in $[0, 1]$ by the clipped-linear normalisation

$$r_k(e_k) = \mathrm{clip}\left(\frac{e_k^{\mathrm{bad}} - e_k}{e_k^{\mathrm{bad}} - e_k^{\mathrm{good}}},\, 0,\, 1\right), \qquad r_{\mathrm{iso}} = \frac{\sum_k w_k\, e^{\alpha\, r_k}\, r_k}{\sum_k w_k\, e^{\alpha\, r_k}}, \tag{7}$$

with $k \in \{I_p, \psi, B_x\}$. Here α is the tunable mixing coefficient: $\alpha = 0$ recovers a plain weighted mean, $\alpha \to -\infty$ collapses to the worst component, and $\alpha \to +\infty$ collapses to the best.

The default hyperparameters are $w_{I_p} = w_\psi = w_{B_x} = 1.0$ and $\alpha = 0$, so the headline reward is the unweighted mean of the three normalised tracking signals. The clipped-linear cutoffs are $e_{I_p}^{\mathrm{good}} = 0$ and $e_{I_p}^{\mathrm{bad}} = 1.0\,\mathrm{kA}$, $e_\psi^{\mathrm{good}} = 0$ and $e_\psi^{\mathrm{bad}} = 0.1$ in normalised flux units, $e_{B_x}^{\mathrm{good}} = 0$ and $e_{B_x}^{\mathrm{bad}} = 0.05\,\mathrm{T}$. A fourth optional channel (per-step action smoothness, mean squared action delta) is exposed by the same combiner with weight 0 and is therefore inactive in the headline configuration. Baselines that opt into action smoothing flip its weight in the scenario config. Source: `plasmabench_src/envs/PCSVP/env/reward.py`.

C.7.2 Profile NTM-suppression reward

The NTM-suppression reward referenced in Section 3.2 takes the explicit form

$$r_{\mathrm{ntm}} = \exp\big(-a \cdot \max(w - w_{\mathrm{thr}}, 0)\big) \cdot \left(\frac{W_{\mathrm{th}}}{W_{\mathrm{th,ref}}}\right)^b, \tag{8}$$

with default hyperparameters $a = 5.0\,\mathrm{m}^{-1}$, $b = 1.0$, $w_{\mathrm{thr}} = 0\,\mathrm{m}$, and $W_{\mathrm{th,ref}} = 60\,\mathrm{kJ}$ on shot 137220. Setting $w_{\mathrm{thr}} = 0$ makes the penalty active for any non-zero island width and removes the soft plateau that the original `NTMStoredEnergyReward` class exposes, which we found gave a tighter learning signal during baseline tuning. The stored-thermal-energy integral evaluates as $W_{\mathrm{th}} = W_e + W_i$ with $W_X = (3/2)E_{\mathrm{eV}} \int n_X T_X V'\, d\hat{\rho}$, the conversion constant $E_{\mathrm{eV}} = 1.602 \times 10^{-19}$ J/eV absorbing k_B for T in eV; the ion side (n_i, T_i) is closed via the per-scenario ratio $n_i/n_e = 0.385$ and $T_i/T_e = 0.581$ taken from the EAST-grounded ion-profile fit. Episodes terminate the first step at which w exceeds the disruption threshold $w_{\mathrm{disrupt}} = 0.15\,\mathrm{m}$, with a fixed terminal penalty of -10 added to that step's reward.

C.8 Per-episode metric definitions

C.8.1 Tracking error

The tracking error referenced in Section 4 is a per-episode scalar that integrates the normalised deviation of every controlled variable from its scenario reference over the rollout length T. For `IsoFlux` on `isoflux_eq_track` the controlled variables are the eight isoflux control-point fluxes ϕ_k^{cp} tracked against the X-point flux ϕ^{x}, the two X-point poloidal-field components B_x^R and B_x^Z tracked against zero, and the plasma current I_p tracked against its scenario target $I_p^\star$,

$$E_{\mathrm{track}} = \frac{1}{8}\sum_{k=1}^{8} \mathrm{RMS}_t\left(\frac{\phi_k^{\mathrm{cp}}(t) - \phi^{\mathrm{x}}(t)}{\tau_\phi}\right) + \mathrm{RMS}_t\left(\frac{B_x^R(t)}{\tau_B}\right) + \mathrm{RMS}_t\left(\frac{B_x^Z(t)}{\tau_B}\right) + \mathrm{RMS}_t\left(\frac{I_p(t) - I_p^\star}{\tau_I}\right), \tag{9}$$

with $\mathrm{RMS}_t(x) = \left(\frac{1}{T}\sum_{t=1}^{T} x(t)^2\right)^{1/2}$ and per-channel reference scales $\tau_\phi = 10^{-3}\,\mathrm{Wb}$, $\tau_B = 2 \times 10^{-4}\,\mathrm{T}$, $\tau_I = 1.0\,\mathrm{kA}$ matching the operator deadbands of Section 3.1. A schedule that follows every reference exactly scores zero. For `Profile` on `ntm_w_energy` the controlled variables are the magnetic-island width $w(t)$ tracked against zero and the stored thermal energy $W_{\mathrm{th}}(t)$ tracked against $W_{\mathrm{th,ref}} = 60\,\mathrm{kJ}$ on shot 137220, integrated under the same RMS aggregation.

C.8.2 MHD-VIOLATION RATE

The MHD-violation rate is the fraction of control steps on which the scenario's stability proxy crosses its threshold,

$$R_{\mathrm{MHD}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}[g(s_t) > g_{\mathrm{thr}}]\,, \tag{10}$$

with the stability proxy g and threshold g_{thr} chosen per scenario. For `IsoFlux` on `isoflux_eq_track`, $g(s_t)$ is the maximum boundary excursion past the operator deadbands, $g(s_t) = \max\big(\max_k |\phi_k^{\mathrm{cp}} - \phi^{\mathrm{x}}|/\tau_\phi,\ \|B_x\|_2/\tau_B,\ |I_p - I_p^\star|/\tau_I\big)$ with the same channel scales as Equation 9, and $g_{\mathrm{thr}} = 1$ marks the deadband edge. For `Profile` on `ntm_w_energy`, $g(s_t) = w(t)$ is the modified-Rutherford island width of Equation 8 and $g_{\mathrm{thr}} = w_{\mathrm{disrupt}} = 0.15\,\mathrm{m}$ is the locked-mode disruption threshold of Section 3.2. Both choices align R_{MHD} with the failure modes of Section 3.1.

C.8.3 TIME-TO-FAILURE

Let $t^\dagger$ denote the first step at which the scenario's terminal failure-mode predicate fires (boundary tracking violation for `IsoFlux`, $w(t) > w_{\mathrm{disrupt}}$ for `Profile`) or $t^\dagger = T_{\max} + 1$ when the episode runs to its truncation horizon without termination. The time-to-failure metric is

$$T_{\mathrm{TTF}} = \min\left(\frac{t^\dagger}{T_{\max}},\ 1\right), \tag{11}$$

with $T_{\max} = 2000$ outer steps for `IsoFlux` (2.0 s of physical time at the 1 kHz control rate) and $T_{\max} = 400$ for `Profile` (4.0 s at the 100 Hz control rate). Non-terminating episodes score one, and a controller that fails halfway through scores 0.5.

C.8.4 ACTUATOR-SATURATION FRACTION

The actuator-saturation fraction is the fraction of control steps on which any actuator command sits on its per-channel hardware rail,

$$F_{\mathrm{sat}} = \frac{1}{T \cdot d_a}\sum_{t=1}^{T}\sum_{c=1}^{d_a} \mathbf{1}[|a_c(t)| \geq 1 - \delta_{\mathrm{rail}}]\,, \tag{12}$$

where $a_c(t)$ is the normalised raw policy command on channel c before env clipping, d_a is the action dimension, and $\delta_{\mathrm{rail}} = 0.01$ defines a 1% rail margin so that policies crowding the limit count as saturated. For `IsoFlux` ($d_a = 12$) the rails are the symmetric ± 1 normalised PF and CS voltage saturations of Section 3.1 that decode through the EAST PF power-supply envelope. For `Profile` ($d_a = 12$, four beams of three channels) the per-channel rails are $r_{\mathrm{dep}} \in [0, 0.85]$, $w_{\mathrm{dep}} \in [0.15, 0.25]$, and $P_{\mathrm{in}} \in [0.3, 0.8]$ MW per beam from Section 3.2, with the bound nearest the raw command treated as the rail. F_{sat} surfaces the hard-safety challenge axis as a controller-level diagnostic and complements the scenario-level E_{track}, R_{MHD}, T_{TTF} triple of Section 4.

C.9 VECTORISED ENVIRONMENT API SPECIFICATION

The two environments expose a common vectorised contract through thin adapters defined in `plasmabench_src/baselines/env_adapter.py`, namely `PCSVPVecEnvAdapter` for `IsoFlux` and `ProfileSimVecEnvAdapter` for `Profile`. Each adapter is built from a scenario YAML through a factory, `make_env_pcsvp(scenario_yaml, **overrides)` for `IsoFlux` and `make_env(scenario_yaml, extra_cli_args=None)` for `Profile`, both of which resolve scenario-relative file paths against the YAML location so callers may launch from any working directory. All step buffers, observations, and actions are GPU-resident PyTorch tensors of shape `[num_envs, ...]`, so a full rollout of B parallel envs proceeds without any per-step CPU round-trip. A shared `BasePolicy` protocol (also in `env_adapter.py`) defines `policy.act(obs, deterministic) -> action` so that a baseline can switch between the two environments by changing only the factory call.

The public method surface of an adapter is the following.

- `reset() -> Tensor`: reset every parallel env and return the observation tensor of shape `[num_envs, num_obs]` in `cfg.dtype`. Per-row resets during a rollout happen automatically inside `step()`.
- `step(actions: Tensor) -> (obs, rew, dones, extras)`: advance every env by one control-rate-aggregated outer step. `actions` has shape `[num_envs, num_actions]` and is internally clamped to $[-1, +1]$, `obs` has shape `[num_envs, num_obs]`, `rew` has shape `[num_envs]` in `cfg.dtype`, and `dones` has shape `[num_envs]` in `int64` with value 1 on the step at which a row was reset.
- `extras["time_outs"]`: `[num_envs] int64` mask that distinguishes truncations from physical failures, used by every baseline's value-bootstrap path.
- `extras["reward_components"]`: optional per-channel reward dict written by combiner-style reward callables.
- `close()`: release the underlying engine (CUDA streams, RAPTOR solver buffers). Idempotent.
- Inner-env attributes such as `num_envs`, `num_obs`, `num_states`, `num_actions`, `max_episode_length`, `observation_space`, `action_space`, and `cfg` reach the user through the adapter's `__getattr__` forwarder.

Scenario selection and reward customisation are not separate methods. The factory's `**overrides` (`IsoFlux`) or `extra_cli_args` (`Profile`) replace any config field, and the reward function is swapped in place by assigning a new callable to `adapter._env.reward_fn` between episodes. The privileged plant state used for the asymmetric-critic baselines is reachable through `adapter._env`.

Termination semantics. An episode for a given parallel env ends either at `progress_buf >= max_episode_length` (truncation) or on the first step at which the env's `_failure_condition` is met. The physical-failure trigger on `IsoFlux` is $|I_p|$, flux, or B_x runaway, and on `Profile NTMStoredEnergyReward` the trigger is the island width exceeding $w_{\text{disrupt}} = 0.15\,\text{m}$. On the step that triggers either condition, `dones` is set to 1, the configured terminal reward is added to that step's `rew`, and the inner env auto-resets the affected rows in place. Truncations and failures are distinguished only through `extras["time_outs"]`, which the GAE bootstrap reads to apply value-target bootstrapping on truncated rows and zero on failed rows.

A minimal training-loop skeleton (one PPO-style outer iteration on `IsoFlux`) reads:

```
from plasmabench_src.baselines.env_adapter import make_env_pcsvp

env = make_env_pcsvp("plasmabench_src/envs/PCSVP/ \
scenario_files/scenario_config.yaml")
obs = env.reset()                                    # [B, num_obs]
buffer = RolloutBuffer(env.num_envs, T, env.num_obs, env.num_actions)

for update in range(num_updates):
    for t in range(T):
        with torch.no_grad():
            action, logp, value = policy.act_and_value(obs)
        next_obs, rew, dones, extras = env.step(action)
        buffer.add(obs, action, logp, value, rew, dones,
        \extras["time_outs"])
        obs = next_obs

    with torch.no_grad():
        last_value = policy.value(obs)
    advantages, returns = buffer.compute_gae(last_value,
    \gamma=0.99, lam=0.95)

    for epoch in range(update_epochs):
```

```
        for mb in buffer.minibatches(num_minibatches):
            loss = ppo_loss(policy, mb, advantages, returns,
            clip=0.2)
            optim.zero_grad(); loss.backward(); optim.step()

env.close()
```

C.10 Repository layout and entry points

The released code is organised as one Python package, one launcher folder, and four artefact folders, all rooted at the repository top level. `plasmabench_src/` holds the source: `envs/PCSVP/` (the `IsoFlux` plant and its CUDA / fused / RK3 backends), `envs/` (the `Profile` entry hooked into the parent Te_profile package at install time), and `baselines/` (one subpackage per controller plus the shared `env_adapter.py` described above). `scripts/` holds the user-facing entry points, and `ckpt/`, `results/`, and `figure/` hold checkpoints, evaluation tables, and rendered figures, and are gitignored except for the headline artefacts shipped with the release. The package is installed editable through `setup.py` at the repository root.

```
PlasmaBench/
|-- plasmabench_src/
|   |-- baselines/          (PPO, PPO_LSTM, PPO_FO, ...)
|   |   |-- env_adapter.py
|   |   `-- eval_baseline.py
|   |-- envs/
|   |   `-- PCSVP/
|   |-- __init__.py         (Gymnasium ID reservations)
|   `-- filepath.py
|-- scripts/                 (training and evaluation launchers,
plot drivers)
|-- ckpt/                    (released headline checkpoints)
|-- results/                 (per-checkpoint eval tables, .json /
.csv)
|-- figure/                  (rendered .pdf used by the paper)
|-- setup.py
`-- README.md
```

The user-facing entry points under `scripts/` are:

- `run_training.sh` starts a single-config training run and is the launcher used to produce the headline checkpoints.
- `run_paper_eval_best_ckpts.sh` and `run_paper_eval_synthetic.sh` drive the headline evaluation sweeps over the released checkpoints and the synthetic-policy baselines respectively.
- `eval_all_top_ckpts.sh` runs the same evaluation across every top-K checkpoint of a controller for the seed-aggregated tables.
- `collect_eval_metrics.py` aggregates per-rollout results into the per-(scenario, controller, seed) summary JSON consumed by the figure scripts.
- `plot_overlay_5panel.py`, `plot_overlay_5panel_compare.sh`, `plot_overlay_4panel.py`, and `plot_pareto_flux_bx.py` produce the rollout overlays and the Pareto figures.
- `test_env.py` is a smoke test that constructs both adapters, runs a few steps with random actions, and prints buffer shapes and dtypes. It is the recommended first check after install.

The `IsoFlux` plant ships with six scenario variants, namely `Simple3`, `Simple18`, `Shape8`, `Shape12`, `PO165`, and `FO537`, each driven by a YAML configuration under `plasmabench_src/envs/PCSVP/scenario_files/`. Headline results in Section 4 use the `Shape12` twelve-actuator shape-control variant whose action and observation dimensions match the in-text `IsoFlux` description. Six matching Gymnasium identifiers

`PlasmaBench/RZIP{Simple3,Simple18,Shape8,Shape12,PO165,FO537}-v0` are reserved in `plasmabench_src/__init__.py`, and the released code uses the `make_env_pcsvp(scenario_yaml)` factory of Section C.9 as the entry point, with `gym.make()` dispatch slated for a future revision.

C.11 Three levels of reuse and porting

In reference-discharge-based controller development, a controller is designed around the operating point, equilibrium, actuator setup, and target trajectory of a reference discharge; a subsequent matched discharge aims to reproduce these conditions as closely as practical. Matched-equilibrium evaluation therefore addresses a relevant local-feedback problem without establishing generalisation across arbitrary equilibria. Configurable targets, initial states, and perturbations permit task variations, but these capabilities are distinct from the variations actually evaluated. The additional equilibrium tests in Appendix E.1 examine one such generalisation boundary.

New task on a fixed plant. Target trajectories and initial and boundary conditions define a scenario on fixed plant dynamics. Changing these through configuration does not require changing the GPU solver or RL pipeline.

New EAST equilibrium. A new equilibrium requires construction and validation of a device-specific plant bundle: the EFIT equilibrium, linearised A, B, C matrices, equilibrium vectors, diagnostic and control-point mappings, and actuator limits. Configuration selects this bundle; it does not replace the modelling and validation needed to produce it. Once the bundle is available, the same environment, vectorised solver, and training and evaluation pipeline can be reused. The shot-137220 instantiation and held-out evaluation in Section E.1 demonstrate this distinction from the original shot-152444 instance.

Another device. Porting additionally requires a device-specific equilibrium response model, coil and passive-structure topology, diagnostic and actuator mappings, and hardware limits. The common solver and controller pipeline are reusable after the corresponding bundle is constructed, not by changing a scenario YAML alone.

D Reference baseline controllers

D.1 Engineering PID reference

The PID controller used for EAST IsoFlux control serves as an engineering reference alongside the learned controllers. Comparisons align the target trajectory, actuator limits, input delay, reward, episode horizon, and evaluation protocol. This alignment does not make the controllers' information access or engineering priors equivalent: PID incorporates control-design knowledge, sensor-only PPO and PPO-LSTM learn from the available observation channel, and PPO-FO receives privileged plant state. The reference therefore tests whether learned policies improve on an engineering controller under the stated evaluation conditions, not whether equally informed learning algorithms outperform one another. For the Profile task considered here, EAST operation remains largely feed-forward, and no readily available operational closed-loop comparator is supplied in this benchmark. This is a limitation of comparator coverage, not a claim that classical feedback is inapplicable. The added IsoFlux PID reference is distinct from the internal stabilisation and PI loops that are part of the environment dynamics.

D.2 PPO

Method. PPO (Schulman et al., 2017) optimises a clipped surrogate objective with generalised advantage estimation. Each update collects a fixed-length on-policy rollout from all parallel environments, computes per-step advantages, and runs several minibatch SGD epochs against the clipped policy objective and a value-function loss. The actor and the critic are two separate two-layer multilayer perceptrons (MLPs) of 256 hidden units with Tanh activations and orthogonal initialisation; continuous actions are sampled from a diagonal Gaussian whose mean is the actor output and whose log standard deviation is a learned vector clamped to $[-2, 2]$.

Table 9: PPO and PPO-LSTM training hyperparameters used to produce the headline checkpoints. Values that span both columns are shared across `IsoFlux` and `Profile`; LSTM-specific rows apply to PPO-LSTM only.

Hyperparameter	`IsoFlux`	`Profile`
Algorithm		
Total environment steps	3×10^7	
Optimiser	Adam, $\epsilon = 10^{-5}$	
Learning rate	3×10^{-4}, linearly annealed to 0	
Discount factor γ	0.99	
GAE λ	0.95	
Clip ratio ϵ	0.2	
Value-loss coefficient	0.5	
Entropy coefficient	0.0	
Maximum gradient norm	1.0	
Rollout length per env	32 steps	
Update epochs	4	
Minibatches per update	4	
Hidden width (MLP)	256, Tanh, orthogonal init	
LSTM hidden size (PPO-LSTM)	128	
LSTM layers (PPO-LSTM)	1	
Environment		
Parallel envs (training)	256	1024
Episode length	2000 steps	400 steps
Control rate	1 kHz	100 Hz
Numerical precision	fp64	fp32

Implementation. Our implementation closely follows the continuous-action variant in CleanRL (Huang et al., 2022) (`ppo_continuous_action.py`). Project-side modifications are limited to running observation normalisation, correct handling of the env's truncation flag so that the generalised-advantage bootstrap at `max_episode_length` does not contaminate failure transitions, and periodic evaluation rollouts under a fixed eval seed for cross-baseline comparison. Both the actor and the critic consume the env's sensor-only actor observation; the privileged plant-state channel of the asymmetric interface is left unused, so PPO and PPO-LSTM share a single observation tensor at every step. Source: `plasmabench_src/baselines/PPO/ppo.py`.

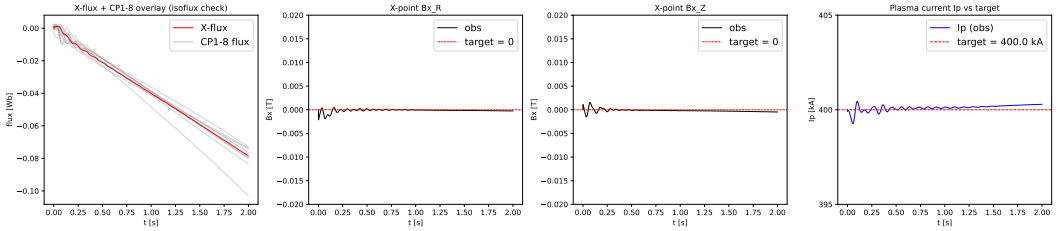
D.3 PPO-LSTM

Method. PPO-LSTM extends PPO with a recurrent actor and critic that share a recurrent trunk. The observation is first projected by a two-layer MLP, fed through a single-layer LSTM, then mapped by a one-layer head MLP into separate actor-mean and critic-value linear projections. Truncated backpropagation through time aggregates gradients over each rollout. The recurrent core is intended to characterise the partial-observation regime exposed by the sensor-only actor channel, where access to past observations may compensate for the limited information available at any single step.

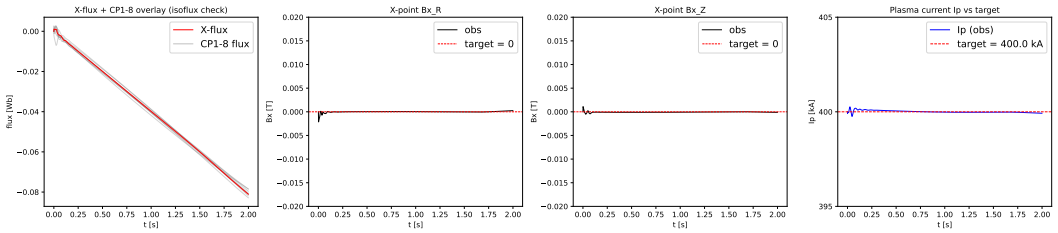
Implementation. The recurrent core mirrors the CleanRL (Huang et al., 2022) reference implementation (`ppo_atari_lstm.py`) ported to continuous control. Hidden states are reset at episode boundaries through a per-env done mask broadcast into the LSTM cell and hidden tensors. All non-recurrent hyperparameters match the feed-forward variant. Source: `plasmabench_src/baselines/PPO_LSTM/ppo_lstm.py`.

D.4 Training hyperparameters

Both controllers share the algorithm-side configuration listed in Table 9; environment-specific differences are confined to the parallel-rollout count, the per-episode horizon set by each environment's control rate, and the numerical precision of the underlying solver.



(a) PPO



(b) PPO-LSTM

Figure 7: Trained-policy closed-loop rollouts on `IsoFlux isoflux_eq_track`. Each 1×4 row shows the X-point flux (red) overlaid with the eight boundary control-point fluxes CP1–CP8 (grey, isoflux co-drift check), the X-point poloidal-field components $B_{x,R}$ and $B_{x,Z}$ around their zero target, and the plasma current I_p on a ± 5 kA window around its target schedule.

D.5 Complete inference results

Figure 7 shows shot-152444 `IsoFlux isoflux_eq_track` examples of the trained PPO and PPO-LSTM policies at their best evaluation checkpoints, with mean evaluation reward and actuator-saturation fraction matching Table 3. These illustrate the current main instance, not the supplemental shot-137220 instance. Each panel set is a 1×4 summary row. The leftmost panel overlays the eight boundary control-point flux traces (CP1–CP8, grey) on the X-point flux (red), so co-drift of the bundle around the X-flux trace is the visual signature of a closed isoflux loop on the magnetic side. The next two panels show the X-point poloidal-field components $B_{x,R}$ and $B_{x,Z}$ regulated near zero on a tight zero-centred axis. The rightmost panel shows the plasma current I_p tracking its target schedule on a ± 5 kA window. Both controllers keep the eight control-point fluxes co-drifting with the X-point flux over the full 2 s episode, regulate $B_{x,R}$ and $B_{x,Z}$ near zero, and hold I_p on its target. PPO-LSTM achieves the higher mean evaluation reward of the two, consistent with the actor's recurrent state aggregating information that the sensor-only observation does not expose at any single step. PPO operates at the higher actuator-saturation fraction on this instance, consistent with the partial-observation policy's reliance on actuator headroom to compensate for missing information (Section 4.3).

D.6 Training compute resources

Each baseline (PPO, PPO-LSTM, PPO-FO) is trained on a single GPU with 80 GB of memory at the recommended batch size of Table 9. A full training run takes approximately 9 hours of wall-clock time on either environment.

E Additional Experimental Results

E.1 Additional IsoFlux instantiation on EAST shot 137220

The shot-137220 `IsoFlux` instantiation replaces the equilibrium bundle and target configuration while retaining the solver and RL pipeline. Policies were newly trained on equilibrium A: this is not zero-shot transfer from shot 152444. The resulting checkpoints were frozen and evaluated on

A and held-out equilibrium B without retraining or retuning, together with the same PID reference. Table 10 reports both cases.

Table 10: `IsoFlux` evaluation on equilibria A and B of EAST shot 137220. Learned checkpoints trained on A and the PID reference are fixed for the held-out B evaluation.

	Equilibrium A		Equilibrium B (held out)	
Controller	Reward	Saturation	Reward	Saturation
PID	0.9539	0.0000	0.9339	0.0000
PPO	0.5952	0.2455	0.5465	0.2636
PPO-LSTM	0.8019	0.8240	0.4795	0.8592
PPO-FO	0.9536	0.6470	0.2722	0.5145

Among the learned controllers, reward ranks PPO-FO $>$ PPO-LSTM $>$ PPO on A, but PPO $>$ PPO-LSTM $>$ PPO-FO on B. All three lose reward on B, whereas PID remains comparatively stable with zero reported saturation. The experiment supports reuse of the environment across plant bundles while revealing limited cross-equilibrium generalisation of these learned policies; it does not establish reliable cross-shot or cross-device transfer.

E.2 Local Profile diffusivity sensitivity

Frozen `Profile` checkpoints were evaluated with a local $\pm 2\%$ perturbation of χ_e, using a cosine window centred at $\hat{\rho} = 0.50$ with half-width 0.15. Table 11 includes the nominal setting and both perturbations.

Table 11: Local `Profile` χ_e sensitivity with frozen policies. Each setting reports evaluation reward and actuator saturation.

	-2%		Nominal		$+2\%$	
Controller	Reward	Saturation	Reward	Saturation	Reward	Saturation
PPO	0.6421	0.0090	0.6420	0.0102	0.6416	0.0138
PPO-LSTM	0.6420	0.0871	0.6412	0.0877	0.6406	0.0886
PPO-FO	0.6340	0.1773	0.6329	0.2094	0.6318	0.2433

Within this tested local range, the reward ordering PPO $>$ PPO-LSTM $>$ PPO-FO is maintained. PPO-FO shows the largest saturation response, so stable reward rankings do not imply insensitive actuator-margin conclusions. This is a local closure-parameter perturbation study, not evidence that the approximately 2% reference-trajectory output discrepancy is harmless: parameter perturbations and trajectory discrepancies are different quantities.

E.3 Parity-error distribution

Figure 8 reports the diagnostic-channel residuals that complement the main-text $T_e(\hat{\rho})$ parity (Figure 4) for scenario `test_data_3`. Snapshots are taken at $t = 0, 1, 5$ ms during the initial transport relaxation, before the EC step. `Profile` and RAPTOR agree on every panel except for a mild mid-radius offset on χ_e that carries through to $\sigma_\parallel$ and the bootstrap current j_{bs}.

F Limitations and future work

The evaluated instances cover a limited set of EAST configurations: the original shot-152444 `IsoFlux` instance, equilibria A and B of shot 137220 for `IsoFlux`, and the shot-137220-derived `Profile` configuration. The `IsoFlux` free-boundary linearisation is accurate inside a deviation neighbourhood of its reference equilibrium and loses fidelity at strong shape excursions or off-equilibrium transients beyond the Taylor-expansion radius. Three-dimensional MHD modes and advanced divertor topologies are absent from IsoFlux, scoping it to axisymmetric magnetic-equilibrium control rather than full-discharge stability arbitration. On `Profile` the electron density n_e is held

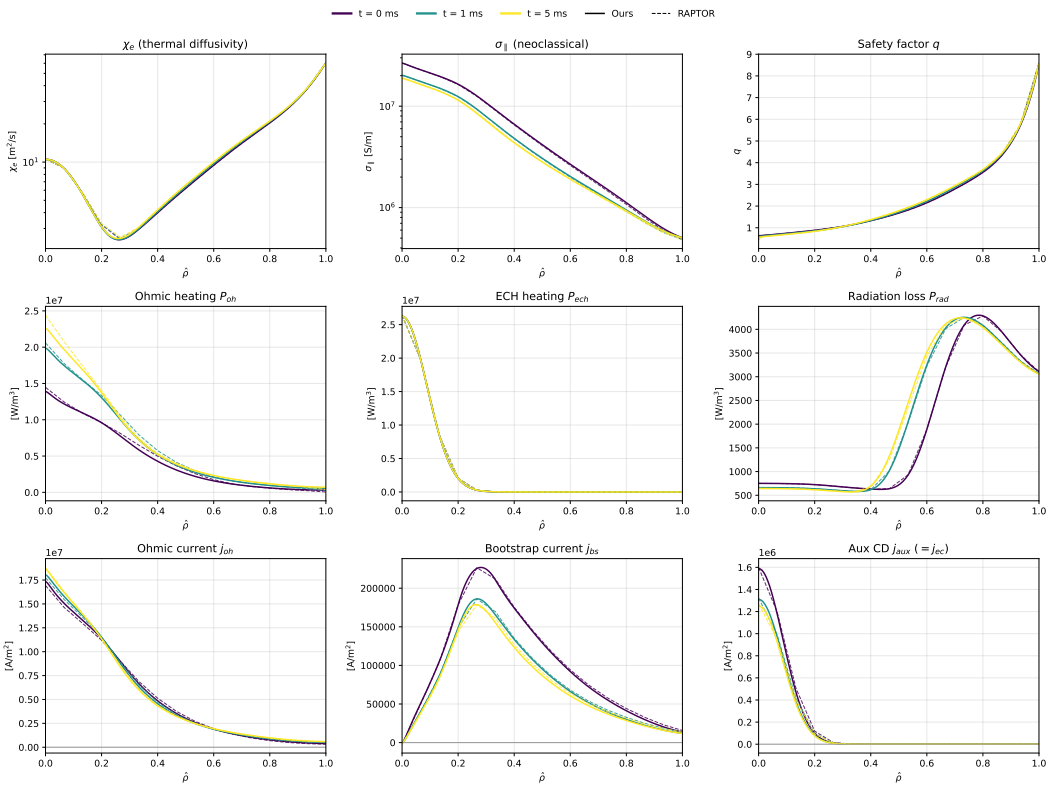


Figure 8: Supporting diagnostic channels for the `Profile` parity check on `test_data_3`, with three early-time snapshots at $t = 0, 1, 5$ ms overlaid in viridis colours. Solid traces are `Profile`, dashed are RAPTOR. Top row: thermal diffusivity χ_e, neoclassical conductivity $\sigma_\parallel$, safety factor q. Middle row: ohmic heating P_{oh}, ECH heating P_{ech}, radiation loss P_{rad}. Bottom row: ohmic, bootstrap, and auxiliary-CD currents j_{oh}, j_{bs}, j_{aux}.

frozen at its initial profile and the heat-diffusivity closure is the Felici four-term ad-hoc form (Felici et al., 2011) or the Bohm-gyro-Bohm alternative (Erba et al., 1997), both calibrated transport models rather than first-principles closures, so density-control studies and quantitative TGLF-level comparisons are out of scope. The synthetic ECE forward model and the modified-Rutherford NTM closure with a multiplicative χ_e short-circuit at the rational surface (Buttery et al., 2000; Sauter et al., 1997) are documented surrogates rather than calibrated diagnostics or first-principles three-dimensional transport. Future releases target whole-discharge ramp-up and ramp-down scenarios, an evolving n_e channel, a quasilinear TGLF-quality transport closure, internal-transport-barrier and pedestal scenarios, and cross-tokamak grounding on KSTAR, MAST-U, and ITER baselines so that cross-machine transfer becomes a quantifiable axis.

G Broader Impacts

The method we propose offers a means to actively interact with the physical world and achieve specific control objectives. This approach presents significant opportunities for various domains, including fluid control, robotic control, controllable nuclear fusion, and more. However, it is imperative to remain vigilant about potential negative consequences that may arise from this technology to prevent its application to unethical or illegal control issues.