Advantage-level Aggregation Reinforcement Learning for X-point Target Magnetic Configuration Control in an EXL-50U Experiment-Calibrated Simulation Environment

Siqi Ding[1,2,3], Xuanhe Wang[1,2,3], Pei Guo[4], Guoyang Shi[1,2,3], Changquan Yu[5], Yiting Wang[6], Xianming Song[1,2,3], Xiang Gu[1,2,3], Zhengyuan Chen[1,2,3], Lei Xing[1,2,3], Yapeng Zhang[1,2,3], Jianguo Chen[1,2,3] and Tianyuan Liu[1,2,3,*]

[1] *Beijing ENN Fusion Energy Science and Technology Co., Ltd., Beijing 101111, China*
[2] *Beijing Key Laboratory of High Magnetic Field Spherical Torus Fusion Energy, Beijing 101111, China*
[3] *Hebei Key Laboratory of Compact Fusion, Langfang 065000, China*
[4] *Key Laboratory of Materials Modification by Beams of the Ministry of Education, School of Physics, Dalian University of Technology, Dalian 116024, China*
[5] *School of Nuclear Science and Engineering, East China University of Technology, Nanchang 330013, Jiangxi, China*
[6] *School of Mechanics and Engineering Science, Peking University, Beijing 100084, China*

** Corresponding authors: Tianyuan Liu (liutianyuan@enn.cn).*

Abstract

Managing divertor heat loads is a central challenge for compact, high-power tokamaks. To increase local flux expansion and create a dissipation volume magnetically decoupled from the core, the compact spherical-torus design EHL-2 adopts the X-point target (XPT) as its reference divertor configuration. Realising these benefits requires the secondary X-point to remain on the divertor leg; its displacement can degrade the intended topology and exhaust geometry. Existing experiments, including recent EXL-50U discharges, still rely mainly on precomputed feedforward waveforms with proportional–integral–derivative (PID) loops on global quantities. The secondary null is not yet under dedicated closed-loop feedback, so XPT operation remains repeatable but not yet routine. We formulate reconstructed-null XPT feedback as a multi-objective reinforcement learning (RL) control problem in a free-boundary environment calibrated to EXL-50U discharge #13906. To address the strong coupling among plasma current, boundary shape, and multiple null constraints—where conventional reward-level scalarisation collapses objective-specific temporal credit into a single advantage—we develop Advantage Aggregation (AdvA), which preserves objective-wise temporal credit assignment before worst-objective-aware nonlinear scalarisation and introduces a controlled residual correction to the policy update. Instantiated with proximal policy optimisation (PPO), AdvA-PPO is evaluated against conventional reward-level PPO and the experiment-derived feedforward-plus-PID baseline under nominal operation, measurement uncertainties, and unseen initial equilibria. On the nominal 500 ms rollout, AdvA-PPO raises the mean worst-channel score from 0.23 to 0.81 relative to Reward-PPO and reduces mean X-point flux root-mean-square error (RMSE) by about 20×. Under combined measurement uncertainties, it is the only learned controller that completes the horizon while retaining a usable XPT shape. Multi-initialization fine-tuning further enables a single AdvA-PPO policy to complete full-horizon operation across both divertor and limiter initial equilibria. These results provide a simulation-based foundation for future real-time XPT validation on EXL-50U.

Keywords: X-point target divertor; magnetic configuration control; artificial intelligence for fusion; multi-objective reinforcement learning; tokamak

1 Introduction

Magnetic-confinement fusion offers a route to abundant, low-carbon energy, and the tokamak is among its most developed concepts. A reactor-scale tokamak, however, must exhaust intense particle and power fluxes without exceeding the steady-state limits of plasma-facing materials. Extrapolations for compact, high-power devices illustrate the severity of this challenge [1]. Advanced divertor configurations (ADCs) address it by reshaping the scrape-off-layer magnetic geometry to increase connection length, flux expansion, wetted area, and the volume available for power and momentum dissipation. Their exhaust benefit therefore depends not only on edge-plasma physics, but also on the ability to establish and maintain the intended magnetic topology.

Experiments have realized several ADC families, including the snowflake (SF), quasi-snowflake (QSF), Super-X, X-divertor (XD), and X-point target (XPT). Their exhaust benefits are well established: dedicated SF control on DIII-D sustained a $\sim 2.5\times$ peak-heat-flux reduction for 2–3 s, NSTX SF experiments reduced the peak heat flux between edge-localised modes (ELMs) by 3–5×, and the three-X-point TCV Jellyfish reduced the peak parallel target heat flux by at least 50% relative to a single null [2, 3, 4]. Control maturity, however, does not follow directly from exhaust performance. As the number of nulls increases, plasma current, the main boundary, divertor legs, and strike points must share a limited set of poloidal-field (PF) actuators, so improving one quantity can degrade another.

Among these ADCs, the XPT is the configuration most directly relevant to the EHL-2 divertor route. It places a secondary magnetic null in the divertor leg, away from the confined plasma, thereby combining a long leg with strong local flux expansion and a dissipation volume that is magnetically decoupled from the core [5, 6]. Experiments on TCV and MAST-U show that this topology improves divertor conditions, sustains an X-point radiator regime, and passively rejects disturbances near the secondary null [7, 8, 9, 10]. These exhaust properties are conditional on the magnetic topology: the secondary null must remain on the divertor leg of the primary separatrix to preserve the intended flux expansion and magnetically decoupled dissipation volume. Secondary-null displacement can therefore degrade the functional geometry even when plasma current and global position remain acceptable. XPT control is consequently not an ancillary shape-control task, but an enabling requirement for converting magnetic access to the configuration into repeatable exhaust capability.

For EHL-2 the XPT is not one option among several but the main reference divertor configuration: integrated poloidal-field-system and free-boundary equilibrium design shows that an outer XPT is magnetically accessible within the coil set, and edge-transport calculations predict the associated heat-load and detachment behaviour [11, 12]. Because the EHL-2 evidence is design and simulation based, it defines both the target topology and the control requirement that motivate the present study.

Motivated by this route, XPT equilibria have now been realized in multiple EXL-50U discharges. Figure 1 shows a representative reconstruction from discharge #13906. These experiments establish that the target topology is magnetically accessible on the device, but access in a reconstructed equilibrium is not equivalent to regulating the performance-critical secondary null throughout a discharge. The experiments still rely primarily on a classical feedforward-plus-PID scheme (FF+PID): precomputed PF-coil feedforward waveforms supplemented by PID feedback on global quantities. Reaching the desired topology still requires shot-to-shot tuning, and the secondary X-point is not yet maintained by a dedicated real-time feedback loop. XPT operation on EXL-50U should therefore be regarded as repeatably demonstrated but not yet routine. A similar commissioning burden was reported on MAST-U, where the first steady XPT required three successive discharges to tune the feedforward virtual-circuit waveform [13]. The EXL-50U experiments provide physically realized target equilibria and a basis for calibrating the control environment, rather than evidence of closed-loop secondary-null control.

This limitation is shared by the small number of XPT experiments worldwide, as summarized in Table 1. On MAST-U the XPT could only be obtained transiently in the 2021 campaign under pure feedforward operation; steady-state XPT became possible once boundary variables were placed under feedback, but the real-time reconstruction did not estimate the secondary X-point. That null was instead created by feedforward virtual-circuit commands driving the local radial and vertical fields toward zero at a user-specified location, which left a finite and slowly growing field offset and a ∼10 cm discrepancy in the achieved null position, and required three successive discharges to converge [13]. On TCV, published XPT campaigns quantify exhaust performance, the X-point radiator regime, and detachment dynamics under the device's divertor shape control; they do not report a feedback loop that treats the secondary null as an explicit real-time control variable together with the main boundary [14, 7, 8, 10]. Offline reconstruction can recover multi-null topologies after a shot, and real-time reconstructors already support selected boundary and primary-X quantities on several devices [13, 15]. For XPT, the enabling control requirement is therefore not reconstruction in isolation, but *closing the loop on the secondary null*: keeping that null on the divertor leg of the primary separatrix so that the intended flux expansion and magnetically decoupled dissipation volume survive disturbances and shot-to-shot drift. Absent such feedback, secondary-null placement remains a feedforward commissioning task even when I_p and the global boundary track well—as illustrated by the MAST-U residual field offset and multi-discharge tuning [13].

Regulating multi-null divertor geometry therefore requires one of two interfaces. The first closes feedback on explicit null states—null positions and related flux constraints obtained from a real-time equilibrium reconstruction or a dedicated null observer—together with I_p and the last closed flux surface (LCFS).

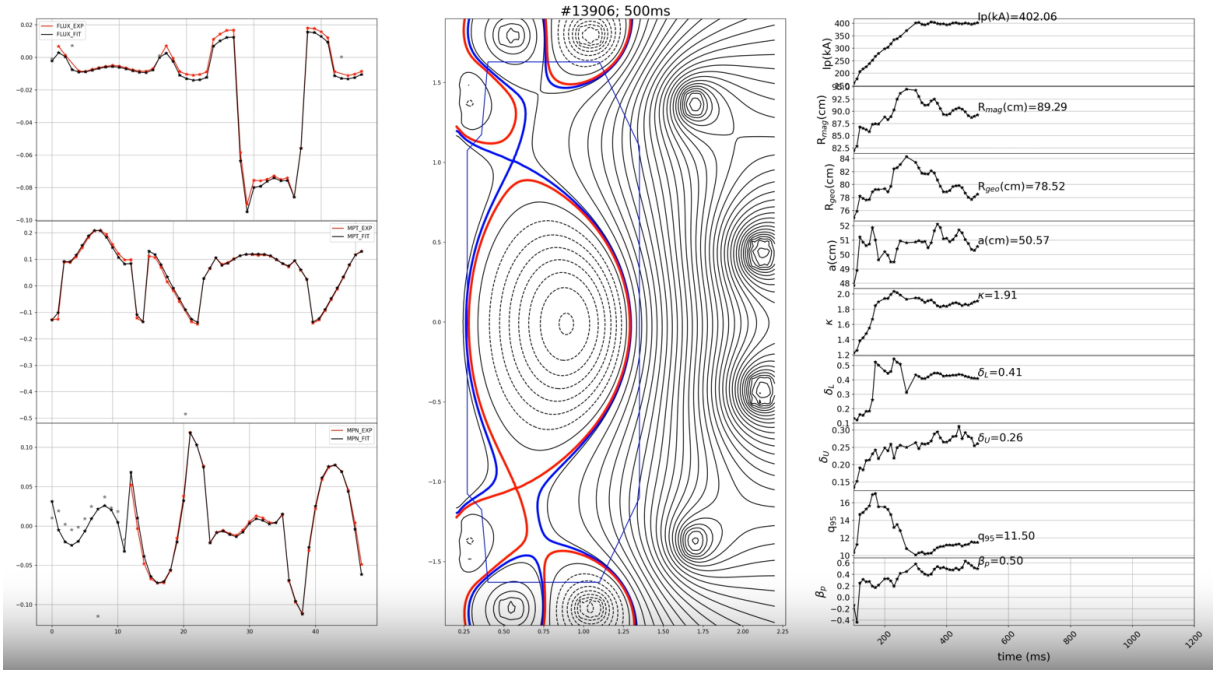

Figure 1: Representative experimental XPT equilibrium in EXL-50U discharge #13906 at $t = 500\,\mathrm{ms}$. The reconstructed magnetic topology demonstrates experimental access to XPT operation.

Table 1: Experimental XPT magnetic control to date. "Real-time secondary null in FB" asks whether a real-time estimate of the secondary null enters the feedback law (as opposed to feedforward coil programming alone).

Device	Method	FB targets	Real-time secondary null in FB
TCV [14, 7, 8, 10]	Divertor shape control	I_p, boundary / shape	No
MAST-U [13, 9]	LEMUR + FF $B_{r,z} \to 0$	R_{OUT}, R_{IN}, Z_X	No
EXL-50U (this exp.)	FF+PID	I_p, R, Z	No

The second is reconstruction-free: a controller maps magnetic probes, flux loops, and coil currents directly to actuator commands, so that multi-null geometry is regulated without feeding a real-time equilibrium into the loop. Deep RL on TCV has demonstrated the latter for snowflake and Jellyfish targets, including a snowflake transition that reduced X-point separation from 34 to 6.6 cm and the three-null Jellyfish configuration, with X-point structure latent in the learned policy rather than exposed as reconstructed feedback channels [16, 17, 4]. Related reconstruction-free magnetic control has also been shown on DIII-D and WEST [18, 19]. Those results establish that multi-null feedback need not be reconstruction-based; they do not, however, close an XPT loop in which a *secondary* null on the divertor leg is an explicit real-time feedback variable. On EXL-50U the reconstruction-free magnetics interface for multi-null divertor control has not yet been validated, whereas the planned plant stack routes equilibrium features from the real-time reconstruction code PTEFIT to the controller [20, 15]. We therefore first develop the explicit, reconstruction-based route: closed-loop XPT control on reconstructed secondary- null position and flux together with I_p and the LCFS—an interface that, to our knowledge, has not been demonstrated in XPT experiments to date. We do not claim superiority of either interface; the present choice matches the near-term EXL-50U control architecture and makes the secondary-null channels auditable for multi-objective learning. Model-predictive control provides a non-learning alternative for shape regulation [21], but a validated control-oriented model of the coupled EXL-50U XPT nulls is not presently available.

XPT control is multi-objective on a single trajectory: plasma current, the LCFS, and multiple null position/flux channels must remain jointly acceptable under shared PF actuators, while the identity of the worst channel can change over the discharge. Fusion RL typically scalarizes these channels in the reward before estimating one advantage [16, 17]. In robotics and multi-task learning, related conflicts are instead attacked inside the learner—notably by projected conflicting gradients (PCGrad) [22, 23] and by preference-conditioned policies with multi-head advantage decomposition [24].

Those methods mainly mitigate gradient interference or support preference trade-offs across tasks.

Reconstructed-null XPT control is harder-coupled: a weak null/flux channel can fail the topology even when the scalar return looks acceptable, and tokamak magnetic RL still rarely keeps objective structure through credit assignment. The relevant design choice is therefore where nonlinear scalarization occurs relative to objective-wise temporal credit.

The gap addressed here is therefore twofold. At the control-system level, existing XPT operation does not close feedback on reconstructed secondary-null position and flux together with the plasma current and boundary. At the algorithmic level, conventional reward-level scalarization discards objective identity before temporal credit assignment. We develop AdvA along the second axis: a shared critic backbone with one value head and one generalised advantage estimation (GAE) [25] per channel, with worst-objective-aware softmin aggregation (implementation name SmoothMax; $\alpha < 0$) applied only after advantage estimation, augmented by a controlled component residual on the aggregated update. On EXL-50U, PTEFIT already provides real-time equilibrium reconstruction for boundary quantities at the centimetre level [20, 15]; whether its secondary-null position and flux estimates are accurate enough for closed-loop XPT feedback remains to be established experimentally. The present study therefore validates the control interface and AdvA in an experiment-calibrated free-boundary simulation, as a precursor to PTEFIT-in-the-loop tests.

The contributions address these gaps as follows:

1. We establish an AI-enabled multi-objective framework for closed-loop XPT magnetic configuration control in an experiment-calibrated EXL-50U free-boundary simulation environment. The framework jointly regulates plasma current, LCFS geometry, X-point positions, and X-point flux constraints through shared poloidal-field actuators.

2. We develop Advantage Aggregation (AdvA) for coupled multi-objective magnetic control. By retaining objective-specific value heads and generalised advantage estimates before nonlinear scalarisation, together with a controlled residual correction, AdvA preserves objective-wise temporal credit assignment and supports coordinated policy optimisation.

3. We evaluate the framework against conventional reward-level RL and the experiment-derived FF+PID baseline across nominal operation, measurement uncertainties, and unseen initial equilibria, together with multi-initial adaptation. The results characterise its control capability, robustness, transferability, and remaining challenges for simulation-based XPT magnetic control.

The present evidence is simulation-based; closed-loop XPT control on EXL-50U with PTEFIT-supplied null states remains future experimental work.

2 Methods

2.1 EXL-50U XPT control problem

Figure 2 summarises the EXL-50U control stack. Twelve active coil circuits—the central solenoid (CS), ten Poloidal Field coils (PF1–PF10), and the vertical-stability coil (VS)—regulate the plasma current and magnetic configuration. Learned policies command CS and PS1–PS10 (11 voltages); VS remains under a separate PID loop due to highe frequency for every controller reported here. The target XPT has two primary nulls near the confined boundary and two secondary nulls in the divertor legs. We use one-based labels X_1–X_4: X_2, X_3 are the upper/lower inner primary pair and X_1, X_4 the upper/lower outer secondary pair. The plant is strongly coupled—each coil simultaneously affects I_p, the LCFS, and all four nulls—so improving one X-point condition can displace another null or deform the boundary.

Training and evaluation close the loop through the calibrated fast free-boundary Grad–Shafranov evolutive solver (FGE) [26] (Sec. 2.2.2): FGE supplies I_p, coil currents, LCFS, and X-point features; the policy returns the 11 coil voltages; the environment advances and scores the next step. The same observation and command interfaces are already wired for machine deployment (Figure 2): diagnostic and coil signals pass through reflective memory (RFM) to real-time equilibrium reconstruction, the policy runs under TensorRT, and coil commands return to the power supplies. With that end-to-end path in place, a controller validated in FGE is ready for on-machine closed-loop tests; the experiments below remain simulation-based and do not yet report PTEFIT-in-the-loop XPT feedback on EXL-50U.

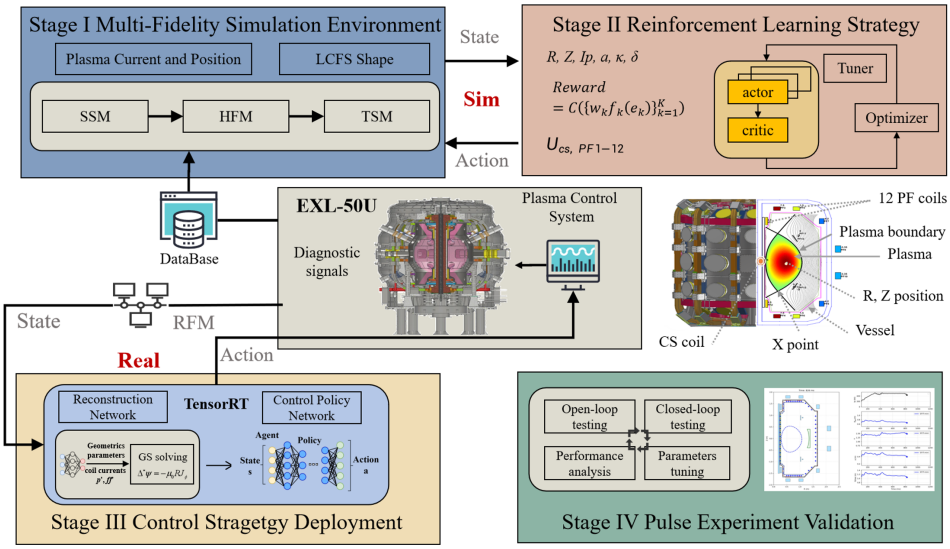

Figure 2: End-to-end architecture for reconstructed-null XPT control. Simulation closes the loop through FGE; deployment swaps FGE observations for real-time reconstruction while keeping the same state and coil-command interfaces (11 policy coils; VS under PID).

2.2 FGE simulation environment

2.2.1 Fast free-boundary Grad–Shafranov evolutive solver

All controllers interact with FGE [26], a control-oriented free-boundary equilibrium evolution solver in the Matlab EQuilibrium suite (MEQ), previously used for RL magnetic control on TCV [16]. Given PF-coil voltages, FGE advances the axisymmetric equilibrium on resistive timescales by coupling three blocks. The poloidal flux satisfies the Grad–Shafranov equation

$$\Delta^*\psi = -2\pi R\mu_0 j_\phi = -4\pi^2 \left(\mu_0 R^2 p'(\psi) + TT'(\psi)\right), \tag{1}$$

with free profiles $p'(\psi)$ and $TT'(\psi)$ fixed by scalar constraints (here β_p and q_0). Active and passive conductor currents I_e obey the circuit equation

$$\begin{pmatrix} V_a \\ 0 \end{pmatrix} = M_{ee}\dot{I}_e + M_{ey}\dot{I}_y + R_e I_e, \tag{2}$$

where V_a are the applied coil voltages, I_y is the plasma current distribution, and M_{ee}, M_{ey}, R_e are the mutual-inductance and resistance matrices. The bulk plasma current is closed by a rigid Ohmic model (OhmTor-rigid in [26]),

$$0 = L_p\dot{I}_p + M_{pe}\dot{I}_e + R_p\left(I_p - I_{ni}\right), \tag{3}$$

with plasma self-inductance L_p, plasma–conductor mutual M_{pe}, bulk resistance R_p, and non-inductive current I_{ni}. Each step returns I_p, coil currents, the LCFS, X-point positions, and fluxes for observations and rewards. On EXL-50U we use a 66×65 (R, Z) grid, 544 passive filaments plus 12 active circuits, and a fixed 1 ms step; β_p and R_p are held at discharge-calibrated values (Sec. 2.2.2). The model is restricted to the magnetic degrees of freedom needed for XPT topology regulation and is fast enough for parallel policy training.

2.2.2 Calibration and validation on #13906

The environment is calibrated to EXL-50U discharge #13906, which sustained an XPT during the I_p flat top near 400–700 ms. Reference equilibria come from the LIUQE tokamak equilibrium reconstruction code [27]; we use the 500–700 ms window. Shot-specific inputs are the reconstructed $\beta_p(t)$ and $q_0(t)$ as profile constraints in Eq. (1), and a constant bulk plasma resistance R_p chosen so that simulated I_p tracks the experiment over that window. Coil and vessel geometry, inductances, resistances, and power-supply parameters follow engineering calibrations (including vacuum shots) and are held fixed.

For validation, FGE starts from the LIUQE equilibrium at $t = 0.5$ s and advances to $t = 0.7$ s under the recorded coil voltages. Figure 3 overlays boundaries and nulls; Figure 4 compares I_p, R, $R_{\max}$, and κ; Table 2 summarises bias, RMSE, and maxima. The XPT topology (all four nulls) is preserved; I_p RMSE

is 0.8% (1.6% at worst), and LCFS/X-point offsets stay at the centimetre scale of the grid cell ($\approx 1.7\,\mathrm{cm}$), which matches the tolerances of the configuration-control task. Residual mismatches concentrate in the divertor legs and in small biases of κ and $R_{\max}$, consistent with the reduced profile model.

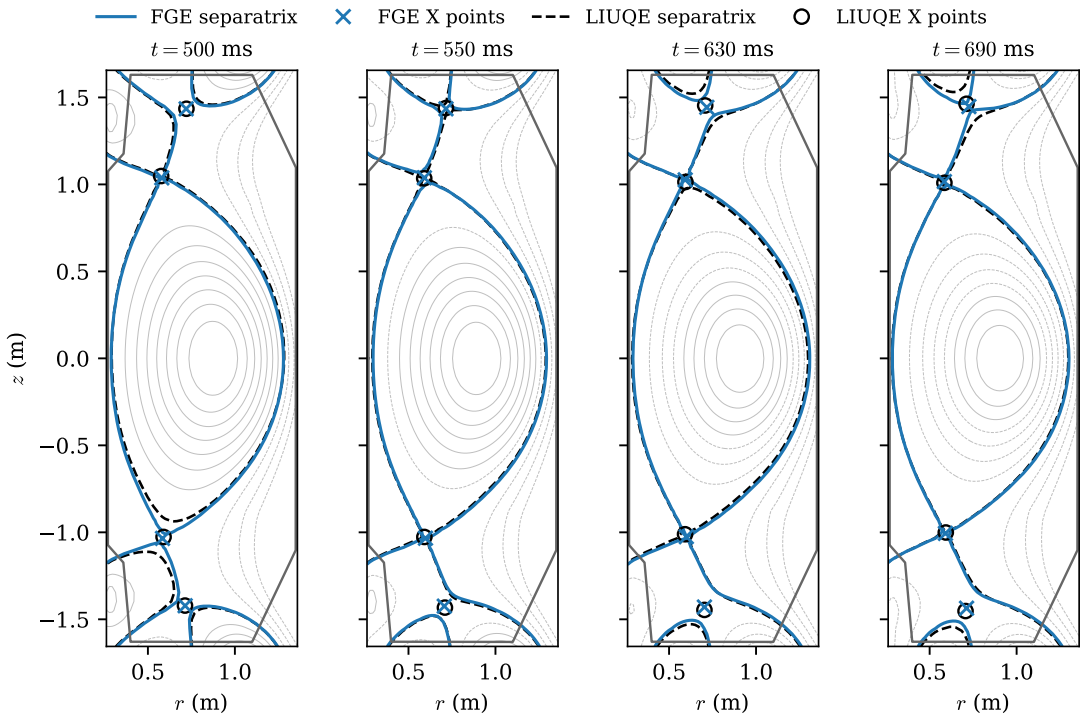

Figure 3: Free-boundary equilibria from the FGE validation run of EXL-50U discharge #13906 at $t = 0.50$, 0.55, 0.63, and 0.69 s (left to right). Each panel overlays the simulated separatrix (solid blue) and X points (blue crosses) on the experimental LIUQE-reconstructed separatrix (dashed black) and X points (open circles); thin gray curves show simulated flux surfaces and the dark gray contour marks the limiter.

Table 2: Deviations between the FGE validation run and the LIUQE reconstruction of discharge #13906 over 0.5–0.7 s: time-averaged signed deviation (bias, simulation minus reconstruction), root-mean-square error (RMSE), and maximum absolute deviation. The LCFS row is based on the symmetrized mean distance between the simulated and reconstructed boundary contours, and the X-point rows on the Euclidean position offsets averaged over the upper and lower X point of each type; these distances are nonnegative by construction, so no bias is reported.

Quantity	Bias	RMSE	Maximum
I_p (kA)	+1.4	3.1	6.5
R (mm)	+4.7	5.3	11.4
$R_{\max}$ (mm)	+7.6	8.7	16.7
κ	+0.045	0.051	0.101
LCFS distance (mm)	—	13.8	30.5
Primary X points X_2, X_3 (mm)	—	18.3	38.0
Secondary X points X_1, X_4 (mm)	—	13.5	21.6

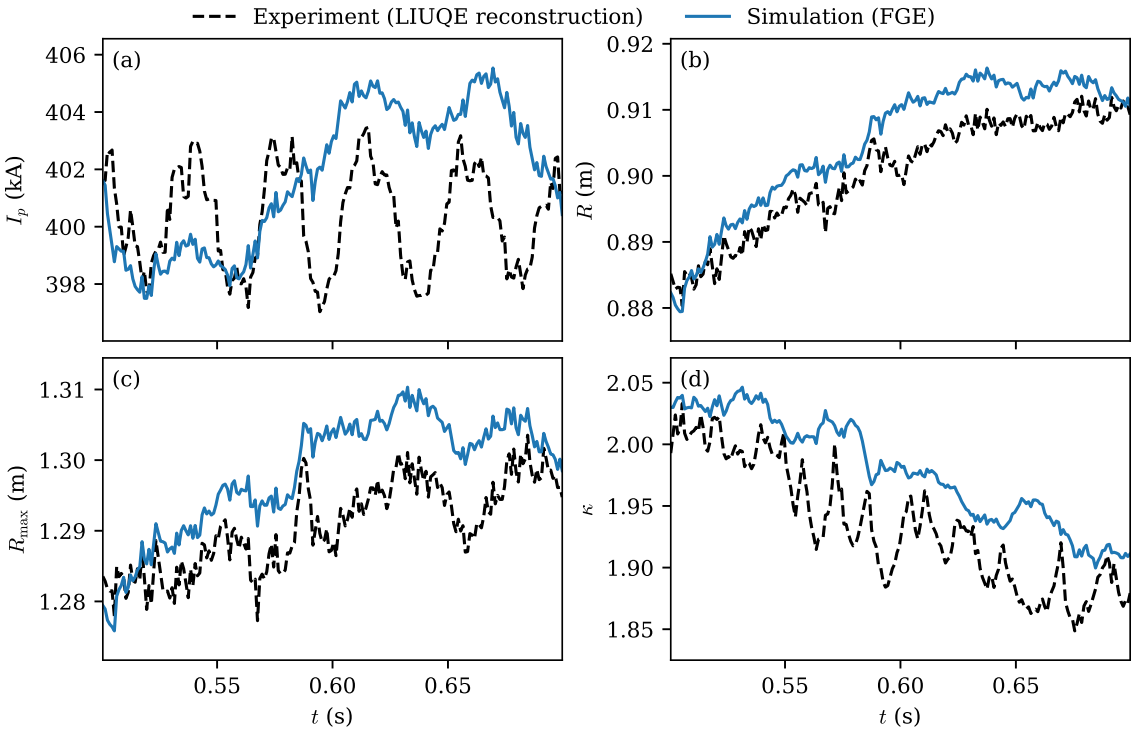

Figure 4: Time traces from the FGE validation run of EXL-50U discharge #13906 (solid blue) compared with the experimental LIUQE reconstruction (dashed black): (a) plasma current I_p; (b) plasma radial position R; (c) outer radial extent $R_{\max}$ of the LCFS; (d) LCFS elongation κ.

2.3 Control methods

2.3.1 Baseline: FF+PID

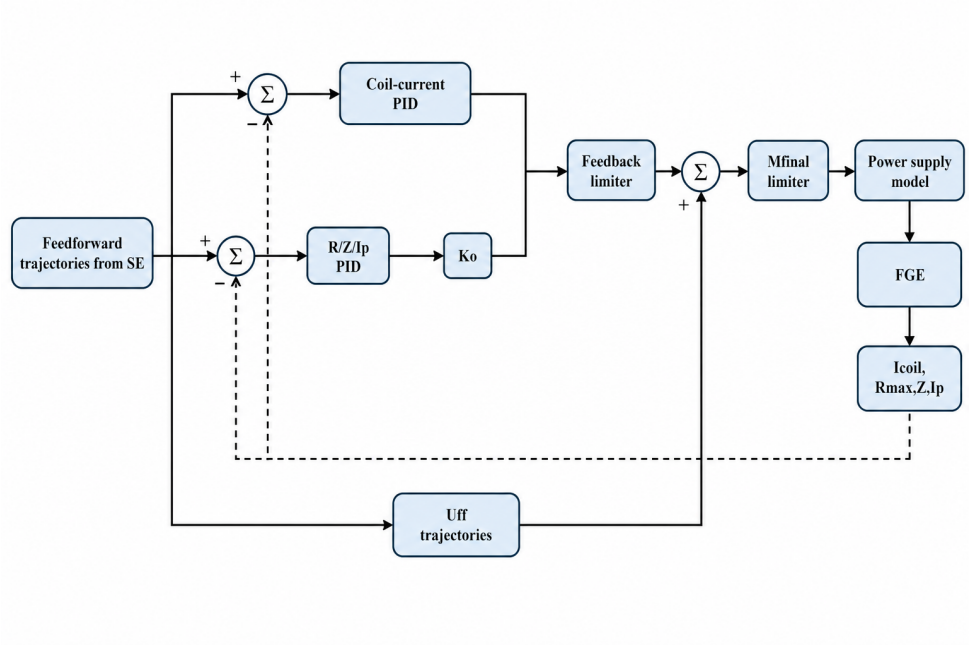

Figure 5: Baseline FF+PID control structure.

The classical baseline is feedforward plus PID (Figure 5). Offline, the MATLAB Shape Editor (SE) designs PF-coil current and I_p references and the associated feedforward voltages [28]. Online, PID loops on PF-coil currents, I_p, and plasma position (R, Z) produce a voltage correction that is added to the feedforward,

$$\boldsymbol{V}_{\mathrm{cmd}} = \boldsymbol{V}_{\mathrm{ff}} + \Delta \boldsymbol{V}_{\mathrm{PID}}. \tag{4}$$

All FF+PID scores in this paper reuse the #13906 plant feedforward and PID gains in a closed-loop FGE rollout, without per-initial redesign of the feedforward table.

2.3.2 RL control

We cast XPT magnetic control as a partially observed decision process: the full FGE equilibrium is the latent state, but the feedforward policy $\pi(\mathbf{a}_t \mid \mathbf{o}_t)$ acts only on a reconstructed observation $\mathbf{o}_t$ and is trained to maximise the discounted return $\mathbb{E}_\pi[\sum_t \gamma^t r_t]$. The interface used throughout is as follows.

Observation: The 45-dimensional observation $\mathbf{o}_t$ concatenates normalised plasma-current error, 12 coil currents (CS, PS1–PS10, and VS), 4 X-point features (validity, R–Z position errors, and flux relative to the LCFS), and radial and vertical LCFS errors at eight boundary points equally spaced in poloidal angle about the magnetic axis. The same observation is used for every learned controller in this paper.

Action: The learned policy commands an 11-dimensional coil-voltage vector (CS and PS1–PS10). The vertical-stability coil (VS) is regulated by a separate PID loop and is excluded from the policy action for every controller in this paper. Policy outputs are passed through tanh squashing and rescaled to the physical voltage limits. Actuation metrics $|\Delta V|$ are computed on these 11 policy-commanded coils.

Reward: Tracking errors for the plasma current, the four X-point positions, the four X-point flux-difference conditions, and the LCFS boundary are extracted from the FGE output and mapped to satisfaction signals $r_{t,i} \in [0, 1]$ by sigmoid or softplus shaping with "good"/"bad" thresholds. How these K signals are aggregated relative to temporal credit assignment is the algorithmic axis of the next subsections.

2.3.3 Actor–critic instantiation with PPO

Advantage Aggregation (AdvA) needs a per-objective advantage signal, so it belongs to the actor–critic family: methods that maintain a critic (or critics) and form advantages for a policy update. Pure value-based algorithms without an explicit advantage pathway, such as the deep Q-network (DQN) [29], are outside this scope. We instantiate both the reward-level baseline and AdvA with PPO [30]; the same aggregation design can be attached to other actor–critic algorithms. In results we write **AdvA-PPO** for the AdvA instance on PPO.

PPO stabilises policy-gradient updates through a clipped surrogate. With probability ratio $\chi_t(\theta) = \pi_\theta(\mathbf{a}_t \mid \mathbf{o}_t)/\pi_{\theta_{\text{old}}}(\mathbf{a}_t \mid \mathbf{o}_t)$ and advantage $\widehat{A}_t$,

$$\mathcal{L}^{\text{CLIP}}(\theta) = -\mathbb{E}_t\Big[\min\Big(\chi_t(\theta)\,\widehat{A}_t,\ \text{clip}\big(\chi_t(\theta), 1-\epsilon, 1+\epsilon\big)\,\widehat{A}_t\Big)\Big], \tag{5}$$

the clip removes the incentive for large probability-ratio steps when they would inflate the objective. The full training loss adds a value-function error and an entropy bonus,

$$\mathcal{L}^{\text{PPO}}(\theta) = \mathcal{L}^{\text{CLIP}}(\theta) + c_1\,\mathcal{L}^{\text{VF}}(\theta) - c_2\,\mathcal{H}[\pi_\theta], \tag{6}$$

where $\mathcal{L}^{\text{VF}}$ fits the critic to empirical returns and $\mathcal{H}$ encourages exploration. Both the reward-level baseline and every AdvA variant share the 45-dimensional observation and 11-dimensional action of Sec. 2.3.2.

2.3.4 Reward-level SmoothMax baseline

The conventional multi-objective baseline scalarises the channel rewards *before* GAE [25]: that is why we call it reward-level. The same worst-objective-aware exponential weighting used later by AdvA defines

$$\omega_{t,i} = \frac{w_i^0 \exp(\alpha r_{t,i})}{\sum_j w_j^0 \exp(\alpha r_{t,j})}, \tag{7}$$

$$u(\mathbf{r}_t) = \bar{r}_t = \sum_i \omega_{t,i} r_{t,i}, \qquad \alpha < 0, \tag{8}$$

so a poorly satisfied channel receives more weight. We keep the implementation name "SmoothMax"; with $\alpha < 0$ the softmin emphasises the worst objective. A single value function and GAE fitted to $\bar{r}_t$ then supply the PPO advantage in Eq. (5). Results and ablations refer to this baseline as **Reward-PPO** (Table 3).

2.3.5 Advantage Aggregation (AdvA)

AdvA moves scalarisation to the *advantage* layer: one value head and one GAE per physical channel on a shared critic backbone, with nonlinear mixing only afterwards. Relative to reward-level SmoothMax, a weak channel no longer collapses the temporal credit of every other channel before advantage estimation. Figure 6 shows the closed loop and update path.

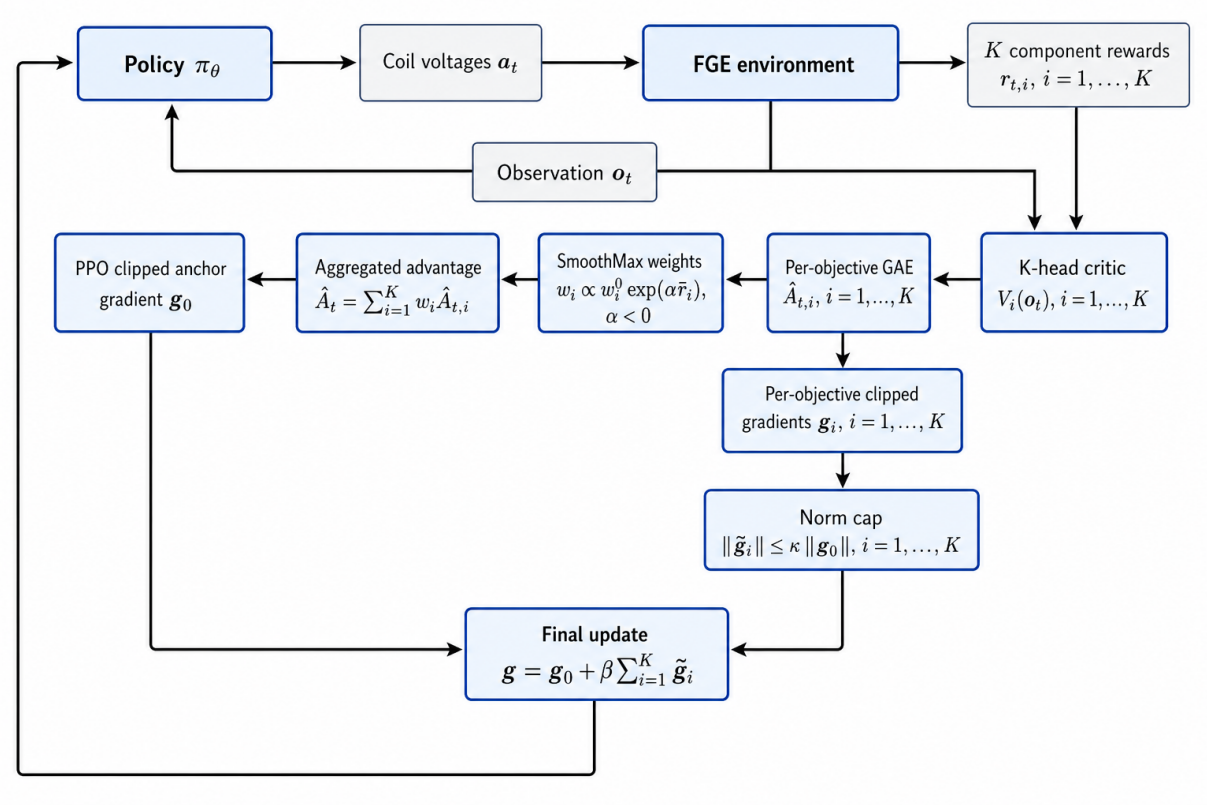

Figure 6: Policy update and environment interaction for AdvA-PPO. The FGE environment provides the reconstructed magnetic observation and component rewards. Per-objective value heads and GAE preserve temporal credit assignment before SmoothMax aggregation; the PPO anchor is then corrected by the norm-capped residual of Eqs. (15)–(16). Reliability gating and gradient projection are ablation options and are off in AdvA-PPO.

Let $r_{t,i} \in [0, 1]$ be the channel reward and w_i^0 a fixed prior weight (inner X-point positions and the LCFS receive larger w_i^0). A shared critic backbone feeds heads $V_i(\mathbf{o}_t)$ with

$$\delta_{t,i} = r_{t,i} + \gamma V_i(\mathbf{o}_{t+1}) - V_i(\mathbf{o}_t), \tag{9}$$

$$\widehat{A}_{t,i} = \sum_{\ell=0}^{T-t-1} (\gamma\lambda)^\ell \delta_{t+\ell,i}, \qquad i = 1, \ldots, K. \tag{10}$$

Each head is trained against $\widehat{G}_{t,i} = \widehat{A}_{t,i} + V_i(\mathbf{o}_t)$.

SmoothMax advantage aggregation. Batch-mean channel rewards $\bar{r}_i$ define worst-case-aware weights

$$w_i = \frac{w_i^0 \exp(\alpha \bar{r}_i)}{\sum_j w_j^0 \exp(\alpha \bar{r}_j)}, \qquad \alpha < 0, \tag{11}$$

and the scalar advantage

$$\widehat{A}_t = \sum_{i=1}^K w_i \, \widehat{A}_{t,i} \tag{12}$$

replaces the usual PPO advantage column. Setting $\alpha = 0$ recovers a fixed-weight mean (Adv-mean in the ablation). The actor uses a single clipped surrogate on $\widehat{A}_t$.

Reliability gate. A head with low explained variance, or a channel already saturated near reward 1, contributes little useful gradient. Exponential moving averages of explained variance $\overline{\mathrm{EV}}_i$ and saturation fraction $\bar{s}_i$ define

$$q_i = \mathrm{clip}\left(\overline{\mathrm{EV}}_i, 0, 1\right)(1 - \bar{s}_i) \in [0, 1], \tag{13}$$

and when enabled multiply the aggregation weights, $w_i \leftarrow w_i q_i$, followed by renormalisation. The gate is an ablation option and is off in AdvA-PPO.

Norm-capped residual. PPO clipping acts once on the aggregated advantage, so a minority channel whose $\widehat{A}_{t,i}$ disagrees with $\widehat{A}_t$ can be silenced. Component surrogates

$$\mathcal{L}_i^{\mathrm{clip}}(\theta) = -\mathbb{E}_t\left[\min\left(\chi_t(\theta)\widehat{A}_{t,i}, \mathrm{clip}\big(\chi_t(\theta), 1-\epsilon, 1+\epsilon\big)\widehat{A}_{t,i}\right)\right] \tag{14}$$

supply gradients $\mathbf{g}_0 = \nabla_\theta \mathcal{L}^{\mathrm{CLIP}}(\widehat{A})$ and $\mathbf{g}_i = \nabla_\theta \mathcal{L}_i^{\mathrm{clip}}$. Each component is capped to the anchor scale,

$$\widetilde{\mathbf{g}}_i = \mathbf{g}_i \min\left(1, \frac{\kappa\|\mathbf{g}_0\|_2}{\|\mathbf{g}_i\|_2 + \varepsilon_g}\right), \tag{15}$$

and the update is

$$\mathbf{g} = \mathbf{g}_0 + \beta \sum_{i=1}^{K} \widetilde{\mathbf{g}}_i, \tag{16}$$

with defaults $\kappa = 1$ and $\beta = 0.2$. Setting $\beta = 0$ recovers pure advantage-level SmoothMax.

Gradient projection. As an optional conflict-resolution step in the spirit of PCGrad [22, 23], the component of $\widetilde{\mathbf{g}}_i$ opposing $\mathbf{g}_0$ can be removed before the sum in Eq. (16). Projection appears only in proxy ablation rows of Sec. 3.2 and is disabled in AdvA-PPO ($\mathbf{p}_i = \widetilde{\mathbf{g}}_i$).

Proposed controller: AdvA-PPO. **AdvA-PPO** is the AdvA instance used in all main experiments: advantage-level SmoothMax ($\alpha = -3$) plus the capped residual ($\beta = 0.2$, $\kappa = 1$), with the reliability gate and projection both off. Sec. 3.2 motivates retaining the residual and omitting gate and projection.

2.3.6 Ablation sequence and controller inventory

Table 3 lists the controllers compared in the Results ablation (Sec. 3.2). Results name the reward-level SmoothMax PPO baseline of Sec. 2.3.4 as **Reward-PPO**. The ablation follows the design decisions in causal order:

1. Start from Reward-PPO (reward-level SmoothMax, scalar GAE).
2. Move to Adv-mean (α=0, multi-head).
3. Replace the mean by Adv-SmoothMax (α= − 3, multi-head).
4. On that SmoothMax anchor, compare + gate (β=0) with + resid (β=0.2) = AdvA-PPO.

Pairwise gate/projection combinations are reported only as temporary proxies from available checkpoints, not as a second method family. Early multi-head runs still carry two radial-extrema rewards ($R_{\min}$, $R_{\max}$), hence $K = 12$; those channels are excluded from every reported score. Clean contrasts in Sec. 3.2 are therefore Adv-mean versus Adv-SmoothMax ($K = 12$), gate versus residual on the SmoothMax anchor ($K = 10$), and the family-level Reward-PPO versus AdvA-PPO comparison; proxy rows are not matched one-variable AdvA-native cells. Within a channel set, matched variants share the 45/11 observation–action interface, capacity, optimizer, and checkpoint rule; FF+PID is evaluated, not retrained.

Training and inference implementation. The actor and shared value backbone are two fully connected layers of 256 units with tanh activations; AdvA controllers attach one scalar value head per reward channel. Policies use a squashed Gaussian action distribution and are trained in PyTorch with Adam at learning rate 3×10^{-4}. The PPO clip is $\epsilon = 0.2$, $\gamma = 0.98$, $\lambda = 0.95$, the entropy coefficient is 5×10^{-3}, and global gradient norm is clipped at 0.3. Twenty-three parallel FGE workers collect one 300-step episode each per iteration, giving a batch of 6900 transitions; each batch is optimised for ten epochs with minibatches of 256. The reported AdvA-PPO checkpoint is the final checkpoint after 300 iterations (2.07×10^6 environment steps). Evaluation uses the deterministic mean action.

Table 3: Controllers in the component ablation, in the same order as the Results ablation (Sec. 3.2). "Locus" is where nonlinear aggregation occurs relative to objective-wise temporal credit. For AdvA rows, Gate/β/Proj. act on advantage weights or the residual of Eqs. (15)–(16). All rows share the same d_a=11 / d_o=45 interface (Sec. 2.3.2; VS under PID). **Reward-PPO** is the reward-level SmoothMax baseline of Sec. 2.3.4.

Name	Locus	α	Gate	β	Proj.	K	d_a/d_o
Reward-PPO	reward	-3	off	0	off	10	11/45
Adv-mean	advantage	0	off	0	off	10	11/45
Adv-SmoothMax	advantage	-3	off	0	off	10	11/45
+ gate	advantage	-3	on	0	off	10	11/45
+ resid (AdvA-PPO)	advantage	-3	off	0.2	off	10	11/45
+ proj	advantage	-3	off	0	on	10	11/45
+ gate+resid	advantage	-3	on	0.2	off	10	11/45
+ gate+proj	advantage	-3	on	0	on	10	11/45

2.4 Evaluation setup

All reported scores come from one harness, one FGE image pool, and one primary evaluation seed (20260715), with a 500-step (500 ms) closed-loop test horizon and deterministic actions at test time. Training episodes are 300 steps (300 ms); finishing the longer test window is a temporal-extrapolation check.

2.4.1 Test scenarios

Four initial equilibria (I1–I4; Table 4) share the training inductance basis and request the same frozen #13906 XPT target. Each control window starts near $t_0 \approx 500$ ms of the parent discharge and runs for 500 ms. I1 is the calibrated XPT training initial; I2–I4 are withheld cross-initialization cells (same geometric target and reward operators; only the FGE initial and I_p reference change).

Robustness cells R1–R3 reuse I1. Disturbances enter the *controller observation* only; the simulator state used for scoring stays clean. R1 redraws an observation delay uniformly in 1–3 ms at every control step (non-monotonic jitter, harder than a fixed mean delay). R2 adds full diagnostic/coil noise: I_p σ=6 kA, per-coil I_{PF} at 1%, and absolute 5 mm on raw X-point and LCFS coordinates. R3 applies R1 and R2 together.

Table 4: Evaluation initials I1–I4. All windows start at $t_0 \approx 500$ ms (CSV start 0.502 s) and use the frozen #13906 XPT geometric target; $I_p^\star$ is the current reference for that cell.

ID	Shot	t_0	Topology	$I_p^\star$
I1	#13906	≈ 500 ms	XPT (training / test)	≈ 400 kA
I2	#13705	≈ 500 ms	divertor (test)	≈ 500 kA
I3	#13844	≈ 500 ms	divertor (test)	≈ 500 kA
I4	#15892	≈ 500 ms	limiter (test)	≈ 500 kA

2.4.2 Metrics and statistical protocol

Scores are recomputed from archived equilibria by one pipeline. Raw FGE X-point candidates are mapped to the four slots X_1–X_4 of Sec. 2.1 before any error is formed. Position error is the Euclidean distance to the frozen target in that slot; flux error is the poloidal-flux difference $|\psi_X - \psi_B|$ in that slot, reported in Wb. Mean d_X or mean flux is the average of the four per-slot RMSE / max values (primary/secondary columns average only the corresponding pair). The LCFS channel is the boundary deviation used by the reward operator,

$$e_{\mathrm{LCFS}} = \tfrac{1}{2}\left(\overline{|(r_B - r_B^{\mathrm{ref}})/(|r_B^{\mathrm{ref}}| + \epsilon)|} + \overline{|(z_B - z_B^{\mathrm{ref}})/(|z_B^{\mathrm{ref}}| + \epsilon)|}\right), \qquad \epsilon = 10^{-6}\,\mathrm{m}, \tag{17}$$

on eight boundary points equally spaced in poloidal angle about the magnetic axis. Tables report RMSE and max absolute error over the survived horizon, survival length (full = 500 ms), and $|\Delta V|$ on the 11 policy coils. On short survivals, errors are secondary to survival; stochastic noise/jitter cells are multi-seed averaged or labelled as a single realisation.

Because physical rows have incompatible units, we also report time means of the shared ten-channel scores (I_p, four positions, four fluxes, LCFS): the reward-level SmoothMax $\bar{u}$ of Eq. (8) ($\alpha = -3$) and

the hard worst-channel mean $\bar{r} = \overline{\min_i r_{t,i}}$. Both are defined for FF+PID as well; evaluation $\bar{u}$ is not the advantage-level SmoothMax inside AdvA. Mechanism narrative emphasises physical RMSE / max; Sec. 3.2 also lists $\bar{u}$ and $\bar{r}$. In comparison tables, the best entry in each column (within a block) is **bold** and the second-best is underlined; lower is better for errors and $|\Delta V|$, higher for survival, $\bar{u}$, and $\bar{r}$.

3 Results

3.1 Main controller comparison

Under the protocol of Sec. 2.4, the nominal cell uses the calibrated shot-#13906 initial (training initial; I1 of Table 4) over a 500 ms rollout at 1 ms per control step. Three controllers are compared: FF+PID, Reward-PPO, and AdvA-PPO (ours). The head-to-head isolates the aggregation locus (reward-level versus advantage-level); the component ablation of Sec. 3.2 isolates the algorithmic factors within AdvA.

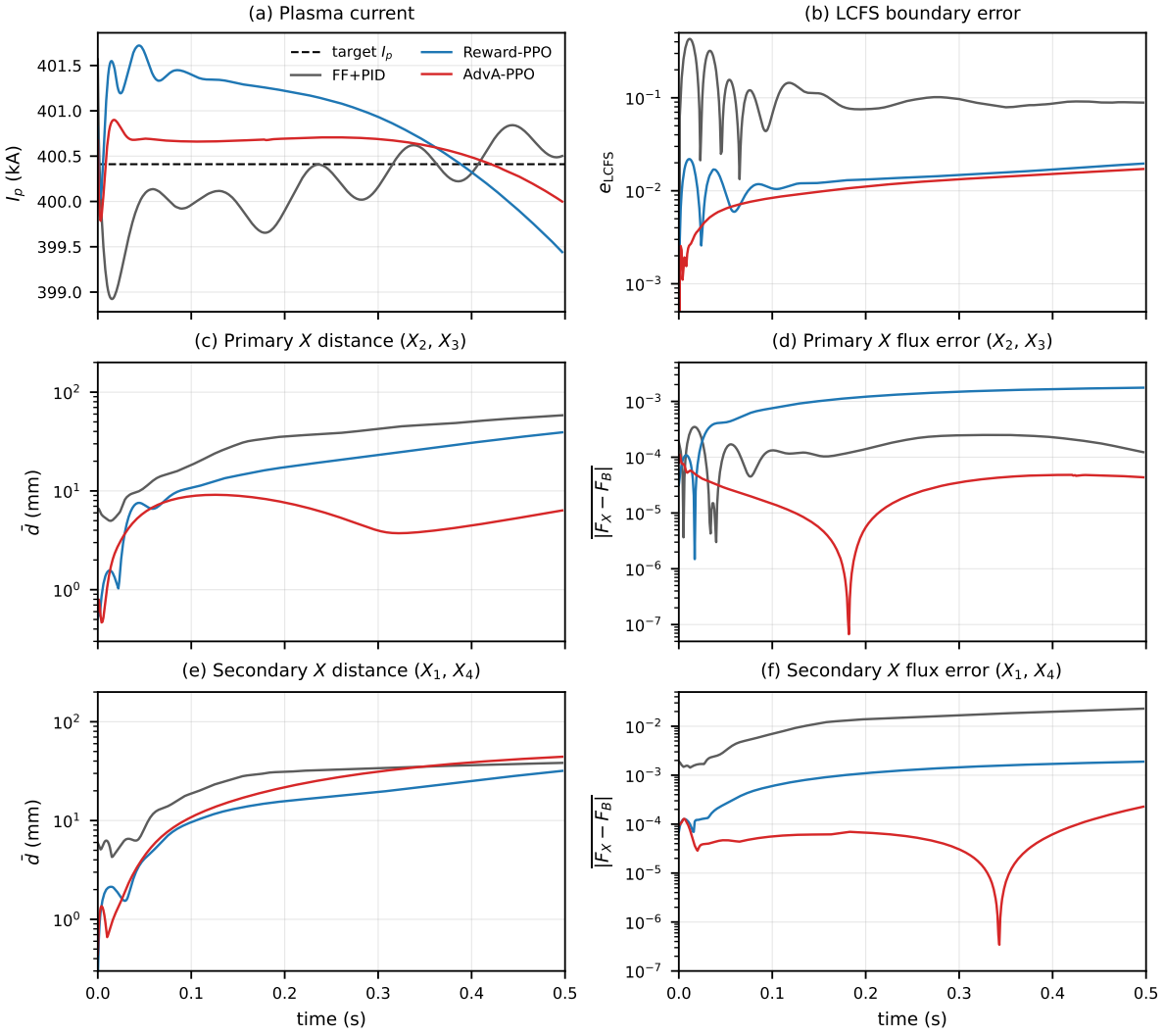

Figure 7: Closed-loop evolution on #13906 (500 ms). Grey: FF+PID; blue: Reward-PPO; red: AdvA-PPO. (a) I_p; (b) LCFS error; (c)–(d) primary X distance and flux (X_2, X_3); (e)–(f) secondary X distance and flux (X_1, X_4).

Figure 7 and Table 5 show that AdvA-PPO is the strongest multi-objective controller on this cell: it leads or matches on I_p, primary X distance, mean X flux, and LCFS, rather than trading one channel for another. All three controllers complete the horizon, so the comparison is tracking quality, not survival. The evolution traces make the channel-wise behaviour visible in time; panel (d) already shows the primary-flux gap that the table later quantifies as about a 20× lower mean X-flux RMSE for AdvA-PPO relative to Reward-PPO (0.66 versus 13.1×10^{-4}).

The decisive contrast is not a single physical row but the shared scoring axis of Sec. 2.4.2. Reward-PPO

reaches $\bar{u} = 0.56$ with worst-channel mean $\bar{r} = 0.23$, limited by the inner primary flux X_2; AdvA-PPO reaches $\bar{u} = 0.93$ with $\bar{r} = 0.81$—a 3.5× lift of the hard worst channel, together with a 2.8× lower I_p RMSE (260 versus 735 A). Reward-PPO is not uniformly weaker: it records the lowest secondary X distance (19.0 versus 28.2 mm), but that gain coincides with the poorest I_p among the three and a much weaker primary-flux channel.

The secondary-null rows deserve a separate reading, because position and flux are not interchangeable descriptors of an XPT. The defining topological condition is that the secondary null sits *on* the divertor leg of the primary separatrix, i.e. $\psi_X \to \psi_B$; where along that leg it sits sets the location of the flux expansion but does not decide whether the configuration is an XPT at all. On this cell AdvA-PPO holds the secondary flux condition an order of magnitude tighter than Reward-PPO (1.07 versus 17.1×10^{-4} Wb on X_1 and 0.61 versus 8.8×10^{-4} Wb on X_4) while allowing the nulls to sit about 10 mm further along the leg. In other words the two policies fail differently: AdvA-PPO keeps the nulls locked to the separatrix and slides them along it, whereas Reward-PPO places them closer to the programmed coordinates but lets them detach from the leg. For the XPT the first failure mode is the benign one, and the residual $\sim$10 mm offset is comparable to the secondary-null reconstruction accuracy of the calibration itself (Table 2, 13.5 mm RMSE), so it should not be read as a resolved ranking. The flux criterion is also intrinsically weak against displacement: $\nabla\psi = 0$ at a null, so $|\psi_X - \psi_B|$ grows only quadratically with the null displacement and a small flux error certifies topology rather than position. The two channels are therefore complementary, and we report both rather than a single secondary-null figure of merit. FF+PID keeps competitive I_p (454 A RMSE) because its loops close on coil currents, I_p, and the geometric centroid (R, Z); it does not feed back X-point position, X-point flux, or a multi-point LCFS, so the shape rows in the table are open-loop observations and the evolution panels (b)–(d) sit systematically above either policy. The AdvA-PPO gains are not bought with larger actuation: its mean per-step $|\Delta V|$ is the lowest of the three (0.36 V against 0.56 V and 0.45 V).

Figure 8 supplies the geometric reading of the same episode. Under AdvA-PPO the separatrix stays on the target boundary and the X points remain close to their targets through the horizon, consistent with the low primary distance and flux errors. Reward-PPO keeps a four-X topology but with a looser core LCFS match, in line with its intermediate boundary and flux scores. Under FF+PID the separatrix bulges outward and the X points drift, as expected from a controller without shape closure. Across all three methods the divertor-leg region remains only loosely constrained: neither the observation nor the reward treats the legs explicitly, so secondary-leg geometry is a shared limitation of the present task rather than a failure unique to one controller.

Table 5: Nominal tracking on I1 (#13906, 500 ms); metrics as in Sec. 2.4.2. FF+PID X-point and LCFS rows are open-loop observations.

Quantity	FF+PID	Reward-PPO	AdvA-PPO (ours)
I_p (A)	454 / 1488	735 / 1310	**260** / **624**
Primary X dist. (mm)	39.0 / 58.3	22.8 / 39.2	**6.3** / **9.5**
Secondary X dist. (mm)	30.1 / 38.3	**19.0** / **31.8**	28.2 / 44.2
Mean X flux (10^{-4} Wb)	77.38 / 117.02	13.06 / 18.89	**0.66** / **1.93**
LCFS ($\times 10^{-3}$)	120.1 / 430.3	14.7 / 21.9	**12.2** / **17.2**
SmoothMax $\bar{u}$ ↑	0.093	0.563	**0.925**
mean worst channel $\bar{r}$ ↑	0.001	0.232	**0.810**
mean / max $\lvert\Delta V\rvert$ (V)	0.45 / 124	0.56 / **50**	**0.36** / 81

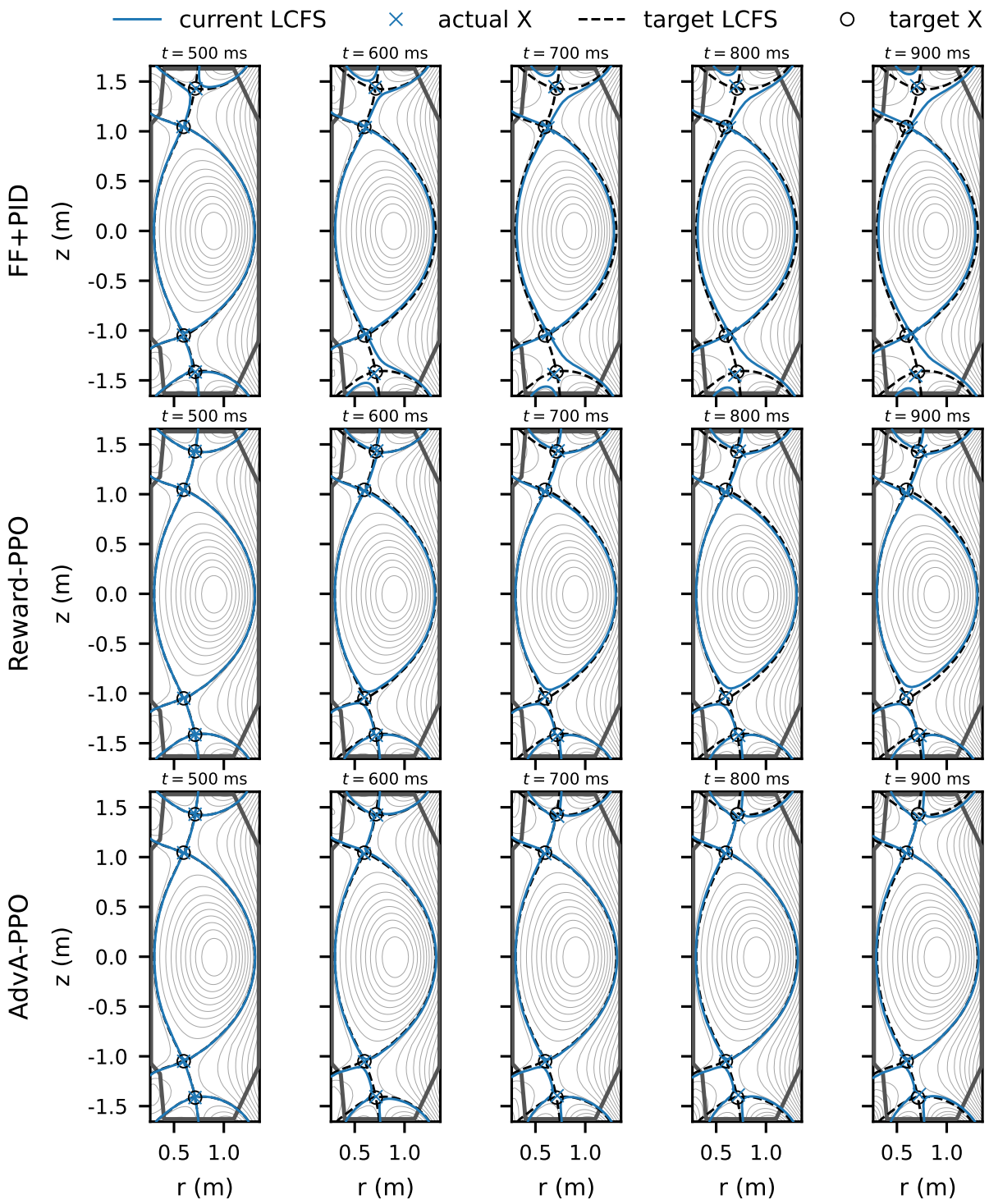

Figure 8: Flux snapshots on #13906 at $t = 500$–$900\,\mathrm{ms}$ (shot clock; control from $t_0 = 500\,\mathrm{ms}$). Rows: FF+PID, Reward-PPO, AdvA-PPO. Blue solid: current LCFS; black dashed: target; open circles: target X; blue crosses: actual X.

3.2 Mechanism and ablation analysis

Table 6: Component ablation on I1 (#13906, 500 ms). Names as in Table 3; metrics as in Sec. 2.4.2.

Variant	I_p (A)	mean d_X (mm)	mean flux (10^{-4} Wb)	LCFS ($\times 10^{-3}$)	$\bar{u}\uparrow$	$\bar{r}\uparrow$
Reward-PPO	735 / 1310	20.9 / 35.5	13.06 / 18.89	14.7 / 21.9	0.563	0.232
Adv-mean	1615 / 2149	34.2 / 45.8	1.77 / 2.70	8.7 / 14.9	0.830	0.665
Adv-SmoothMax	1630 / 2545	17.6 / 27.3	2.44 / 5.22	13.4 / 18.0	0.879	0.760
+ gate	2779 / 3384	18.5 / 29.0	1.30 / 2.76	15.3 / 21.7	0.899	0.760
+ resid = **AdvA-PPO**	**260** / **624**	17.3 / **26.9**	**0.66** / **1.93**	12.2 / 17.2	**0.925**	**0.810**
+ proj	1601 / 1916	18.9 / 29.7	1.27 / 2.10	8.5 / 11.3	0.894	0.765
+ gate+resid	1137 / 1445	**16.8** / 27.3	1.70 / 2.85	7.5 / **8.8**	0.894	0.748
+ gate+proj	925 / 1187	23.3 / 36.7	2.01 / 3.28	**6.1** / **8.8**	0.870	0.726

Table 6 decomposes the nominal AdvA-PPO gain of Sec. 3.1 along the design sequence of Table 3 (metrics in Sec. 2.4.2). Reward-PPO already holds I_p reasonably (735 A RMSE) but leaves a large mean flux error (13.1×10^{-4}), so the XPT is not flux-tight. Moving credit assignment before scalarisation (Adv-mean, then Adv-SmoothMax) cuts that flux error by about an order of magnitude, yet both raise I_p RMSE above 1.6 kA. The two are not interchangeable: Adv-mean reaches a slightly lower flux (1.77 versus 2.44×10^{-4}) but a much worse mean d_X (34.2 versus 17.6 mm), so the softmin temperature alone does not select the operating point—geometry and current still trade. On the Adv-SmoothMax anchor, + gate further lowers mean flux (1.30×10^{-4}) while driving I_p to 2.78 kA RMSE: attenuating saturated channels sharpens the shape update and starves the current loop. Adding only + resid = AdvA-PPO recovers both sides of that trade-off and is the strongest balanced cell: relative to Adv-SmoothMax it reduces I_p RMSE from 1.63 to 0.26 kA and mean flux from 2.44 to 0.66×10^{-4}, with mean d_X essentially unchanged (17.3 versus 17.6 mm), and it also leads the shared scoring axis ($\bar{u} = 0.925$, $\bar{r} = 0.810$). That joint recovery is what links the family-level AdvA-PPO versus Reward-PPO contrast of Sec. 3.1 to a concrete component choice. The remaining proxy combinations do not overturn it: + gate+proj records the best LCFS (6.1×10^{-3}) but is weaker on I_p (925 A) and mean flux (2.01×10^{-4}) than residual alone, so the boundary looks tidy while the XPT flux condition and current loop do not; + proj and + gate+resid likewise fail to match AdvA-PPO on I_p or mean flux. Among gating, projection, and residual, the capped residual is therefore retained: it is the most balanced correction and the one that keeps the four-null flux lock that defines the XPT here.

3.3 Robustness evaluation

This subsection tests whether the same frozen controllers remain usable under two deployment stresses. First, diagnostic noise and observation delay of the kind present in the experimental plant (R1–R3). Second, a change of the magnetic configuration at flattop hand-over: the controller takes over near $t_0 \approx 500$ ms from a different parent discharge while the #13906 XPT geometric target stays fixed (I2–I4). FF+PID, Reward-PPO, and AdvA-PPO are evaluated without retraining; disturbance magnitudes, initials, and the clean-state scoring rule follow Sec. 2.4.1.

3.3.1 Zero-shot diagnostic noise and observation delay

R1–R3 of Sec. 2.4.1 stress the three controllers on I1 (#13906). Table 7 reports survival and RMSE over the survived horizon; Figure 9 shows the channel-wise evolution under the combined delay+noise stress (R3).

Table 7: Zero-shot noise/delay on I1 (#13906); RMSE over the survived horizon (Sec. 2.4.2).

Method	Surv. (ms)	I_p (A)	Prim. X (mm)	Prim. flux (10^{-4} Wb)	Sec. X (mm)	Sec. flux (10^{-4} Wb)	LCFS ($\times 10^{-3}$)
R1: delay 1–3 ms							
FF+PID	500	**637**	38.1	**2.16**	29.6	149.81	160.8
Reward-PPO	500	11997	**23.0**	28.38	41.1	36.88	229.6
AdvA-PPO	500	4521	26.6	13.04	**24.8**	**17.85**	**18.8**
R2: full diagnostic noise							
FF+PID	500	74817	46.1	23.02	22.3	55.54	222.0
Reward-PPO	500	**2183**	**12.6**	**5.37**	**21.0**	**8.31**	53.0
AdvA-PPO	500	2394	26.4	12.88	31.9	12.60	**19.3**
R3: delay + noise							
FF+PID	**500**	68752	48.6	368.72	**24.9**	356.89	422.3
Reward-PPO	322	**5222**	30.8	15.17	28.5	21.39	94.6
AdvA-PPO	**500**	5521	**29.0**	**12.79**	41.7	**17.82**	**26.3**

The three conditions expose different failure modes rather than a single winner. On R1 (delay only) all three finish the horizon: FF+PID keeps the best I_p tracking (637 A RMSE), consistent with a strong

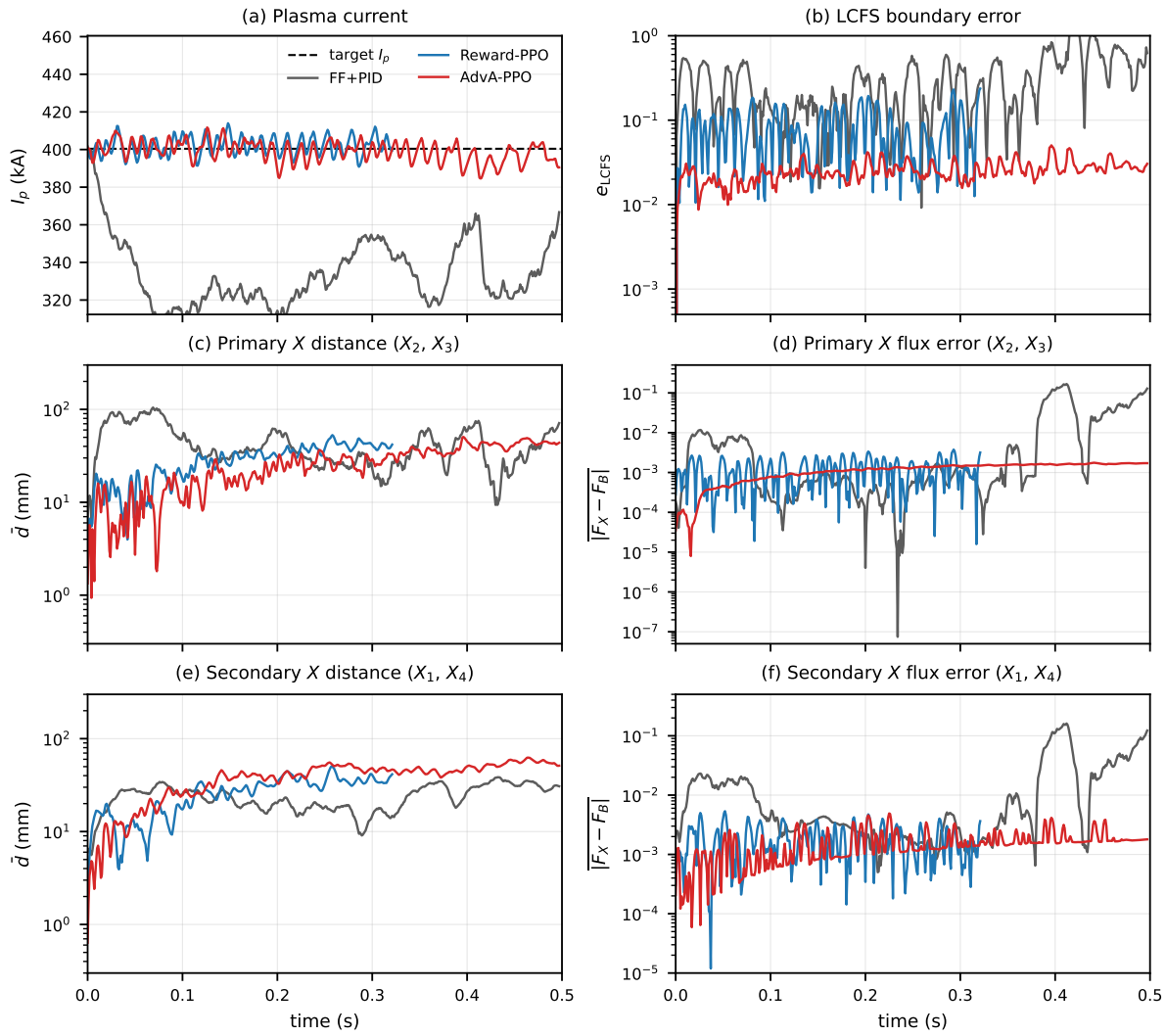

Figure 9: R3: delay + noise on I1 (#13906). Panel layout as Figure 7.

feedforward backbone that is only weakly disturbed by a few milliseconds of observation lag, while AdvA-PPO leads on LCFS and Reward-PPO suffers a large I_p RMSE under the per-step delay jitter. On R2 (noise only) the ranking reverses on the current channel: both learned controllers hold I_p near 2–3 kA RMSE, whereas FF+PID collapses to > 70 kA. That contrast is expected from the classical loop structure (Sec. 2.3.1): derivative action on noisy I_p, coil-current, and R/Z measurements amplifies high-frequency sensor error into coil-voltage corrections, so the same PID that is benign under clean delayed observations becomes harmful under full diagnostic noise. Within the learned pair on R2, Reward-PPO is honestly the stronger cell—best I_p, primary/secondary X distance, and both flux channels—with AdvA-PPO retaining the best LCFS. R3 combines both stresses and is the decisive joint test: Reward-PPO's R2 advantage does not carry over—it terminates at 322 ms—while AdvA-PPO and FF+PID complete the horizon (Figure 9). Among survivors AdvA-PPO keeps the lowest boundary and primary-flux errors; FF+PID again survives but with the same noise-dominated current/flux/boundary degradation seen on R2. Taken together, the single-factor cells are informative but not conclusive: FF+PID resists delay yet fails under noise, and Reward-PPO wins the noise-only comparison yet breaks under the joint load. On the combined R3 protocol that matches concurrent plant delay and diagnostic noise, AdvA-PPO is the strongest controller—the only learned policy that both survives and retains a usable XPT shape.

3.3.2 Zero-shot cross-initialization

I2–I4 of Table 4 hold the #13906 XPT target, observation map, and reward operators fixed, and change two quantities at once: the FGE initial equilibrium and the I_p reference (training-scale ~ 400 kA to ~ 500 kA). That joint shift—new shape or topology at hand-over plus a higher current setpoint—is a hard zero-shot test for a policy trained on a single initial. Each controller keeps the checkpoint (or feedforward table) already used on I1, with no per-initial redesign. In particular, FF+PID does not

receive a newly computed feedforward for each withheld discharge—that would amount to redesigning the classical controller and is outside the present scope—so the classical baseline is the #13906 plant feedforward and PID gains of Sec. 2.3.1, only re-anchored at reset to remove the trivial CS/PF current offset (and with the CS current guard removed). We report survival and whether the frozen XPT shape is reached (Figure 10, Table 8).

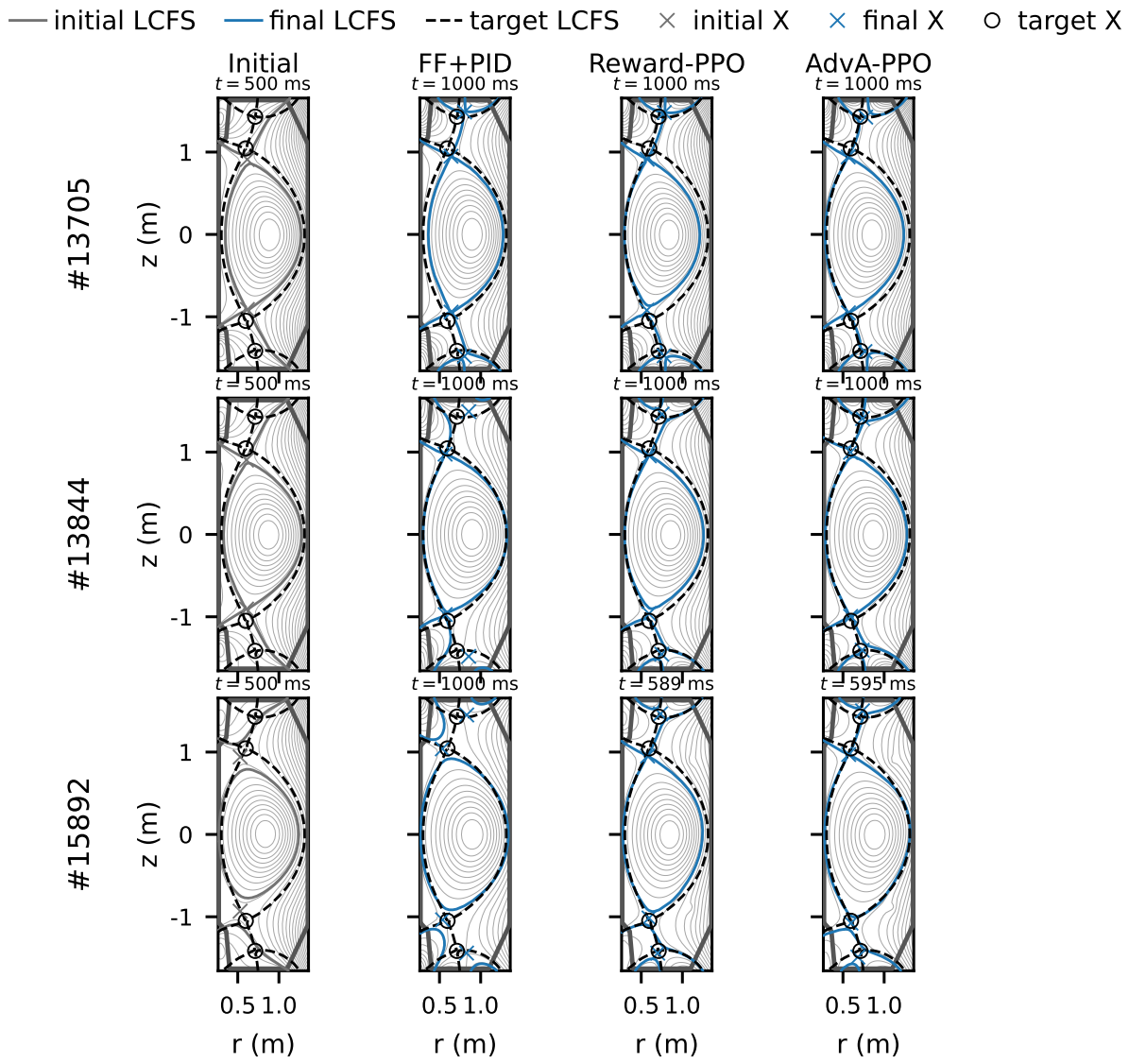

Figure 10: Zero-shot cross-init flux under frozen #13906 XPT. Rows: #13705, #13844, #15892. Columns: initial equilibrium and FF+PID / Reward-PPO / AdvA-PPO at episode end. Grey: initial LCFS/X; blue: final; black dashed/circles: target.

Table 8: Zero-shot cross-init to frozen #13906 XPT (I2–I4); metrics as in Sec. 2.4.2 (full survival = 500 ms). FF+PID feedforward is re-anchored as described in the text.

Init.	Controller	Surv.	I_p (kA)	mean d_X (mm)	mean flux (10^{-4} Wb)	LCFS ($\times 10^{-3}$)	$\bar{u}\uparrow$	$\bar{r}\uparrow$
I2 #13705 divertor	FF+PID	full	**9.4** / **42.3**	125 / **178**	1315 / 5054	232 / 596	0.02	0.00
	Reward-PPO	full	32.8 / 79.3	85 / 179	**889** / **5013**	**46** / **70**	0.07	0.00
	AdvA-PPO	full	31.8 / 75.6	**69** / 179	909 / 5187	63 / 129	**0.11**	**0.01**
I3 #13844 divertor	FF+PID	full	**2.8** / **6.1**	126 / 174	1305 / **5009**	97 / 119	0.02	0.00
	Reward-PPO	full	19.6 / 56.8	72 / **168**	**990** / 5333	**50** / **110**	0.10	0.00
	AdvA-PPO	full	17.6 / 57.3	**60** / 176	1061 / 5781	79 / 241	**0.14**	0.00
I4 #15892 limiter	FF+PID	**full**	**3.4** / **6.9**	**93** / **167**	**1168** / 5082	139 / **150**	0.02	0.00
	Reward-PPO	90	31.4 / 43.2	103 / 173	1697 / **5074**	**133** / 273	**0.04**	0.00
	AdvA-PPO	96	20.7 / 26.5	99 / 170	1714 / 5207	182 / 272	**0.04**	0.00

Under this shift the classical baseline is the more transferable stabiliser. The same #13906 FF+PID gains and feedforward, only re-anchored, complete the horizon on every withheld initial and hold I_p tightly

(2.8–9.4 kA RMSE against 17–33 kA for the policies), spanning the training-scale 400 kA plant and the 500 kA test setpoints. That stability does not buy shape precision: mean d_X stays ∼93–126 mm, mean flux errors remain large, and the limiter initial is held without converting to the target XPT (Figure 10).

On the two divertor initials both policies survive and improve the geometric channels relative to FF+PID (mean d_X 60–85 mm, lower LCFS and flux RMSE), with AdvA-PPO leading on d_X and $\bar{u}$ while Reward-PPO keeps the lower LCFS error—so the advantage-level update is not uniformly better off-distribution. The current and boundary channels trade against each other on #13705, whose CS starts at −15.7 kA and therefore has about half the volt-second headroom of the other initials: FF+PID does hold I_p there, but only by drawing 54 kA of CS current against 43–45 kA on the other two.

The limiter I4 (#15892) fails both learned controllers: Reward-PPO terminates at step 90 and AdvA-PPO at step 96, each with mean X-point distance still ∼100 mm, while FF+PID survives without converting to the target XPT (Table 8). The required limiter-to-divertor topology change is absent from the AdvA-PPO training distribution; the matched early stop of the Reward-PPO slot shows the difficulty is shared. FF+PID already travels across I_p and initial as a stabiliser, while the learned controller can still be adapted for the missing conversion: that is the motivation for the multi-initialization fine-tuning of Sec. 3.3.3.

3.3.3 Fine-tuning across magnetic initials

Zero-shot AdvA-PPO covers the two divertor initials but fails on the limiter I4 (#15892). A short multi-initialization fine-tune of the frozen checkpoint asks whether one adapted policy can span divertor and limiter resets under the same frozen #13906 XPT target. The protocol widens only the reset distribution: 100 further iterations with a Kullback–Leibler (KL) divergence-capped update [30], unchanged reward operators and observation map, and evaluation under the harness of Sec. 3.1. Because survival differs before and after adaptation, Figure 11 and Table 9 summarise the before/after comparison; every error and objective entry in the table is taken on the common window of its before/after pair, with survival reported separately.

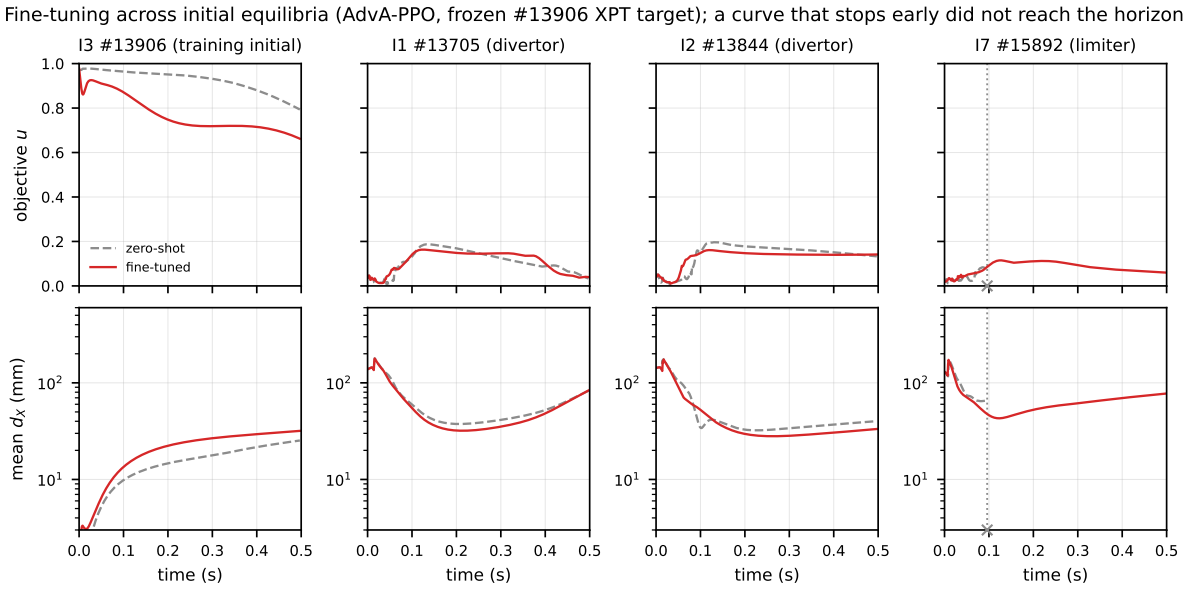

Figure 11: AdvA-PPO multi-initialization fine-tune (100 iterations, frozen #13906 XPT). Grey dashed: zero-shot; red: fine-tuned. Top: $\bar{u}$; bottom: mean d_X. Early stop marked by a dotted vertical line.

The fine-tuned policy covers all four initials (Figure 11). On #15892 survival rises from 96 steps to the full horizon; over the shared 96 steps the objective is essentially unchanged (0.04 against 0.05) while the boundary error halves (182 to 88×10^{-3}), so the budget buys the remaining horizon and a cleaner boundary rather than a uniformly stronger tracker. On the two divertor initials, already full-horizon zero-shot, the objective stays flat (0.11 to 0.11 on #13705, 0.14 to 0.13 on #13844), with a modest I_p gain (31.8 to 26.2 and 17.6 to 13.7 kA RMSE) and geometric channels essentially held. The inevitable compromise appears on the training initial #13906: $\bar{u}$ drops from 0.92 to 0.77 and the mean worst channel from 0.81 to 0.58. Multi-initialization coverage and nominal specialisation are therefore not free jointly

Table 9: AdvA-PPO before/after multi-initialization fine-tuning (100 iterations, frozen #13906 XPT). Errors and $\bar{u}/\bar{r}$ on the common window of each pair; survival separate. On #15892 the common window is the 96 zero-shot steps; over its own 500 ms the fine-tuned run reaches $\bar{u} = 0.08$, I_p RMSE 20.8 kA, mean $d_X = 74$ mm.

Init.	Controller	Surv.	I_p (kA)	mean d_X (mm)	LCFS ($\times 10^{-3}$)	$\bar{u} \uparrow$	$\bar{r} \uparrow$
I1 #13906 (XPT / training)	zero-shot	full	**0.3**	**17**	**12**	**0.92**	**0.81**
	fine-tuned	full	0.7	24	17	0.77	0.58
I2 #13705 (divertor)	zero-shot	full	31.8	69	**63**	**0.11**	**0.01**
	fine-tuned	full	**26.2**	**67**	64	**0.11**	0.00
I3 #13844 (divertor)	zero-shot	full	17.6	60	79	**0.14**	**0.00**
	fine-tuned	full	**13.7**	**58**	**66**	0.13	**0.00**
I4 #15892 (limiter)	zero-shot	96	**20.7**	99	182	0.04	**0.00**
	fine-tuned	**full**	25.0	**92**	**88**	**0.05**	**0.00**

under this short budget—the same weights that span divertor and limiter give up part of the #13906 solution.

Each cell is a single trajectory, so survival figures are attempted rollouts rather than success rates. Fine-tuning under the noise and delay conditions of Sec. 3.3.1 is left for future work.

4 Conclusion

We established an AI-enabled multi-objective framework for EXL-50U XPT magnetic control, formulated as reconstructed-null feedback through shared poloidal-field actuators. Its ten physical objectives jointly represent I_p, the LCFS, four X-point positions, and four X-point flux constraints. The FGE environment reproduces the experimental #13906 equilibrium at centimetre-scale boundary and null accuracy, providing an experiment-calibrated free-boundary test bed for closed-loop evaluation.

AdvA addresses multi-objective temporal credit assignment by retaining one value head and one GAE per channel on a shared critic backbone, applying worst-objective-aware SmoothMax only after advantage estimation, and adding a norm-capped component residual to the aggregate policy update. On the nominal 500 ms rollout, AdvA-PPO raises the shared evaluation score $\bar{u}$ from 0.56 to 0.93 and the mean worst-channel score from 0.23 to 0.81 relative to Reward-PPO, while reducing mean X-point flux RMSE by about 20× and using the lowest mean per-step voltage change. Among the evaluated corrections, the capped residual gives the most balanced recovery of I_p and X-point flux accuracy; the result favours a controlled residual over stacking every available multi-objective correction.

The broader evaluation reveals complementary, regime-dependent strengths. Under combined measurement noise and observation delay, AdvA-PPO is the only learned controller that completes the horizon while retaining a usable XPT shape, although Reward-PPO is stronger under noise alone and the experiment-derived FF+PID baseline tracks I_p well under delay. Across withheld initial equilibria, the same FF+PID controller is the more transferable stabiliser but remains shape-imprecise and does not produce the limiter-to-XPT conversion; the learned controllers improve geometric regulation on the diverted initials but both fail that conversion zero-shot. Multi-initial adaptation extends AdvA-PPO to all four initials and exposes a coverage–specialisation trade-off. Together, these results establish the capability and current boundaries of reconstructed-null XPT feedback in an experiment-calibrated magnetic simulation; PTEFIT-in-the-loop and closed-loop machine performance remain to be tested experimentally.

5 Future outlook

The immediate next step is staged on-machine validation on EXL-50U. The near-term campaign will first quantify the accuracy, latency, and stability of PTEFIT secondary-null position and flux estimates, and then evaluate the present RL controller with PTEFIT in the loop. Conservative actuator limits, online monitoring, and a deterministic FF+PID fallback will allow the effects of real-time reconstruction and learned control to be assessed separately before extending the tests to disturbances and broader operating conditions.

Algorithmic development will target robustness and generalisation beyond the single-initial training regime. Multi-initial and uncertainty-aware training should span plasma-current targets, measurement uncertainties, actuator errors, and FGE calibration residuals. Domain randomisation, constrained policy

optimisation, and stage-dependent objectives provide complementary routes to wider coverage while limiting the loss of nominal precision exposed by the present fine-tuning study.

The control task should also expand from a fixed XPT target to broader divertor configurations and controlled topology transitions. Strike-point and divertor-leg targets, richer LCFS descriptors, and explicit coil-current, voltage, and slew constraints are natural next objectives. Coordinating these tests with auxiliary-heating experiments will probe magnetic-configuration control under more relevant plasma conditions and prepare later integration with radiation, detachment, and target-heat-flux control as suitable models and diagnostics become available. Throughout this progression, hard safety limits, interlocks, fallback control, and staged validation must make machine protection part of the controller design rather than an external safeguard.

References

[1] A. Q. Kuang et al. Divertor heat flux challenge and mitigation in sparc. *Journal of Plasma Physics*, 86, 2020.

[2] E. Kolemen et al. Initial development of the diii-d snowflake divertor control. *Nuclear Fusion*, 58:066007, 2018.

[3] V. A. Soukhanovskii et al. Snowflake divertor experiments in the diii-d, nstx, and nstx-u tokamaks aimed at the development of the divertor power exhaust solution. *IEEE Transactions on Plasma Science*, 44(12):3445–3455, 2016.

[4] S. Gorno, O. Février, C. Theiler, T. Ewalds, F. Felici, T. Lunt, A. Merle, J. Degrave, B. P. Duval, K. Lee, H. Reimerdes, B. Tracey, M. Wischmeier, and C. Wüthrich. X-point radiator and power exhaust control in configurations with multiple x-points in tcv. *Physics of Plasmas*, 31:072504, 2024.

[5] B. LaBombard et al. Adx: a high field, high power density, advanced divertor and rf tokamak. *Nuclear Fusion*, 55:053020, 2015.

[6] M. V. Umansky et al. Assessment of x-point target divertor configuration for power handling and detachment front control. *Nuclear Materials and Energy*, 2017.

[7] H. Raj et al. Improved heat and particle flux mitigation in high core confinement, baffled, alternate divertor configurations in the tcv tokamak. *Nuclear Fusion*, 2022.

[8] K. Lee et al. X-point target radiator regime in tokamak divertor plasmas. *Physical Review Letters*, 134:185102, 2025.

[9] N. Lonigro et al. Initial observations in x-point target divertor discharges on mast-u, 2026.

[10] M. Winkel, K. Verhaegh, B. Kool, K. Lee, M. Carpita, A. Perek, R. Morgan, G. Derks, O. Février, C. Theiler, D. Brida, and M. van Berkel. Detachment dynamics and disturbance rejection in the TCV x-point target divertor, 2026.

[11] Xiang Gu et al. Poloidal field system and advanced divertor equilibrium configuration design of the ehl-2 spherical torus. *Plasma Science and Technology*, 27:024011, 2025.

[12] Fuqiong Wang et al. Divertor heat flux challenge and mitigation in the ehl-2 spherical torus. *Plasma Science and Technology*, 27:024009, 2025.

[13] H. Anand et al. Real-time plasma equilibrium reconstruction and shape control for the mast upgrade tokamak. *Nuclear Fusion*, 64:086051, 2024.

[14] C. Theiler et al. Results from recent detachment experiments in alternative divertor configurations on tcv. *Nuclear Fusion*, 57:072008, 2017.

[15] Yuejiang Shi et al. Overview of exl-50u experiments: addressing key physics issues for future spherical torus reactors. *Nuclear Fusion*, 66:116009, 2026.

[16] Jonas Degrave et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602:414–419, 2022.

[17] B. D. Tracey, A. Michi, Y. Chervonyi, I. Davies, C. Paduraru, N. Lazic, F. Felici, T. Ewalds, C. Donner, C. Galperti, J. Buchli, M. Neunert, A. Huber, J. Evens, P. Kurylowicz, D. J. Mankowitz, and M. Riedmiller. Towards practical reinforcement learning for tokamak magnetic control. *Fusion Engineering and Design*, 200:114161, 2024.

[18] Artem Subbotin et al. Demonstration of reconstruction-free static magnetic control of diii-d plasma with deep reinforcement learning. *Nuclear Fusion*, 2026.

[19] S. Kerboua-Benlarbi, R. Nouailletas, B. Faugeras, E. Nardon, and P. Moreau. Magnetic control of

WEST plasmas through deep reinforcement learning. *IEEE Transactions on Plasma Science*, 2024.

[20] G. H. Zheng, S. F. Liu, X. Gu, Y. P. Zhang, J. Li, Y. Liu, X. C. Lun, L. Xing, J. G. Chen, Z. Y. Chen, Y. Yu, D. Guo, Z. Y. Yang, H. S. Xie, X. M. Song, Y. J. Shi, and EXL-50U Team. A novel numerical algorithms optimization method with machine learning frameworks: Application on real-time plasmas equilibrium reconstruction in EXL-50U spherical torus, 2026.

[21] A. Mele, M. A. Topalova, C. Galperti, and S. Coda. First experimental demonstration of plasma shape control in a tokamak through model predictive control, 2025.

[22] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[23] Humphrey Munn, Brendan Tidd, Peter Boehm, Marcus Gallagher, and David Howard. Scalable multi-objective robot reinforcement learning through gradient conflict resolution, 2025.

[24] Tanmay Ambadkar, Sourav Panda, Shreyash Kale, Jonathan Dodge, and Abhinav Verma. Preference conditioned multi-objective reinforcement learning: Decomposed, diversity-driven policy optimization, 2026.

[25] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation, 2016.

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

[27] Jean-Marc Moret, B. P. Duval, H. B. Le, S. Coda, F. Felici, and H. Reimerdes. Tokamak equilibrium reconstruction code liuqe and its real time implementation. *Fusion Engineering and Design*, 91:1–15, 2015.

[28] X. M. Song, J. X. Li, J. A. Leuer, J. H. Zhang, and X. Song. First plasma scenario development for HL-2M. *Fusion Engineering and Design*, 147:111254, 2019.

[29] Volodymyr Mnih et al. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

[30] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.