

CONFERENCE PRE-PRINT**PLASMA INSTABILITY EVENTS DETECTION
AND DISRUPTION PREDICTION IN EAST TOKAMAK
VIA HETEROGENEOUS-FEATURE MULTI-TASK LEARNING**

YUNHU JIA, HAOYU WANG

University of Science and Technology of China
Institute of Plasma Physics, HFIPS, Chinese Academy of Sciences
Hefei, ChinaWENHUI HU, BIHAO GUO, QINGQUAN YANG, YAO HUANG, YUEHANG WANG, KAI WU,
ZHENGPIING LUO, RUIRUI ZHANG, JIANQIU ZHU, QIPING YUANInstitute of Plasma Physics, HFIPS, Chinese Academy of Sciences
Hefei, China
Email: huwh@ipp.ac.cn

ZIJIE LIU

Institute of Energy, Hefei Comprehensive National Science Center
Hefei, China

BINGJIA XIAO

Institute of Energy, Hefei Comprehensive National Science Center
Institute of Plasma Physics, HFIPS, Chinese Academy of Sciences
University of Science and Technology of China
Hefei, China**Abstract**

A heterogeneous-feature multi-task learning (HFMTL) framework is proposed for simultaneous tokamak plasma disruption prediction (DP), edge localized mode (ELM) detection, multifaceted asymmetric radiation from the edge (MARFE) detection, and H-mode/L-mode (H/L) identification. Due to the significant heterogeneity of the key signal features underlying these four tasks, conventional multi-task learning (MTL) underperforms single-task learning (STL). To address this limitation, HFMTL incorporates a specialized gated mixture-of-experts neural network, enabling each task branch to adaptively retain highly correlated signal features while suppressing interference from weakly correlated ones. Additionally, HFMTL leverages physics-informed priors in data preprocessing, loss function design, and parameter initialization to further enhance performance. The advantages of HFMTL are demonstrated using ~ 10000 historical discharges from the EAST tokamak, spanning two wall configurations: pre-upgrade non-full metal wall (non-FMW) and post-upgrade full metal wall (FMW). Trained and tested on non-FMW dataset, HFMTL achieves $AUC(DP)=0.986$, $AUC(ELM)=0.996$, $AUC(MARFE)=0.975$, $AUC(H/L)=0.999$. Then, tested directly on FMW dataset without any retraining, HFMTL still achieves $AUC(DP)=0.965$, $AUC(ELM)=0.973$, $AUC(MARFE)=0.935$ and $AUC(H/L)=0.985$. These HFMTL's AUC values are superior to those of conventional MTL and STL. Compared to previous EAST DP model (achieving $AUC(DP)=0.97$ on non-FMW and 0.72 on FMW), HFMTL demonstrates a significant improvement in zero-shot cross-wall DP performance. Notably, Pearson correlation analysis shows HFMTL's MARFE detection output correlates positively with DP output ($r=0.437$), whereas ELMs output correlate negatively with DP ($r=-0.322$). These correlations emerge naturally from the dataset without prior-restriction, and in agreement with an independent investigation at JET. The paper presents a promising strategy for the integrated monitoring of tokamak plasma states, with potential applications in targeted disruption avoidance for different precursor phenomena.

1. INTRODUCTION

Automatic monitoring and prediction of key plasma phenomena are crucial for the safe, high-performance operation of tokamaks. Plasma disruptions pose a significant risk to next generation high-power tokamaks, making reliable disruption prediction (DP) essential for implementing mitigation strategies[1]. Since disruptions are often

preceded by precursor phenomena like instabilities (e.g., MARFE, ELM) or confinement degradation, integrating DP with precursor monitoring can enable deeper causal analysis and facilitate targeted interventions[2].

Machine learning (ML), particularly Multi-Task Learning (MTL), offers a promising path toward an integrated monitoring solution. By training multiple related tasks within a unified model, MTL can improve generalization, reduce deployment costs, and leverage inter-task correlations to enhance performance[3]. Indeed, MTL is increasingly applied in tokamak research, with studies showing improved performance and interpretability for tasks like joint disruption and precursor identification[4], as well as simultaneous ELM detection and H/L mode classification[5].

However, MTL is susceptible to "negative transfer," where inappropriate information sharing between tasks degrades performance, sometimes falling below that of single-task learning (STL) models. This issue has been observed in fusion research, for instance, when low-accuracy precursor detection harmed DP performance on DIII-D[4], and has been reported in other predictive modeling tasks for tokamaks[6, 7]. Despite its prevalence, strategies to mitigate negative transfer in fusion applications remain underexplored.

To address this challenge, we propose a Heterogeneous-Feature Multi-Task Learning (HFMTL) framework, motivated by the insight that feature space heterogeneity across tasks is a primary driver of negative transfer. We validate HFMTL on the EAST tokamak using four physically coupled yet feature-heterogeneous tasks: disruption prediction (DP), ELM detection, MARFE detection, and H/L confinement mode identification[8, 9].

The HFMTL model was trained exclusively on data from EAST's pre-upgrade non-full-metal-wall (non-FMW) configuration and tested on both unseen non-FMW and post-upgrade full-metal-wall (FMW) datasets[10]. On the non-FMW test set, HFMTL outperformed both STL and conventional MTL, achieving AUC scores of 0.986 (DP), 0.996 (ELM), 0.975 (MARFE), and 0.999 (H/L). Crucially, without retraining, the model demonstrated excellent zero-shot generalization to the FMW dataset, with AUCs of 0.965 (DP), 0.973 (ELM), 0.935 (MARFE), and 0.985 (H/L). This marks a significant improvement over a previous model whose zero-shot DP performance on the FMW set was notably lower (AUC of 0.72)[11].

This paper is organized as follows: Section 2 details the database and tasks, Section 3 introduces the HFMTL framework, Section 4 presents the results.

2. DATABASE AND PRE-PROCESSING

This study utilizes 45 signals from the EAST tokamak [12, 13, 14, 4, 11], summarized in Table 1. All signals are interpolated to 1 kHz and normalized for training stability. To preserve relative distributions, signals within the same group (e.g., D_α channels) share normalization parameters. Crucially, different tasks rely on heterogeneous subsets of these signals, as detailed in Table 2. For instance, while I_{ic} is vital for DP, it is weakly correlated with H/L identification. Including such weakly correlated signals degrades single-task performance. Therefore, these key signal subsets are used as inputs for single-task learning (STL) models and to provide physics-informed priors for the proposed HFMTL framework (as described in Section 3).

TABLE 1. Input Signals

Group No.	Symbol	Description	Signal Count
1	^a $PXUV_{main}$	Fast bolometers XUV channel of main-region	19
2	^b $PXUV_{edge}$	Fast bolometers XUV channel of edge-region	4
3	^c D_α	Deuterium Balmer- α line emission spectrum	4
4	POWERS	P_{NBI} , P_{LH} , P_{ICRF} , P_{OHM} , P_{ECRH} , P_{RAD}	6
5	V_{loop}	Loop voltage	1
6	Z_{error}	The Z error after the high-pass filter in fast Z control	1
7	I_{ic}	Current of in-vessel coil in fast Z control	1
8	I_p	Plasma current	1
9	$I_{perror\ norm}$	$(I_p - \text{Programmed } I_p) / \text{Programmed } I_p$	1
10	f_{GW}	Electron density/Greenwald density	1
11	kappa	Plasma elongation	1
12	li	Internal Inductance	1
13	W_{mhd}	Plasma stored energy	1
14	q95	Safety factor at the 95% flux surface	1
15	B_{center}	Magnetic Field at the toroidal magnetic axis	1
16	a_{minor}	Minor radius of plasma	1

^a $PXUV_{main}$ channels: 54,9,52,11,48,13,46,17,44,19,42,22,39,24,36,26,34,29,32; The purpose of this staggered arrangement is to make it easier for CNNs to learn the Z-symmetry feature of the radiation profile

^b $PXUV_{edge}$ channels: 2,6,56,58

^c D_α channels: L1,L2,U2,U3

TABLE 2. Key Signal Subsets Corresponding to Each Task

Task	Signal
DP	ALL signals
ELM	$D_{\alpha}, PXUV_{edge}, f_{GW}, W_{mhd}, I_p, a_{minor}, V_{loop}, B_{center}$
MARFE	$PXUV_{main}, W_{mhd}, f_{GW}, \text{POWERS}$
H/L	$D_{\alpha}, PXUV_{edge}, f_{GW}, W_{mhd}, I_p, a_{minor}, V_{loop}, B_{center}$

The database comprises 9833 discharges from EAST (2015–2024), spanning both non-full-metal-wall (non-FMW) and post-upgrade full-metal-wall (FMW) configurations. Non-FMW data was partitioned by shot number into training, validation, and testing sets. The FMW data was reserved exclusively as a test set to evaluate zero-shot generalization. All tasks use supervised learning with expert-annotated labels applied only during the plasma flattop phase. For the DP task, training/validation data is labeled with "safe" and "unsafe" phases (200 ms pre-disruption), while test data uses only shot-level "disruptive" vs. "non-disruptive" labels. For ELM, MARFE, and H/L identification, specific time intervals of their occurrence are annotated on subsets of discharges. Table 3 provides a detailed summary of the dataset partitioning and annotation.

TABLE 3. Dataset Partitioning and Annotation Overview

2*Task Type	Non-FMW			FMW
	Training(6545)	Validation(1417)	Testing(1325)	Testing(546)
DP	Disrupt Shots ^a : 1984 nonDisrupt Shots: 4561 unsafe ^b : 396.8s safe: 31635.9s	Disrupt Shots: 437 nonDisrupt Shots: 980 unsafe: 87.4s safe: 6913.8s	Disrupt Shots: 477 nonDisrupt Shots: 848 unsafe: N/A safe: N/A	Disrupt Shots: 212 nonDisrupt Shots: 334 unsafe: N/A safe: N/A
ELMs	Shots: 334 ELM: 492.2s nonELM: 392.2s	Shots: 66 ELM: 98.3s nonELM: 234.1s	Shots: 81 ELM: 98.8s nonELM: 252.2s	Shots: 138 ELM: 531.3s nonELM: 696.5s
MARFE	Shots: 55 MARFE: 27.4s nonMARFE: 202.7s	Shots: 82 MARFE: 20.4s nonMARFE: 207.1s	Shots: 41 MARFE: 11.2s nonMARFE: 106.4s	Shots: 69 MARFE: 42.4s nonMARFE: 83.4s
H/L	Shots: 132 H-mode: 51.9s L-mode: 100.8s	Shots: 73 H-mode: 105.7s L-mode: 47.4s	Shots: 75 H-mode: 108.0s L-mode: 86.5s	Shots: 133 H-mode: 542.8s L-mode: 119.7s

^a **Disrupt/NonDisrupt Shots**: Count of plasma discharges with/without disruption.

^b **Safe/Unsafe**: Cumulative duration (in seconds) annotated as safe/unsafe phases. For the testing dataset, disruption prediction only requires shot-level performance metrics, so detailed safe/unsafe phase annotations are omitted (N/A). For all other tasks, the definition follows analogously, denoting the total annotated duration for "MARFE occurrence/non-occurrence, ELM occurrence/non-occurrence, H-mode occurrence/L-mode occurrence".

Using a "sequence-to-label" approach[14, 4], we extract 100 ms input sequences (X , shape: 45×100) via a sliding window. A corresponding 4D label vector (Y) is generated from the annotations at the final timestep of each sequence, encoding the status of DP (1 for unsafe), ELMs (1 for active), MARFE (1 for active), and H/L mode (1 for H-mode). Due to limited annotation resources, not all labels exist for every sample. These unannotated labels are marked as "missing" and are ignored during loss computation. This "missing label" mechanism is also used to address class imbalance by randomly masking labels from majority-class samples in the training set.

3. METHODS

3.1. Model Architecture

To address the heterogeneous signal dependencies of four distinct tasks—disruption prediction (DP), ELM detection, MARFE detection, and H/L-mode identification—we developed a specialized Multi-Task Learning (MTL) model. Inspired by the Multi-gate Mixture-of-Experts (MMoE) paradigm[15], our architecture (Fig. 1) is designed to prevent performance degradation from irrelevant signals. It comprises three main components: expert modules, gating modules, and task heads.

Unlike standard MMoE, input signals are first partitioned into physically meaningful groups (Table 1), with each group processed by a dedicated convolutional expert module. This prevents information confusion during feature extraction. The features from all experts are then adaptively weighted by a task-specific gating module before being passed to the corresponding task head. The gate calculates a weight vector $\mathbf{w}$ using a softmax

function, which amplifies relevant features while suppressing other irrelevant features:

$$w_i = \frac{\exp(v_i/T_c)}{\sum_{j=1}^{16} \exp(v_j/T_c)}, \quad (1)$$

where $\mathbf{v}$ is a trainable parameter vector and T_c is a temperature coefficient. To incorporate prior knowledge, $\mathbf{v}$ is initialized based on the key signal subsets for each task (Table 2), providing a physics-informed starting point. Finally, each task head uses linear layers to produce a prediction from the weighted features. Although experts operate on separate inputs, they are updated by gradients from all tasks, enabling them to learn generalizable representations.

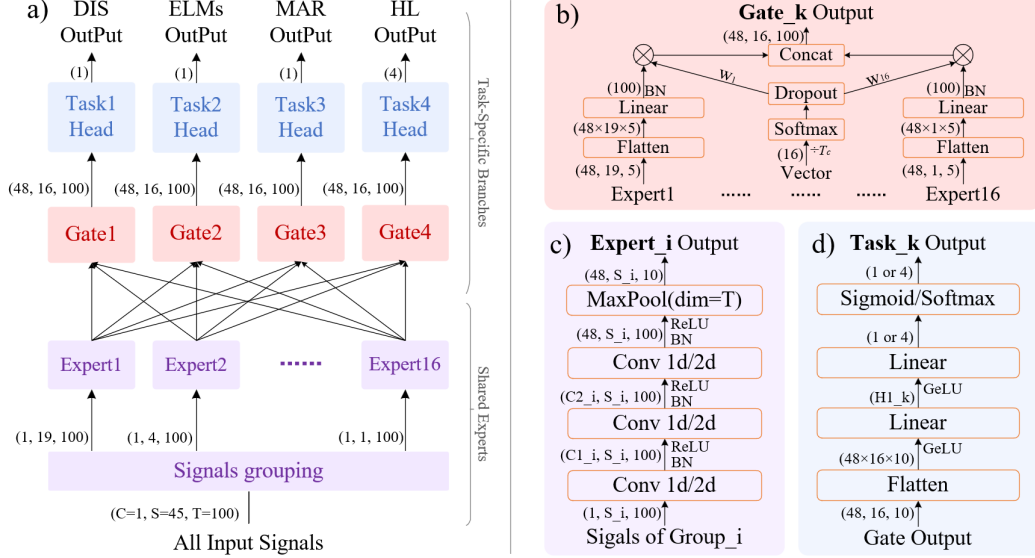


FIG. 1. Schematic diagrams of the model structure. a) shows the overall model structure, b) details of the gating layer, c) details of an expert module in the feature extractor, and d) details of a task branch.

3.2. Loss Function

The total loss L_{total} combines a main task loss L_{main} and a physics-informed regularization term L_{phys} , where samples with missing labels are excluded from the loss calculation of the corresponding task.

$$L_{\text{total}} = L_{\text{main}} + L_{\text{phys}} \quad (2)$$

The main loss is a weighted sum of the binary cross-entropy (BCE) losses for the four tasks:

$$L_{\text{main}} = \sum_{t \in \text{Tasks}} \alpha_t L_t, \quad \text{where} \quad L_t = \frac{1}{N_t} \sum_{n=1}^{N_t} \ell_{\text{BCE}}(y_t^{(n)}, \hat{y}_t^{(n)}) \quad (3)$$

The L_{phys} term, defined as $L_{\text{phys}} = \sum_{k=1}^2 \alpha_k \cdot L_{\text{phys},k}$, enforces physical consistency by penalizing predictions that violate two known priors:

1. **ELMs occur only in H-mode:** Penalizes low H-mode probability for ELM-positive samples.

$$L_{\text{phys},1} = \frac{1}{N_{\text{ELM}}} \sum_{n=1}^{N_{\text{ELM}}} y_{\text{ELM}}^{(n)} \cdot (-\ln(\hat{y}_{\text{H/L}}^{(n)})) \quad (4)$$

2. **L-mode is ELM-free:** Penalizes high ELM probability for L-mode samples.

$$L_{\text{phys},2} = \frac{1}{N_{\text{H/L}}} \sum_{n=1}^{N_{\text{H/L}}} (1 - y_{\text{H/L}}^{(n)}) \cdot (-\ln(1 - \hat{y}_{\text{ELM}}^{(n)})) \quad (5)$$

3.3. Model Training and Hyperparameter Optimization

Models were trained using the Adam optimizer with an "asynchronous early-stopping mechanism," where a task's loss weight is halved if its validation loss stagnates, and global training terminates if all tasks stagnate. We optimized hyperparameters using a genetic algorithm to maximize the sum of AUCs on the validation set. The final optimized configuration, achieves an inference time under 1 ms on an NVIDIA RTX 4090D GPU, satisfying real-time processing requirements.

4. RESULTS AND DISCUSSION

4.1. Evaluation and Performance

We evaluated performance using shot-level metrics (AUC, warning time) for disruption prediction (DP) and sample-level AUC for ELM detection, MARFE detection, and H/L identification. The HFMTL model was trained on non-FMW data and tested on both unseen non-FMW and FMW datasets, with all results averaged over 20 independent trials.

For the DP task, HFMTL achieved a state-of-the-art AUC of 0.986 ± 0.002 on the non-FMW test set and a robust zero-shot AUC of 0.965 ± 0.009 on the FMW set, significantly outperforming previous benchmarks[11] (Fig. 2). At a fixed FPR of 0.08, it provided ample warning times on both datasets (Fig. 3). The framework also showed excellent performance on the other tasks (Figs. 4, 5, 6). Specifically, trained and tested on the non-FMW dataset, HFMTL achieved $\text{AUC}(\text{DP})=0.986$, $\text{AUC}(\text{ELM})=0.996$, $\text{AUC}(\text{MARFE})=0.975$, and $\text{AUC}(\text{H/L})=0.999$. When tested directly on the FMW dataset without retraining, the model maintained high performance with $\text{AUC}(\text{DP})=0.965$, $\text{AUC}(\text{ELM})=0.973$, $\text{AUC}(\text{MARFE})=0.935$, and $\text{AUC}(\text{H/L})=0.985$. Example operational traces confirm the model's real-time accuracy in identifying precursors and predicting disruptions (Figs. 7).

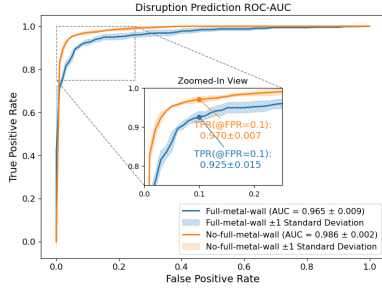


FIG. 2. Disruption Prediction ROC-AUC.

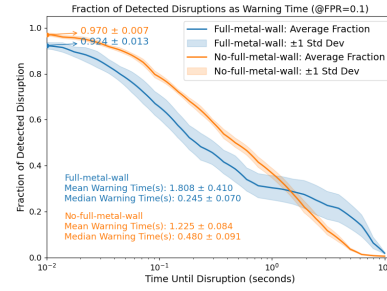


FIG. 3. Fraction of Detected Disruptions vs. Warning Time.

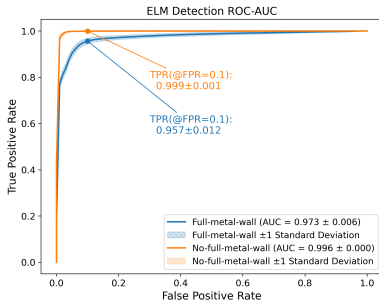


FIG. 4. ELM ROC-AUC.

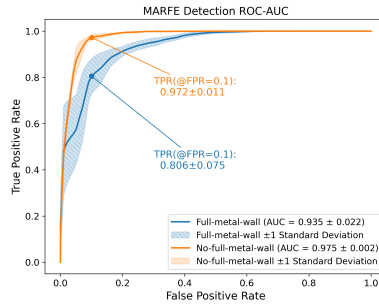


FIG. 5. MARFE ROC-AUC.

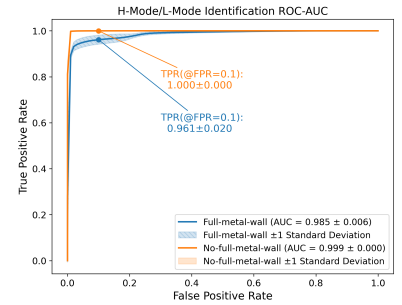


FIG. 6. H/L ROC-AUC.

4.2. Model Comparison and Ablation

To validate our approach, we compared HFMTL against three architectural variants: a conventional MTL model, an STL model with all signals, and an STL model with task-specific signals (Fig. 8). Results from these ablation studies (Fig. 9) confirm that weakly correlated signals introduce noise that degrades performance in STL models. HFMTL consistently outperformed all variants, demonstrating that its gating mechanism effectively mitigates negative transfer by adaptively managing feature heterogeneity, leading to significant performance gains and reduced variance.

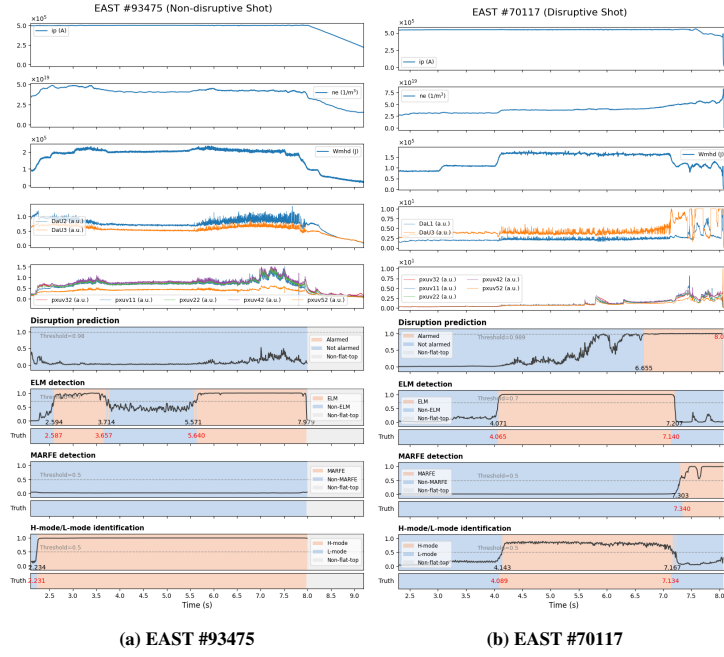


FIG. 7. Schematic diagram of HFMTL model operation on non-FMW disruptive and non-disruptive shots data.

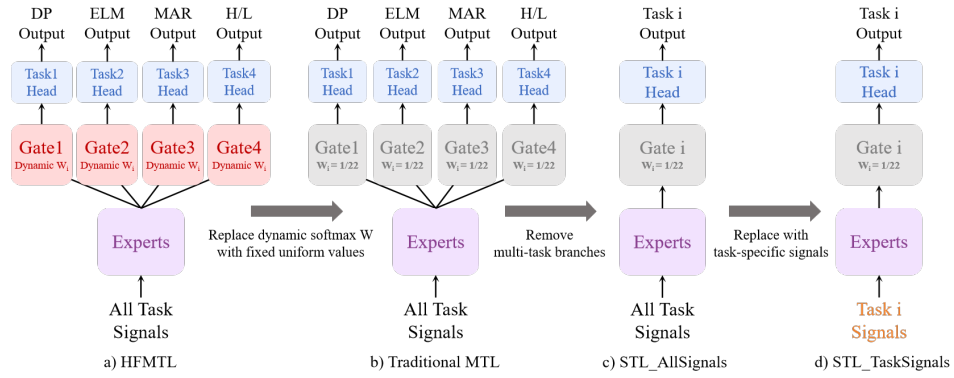


FIG. 8. Schematic diagrams of the comparative models for ablation.

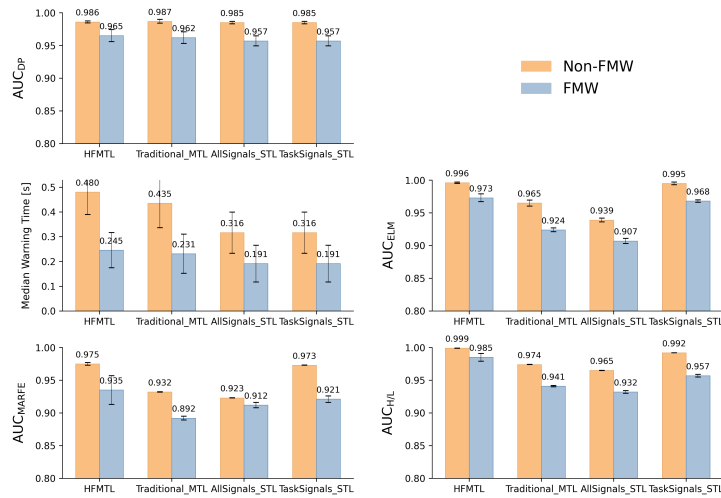


FIG. 9. Ablation experiment results comparing HFMTL with variants.

4.3. Interpretability and Physics Consistency

Analysis of the learned gating module weights provides model interpretability (Fig. 10). The final weights, adapted from their physics-informed initializations, align with established physics: the DP task emphasizes main-region radiation and vertical control signals; ELM detection emphasizes top and bottom radiation; MARFE detection focuses on abnormal bulk radiation; and H/L identification relies on confinement and ELM activity signals. This confirms the model learns to prioritize physically relevant features for each task.

Furthermore, we computed Pearson correlations between the model’s task outputs to test for physical consistency (Fig. 11). The results reveal physically meaningful relationships: a negative correlation between DP and ELM ($r = -0.322$), a positive one between DP and MARFE ($r = 0.437$), and a strong positive link between ELM and H/L ($r = 0.844$). These correlations align with findings from other devices[sieglin2024disruption, sieglin2025hmode, 8] and, crucially, emerged spontaneously from the data without being explicitly encoded in the loss function (except for the ELM-H/L link). This demonstrates the model’s ability to uncover underlying physics knowledge.

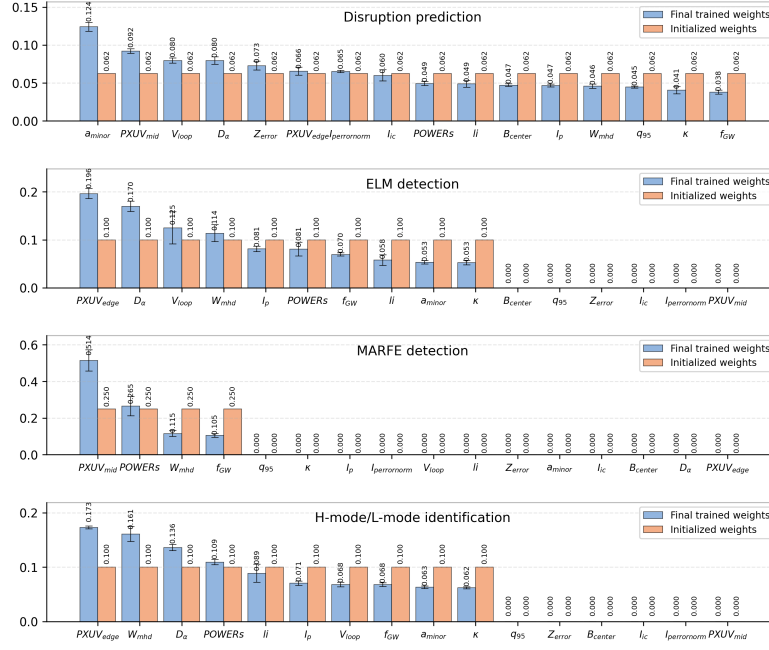


FIG. 10. The weight distributions of gating modules in HFMTL for four key tasks. Orange bars denote the initialized weights, while blue bars represent the final weight statistics from models trained with 20 different random seeds.

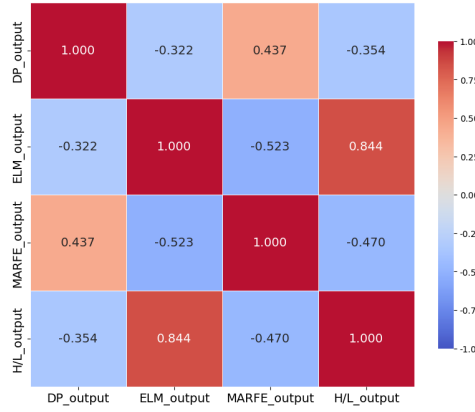


FIG. 11. Pearson correlation matrix of task outputs.

ACKNOWLEDGEMENTS

This work is supported by the National MCF Energy R&D Program of China (Grant Nos. 2022YFE03010002 and 2024YFE03020003), the National Natural Science Foundation of China (Grant Nos. 12575246, 12575247, U24A20342, 12205195, and 12205336), the Institute of Energy, Hefei Comprehensive National Science Center under Grant No. 24KZS304, and the funding support from International Atomic Energy Agency with Research Contract Number of 26478.

REFERENCES

- [1] PC De Vries et al. “Requirements for triggering the ITER disruption mitigation system”. In: *Fusion Science and Technology* 69.2 (2016), pp. 471–484.
- [2] PC De Vries et al. “Survey of disruption causes at JET”. In: *Nuclear Fusion* 51.5 (2011), p. 053018.
- [3] Sebastian Ruder. *An Overview of Multi-Task Learning in Deep Neural Networks*. 2017. arXiv: 1706.05098 [cs.LG]. URL: <https://arxiv.org/abs/1706.05098>.
- [4] JX Zhu et al. “Integrated deep learning framework for unstable event identification and disruption prediction of tokamak plasmas”. In: *Nuclear Fusion* 63.4 (2023), p. 046009.
- [5] Guo Hong Deng et al. “Automatic identification of tokamak plasma confinement states (L-mode, ELM-free H-mode, and ELMy H-mode) with Multi-Task Learning Neural Network”. In: *Nuclear Fusion* (2025).
- [6] Xuan Sun et al. “Impact of various DIII-D diagnostics on the accuracy of neural network surrogates for kinetic EFIT reconstructions”. In: *Nuclear Fusion* 64.8 (2024), p. 086065.
- [7] Andreas Gillgren et al. “Enabling adaptive pedestals in predictive transport simulations using neural networks”. In: *Nuclear Fusion* 62.9 (2022), p. 096006.
- [8] Diogo R Ferreira et al. “Investigating the physics of disruptions with real-time tomography at JET”. In: *Plasma Science and Technology* 24.3 (2022), p. 035103.
- [9] Wenhui Hu et al. “Prediction of multifaceted asymmetric radiation from the edge movement in density-limit disruptive plasmas on Experimental Advanced Superconducting Tokamak using random forest”. In: *Chinese Physics B* 32.7 (2023), p. 075211.
- [10] DL Chen et al. “Characterization of disruption halo current between ‘W-Like’ graphite divertor and ‘ITER-Like’ divertor structure on EAST tokamak”. In: *Plasma Physics and Controlled Fusion* 62.9 (2020), p. 095019.
- [11] BH Guo et al. “Disruption prediction on EAST with different wall conditions based on a multi-scale deep hybrid neural network”. In: *Nuclear Fusion* 63.9 (2023), p. 094001.
- [12] David Orozco et al. “Neural Network-Based Confinement Mode Prediction for Real-Time Disruption Avoidance”. In: *IEEE Transactions on Plasma Science* 50.11 (2022), pp. 4157–4164.
- [13] Jesús Vega et al. “Disruption prediction with artificial intelligence techniques in tokamak plasmas”. In: *Nature Physics* 18.7 (2022), pp. 741–750.
- [14] JX Zhu et al. “Hybrid deep-learning architecture for general disruption prediction across multiple tokamaks”. In: *Nuclear Fusion* 61.2 (2020), p. 026007.
- [15] Jiaqi Ma et al. “Modeling task relationships in multi-task learning with multi-gate mixture-of-experts”. In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2018, pp. 1930–1939.